Extreme meteo-oceanic events

Franck Mazas

UNIVERSITÉ PARIS-EST

École Doctorale SIE
Sciences, Ingénierie et Environnement

Thèse

présentée pour l'obtention du grade de

Docteur de l'Université Paris-Est

par

Franck Mazas

Évènements météo-océaniques extrêmes

Spécialité : Sciences et techniques de l'environnement

Soutenue le 17 novembre 2017 devant le jury composé de :

Rapporteur	Dr. Xavier Bertin	LIENSs, CNRS - Université de La Rochelle
Rapporteur	Pr. Éric Gaume	IFSTTAR
Présidente	Pr. Liliane Bel	AgroParisTech
Examinateur	Pr. Michel Benoit	École Centrale Marseille, Irphé
Examinateur	Dr. Pietro Bernardara	CEREA, EDF R&D
Examinateur	Dr. Ivan Haigh	University of Southampton
Directrice de thèse	Dr. Nicole Goutal	LHSV, EDF R&D
Co-encadrant de thèse	Dr. Luc Hamm	ARTELIA

Thèse préparée au sein d'ARTELIA Eau et Environnement

Université Paris-Est

École Doctorale Sciences, Ingénierie et Environnement (SIE)

EXTREME METEO-OCEANOGRAPHIC EVENTS

PH.D. BY PUBLISHED WORKS – FRANCK MAZAS

SUPPORTING STATEMENT

ARTELIA Eau & Environnement
6 rue de Lorraine
38130 - Echirolles
Tel. : +33 (0) 4 76 33 40 00
Fax : +33 (0) 4 76 33 43 33

DATE : NOVEMBER 2017 **REF :** 171 5617

Sources of illustrations

Cover: Southern Ocean onboard Jérôme Poncet's *Golden Fleece*, January 2014, © Nelly Meignié Huber

Part 1: Emmanuel Lepage, *Voyage aux Îles de la Désolation*

Part 2: Marin-Marie, *Le Pourquoi Pas ? au large de l'Islande*

"Ce sont les évènements qui commandent aux hommes, et non les hommes aux évènements."
Hérodote

"Watch therefore, for ye know neither the day nor the hour."
Matthew, 25:13

"Prognostics do not always prove prophecies; at least the wisest prophets make sure of the event first."
Horace Walpole

TABLE OF CONTENTS

FIGURES

TABLES

CONTEXT AND ACKNOWLEDGEMENTS

The present document is the supporting statement forming part of the PhD by published works submitted to the École Doctorale SIE (Sciences, Ingénierie et Environnement) of the University of Paris-Est.

The aim of part one is to highlight the consistency of the research work that was undertaken, but also its evolution over the almost ten years it covers. A copy of the main publications is provided in part two.

This work of research began in 2007 during a scientific internship at Sogreah while I was a first-year student at the École Nationale des Ponts et Chaussées (ENPC). The context in which it took place, namely a private engineering company that understood the interest of science and was willing to allocate a fair amount of time and resources to it, was not unrelated to my decision to join the company a couple of years later as an engineer. The research hence steadily kept going, both feeding everyday projects and being fed by them and their ever-increasing technical requirements.

This dual dimension of my activity allowed me to be in touch with widely differing communities: colleagues and clients of course, but also members of the national and international scientific community as well as scholars and teachers in the academic world. It was definitely personally and professionally enriching to meet young doctoral students and white-haired scholars during coffee breaks or gala dinners, learn from elder colleagues and teach students from the ENPC and the ESITC, travel the world, and enjoy the life in the Alps.

When it comes to this PhD, I must acknowledge many people. These of course include Dr Luc Hamm who oversaw years of research, enriched it with his scientific mind and advice, and granted me support in the company. And support there was, notwithstanding the changes in organisation and managers: Jacques Viguier, Christophe Peronnard, Sophie Ancel, Alain Deforche and others helped me and allocated resources to enable me to achieve this goal.

In addition to the support he gave to all of his students at the Génie Mécanique et Matériaux department of the ENPC, Dr Alain Ehrlacher was also an early and convincing supporter of developing the research and transforming into a PhD, a position that is not very common in the academic community but that reflects his deep humanity. This warm support was also expressed by many fellow researchers, many of whom are members of the jury. I think particularly of the members of the OSSË working group: Pr Michel Benoit, Dr Pietro Bernardara, Dr Xavier Kergadallan, Marc Andreewsky, Dr Jérôme Weiss, Roberto Frau. I won't name them here but I think also of many others met at conferences, in particular the "extremers" group of EVAN conferences.

Lastly, personal support and encouragement were also unwavering from my wife Apolline and many friends who know I think of them while writing these words.

oOo

PART 1
PRESENTATION OF THE RESEARCH WORK

PRESENTATION OF THE DATASETS

The datasets and results presented in this supporting statement are for illustrative purpose and do not aim at providing quantitative results. Several datasets are also presented in detail in the publications in the second part of this document.

- Figure 6, Figure 21 to Figure 27, Figure 29, Figure 32, Figure 34 to Figure 36, Figure 39:
  - atmospheric fields (sea level pressure and 500 hPa geopotential height): archives of ERA-Interim atmospheric reanalyses 0.75°,
  - wind speed: measurements at Météo-France's weather station of Guipavas (29),
  - significant wave height H_s: output point (48.5°N, 5°W, Figure 1) of a 6-hourly database of sea states over the period 1948-2012 from a numerical model built with the WaveWatch III code forced over the Atlantic Ocean by NCEP wind fields and run with the European "Cycle 4" parameterization (***Bertin et al., 2013***), linearly interpolated every 1 hour,
  - sea level: hourly measurements at SHOM's Brest tide gauge (4.4950°W, 48.3829°N, Figure 1), referenced to the local Chart Datum *Zéro Hydrographique 1996*, from 1953/01/01 to 2010/12/31 (58 years), corrected of the eustatic trend of + 1.48 mm/yr so as to get a mean sea level of + 4.14 m ZH,
  - residual: non-tidal residual (considered as the meteorological surge) after removal of the astronomical retro-predictions computed by SHOM's software SHOMAR;

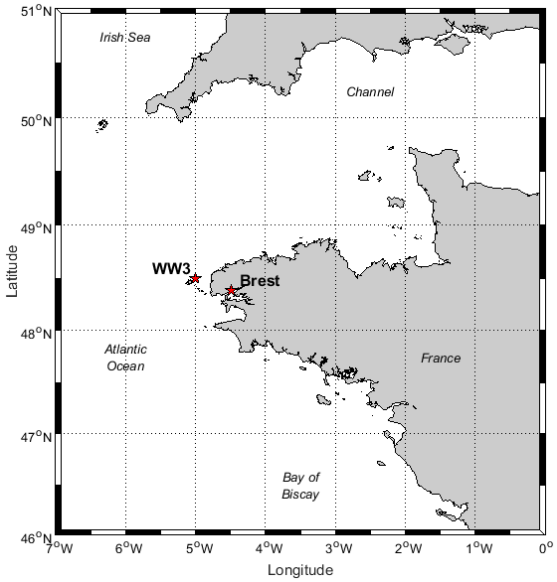

Figure 1. Output point of the WW3 model of Bertin et al. (2013) and location of Brest tide gauge

- Figure 7:
  - atmospheric fields (sea level pressure and 500 hPa geopotential height): GFS 1° Europe, run 18Z of 2010/02/27,
  - sea level: measurements at SHOM's tide gauge of La Rochelle – La Pallice, referenced to the local Chart Datum *Zéro Hydrographique*,
  - tidal level: astronomical retro-predictions computed by SHOM's software SHOMAR,
  - residual: non-tidal residual (considered as the meteorological surge) after removal of the astronomical component from the sea level;
- Figure 13:

- 3-hourly numerical modelling of sea states by a WaveWatch III / SWAN coupled model off Bastia's Carbonite port (~ 450 m deep), run by GlobOcean for ARTELIA;

- Figure 14, Figure 17, Figure 19, Figure 20:
  - Haltenbanken's dataset of storm peaks of H_S provided by the IAHR Working Group on Extreme Wave Analysis (***van Vledder et al., 1994***), issued from the analysis of 3-hourly wave measurements on the Norwegian continental shelf at the deep water locations of 65°05'N, 7°34'E (280 m deep, March 1980 to October 1987) and 65°11'N, 7°15'E (290 m deep, November 1987 to March 1988);
- Figure 40:
  - 1992-2007 3-hourly numerical modelling of sea states by a WaveWatch III / SWAN coupled model off Cotonou (Benin) port, run by GlobOcean for ARTELIA (2008);
- Figure 42 to Figure 47:
  - sea states: 1996-2015 hourly modelling of sea states by a WaveWatch III / SWAN coupled model off Groix island (Figure 2), run by GlobOcean for ARTELIA (2016),
  - wind: output point off Groix island of CFSR wind fields calibrated against satellite measurements by GlobOcean for ARTELIA (2016).

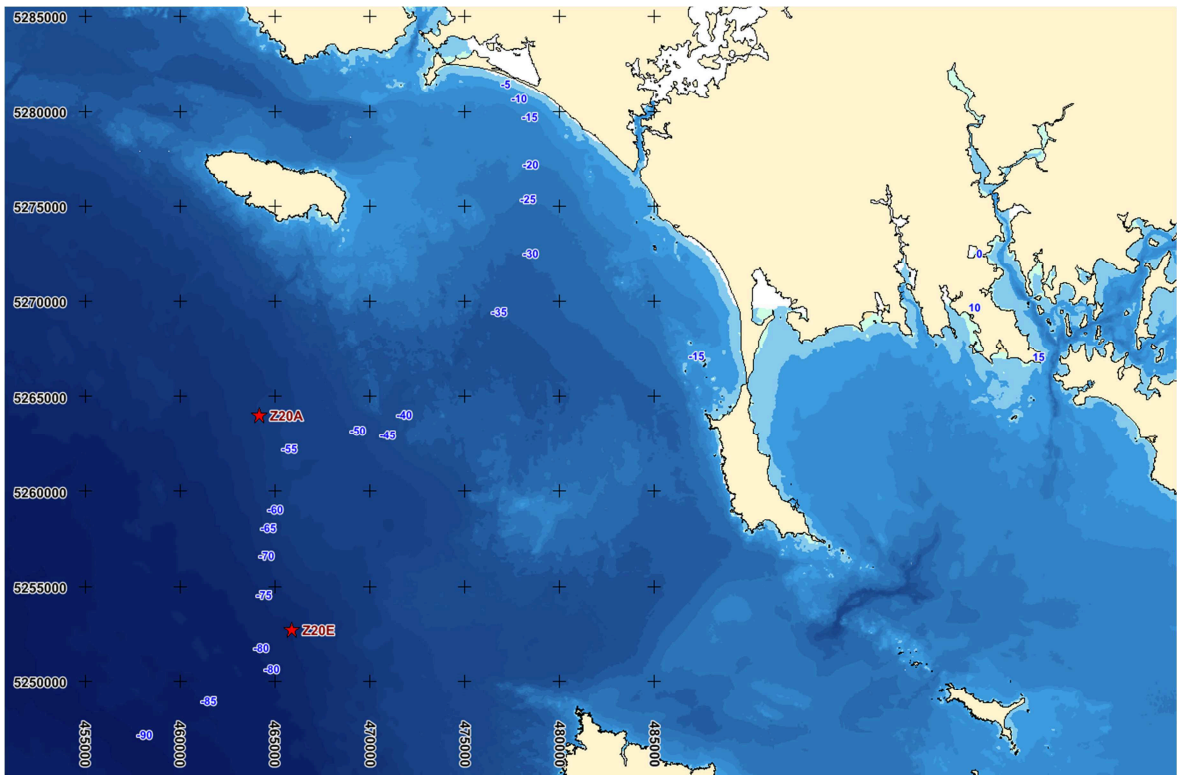

Figure 2. Output points Z20A and Z20E for the Groix meteo-oceanic study

oOo

1. AN INTRODUCTION TO METOCEAN EVENTS

1.1. WHAT IS METEO-OCEANOGRAPHY?

Meteo-oceanography, abbreviated to metocean, is a field of research and knowledge that comes from the gradual connection of two scientific disciplines that have long evolved both in parallel and somewhat separately:

- physical oceanography (sea levels, currents, salinity, temperature, etc.);
- marine meteorology (wind, waves, pressure, etc.).

Offshore engineers, later followed by coastal engineers, are gradually having to address in increasing detail the interactions between the physical quantities of these two once distinct domains. A very simple illustration of such an interaction is the importance of the sea level when considering the propagation of waves in shallow water.

1.1.1. Spatial variability: a useful distinction in geographical domains

An important characteristic of meteo-oceanography is the distinction that can (or even should) be made in several geographical domains from the ocean to the shore. This distinction is based upon the physical processes involved in the generation and propagation of waves:

- the **deep sea domain** (*domaine hauturier*): when the water depth exceeds 200 m or so, i.e. off the continental shelf; waves do not interact with the seabed;
- the **coastal domain** (*domaine côtier*): in the continental shelf from the boundary of the continental slope to the seaward boundary of the surf zone; waves interact quasi-linearly with the seabed (refraction);
- the surf zone or **littoral domain** (*domaine littoral*): wave interaction with the seabed becomes highly non-linear (shoaling, breaking, low-frequency waves);
- the **estuarine domain** (*domaine estuarien*): strong tidal currents, freshwater discharges from rivers and the associated turbidity, and friction must be accounted for;
- the **harbour domain** (*domaine portuaire*): ports, harbours and coastal structures, including access channels; waves interact with structures (reflection, diffraction).

It is interesting to note that the definition of these domains is dynamic as it depends on the wavelength, especially with the first three.

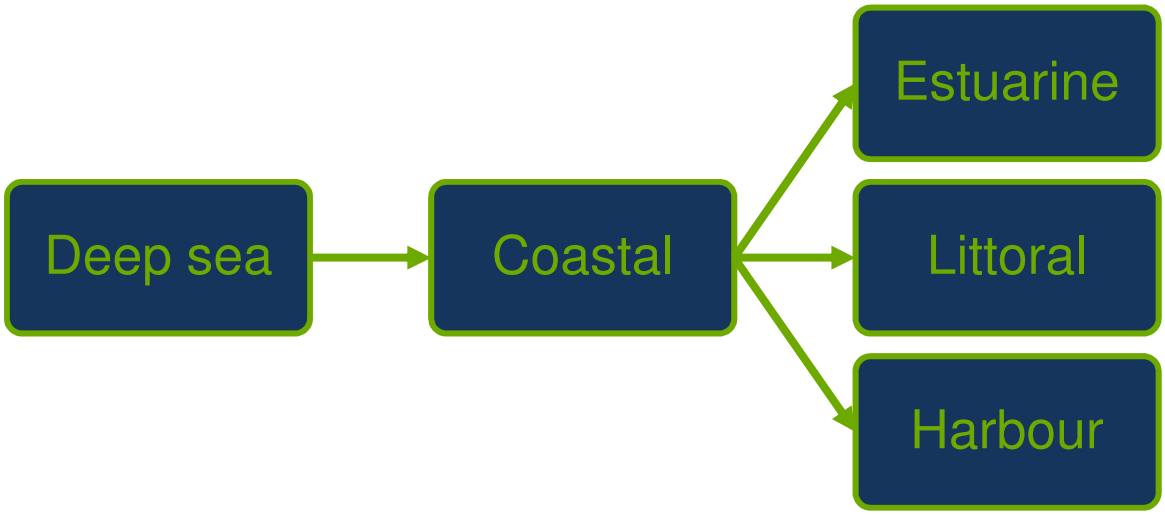

Figure 3. Geographical domains of meteo-oceanography

1.1.2. A far-reaching variety of time scales

We have just mentioned wavelength and water depth. For a scientist, it is thus natural to consider the frequency of the phenomena immediately afterwards. Here again, meteo-oceanography is characterized by a very wide variety of time scales, in particular when it comes to variations in the surface of the sea. The following time scales may be identified:

- geological eras: the drift of continental plates modifies the space available for seawater (isostasy);
- millennia: the alternation of glacial and interglacial periods over the Quaternary period (roughly every 105,000 years over 2.5 million years), driven in particular by the Milanković cycles[1], led to wide variations in sea level (~ 100-200 m);
- centuries: shorter variations in the climate (including current global warming) also yield variations in mean sea level (~ 1 m);
- years to decades: the climate has a natural variability illustrated by oscillations in the ocean-atmosphere coupling over a few years to a few decades (e.g. North Atlantic Oscillation, Arctic Oscillation, El Niño Southern Oscillation, etc.);
- seasons: the tilting of the Earth's axis of rotation relative to the ecliptic plane induces seasonal variations of environmental conditions that increase with latitude;
- days: meteorological patterns such as storms or heatwaves typically last a few days;
- hours: the main tidal components and storm surges have periods of several hours;
- minutes: low-frequency waves that cause harbour or coastal resonance have periods ranging from ~ 1 minute to ~ 1 hour;
- seconds: short waves (wind waves and swells);
- less than 1 second: capillary waves.

[1] Milanković cycles describe the effects of the variations in the Earth's orbital parameters on the global climate over thousands of years. These parameters are the eccentricity of the Earth's orbit (it varies between 0.000055 and 0.0679 with major component periods of 95,000, 125,000 and 413,000 years), the obliquity or axial tilt (it varies between 22.1° and 24.5° with a period of 43,000 years) and the precession of the Earth's axis of rotation (period of 25,760 years).

Figure 4 (***Kinsman, 1965***) illustrates the time scales associated with variations in sea level and their relationship to the different generating forces (atmospheric, astronomic).

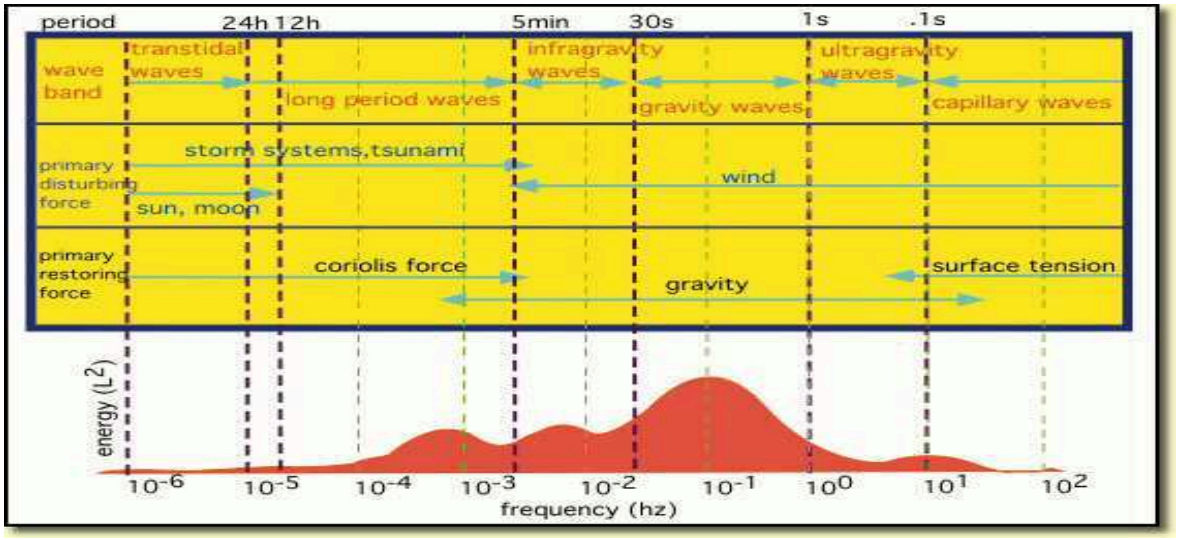

Figure 4. Frequency spectrum of the variations of sea level

source: Kinsman, 1965

An important distinction to be made consists of the difference between short waves and long waves. Although a given, precise limit at any particular frequency is not very meaningful, the period of 30 s is generally accepted as the limit between short and long waves. Below, wave celerity c varies with both period T and water depth h while above it tends to depend solely on water depth, as formulated by Airy's linear wave theory (1841).

$$c = \frac{gT}{2\pi}\tanh\left(\frac{2\pi h}{cT}\right) \xrightarrow[T\to+\infty]{} \sqrt{gh} \qquad (1)$$

Historically, various measurement networks have been set up to measure sea level fluctuations: wave buoys on the one hand for short waves, and tide then pressure gauges on the other hand for long waves.

1.1.3. Input data: measurements and model databases

It is also worth mentioning the data available to the engineer to be used as inputs to his analyses. Roughly speaking, it can be divided in two categories: *in situ* measurements and outputs from numerical models. Both have their pros and cons.

Provided that the measuring device is reliable, measurements provide very accurate data and can be regarded as a faithful reporter of what really happened in nature. In France, the main measurement networks for meteo-oceanic data include the network of coastal wave buoys CANDHIS (*Centre d'Archivage National de Données de Houle In Situ*) operated by CEREMA, the offshore wave buoys operated by Météo-France and the REFMAR network of tide gauges operated by the SHOM (***Pouvreau, 2014***). Radar devices have also been recently set up for measuring sea level (***UNESCO-IOC, 2016***). At the global scale, satellite altimetry provides measurements of various meteo-oceanic quantities including mean sea level and wave heights since TOPEX/Poseidon was launched in 1992 (***Chelton et al., 2001***). But measurements are of course scarce and expensive. In hydrology, one stream gauge measuring the water level (from which the discharge can be deduced) will suffice to give very good knowledge of the river. Along the coastline the density of tide gauges does not need to be very high, but waves present a very wide spatial variability that cannot be reasonably handled by wave buoys. Satellites can cover the whole globe but at any given area the monitoring is only periodic and not continuous. Lastly, measurement devices can neither explore the past nor predict the future but simply report the present: if a port is

to be built in a remote part of Africa, it is inconceivable to set up a few buoys and wait for ten years before beginning the studies.

In some cases, this last difficulty can be partly attenuated by using historical information. This consists of the marks and prints left by the largest observations, either on the environment (high water marks, sediments, etc.) or in the collective memory (archives of local newspapers, memories of elders, etc.). This is particularly useful in hydrology (***Payrastre, 2005***, ***Payrastre et al., 2012***); in coastal engineering, however, the difficulty is that only the highest sea levels can be retrieved whereas the largest meteorological surges would be much more useful (***Bulteau et al., 2015***, ***Hamdi et al., 2015***). As for waves, historical data is generally not available.

Numerical models, on the other hand, can time-travel, either backwards (*hindcast*) or forwards (*forecast*). Output points can be defined anywhere in the computational domain and the marginal cost of additional locations is generally near zero. Depending on the equations, it can be easy to separate the different components and phenomena (see section 1.3.1). Their two main limitations are progressively being pushed back, but can still be major constraints for the analysis. The first one consists of computing power and storage space. The number of nodes can reach 7- to 8-digit figures, and both inputs and outputs could require tera-octets of storage, even when the outputs are not archived at each grid node or at each time step. A clever use of nested models or unstructured meshing becoming progressively more refined can help to overcome this difficulty. The second is more fundamental: this is the translation of the physical reality into a set of mathematical equations; more precisely, equations that can be numerically resolved in a reasonable amount of time. In particular, the non-linearity of the fluid equations, particularly the Navier-Stokes equation, requires several simplifications, such as spatial and/or temporal integrations. The consequence is that numerical models can only predict what lies within their equations. Even when these equations are sufficient for the problem considered, the issue of parameterization can be arduous. As a consequence, the accuracy of the results obtained is sometimes questionable.

The most commonly used databases covering the French coasts include ANEMOC (***Benoit et al., 2008***, ***Laugel et al., 2014***), the databases from Ifremer's work for the IOWAGA and PREVIMER projects and, in particular, the HOMERE database (***Boudière et al., 2013***), and the BRGM's BOBWA-H database (***Charles et al., 2012***).

Today's best practice consists in combining the advantages of both measurements and numerical models by using the former to calibrate and validate the latter. For instance, satellites perform both wind measurements to be used first to calibrate the input wind fields over the generation area, and altimetry measurements to be used next to improve wave generation and propagation by the numerical models through assimilation and calibration. Pressure sensors, wave buoys, tide gauges, wind masts, etc., provide local information regarding either the forcing or the output of the models, allowing calibration at the site of interest. The complementarity of satellite measurements, which offer global coverage in space but seldom cover a precise spot, and local measurement devices, which can provide extensive coverage in time, should be noted. This approach was used in particular by Artelia and its technical and commercial partner GlobOcean to model sea states in the Gulf of Lion (Mediterranean Sea): the use of satellite measurements significantly improved the numerical modelling, particularly in the case of storm events (***Mazas et al., 2015***).

As a result, typical analyses in coastal engineering nowadays rely on databases of waves, winds, sea levels and/or currents, at a time step varying from 10 minutes (for hydrodynamic processes) to 3 hours, over 20 to 25 years. The phenomena can be split up into many components (see section 1.3.1), including astronomical tide, meteorological surge and eustatic rise for sea levels, and wind waves and primary to third swells for sea states. In some cases, the duration of the database may be extended to 50 or 60 years in order to account for decadal variability (e.g. the North-Atlantic oscillation).

1.2. METEO-OCEANIC EXTREMES IN ENGINEERING, RISK AND SOCIETY

1.2.1. Analyses for engineering

A coastal engineer in charge of determining meteo-oceanographic conditions on a project is expected to provide:

- **frequent conditions**, or **operational conditions**, that affect the daily operation of the facility: e.g. wave disturbance at mooring stations for downtime assessment, and wind speed or current velocity values for power production by offshore wind turbines or tidal current turbine farms;
- **extreme conditions**, or **design conditions**, that concern hazard assessments for designing structures: e.g. extreme wave heights for a breakwater, extreme still water levels for coastal flooding, etc.

The engineer has thus to deal with:

- a varying number of physical phenomena, that can themselves be described by a certain number of quantities in constant interaction;
- various geographical domains, each being characterized by predominant physical processes;
- variability at multiple spatial and temporal scales.

1.2.2. A simple definition of risk

Recent decades have seen exponential growth in the use of coastal areas by humans: population settlement, houses, buildings, facilities, etc. - in other words, **assets**, or **elements-at-risk**. These elements-at-risk are:

- **exposed** to **natural hazards**, i.e. natural processes or phenomena that may cause **damage** such as loss of life or injury, property damage, loss of livelihoods and services, social and economic disruption or environmental damage and, more precisely, exposed to meteo-oceanographic hazards;
- **vulnerable** to varying degrees to these hazards, i.e. they present characteristics that make them susceptible to their damaging effects.

The combination of **vulnerability** and **hazard** is the **risk** to which these **assets** are **exposed**. As developed by ***Breysse (2009)***: "*The risk therefore integrates two dimensions: that of hazards and that of loss, both being probabilised. Therefore, a risk is characterised by two components: the level of danger (likelihood of occurrence of a given event and intensity of the hazard), and the severity of the effects or consequences of the event that could have an impact on the assets.*"

Figure 5. Definition of risk
source: UN-SPIDER

1.2.3. Illustrative examples, at home

In western Europe and in particular in France, the risk related to meteo-oceanographic extremes has been rediscovered in recent times after a long period during which it somehow ceased to be a primary concern, or at least was considered as something that human industry and genius would soon be able to tackle once and for all.

Indeed, the period of stunning economic growth during the 1950s-1960s when the rush to the sea began was quite calm in that regard, as were the following decades, the 1970s to the 1990s. While the old villages of the Atlantic coasts had safely been settled upon ancient islets on elevated ground, or a few kilometres away from the coastline, by the elders, new cities and facilities began to sprout right behind (and sometimes over!) the dunes. No major storms occurred during this time and the memories of the elders faded away (***Sauzeau and Acerra, 2012***, ***Péret and Sauzeau, 2014***). Hazards turned out not to occur and vulnerability thus steadily increased.

A first warning shot across the bows did come, however, in December 1999. The two major storms Lothar and Martin hit the French coasts a single day apart. While the winds caused major damage across the country, disaster nearly struck when the wind waves generated over the Gironde estuary, in combination with a large storm surge (above 2 m in the estuary according to ***Salomon, 2002***), overtopped the dykes protecting the Blayais nuclear power plant and threatened to flood it (***Aelbrecht et al., 2004***). But an accident was avoided and in the collective memory the storms of 1999 would be remembered as meteorological events, just like the Great Storm of October 1987 (Figure 6). Coastal communities could keep building houses and camp sites on lowlands.

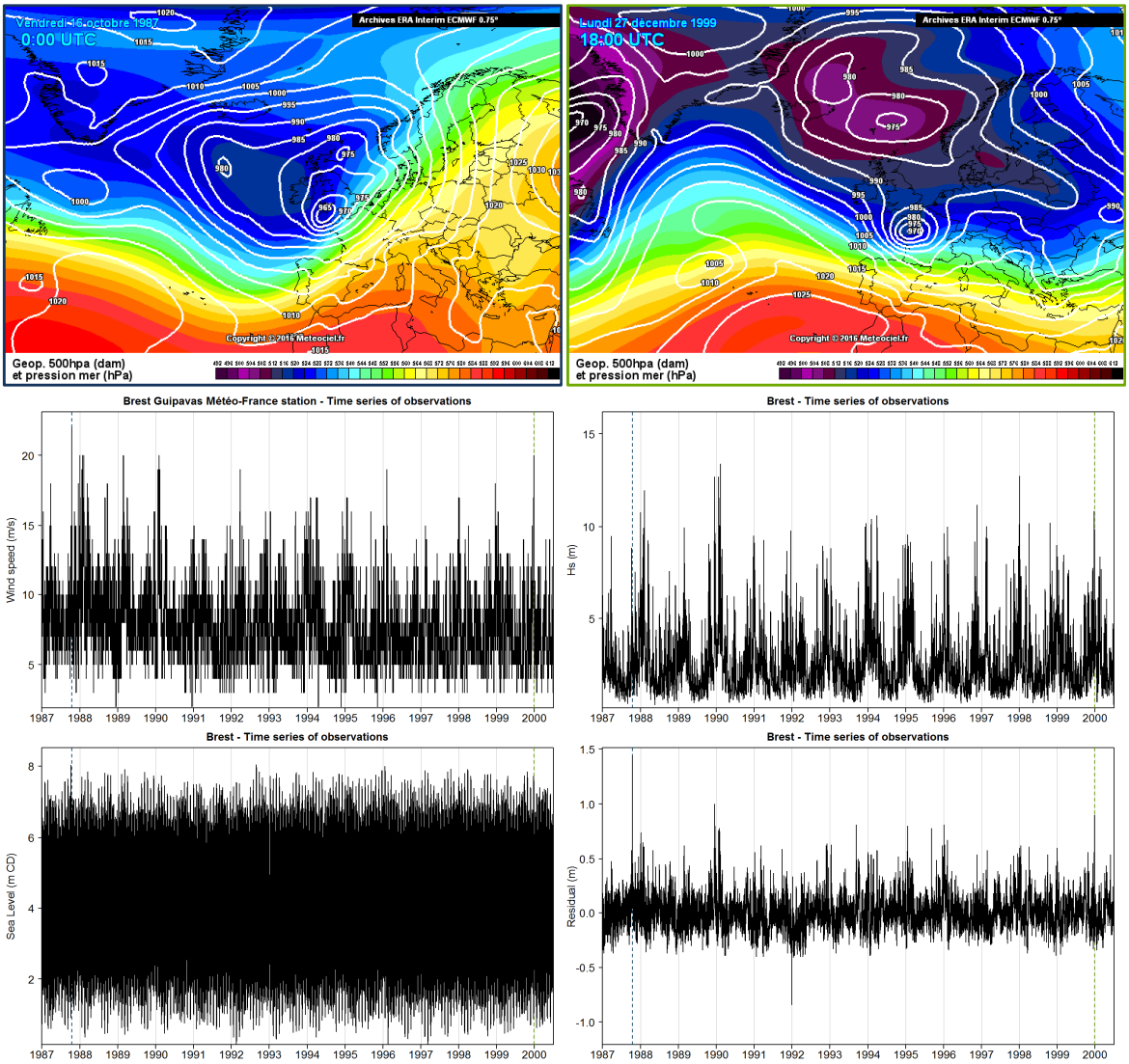

Figure 6. Storms of October 1987 and December 1999 (Martin). Top: 500 hPa geopotential height on 16/10/1987 0Z and 27/12/1999 18Z, middle: time series of wind speed (left) and wave height (right) at Brest, bottom: time series of sea level (left) and non-tidal residual (right) at Brest

Sources: ERA-Interim atmospheric reanalyses, wind measurements at Météo-France's weather station of Guipavas (29), sea state hindcast offshore Brest (48.5°N, 5°W) from Bertin et al. (2013) WW3 model, sea level measurements at Brest tide gauge (SHOM) and non-tidal residual at Brest tide gauge after removal of astronomical retro-predictions and correction of the eustatic trend (see Mazas et al., 2014)

Then storm Xynthia struck in February 2010. Its singular track, from Portugal to the Vendée coast of France, generated short, rough wind waves in the Bay of Biscay that favoured the development of a huge storm surge on the continental shelf (***Bertin et al., 2012***). This surge reached the shallow end of the Bay, known as the Pertuis Breton, in the darkest hour of a cold February night, just coinciding with the high water of a strong spring tide (Figure 7).

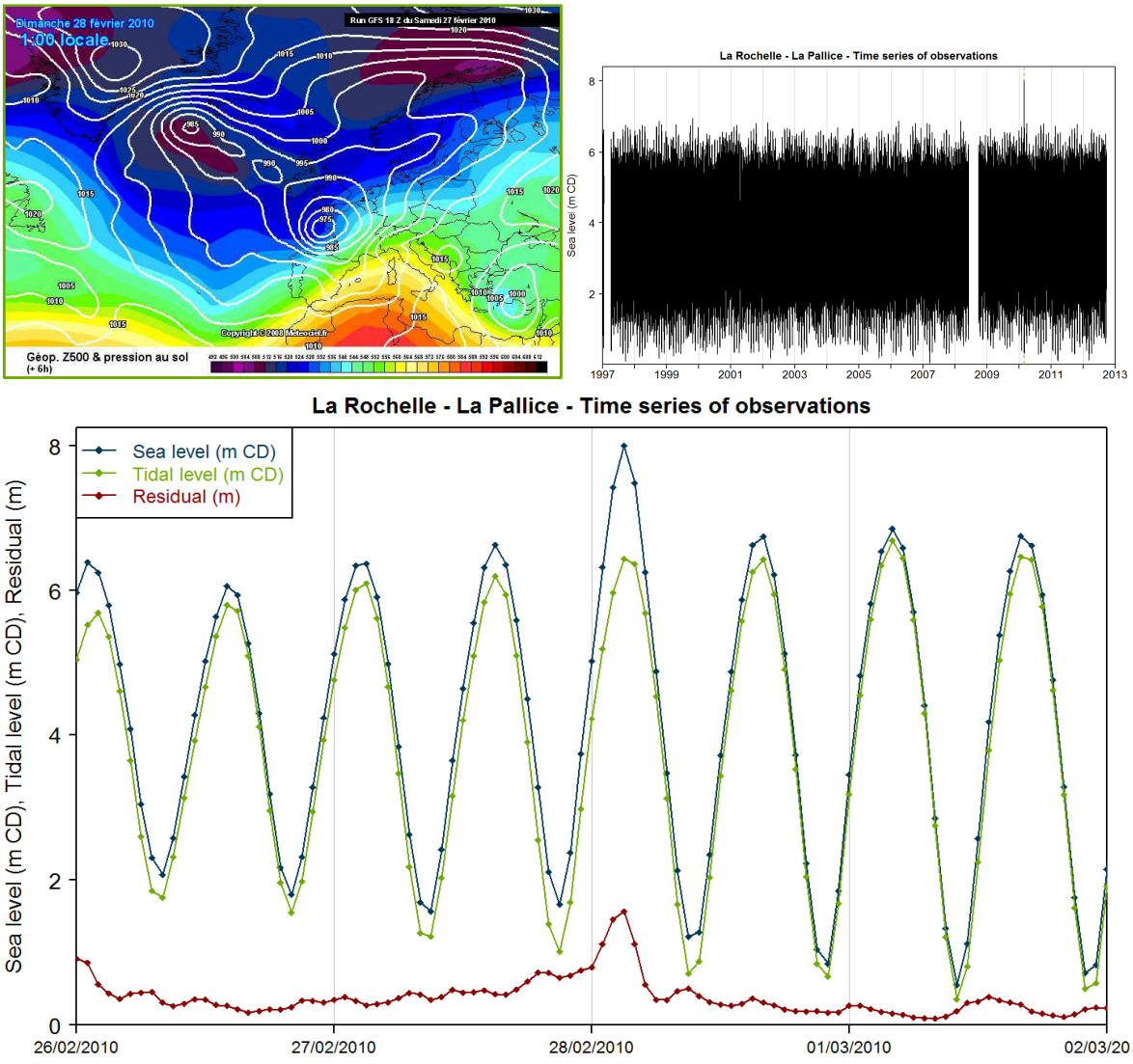

***Figure 7.* Storm Xynthia. Top: 500 hPa geopotential height on 28/02/2010 0Z (left), time series of sea level at La Rochelle-La Pallice tide gauge (right), bottom: time series of sea level, tidal level and non-tidal residual around Xynthia**

Sources: ERA-Interim atmospheric reanalyses, sea level measurements at La Rochelle-La Pallice tide gauge (SHOM), astronomical retro-predictions from SHOMAR software (SHOM) and non-tidal residual

29 people drowned in La-Faute-sur-Mer, in a lowland area surrounded by the Ocean and a coastal river protected by dykes that were submerged by the sea (Figure 8).

***Figure 8.* La Faute-sur-Mer, hours after the storm peak. In the foreground: the Lay coastal river; in the background: the Atlantic Ocean**

source: © Photo PQR/OuestFrance

This disaster was a brutal reminder of the coastal risks for French society. During the winter of 2013-2014, an exceptional succession of Atlantic storms caused dramatic coastal erosion in Biscay (***Masselink et al., 2016***). Scientists and historians have charted the evolution of society's perception of hazard and risk across the centuries. The long period of respite may have been linked to natural decadal climate variability, or simply a result of mere luck, but the risk related to meteo-oceanographic extremes had been forgotten by society, because the hazard simply did not materialise (***Garnier and Surville, 2010***).

Things had been different in the Netherlands. The coastal flooding of 1953 caused by the same joint occurrence of spring tide and storm surge triggered a gigantic plan to protect land against sea flooding, the Delta Plan. A culture of risk had been deeply embedded in Dutch society (***Stive, 2012***).

Today, fresh memories of these events and, in parallel, ever-growing concerns over the sea level rise induced by climate change are bringing about a change in mentalities (***IPCC, 2012***). Old habits are deeply rooted but it is becoming ever more clear to ever more people that the coastal strip is a place where human settlements and activities cannot be taken for granted as previously imagined. New approaches and strategies are being considered and applied, sometimes painfully (***Cousin, 2011***).

In this context, the assessment of extreme meteo-oceanographic events has become more than a scientific challenge: it is now a societal issue (***Sauzeau, 2011***).

1.3. PHYSICS AND STATISTICS: A MATTER OF TERMINOLOGY

1.3.1. Physical definitions and... non-definitions

As will be shown subsequently, statistical methods (or, to be more precise, probability methods) are used to estimate extreme values of physical quantities. It is hence necessary to define some basic concepts that take a different meaning in these two fields.

A **phenomenon**, from the Greek *phainómenon*, itself from the verb *phainein*, "to show, shine, appear, to be manifest or manifest itself", can be defined as any thing which manifests itself. In science, it may be described as a system of information related to matter, energy or spacetime. In other words, a **physical phenomenon** is a natural phenomenon that involves the physical properties of matter and energy.

Wind: a flow of gases; currents: a flow of liquid; sea state: fluctuations of free surface in a certain range of frequencies caused by the wind; astronomical tide: fluctuations of free surface at low-frequency induced by astronomical forcing; these are examples of physical phenomena.

The International Vocabulary of Metrology (VIM) defined by the Joint Committee for Guides in Metrology (***JCGM, 2012***) specifies that a **physical quantity** is a "*property of a phenomenon, body or substance, where the property has a magnitude that can be expressed as a number and a reference*". In other words, a property that can be quantified by measurement. Wind speed, current direction, spectral significant wave height, peak period and atmospheric pressure are thus examples of physical quantities.

Interestingly, the VIM also specifies that "*in some definitions, the use of non-defined concepts (also called "primitives") is unavoidable. In this vocabulary, such non-defined concepts include: system, component, phenomenon, body, substance, property, reference…*"

Things hence become harder because we also have to consider *phenomena* made of several *components* that happen on the same *body*: three concepts that are non-defined in this system of definitions…

For instance, sea level fluctuations are the result of the superposition of many **components** associated with distinct physical phenomena: long-term variations in mean sea level (eustatism, isostasy, seasonal heat variation, etc.), astronomical tide, meteorological surge, low-frequency waves, tsunamis, short waves. Each component corresponds to a particular phenomenon and can be described by physical quantities such as height and period (for periodic phenomena); their sum can be described by the physical quantity "sea level", i.e. the vertical position of the upper limit of a physical body: ocean, sea or lake. Following the VIM, we can say that these components are all **of the same kind**, the **kind of quantity** being the aspect common to mutually comparable quantities.

It should also be stressed that a sea state (a general phenomenon described by physical quantities such as the spectro-angular density and the associated reduced parameters) can be split up into different wave systems (wind sea, swells), each of which corresponds to a more specific phenomenon. Here again, the spectro-angular densities of a wind sea and a swell are of the same kind and can be summed.

However, when considering waves propagating on a vein of strong tidal current, these two phenomena are not of the same kind and cannot be considered as components of a general phenomenon[2].

1.3.2. Statistics: probabilities of… what exactly?

In probability and statistics, there is generally no such distinction between phenomena, quantities, bodies, etc. A **random variable** can be quite simply defined as "*a variable quantity whose possible values are numerical outcomes of a random phenomenon*" (***Blitzstein and Hwang, 2014***), and a **covariate** is a variable that may be predictive of the outcome being studied.

Random is itself a lack of predictability in these outcomes, be it by nature (quantum phenomena) or due to a lack of information (see Laplace's demon). Let us consider a random experiment that leads to the realization of a single outcome among a set of possible outcomes Ω (called the universe). A subset A of Ω is called an **event** and its **probability** is a measure of the likelihood that

[2] Strictly speaking, the mathematicians who unified these two processes in a same equation within the theory of free surface gravity waves could dispute this assertion for this particular example…

it will occur, defined as a weighting of this event relative to Ω. The axioms proposed by Kolmogorov (positivity, unit mass, additivity) allowed a mathematical theory to be built on that concept.

The different interpretations of this measure of likelihood are a broad topic in themselves. At this stage, we simply need to mention the two main categories of probability interpretation:

- the **frequentist interpretation** considers that this concept describes some objective or physical state of affairs and that the probability of an event is its relative frequency of occurrence over time;
- the **subjective interpretation** views probability as a measure of a "degree of belief" of the individual assessing the uncertainty of a particular situation; in particular, Bayesian probability is based on such an interpretation.

The probabilities of occurrence of the different possible outcomes (events) of a random experiment can be provided by a mathematical function called a **probability distribution**. In other words, a probability distribution is description of a random phenomenon in terms of the probabilities of an event. Probability distributions of continuous random variables can be described in several ways, in particular the **probability density function** f that describes the infinitesimal probability of any given value (probability of occurrence) and the **cumulative distribution function** F that describes the probability that the random variable is no larger than a given value (probability of non-exceedance). The probability of exceedance $p = 1 - F$ is also often convenient.

Another useful definition is as follows: a collection of random variables indexed by a set of numbers, such as points in time, is called a **stochastic process**. In our case, we can consider a stochastic process as the variation of a random variable over time, as opposed to a **deterministic process** in which random is totally excluded: the outcome of the variable can be determined if a finite number of conditions are known. In particular, stochastic processes include **Poisson point processes**, a counting process that represents the random number of points or events up to a given time. In meteo-oceanography, the astronomic tide is considered to be a deterministic process because we have very good knowledge of the conditions, and hence of the outcome. This is a typically Newtonian situation in which Laplace's demon would be almost omniscient. In contrast, atmospheric events require such a huge amount of prior knowledge that they very quickly become fully stochastic: a surge is handled as a stochastic variable.

Yet ambiguous or ill-defined concepts are not alien to probabilities. A key figure of extreme value studies is the **return period** $\mathcal{T}$, first introduced by ***Fuller (1914)*** (in hydrology) more than a century ago. Still it continues to be particularly prone to misleading interpretations, and this is particularly damaging because it is widely used to popularize the results of studies and help the public understand[3].

When speaking with stakeholders, inhabitants or even fellow engineers who are not probability theory specialists, the return period is generally heard as "the average period between two occurrences" of the event (***Fleming et al., 2002***). So, taking a record covering 3,000 years, the 100-year quantile will be observed or exceeded roughly 30 times. This is not incorrect, but in most cases it is completely useless. We do not have such long records and, furthermore, the phenomenon is certainly not stationary over such a long period. The return period must hence be understood as a *probability of exceedance* and, more precisely, a *yearly probability of exceedance*: every year, the probability that the $\mathcal{T}$-year value is reached or exceeded at least once is $1/\mathcal{T}$. This implies that several events may occur in the same year!

In the univariate and stationary case, considering a random variable X, we can easily link the return period $\mathcal{T}$, the cumulative distribution function F_X or its complement, the probability of exceedance p, and either the mean number of occurrences per year λ or its reciprocal, the average inter-arrival time between two realizations of the process μ:

[3] See for instance the public hearing at the French Senate following Xynthia (in French): http://www.senat.fr/rap/r09-647-2/r09-647-28.html

$$\mathcal{T} = \begin{cases} \dfrac{\mu}{\mathbb{P}[X > x]} = \dfrac{\mu}{p_X} = \dfrac{\mu}{1 - F(x)} \\ = \dfrac{1}{\lambda \mathbb{P}[X > x]} = \dfrac{1}{\lambda p_X} = \dfrac{1}{\lambda[1 - F(x)]} \end{cases} \tag{2}$$

The return period is clearly a frequentist interpretation of the probability of an "event". The definition of Equation (2) can even be generalized by introducing this probability $\mathbb{P}_A$:

$$\mathcal{T} = \frac{\mu}{\mathbb{P}_A} = \frac{1}{\lambda \mathbb{P}_A} \tag{3}$$

A second mistake consists in forgetting the yearly aspect of this probability and the cumulative effect. When inhabitants are told after a storm that wreaked havoc on their houses that its return period was 50 years, not only will they feel safe for the next decades, but they may also consider that a probability of occurrence of 2% is not something to worry about. But if they plan to live in their house for many years to come, this yearly probability is to be accounted for every year: this cumulative effect is described by the **encounter probability** p_E associated with $\mathcal{T}$ that depends on the duration or lifetime D.

$$p_E = 1 - \left(1 - \frac{1}{\mathcal{T}}\right)^D \tag{4}$$

The encounter probability p_E is much larger; if the aforementioned inhabitants stay 25 years, the probability of encountering a 50-year storm rises to 40% (Table 1). Last, this cumulative effect is also to be considered within one single year: several events can occur during the same year.

Table 1 – Relationships between return period, lifetime and encounter probability

		Lifetime D (years)						
		1	**5**	**10**	**25**	**50**	**75**	**100**
Return period $\mathcal{T}$ (years)	**5**	0.20	0.67	0.89	1.00	1.00	1.00	1.00
	10	0.10	0.41	0.65	0.93	0.99	1.00	1.00
	50	0.02	0.10	0.18	0.40	0.64	0.78	0.87
	100	0.01	0.05	0.10	0.22	0.39	0.53	0.63
	200	0.01	0.02	0.05	0.12	0.22	0.31	0.39
	500	0.00	0.01	0.02	0.05	0.10	0.14	0.18
	1000	0.00	0.00	0.01	0.02	0.05	0.07	0.10

When these two properties are understood, the return period seems to be unequivocal: the probability of an event exceeding a value over a given duration. However, we will see later that it becomes highly ambiguous in the multivariate case.

But what is this event, this subset of all the possible outcomes of a random experiment? How can we interpret this concept? Do we have a choice to make that will affect the meaning of the return period?

1.3.3. A first approach to events: etymology and definitions

The following sections will explain how the event concept arose as the core principle in the research carried out to refine the estimation of extreme meteo-oceanic conditions. But first of all, it is necessary to discuss the meaning of this word beyond the cold probabilistic definition.

In such a case, it is always wise to go back to the etymology. According to the Online Etymology Dictionary, the English word *event* comes from the Middle French *event*, itself from the Latin

eventus "occurrence, accident, event, fortune, fate, lot, issue", itself from past participle stem of *evenire* "to come out, happen, result", itself from the assimilated form of *ex-* "out" + *venire* "to come".

It is worth noting first that two meanings appear: either what happens, or what results from what has happened (as in *eventually*). The latter sense is mentioned to have appeared in English in the 1570s, as "the consequence of anything", just prior to the former meaning in the 1580s, "that which happens".

The dictionary of the Académie française[4] agrees on the etymology and also on the two meanings. Although it is in French, it is worth being quoted:

ÉVÈNEMENT ou ÉVÉNEMENT n. m. XVe siècle.

Dérivé savant, sur le modèle d'avènement, du latin evenire, « sortir, se produire », de venire, « venir ».
1. Vieilli. Issue, conséquence bonne ou mauvaise d'une action ou d'une situation. S'emploie encore dans quelques expressions
2. Ce qui survient, ce qui arrive, en un temps et en un lieu déterminés. PHYS. Tout phénomène se produisant en un point et à un instant donnés. - MATH. En calcul de probabilités, résultat éventuel d'un tirage au sort, d'un jeu de hasard, d'un pronostic, etc.

It can be translated as follows:

1. (Old) Issue, outcome, bad or good consequence of an action or situation. Is still used in a few expressions.
2. What occurs or happens at a given time and a given place. PHYS. Any phenomenon occurring at a given point and a given time. MATH. In probabilities, result of a random draw, a game of chance, a forecast, etc.

In the famous Encyclopaedia of Diderot and d'Alembert, Louis-Jacques Goussier[5] wrote the following definition in 1756:

S. m. (Grammaire) terme par lequel on désigne, ou la production, ou la fin, ou quelque circonstance remarquable et déterminée dans la durée de toutes les choses contingentes. Mais peut-être ce terme est-il un des radicaux de la langue : et servant à définir les autres termes, ne se peut-il définir lui-même ? Voyez l'article DICTIONNAIRE. Voyez aussi à l'article ENCYCLOPEDIE, la manière de fixer la notion des termes radicaux.
ÉVENEMENT, eventus, (Médecine) ; ce terme est employé pour signifier la fin d'une maladie, l'issue qu'elle a, bonne ou mauvaise.

Once again, the distinct meanings of the outcome of "something" or this "something" itself appear. But the second one is interestingly put. First, the event is defined as any "remarkable" circumstance; it depends on "all contingent things", i.e. things that can occur or not, or may depend on chance; and it is determined "over time". This is somewhat different from the previous definitions based on a fact that happens at a given time and place. Second, the author raises the possibility

[4] French Academy: the pre-eminent French council for matters pertaining to the French language.

[5] Louis-Jacques Goussier, 1722-1799, French illustrator and encyclopaedist.

that it is a radical of the language itself, so necessary for defining the other words that it is hardly definable itself.

It is worth noting that the core concept of the work presented here has been acknowledged to be almost un-definable in one of the major works of the Age of Enlightenment!

In addition, ISO/Guide 73:2009 provides a vocabulary relative to risk management with definitions of events that are rather more arid than those given above, but interesting in terms of their approach to risk. If risk is defined as the *"effect of uncertainty on objectives"*, the notes below specify that *"risk is often characterized with reference to potential events and consequences, or a combination of these"* and that *"risk is often expressed in terms of a combination of the consequences of an event (including changes in circumstances) and the associated likelihood of occurrence"*. An event is then defined as an *"occurrence or change of a particular set of circumstances"*, with the following precisions: an event *"can be one or more occurrences, and can have several causes"* and *"can consist of something not happening"*, and *"can sometimes be referred to as an 'incident' or 'accident'"*. In modern times, the meaning of "something occurring" is definitely accepted and the oldest sense of the outcome or consequence no longer holds.

In this spirit, this thesis proposes a better understanding of meteo-oceanic events, and more generally environmental events, which is summarised in particular in sections 2.4 and 3.2.1.

1.4. BRIEF DESCRIPTION OF PUBLICATIONS

This section provides a very brief description of the publications of the last ten years. It gives the chronology and the material context for each topic of research; the interest of these topics is described in detail in upcoming sections 2 and 3.

1.4.1. A multi-distribution adaptation of the existing POT framework

During a first three-month internship in 2007, the methodology for determining extreme wave heights used at ARTELIA (SOGREAH at that time) was updated in line with the international state of the art, as described in section 2.1.

This work earned the prize for best scientific internship of the Ecole Nationale des Ponts et Chaussées, awarded by Ponts Alliance, and was presented at:

[A] the *X*^{èmes} *Journées Nationales Génie Civil Génie Côtier* at Sophia-Antipolis in 2008.

This work was resumed after I was hired by SOGREAH in 2009. The multi-distribution methodology was made more sophisticated, applied to case studies and published:

[B] a paper in the *European Journal of Environmental and Civil Engineering* in 2010 for a case study in Tangiers;

[C] a paper in two parts in *La Houille Blanche* in 2010 (part 1: theory, part 2: application);

[D] a presentation at the WISE (Waves In Shallow water Environment) conference in Brest in 2010;

[E] a paper in *Coastal Engineering* in 2011;

[F] a reply to discussion about this last paper in *Coastal Engineering*.

The paper in *Coastal Engineering* is reproduced as part of the present statement.

1.4.2. A two-step framework for over-threshold modelling

Following publication of the updated methodology in *Coastal Engineering* in 2011, a working group was set up between ARTELIA (Luc Hamm, Franck Mazas), the LHSV[6] (Michel Benoit), EDF R&D (Pietro Bernardara, Marc Andreewsky, Jérôme Weiss) and the CEREMA/DTecEMF, then CETMEF (Xavier Kergadallan). It was later christened OSSË (*Ocean and Sea Statistics for Extremes*) and was a privileged place for in-depth reflection on and discussion of meteo-oceanic extremes. At the beginning of his PhD, Roberto Frau also joined the group and became a valuable contributor.

Figure 9. Logo of the OSSË group

The double threshold introduced in the methodology of univariate extremes was identified as a topic that could be further discussed and justified. Dr Pietro Bernardara and I were the leaders on that topic and we refined and strengthened the approach that was presented in two conferences and then published in a peer-reviewed journal with the broadest possible scope in the field of environmental sciences:

[G] a presentation at the ICCE (International Conference on Coastal Engineering) at Santander in 2012;

[H] a presentation at the EGU (European Geosciences Union General Assembly) in Vienna in 2013;

[I] a paper in *Natural Hazards and Earth System Sciences* in 2014.

The paper in *Natural Hazard and Earth System Sciences* is reproduced as part of the present statement.

1.4.3. Maximum Likelihood Estimator and its *virgae*

In parallel, the use of the Maximum Likelihood Estimator for over-threshold modelling was closely examined. The refinement in the sensitivity study to determine the threshold that had been allowed by the double-threshold approach had made strange patterns appear in the estimation of the distribution of the parameters and hence of the quantiles.

A question that was asked by Belgian scientists at the WISE congress in Brest in 2010 made me take a close look at that topic and its relationship with the threshold. A collaboration began with Dr Philippe Garat from the Jean-Kuntzmann Laboratory at the University of Grenoble-Alpes, back then called the University Pierre-Mendès-France.

This work led to the following publications:

[J] a presentation at the EVA (Extreme Value Analysis) conference in Lyons in 2011;

[K] a presentation at the ICCE (International Conference on Coastal Engineering) in Santander in 2012;

[L] a paper in *Ocean Engineering* in 2014.

The paper in *Ocean Engineering* is reproduced as part of the present statement.

[6] Laboratoire d'Hydraulique de Saint-Venant, Université Paris-Est

1.4.4. Extreme sea levels: a first approach to bivariate analysis

In February 2010, the surge generated by storm Xynthia occurred quasi-simultaneously with the high water of a spring tide (tide coefficient of 103[7], corresponding to a tidal range of 5.86 m and a high water level of + 6.49 m CD). This caused widespread coastal flooding in the *pertuis charentais*, along the French coast of the Bay of Biscay. The need for an accurate estimation of extreme sea levels was abruptly highlighted.

We applied the results established in the univariate case to the so-called Joint Probability Method. The site of Brest was chosen as a case study to be presented in the publications because of the length of the dataset of sea level measurements there, but the method was applied to sea level measurement at La Rochelle to estimate the return period associated with Xynthia for an appraisal ordered by the district court of Les Sables-d'Olonne.

The methodology was presented in the following publications:

[M] a first paper presenting an overview of the state of the art in *La Houille Blanche* in 2011;

[N] a presentation at the Journées REFMAR conference in St-Mandé (France) in 2013;

[O] a presentation at the EGU (European Geosciences Union General Assembly) in Vienna in 2013;

[P] a paper in *Coastal Engineering* in 2014.

The paper in *Coastal Engineering* is reproduced as part of the present statement.

1.4.5. Joint occurrence of extreme waves and sea levels: from bivariate to multivariate

During coastal flooding events, waves often play an important role by adding an extra component at the coastline related to their breaking: wave set-up. Waves can also influence the surge through complex mechanisms. Moreover, many coastal structures such as breakwaters, seawalls and quays are very sensitive to the joint effect of waves and sea levels that results in overtopping or structural damage.

Studying the joint occurrence of waves and sea levels was hence the logical next step and an intern, Vincent Auger, was hired in 2014 for six months under the joint supervision of Dr Luc Hamm and myself to extend our methodologies to this case. The internship benefited from the help and advice of Pr. Clémentine Prieur (University of Grenoble-Alpes – LJK) and Pr. Anne-Catherine Favre (ENSE3/INPG, LTHE). Pr. Peter Hawkes (HR Wallingford) also kindly provided assistance.

The work carried out during the internship was then extended and continued over the couple of years that followed.

As a result, the methodologies were presented in the following publications:

[Q] a presentation at the EVAN conference (International Conference on Advances in Extreme Value Analysis and Application to Natural Hazards) in Santander in 2015;

[R] a presentation at the ICCE (International Conference on Coastal Engineering) in Istanbul in 2016;

[7] In France, the amplitude of the tidal range is expressed in terms of tidal coefficients, introduced by Laplace in 1799. The coefficient can take values from 20 (lowest astronomical tidal range) to 120 (highest astronomical tidal range). The mean spring tides, respectively mean neap tides, correspond to coefficients of 95, respectively 45. The mean tidal range corresponds to a coefficient of 70.

[S] a paper in *Coastal Engineering* in 2017;

[T] a submission (accepted) for the ICE conference to be held in Liverpool in 2017;

[U] a submission for the EVAN conference to be held in Southampton in 2017.

The paper in *Coastal Engineering* is reproduced as part of the present statement.

oOo

2. FROM STORM PEAKS TO EXTREME UNIVARIATE EVENTS

2.1. A MULTI-DISTRIBUTION ADAPTATION OF THE EXISTING POT FRAMEWORK

The works presented in section 1.4.1 lasted from 2007 to 2014. They began with a simple updating of a methodology for determining extreme wave heights to bring it in line with the state of the art.

The methodology used at the time by SOGREAH consisted of the following four steps, as per the recommendations of the IAHR Working Group (***Mathiesen et al., 1994***):

- processing of the time series and identification of directional wave sectors;
- selection of storm peaks using the Peaks-Over-Threshold (POT) approach;
- fitting of a Weibull distribution to the peaks using the least square method;
- computation of quantiles (extreme wave heights).

This approach was updated by incorporating the Extreme Value Theory (EVT) described in detail in ***Coles (2001)***. At the end of the internship, the following methodological improvements had been implemented:

- introduction of the Generalized Pareto Distribution (GPD);
- use of the Maximum Likelihood Estimator (MLE);
- extension to a multi-distribution framework by considering other distributions such as Weibull and Gamma/Pearson-III;
- goodness-of-fit assessment using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

These improvements did not significantly alter the general framework of the methodology. In particular, the POT approach was kept, as it is particularly well suited to the maritime field where the number of significant storms per year is generally large enough for it to be deemed preferable to the annual maxima method (***Cunnane, 1973***). However, almost every step of the methodology was closely examined and new ideas arose from the discussions with fellow scientists and engineers. It must also be said that the dual approach of academic research on one hand and everyday coastal engineering on the other hand was particularly fruitful for setting up methods capable of blending the rigor of the theory and the flexibility required to deal with real-world projects.

This dual approach appeared right at the beginning of the work. On one side, the literature had already been proposing many very rigorous references to the Extreme Value Theory for a number of years. The GPD-Poisson model in particular was justified and detailed. On the other side, the literature more specific to the coastal community included the IAHR Working Group and the works by Pr. Yoshimi Gōda (***Gōda, 1988***, ***Gōda and Kobune, 1990***). Decades of experience had taught the engineers to keep several options at hand to deal with wave datasets, which sometimes present strange behaviours (***Gōda, 2011***). In particular, the need to deal with homogeneous populations, or identically distributed datasets, was highlighted through the recommendation to consider wave directional sectors, associated with distinct fetches, wave ages and / or meteorological phenomena. Figure 10 illustrates a case study off Bastia, where it appears clearly that waves from the NE or SSW are limited by a short fetch compared with sea states from the SE.

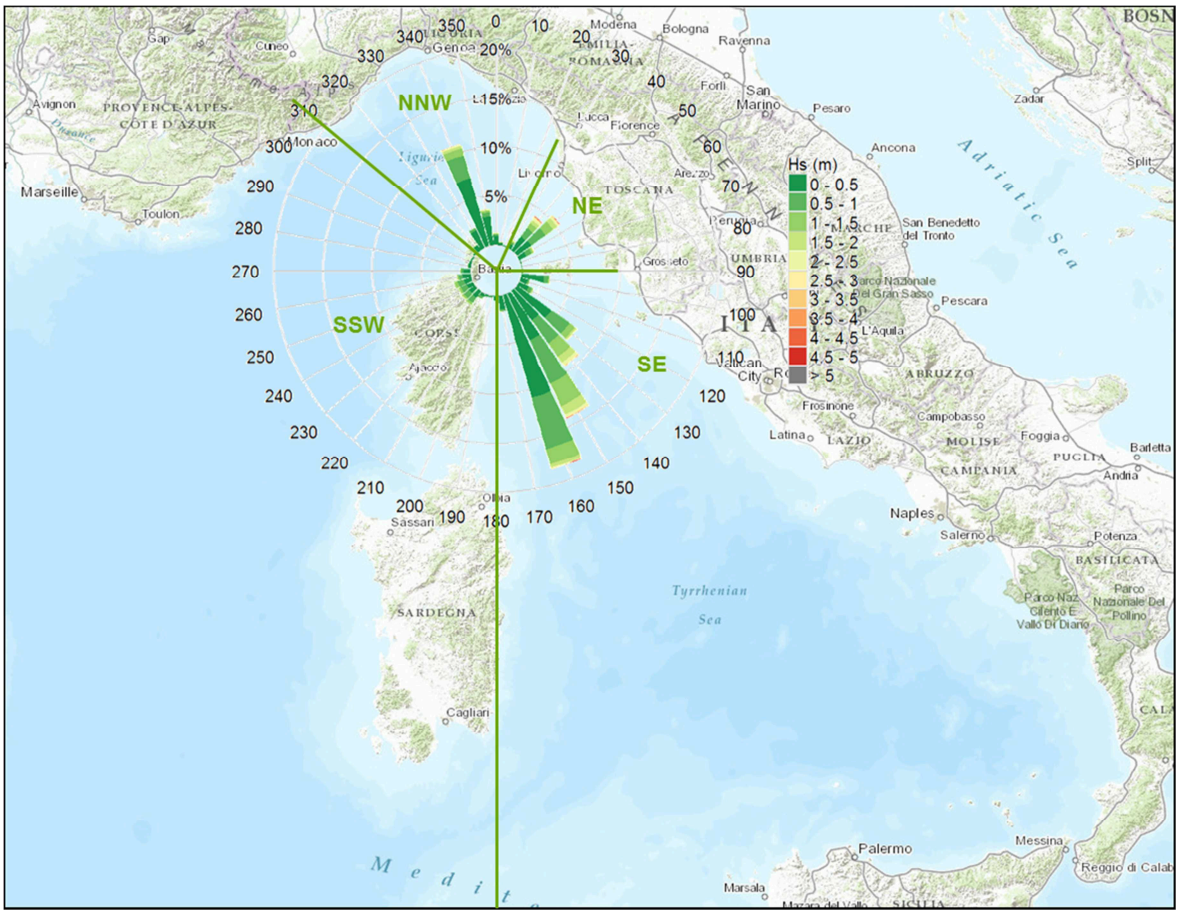

Figure 10. Identification of homogeneous wave populations through the use of directional sectors off Bastia

Another meaningful example takes place at Réunion island, where three very distinct wave populations can be observed (Figure 11):

- wind waves generated by trade winds from the south-east; these are very regular winds that cause very regular and gentle waves;
- southerly swell generated in the roaring fifties by powerful storms and that propagate for thousands of kilometres in very marked wave groups with very large periods and very small steepness and directional spreading;
- cyclonic waves generated by tropical cyclones in the vicinity of the island, with a very large steepness and directional spreading.

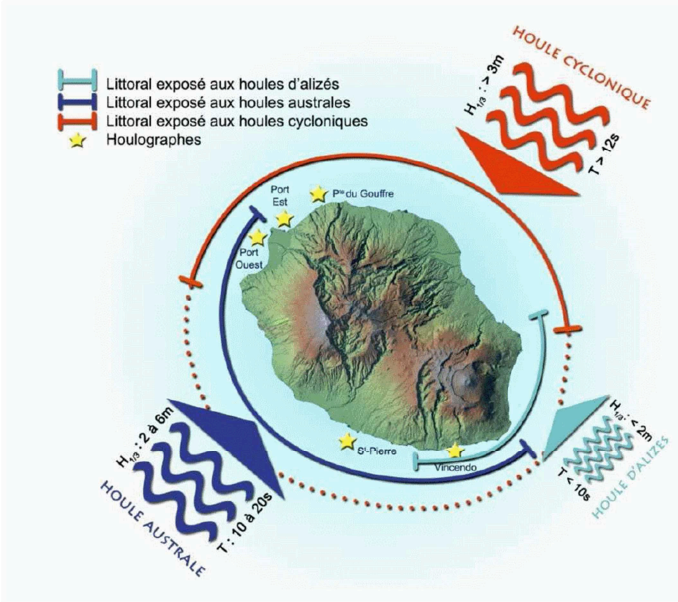

Figure 11. Wave populations at Réunion island

source: Météo-France

The diversity in the meteorological phenomena (trade winds, mid-latitude lows, tropical cyclones) and in the wave characteristics (spreading, steepness) is such (Figure 12) that the assumption of identical distribution cannot hold: these populations must be analysed separately (it is always possible to eventually combine the result of the analyses).

Figure 12. Wave populations at Réunion island. Top: southerly swell in July 2017; bottom: cyclonic waves from tropical storm Chezda in January 2015

source: Cyril Giraudel, http://reunion-extreme.re

Regarding the choice of threshold, we attempted to reconcile:

- the engineer-based recommendations of the IAHR Working Group: choose the threshold on the basis of physical and meteorological information (e.g. using weather charts) and set its value so as to obtain a mean number of peaks per year corresponding to the average number of (significant) storm events per year;
- the tools based on the EVT provided by ***Coles (2001)***, such as the mean excess life plot and the stability plots of the GPD parameters.

The latter approach requires sensitivity studies to be performed on the threshold, which creates a practical difficulty: environmental time series often present fluctuations, and a sequence of large wave heights that are known to have been generated by a single storm can present secondary peaks that are not independent of the largest peak, as illustrated in Figure 13 below.

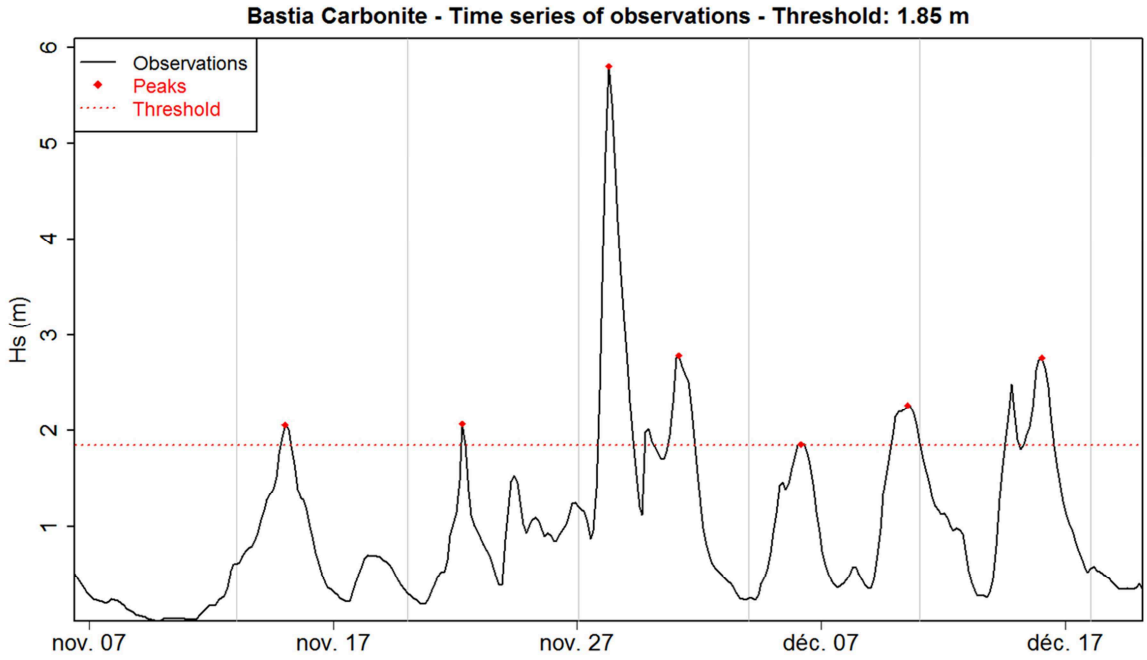

Figure 13. Fluctuations of the time series of significant wave height

It can be clearly seen in this illustration that if the threshold varies between 1 and 3 m, some of the storms that can be easily identified visually (such as the last one) will exhibit first one threshold exceedance, then two, then another, and will eventually fall below the threshold. If we wish to select independent peaks as required by the EVT, sub-parameters have to be set, such as minimal duration between two peaks, minimal duration between the end of a storm and the beginning of the next one, minimal storm duration, or the value falling back below a fraction of the threshold. All these sub-parameters, which may be combined, aim to ensure that peaks are independent, but most of them also have in common the fact that their value depends on the threshold value. So if we wish to carry out a broad sensitivity study on the threshold, we should constantly adapt these parameters, which is an obvious practical difficulty.

It thus seemed quite convenient to propose an approach that could at the same time account for the recommendations of both the engineers and the statisticians and avoid the fluctuation problem: the double threshold. A first threshold u_1 is used to decluster the time series, identify storms and extract their peaks. The sub-parameters ensuring that the selected storms are independent can be set once and for all, accounting for the oceano-meteorological characteristics of the site. This step yields a small i.i.d. sample on which a sensitivity study can be carried out in order to determine a second threshold u_2. A statistical distribution can then be fitted to the peak exceedances over u_2.

In the paper of 2011, we presented this approach with additional recommendations based on the mean number of peaks per year. Based on our experience, setting u_1 so as to obtain 5 to 10 storms per year and u_2 so as to obtain 2 to 5 storms per year seemed reasonable, in particular in mid-latitude areas. Of course it had to be adapted to the local wave climate, particularly in cyclonic areas. The plots of the stability of the GPD shape and modified scale parameters with regard to the threshold, proposed by Coles, were adapted in order to include the variation in sample size when u_2 increases (Figure 14). Lastly, we developed the logic of the plot by suggesting identifying "domains of stability", i.e. threshold ranges within which the parameters do not vary, and selecting the lowest threshold of the highest domain: since the GPD is the asymptotic distribution towards which the law of the exceedances over a threshold tends when u is large, choosing the highest

domain should minimize the bias; then, within this domain, selecting the lower bound should not change the estimation (since the parameters do not vary). It should minimize the variance and uncertainties, however, since it maximizes the sample size. This is a classic case of the bias-variance dilemma.

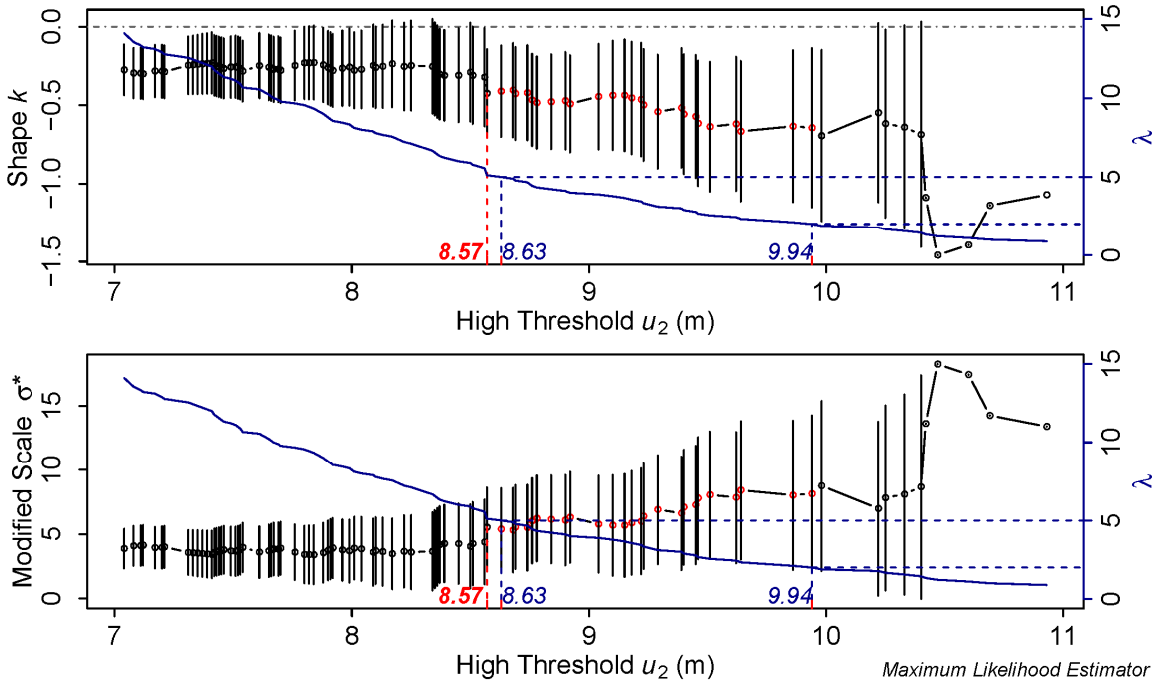

Figure 14. Stability plot for choosing the high threshold proposed in Mazas and Hamm (2011)

These plots are a useful tool for choosing the statistical threshold, but they are to be used in conjunction with other tools (evolution with respect to the threshold of the extreme quantiles, of goodness-of-fit criteria…) discussed in particular in ***Bernardara et al. (2014)***, as presented in section 2.2 below. Nevertheless, it should not be forgotten that the choice of the threshold remains somehow arbitrary, and choosing a threshold means choosing a model. In that sense, considering a wide number of distributions and criteria is better than relying on a single statistical optimization (goodness-of-fit between a sample and one pre-chosen statistical distribution) that potentially can introduce bias. This is why distributions such as the Weibull, Gamma or exponential distributions are considered in addition to the theoretical law, namely the GPD (Figure 15).

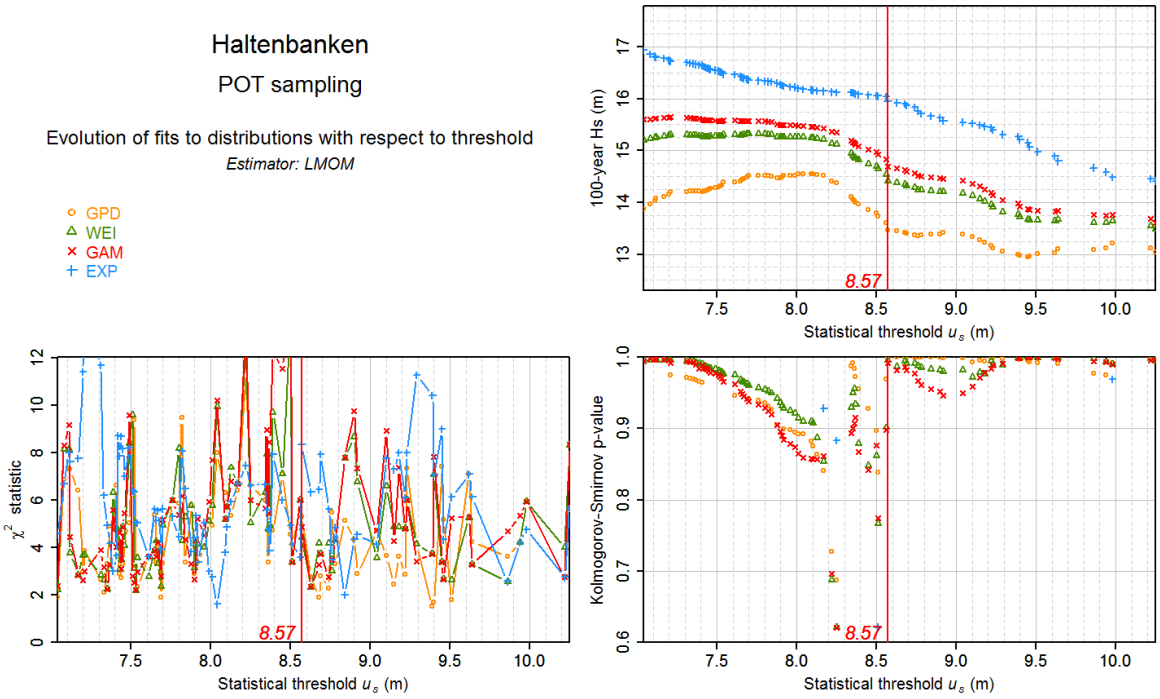

Figure 15. Plots showing the evolution of the fits to the GPD, Weibull, Gamma and exponential distributions with respect to the threshold: 100-yr Hs quantile (top right), Chi2 statistic (bottom left), p-value of the Kolmogorov-Smirnov test (bottom right)

Lastly, it was proposed to compute the confidence intervals using the parametric bootstrap method rather than the conventional delta method. This latter method depends on an assumption that the estimated parameters have a normal distribution, which yields symmetrical confidence intervals around the most probable value of the quantiles, and in some cases a lower bound for the interval that decreases as the return period increases. The parametric bootstrap method eliminates this Gaussian assumption and provides more credible, asymmetrical confidence intervals for the extreme quantiles.

2.2. A TWO-STEP FRAMEWORK FOR OVER-THRESHOLD MODELLING

The multi-distribution model presented above, and in particular this double threshold approach, seemed very convenient, and it appeared to be rather popular in the literature in the following years. However, at the time it was merely considered practical, with no conceptual consequences.

Nevertheless, discussions within the OSSË working group and other engineers and researchers led to work focusing on that double-threshold approach. The related concepts were also examined.

First of all, it was obvious that this approach was not restricted to wave heights, or to the meteo-oceanic field, but could be applied to hydrology and to all kind of environmental data.

Secondly, an in-depth literature review showed that the choice of threshold was based sometimes on physical arguments, sometimes on statistical ones, and sometimes on both (see for instance the paper of ***Lang (1999)*** for a literature review). The need for clarification became obvious.

We based our reasoning on the observation that the analyst deals with a time series of observations of the variable (from measurements or modelling) at a given time step, while the conventional tools provided by the EVT assume that the dataset is independent and identically distributed (i.i.d.). Consequently, a step is needed to go from the former to the latter. This step is a

declustering process (***Smith, 1989***). Once the i.i.d. sample has been set up, a new step of statistical optimization can begin in order to find the best statistical model to fit to the data.

At this point, the solution becomes clear: independent and identically distributed peaks must be extracted from the original time series using physical considerations, and statistical tools may then be used in order to determine the optimal threshold above which a statistical law is fitted to the exceedances.

The first step thus consists in extracting i.i.d. peaks from the time series. Of course, considering the exceedances over a threshold is very practical tool to achieve this goal. But it is necessary to think further about the significance of this operation. The time series provides discrete values, at a given time step, of a continuous environmental variable, i.e. discrete observations of a physical quantity describing a physical phenomenon: wave height, temperature, wind speed, river discharge, rainfall, etc. The basic laws of physics (ultimately the conservation of mass and energy) are such that these quantities are temporally autocorrelated, and their temporal rate of variation is usually bounded. The finer the time step of the series, the stronger the correlation. So setting a threshold will not extract individual values: it will identify time intervals within which the observation is far from its average value. In other words, it will identify a storm, a flood, a heatwave, etc. These anomalies have a duration and a magnitude, and the peak is a very partial description of these.

We now see that we can no longer reason only in terms of discrete values. We have identified "something that happens" to a physical quantity, linked to physical phenomena: in other words, as seen in the introduction, a physical event. And so far we have chosen, very simply, to describe this event in terms of the maximum value or the associated quantity: the peak value.

This yields another consequence: the peak value is a different kind of variable than the variable observed at each time step. This is no longer the quantification of a physical quantity but a description of a physical event. Besides, other descriptions could be imagined, such as the average value over the duration of the event, or temporal integration (the temporal integration of river discharge during a flood event, for instance, will give the flood volume). So at this point we introduce some new terminology:

- the time series is made up of **sequential variables**, i.e. discrete observations of the temporal variations in the environmental variable;
- the i.i.d. sample is made up of **event-describing variables** (most of the time, event peaks).

Another point is that the sub-parameters for ensuring independence must be set according to the comprehension of the physical event.

To sum up, this first-step processing is a physical declustering with the aim of identifying and describing independent events, by extracting an i.i.d. sample of the event-describing variable from the autocorrelated time series of sequential values. This is why the "low threshold u_1" introduced in ***Mazas and Hamm (2011)*** was renamed "physical threshold u_p".

As a matter of fact, the use of a threshold is not always necessary. For instance, cyclonic studies can be carried out by modelling the sea states generated by the cyclones, each of which is archived by the cyclonic warning centres. There is no continuous time series but a set of already clearly identified events. Selecting the maximum wave height generated by the cyclonic atmospheric fields is enough to set up the i.i.d. sample.

Once this i.i.d. sample has been set up, the EVT can be applied and the reasoning becomes purely statistical. In the framework of extreme estimation through over-threshold modelling, the threshold can thus be called "statistical threshold u_s". Physical arguments are of no use here, with the possible exception of recommendations about the mean number of events per year: for instance, when considering the case of extreme discharge of large plain rivers, selecting an average of four floods per year seems contradictory to the physics of the phenomenon.

The conceptual framework exposed in the NHESS paper of 2014 can be summed up as follows:

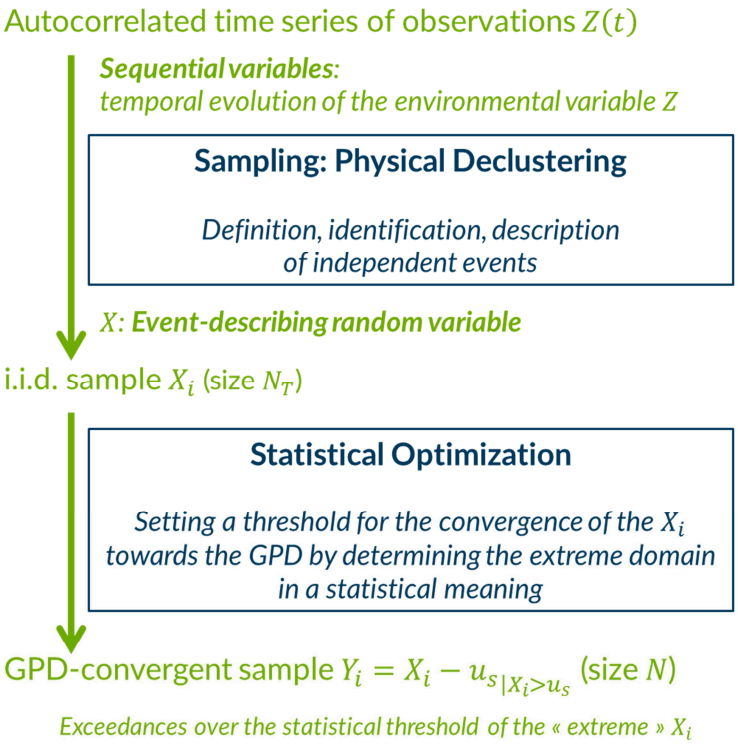

Figure 16. Two-step framework for over-threshold modelling, after Bernardara et al. (2014)

2.3. MAXIMUM LIKELIHOOD ESTIMATOR AND ITS *VIRGAE*

A technical point also arose in relation to over-threshold modelling. The use of a double threshold allows for a detailed sensitivity study with respect to the statistical threshold, since it is no longer necessary to decluster the time series at each threshold value, a process that can be rather time-consuming.

If we consider the classic case study of Haltenbanken wave height peaks provided by the IAHR Working Group on Extreme Wave Analysis (***van Vledder et al., 1994***). The accuracy of the data is 0.01 m and the dataset is such that the range of thresholds can be between 7 and 10 m. Instead of letting the threshold value vary with an approximate step of, say, 0.2 m, we can now make the step match the accuracy of the data and let the threshold vary between 7 and 10 m every 0.01 m, i.e. a total of 301 values to be examined. More specifically, for each value, a GPD is fitted to the peak excesses over this value and the changes in the shape and modified scale parameters and the quantile value (e.g. the 100-yr H_s peak) with respect to u_s are analysed. In accordance with the literature (in particular ***Coles, 2001***), in ***Mazas and Hamm (2011)*** we had used the Maximum Likelihood Estimator (MLE), notably for its asymptotic properties of robustness, consistency and efficiency. But using MLE for such refined sensitivity studies led to the surprising plots shown in Figure 17.

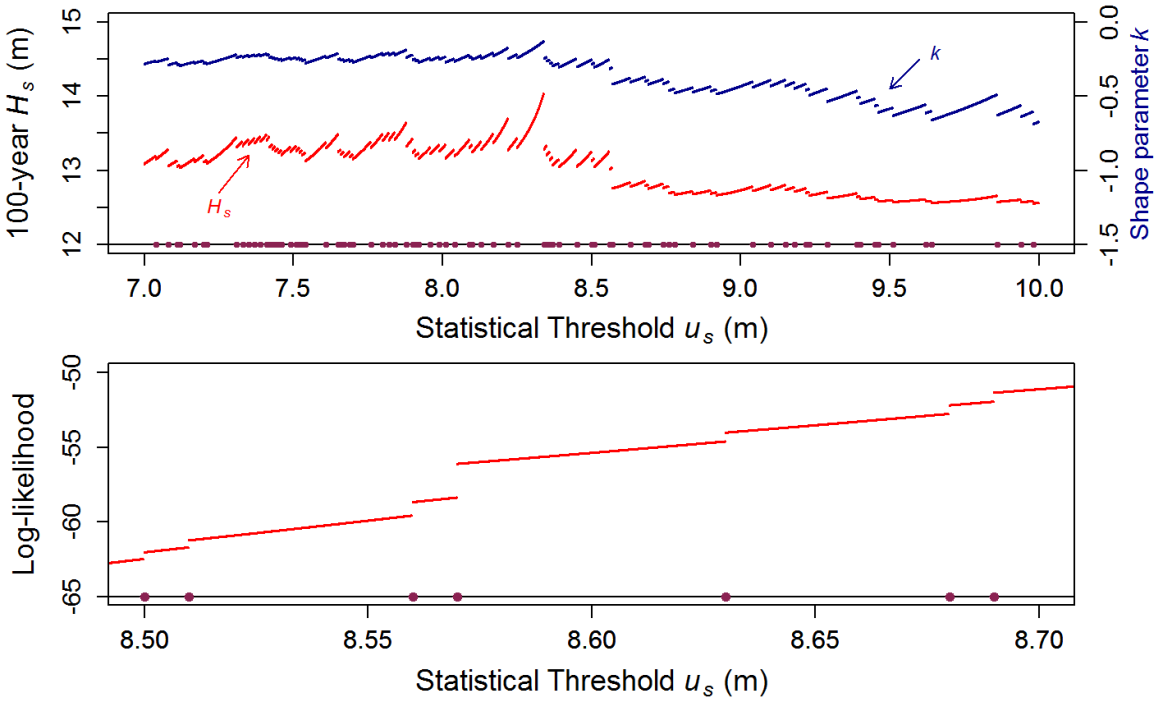

Figure 17. Haltenbanken dataset: change in the ML-estimated GPD shape parameter (top plot, above curve), in 100-yr Hs (top plot, below curve) and in the log-likelihood (down plot, zoom) with respect to the statistical threshold. Dots represent the peak values

The results of the parameter estimation, and hence of the quantile estimation, turned out to be quite unstable, with a distinctive pattern that we named *virgae*, after the shafts of precipitation falling from a cloud that evaporate or sublimate before reaching the ground (Figure 18). A close look shows that a *virga* occurs between two successive values of the dataset of peaks. This means that when u_s increases between two peak values, the estimation of the quantile can vary widely, between 0.1 and 1 m in the example above! In other words, the 100-year H_s can be very different whereas a GPD is fitted to the very same peaks.

Figure 18. Meteorological virgae
source: © Roger Edwards

This seems very counter-intuitive. But it should be borne in mind that it is not to the peaks X_i that the distribution is fitted but to the peak excesses over the threshold $Y_i = X_i - u_s$. So when u_s varies between two peak values, the dataset of X_i exceeding u_s does not change whereas the dataset of Y_i does.

This means that we have a problem of model choice: within such an interval, which value of u_s provides the best model? Looking at the model likelihood would appear to be the natural choice. But as illustrated by the bottom plot in Figure 17, we can see that the likelihood increases constantly and monotonically between two successive peaks. The maximum value is reached at the open upper bound of the interval $[X_i; X_{i+1}[$, i.e. for $u_s = X_{i+1} - \epsilon$. So should the value of ϵ be 0.1, 0.01, 0.001…? Furthermore, at this maximum, the derivative of the log-likelihood is not nil. Yet proving the asymptotic properties (consistency, efficiency) of the Maximum Likelihood Estimator requires that the true (and unknown) vector of parameters be an interior point of an open set (***Lehmann, 1983***, chapter 6), and that the likelihood function converges towards a global (and finite) maximum on this point with locally nil derivatives. This behaviour occurs because the density of distributions such as the GPD is maximal at the lower bound of their domain of validity. Necessarily, the closer the threshold is to the first value of exceedance, the higher the density associated with this value and, thus, the higher the model likelihood.

We can also see that when we set a threshold value we do not only select peaks; we also set the origin of the distribution: if the random variable $Y > 0$, then the random variable $X > u_s$.

To solve this issue we propose to introduce a location parameter μ whose function is to set the origin of the statistical distribution while u_s is restricted to peak selection. For instance, the cumulative distribution function of the 3-parameter GPD is:

$$F_{Y;k,\sigma,\mu}(y) = 1 - \left(1 + k\frac{y-\mu}{\sigma}\right)^{-\frac{1}{k}} \quad (5)$$

with $y = x - u_s$.

The three-parameter version of these distributions cannot be estimated using the MLE: because of the aforementioned maximum density at the origin, the location parameter necessarily tends to the minimum exceedance: $\mu \to \min Y_i = Y_{1:N}$. In parallel, the estimated vector of GPD parameters $\hat{\theta} = (\hat{k}, \hat{\sigma})$ does not converge towards any "target point" when $\mu \to Y_{1:N}^-$, or equivalently when $\epsilon \to 0^+$, as shown by Figure 19 (the plot is drawn for a dataset of simulated data from a GPD with known parameters in order to compare the estimated parameters with the true values).

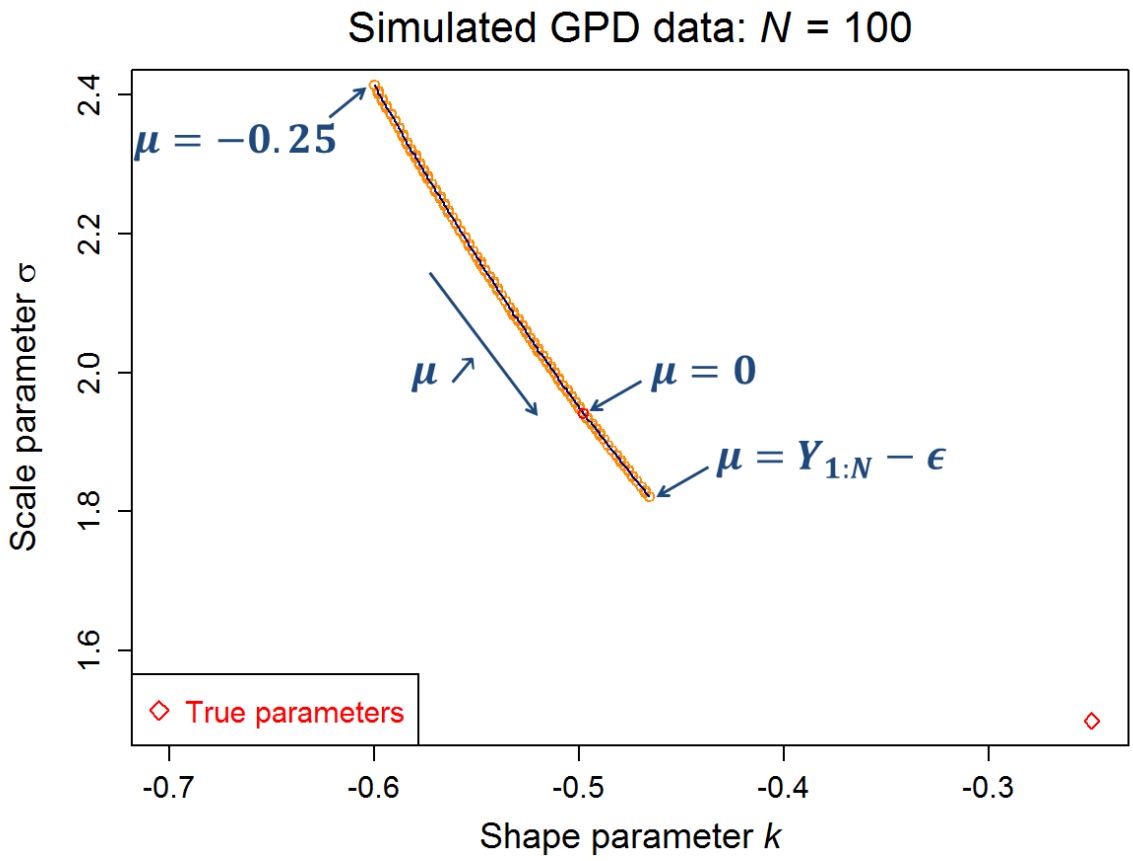

Figure 19. Change in the ML-estimated vector of GPD parameters $\hat{\theta} = (\hat{k}, \hat{\sigma})$ for a simulated dataset of size 100. True vector of parameters indicated in red.

We hence propose using the L-moments estimator proposed by ***Hosking (1990)***. Interestingly, any shift of u_s between two peak values is strictly offset by an opposite translation of μ, so the sample $(Y_i - \mu)$ always remains constant.

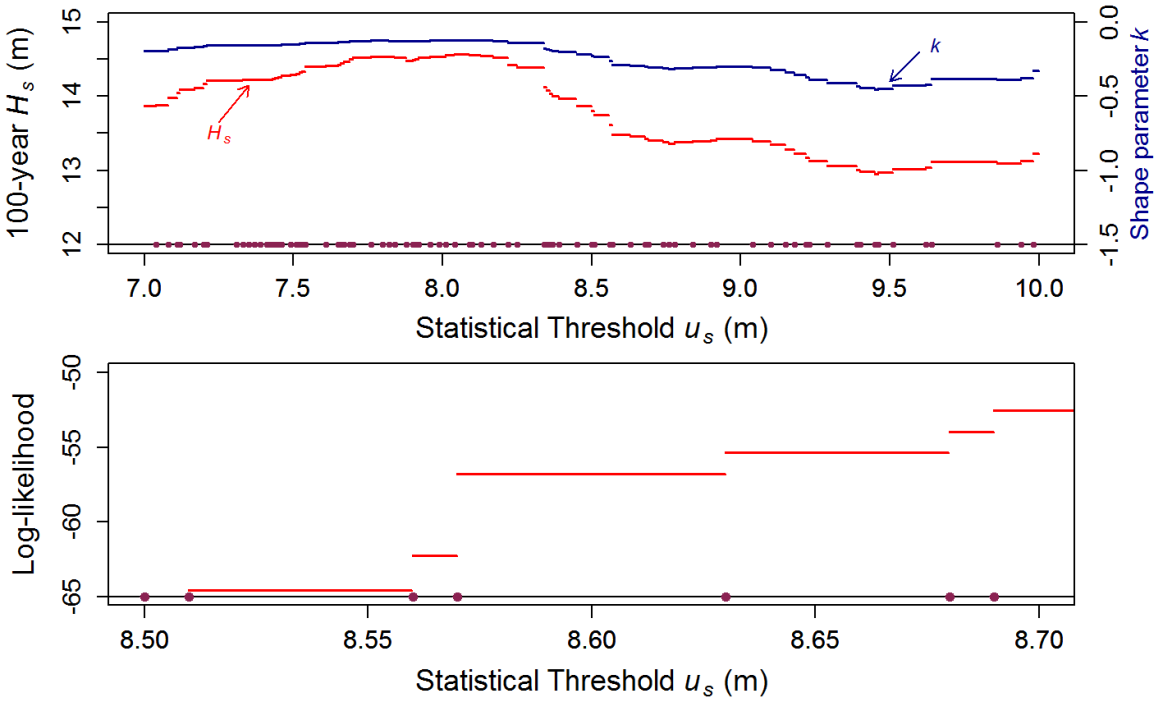

Figure 20. Haltenbanken dataset: change in the L-moments estimated GPD shape parameter (top plot, above curve), in 100-yr Hs (top plot, below curve) and in the log-likelihood (down plot, zoom) with respect to the statistical threshold.

Figure 20 shows that when fitting a 3-parameter GPD estimated using the L-moments, changes in the parameter estimation and hence in extreme value (quantile) estimation are only caused by a change in the sample of the peaks X_i. The conclusion is the same with other distributions such as Weibull or Gamma (named log-Pearson III in its 3-parameter formulation). The log-likelihood remains constant between two data points, but this is no longer a maximized value.

2.4. CONCLUSIONS

The works performed regarding the estimation of univariate extremes have progressively revealed important distinctions, namely the difference between the sequential observations from the event-describing variables and the complementary role of physical and statistical approaches.

At this stage, we can define univariate events as an anomaly of a physical quantity that takes values far from the average over a certain period of time.

Although the physical quantity, or environmental variable, describes a physical phenomenon (e.g. wave height for sea states, wind speed for wind, etc.), it is also quite legitimate to consider the event itself as a physical phenomenon (a storm, flood, etc.). Therefore, even in the simple univariate case, the non-definability of the concepts of event, phenomenon etc. referred to in the introduction can be experienced.

oOo

3. EXTREME MULTIVARIATE EVENTS: FROM SAMPLING TO RETURN PERIOD, A MATTER OF POINT OF VIEW

3.1. EXTREME SEA LEVELS: A FIRST APPROACH TO BIVARIATE ANALYSIS

Extreme sea level determination is a very good example of a "simple bivariate analysis": it can be easily assumed that it is only necessary to consider two physical *phenomena* described by two *physical quantities* of *the same kind*:

- astronomical tide T, a deterministic, predictable variation in sea level generated by astronomical forcing;
- meteorological surge S, a stochastic variation in sea level generated by atmospheric fields (wind and pressure).

The analysis may be made more sophisticated by accounting for the eustatic trend in mean sea level, another quantity of the same kind, or in wave set-up, which is also of the same kind but requires an analysis of sea states. However, we will consider the simple approach above and can write:

$$Z = T + S$$

where Z is sea level, T is astronomical tide and S is surge (in m). Note that one can either consider Z and T as fluctuations (in m) around the mean sea level (a level relative to a given datum), or include the mean sea level both in Z and T, which become a level relative to a datum.

It is easy to understand that in macrotidal environments, in which the range of variation in the astronomical tide (6 to 14 m) is far larger than the range of variation in the meteorological surge (2 to 3 m), a direct extrapolation of the sea levels observed is meaningless. ***Haigh (2010)*** demonstrated that the extreme sea levels induced by this direct approach in macrotidal environments are underestimated. This is why the so-called indirect approaches were developed as early as 1979 (***Pugh and Vassie, 1979***) and made more sophisticated in the decades that followed (***Tawn and Vassie, 1989***, ***Tawn, 1992***, ***Dixon and Tawn, 1994***, ***1999***). The method was called the Joint Probability Method (JPM), in its various forms: Revised JPM (RJPM), Spatial Revised JPM (SRJPM), etc. The basic idea consists in extrapolating the extreme meteorological surge values and then combining this extrapolation with the (known) distribution of astronomical tide to obtain the distribution of sea level.

A first key point of the methodology is the probabilistic model used for extrapolating extreme surges. The models proposed in the RJPM are based on the annual maxima or r-largest declustering methods combined with a GEV distribution. Applying the methodology for univariate extremes presented in section 2, namely a combination of two-step POT declustering along with a multidistribution approach, was a first rationale to update the methodology for determining extreme sea levels.

A second difficulty was quite interesting as it was linked to the concept of events. The observations of sea level (and of its components) exhibit strong temporal dependence (auto-correlation). That is what the distinction made between sequential values and event-describing variables (storm peaks) is all about. In order to obtain the full distribution of extreme sea levels, we have to:

1. extrapolate the extreme surge peaks;

2. determine the distribution of sequential values of surge;
3. determine the distribution of sequential values of astronomical tide;
4. determine the distribution of sequential values of sea level by combining the distributions of sequential values of surge and tide;
5. derive the distribution of extreme sea level peaks.

Step 1 is easily dealt with thanks to the univariate methodology. Step 3 does not present any difficulty provided that the astronomical tide is known: an empirical distribution (estimated by a kernel density estimator) of tide levels over a saros period (6585.32 days, i.e. 18 years and 10 or 11 days[8]) can be assumed to represent almost perfectly the distribution of tide between LAT (Lowest Astronomical Tide) and HAT (Highest Astronomical Tide). Step 4 is also straightforward: the probability theory tells us that the distribution (probability density function) of the sum of two independent variables is the convolution of their individual densities.

This leaves us with two main difficulties:

1. transforming the distribution of the event-describing variable (peak) into the distribution of the sequential value, and vice-versa;
2. modelling accurately both the frequent values (bulk of the distribution) and the extreme values (tails of the distribution).

The literature provided help with these two difficulties. The extremal index, introduced in the JPM by ***Tawn and Vassie (1989)***, accounts for the clustering of sequential values around the peak. It can be seen as the reciprocal of the mean number of sequential values per peak, or as the mean duration of events if the step δt of the time series is accounted for. However, it had to be adapted to a POT approach. With the invaluable help of Dr Xavier Kergadallan, it was possible to derive the distribution F_S of the sequential surges S above the statistical threshold u_s (the limit between the empirical and the parametric domains) from the distribution G_{PS} of the extreme surge peaks P^S:

$$F_S(s) = 1 + \frac{N}{\nu K} d_S(s)\left[G_{PS}(s) - 1\right], \text{ for } s > u_s \quad (6)$$

where N is the number of surge peaks, K the duration of the time series, ν the number of sequential values per year and $d_S(s)$ the mean number of sequential surges per event, a number that varies with s. Note that G_{PS} is a conditional distribution: it is first transformed into the conditional distribution of sequential surges above u_s given that $S > u_s$ (the integral of its density is 1), and then it is transformed into the distribution of sequential surges above u_s by assuming that the probability of $S > u_s$ is given by the ratio of the number of sequential values in the parametric domain over the total number of observations.

Last, a model is needed for the variations in $d_S(s)$ with s, i.e. the mean number of values per event when the threshold defining the event varies. We identified three distinct regimes in $d_S(s)$: first a decrease from $d_S(u_s)$ to 1, though not generally monotonic; then a range in which it stays equal to 1 (each event has a single value); and lastly, a range in which the maximum observed peak is exceeded and $d_S(s)$ can no longer be defined. We proposed to fit a linear regression model for the first regime in log-log scale, while for the second and third regimes $d_S(s) = 1$, as illustrated in Figure 21 below.

[8] depending on the number of leap years in the 18-year period

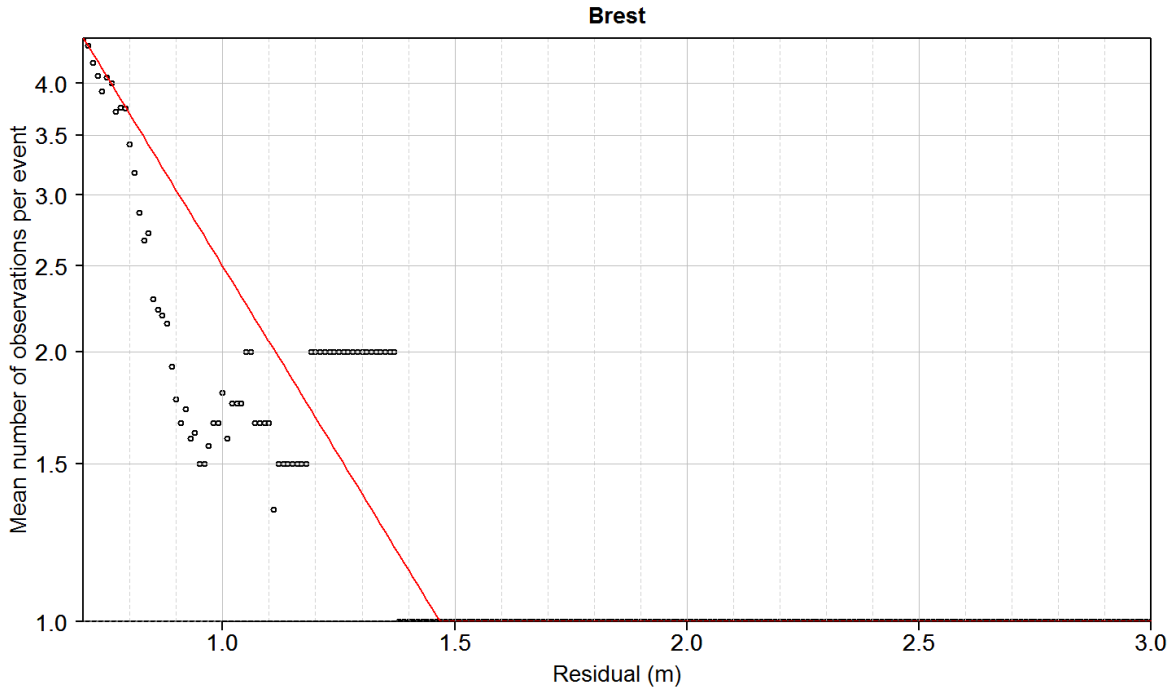

Figure 21. Mean number of sequential values per event with respect to surge height: observations (circles), model (lines).

Source of data: cf. Figure 6

Figure 22 compares the tail of the empirical distribution of hourly surges with the parametric tail distribution $F_S(s)$, for $s > u_s$. The interest of the extrapolation and the accuracy of the limit value at the boundary between the empirical and parametric domains are obvious from this plot.

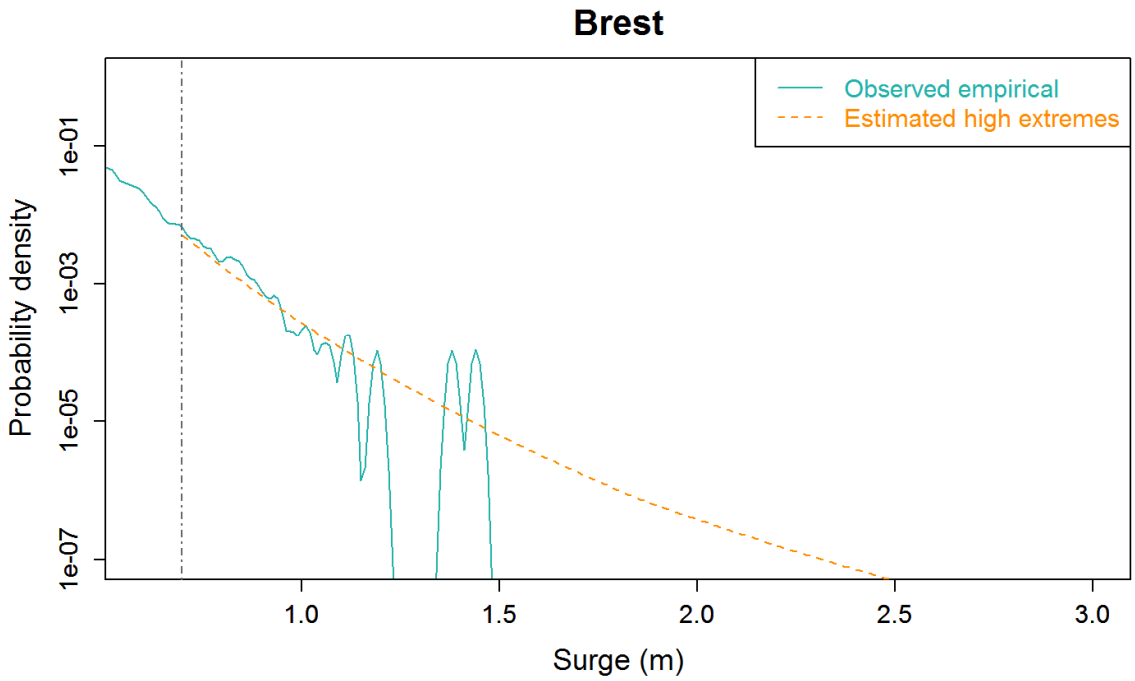

Figure 22. Upper tail of the hourly surge values probability density function.

The second difficulty was easily overcome by using a mixture model that connects a kernel-density estimated empirical distribution for the bulk values (frequent observations) to the parametric tail derived from the extrapolation of surge peaks as presented above. Furthermore, extreme negative surges may also be extrapolated and a lower parametric tail included in the surge distribution. This is of great interest for estimating extreme low sea levels, an item of information that is needed in regard to the toe stability of rubble mound breakwaters or to the cooling circuits of nuclear power plants built by the sea.

The distribution of sequential surges can then be convoluted with the distribution of sequential tidal levels in order to obtain the distribution of sequential sea levels, as illustrated in Figure 23.

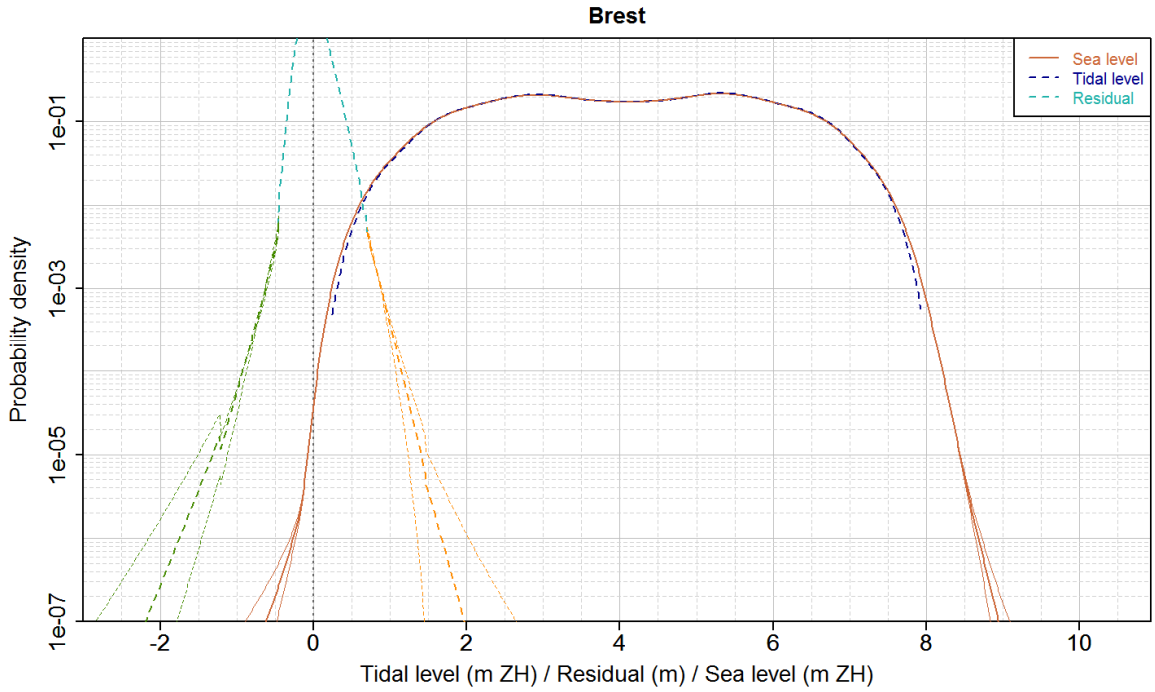

Figure 23. Probability density functions of hourly surge (empirical bulk and parametric tails), astronomical tide and sea level.

Last, the extreme values of sea level for return periods of 1, 10, 100, 1,000 years, etc. are usually requested. This means that we have to define events again, once more by means of POT declustering, but not followed by a fit to the identified peak values. The event identification is used to define an extremal index for sea levels or its reciprocal, the mean number of sequential values per event, $d_Z(z)$. The inverse of the transformation from peaks to sequential values presented above then yields the distribution of extreme sea level peaks.

Temporal autocorrelation is not the only type of dependence to be accounted for; in some places the surge value depends on the tide value, and vice-versa. This is especially true of shallow locations because the celerity of long waves depends on water depth (see equation (1)). Generally speaking, larger surges are expected at low tide because they tend to rise in shallower waters. In parallel, a higher sea level caused by a large surge can modify both the phase and the amplitude of the astronomical component (***Flather, 2001***). Moreover, this interaction can step into the equations of momentum conservation in several terms, namely the advection term, the bottom friction term and the surface stress term (***Jones and Davies, 2008***, ***Zhang et al., 2010***, ***Idier et al., 2012***).

Dixon and Tawn (1994) have proposed a statistic based on a χ^2 test statistic to assess tide-surge dependence, along with a normalization to account for it in the extreme analysis. In our paper we proposed an alternative normalization, based on the average of the surge values exceeding the 95% quantile of tidal bands. This was found to perform better at several sites along the French Atlantic / Channel coasts.

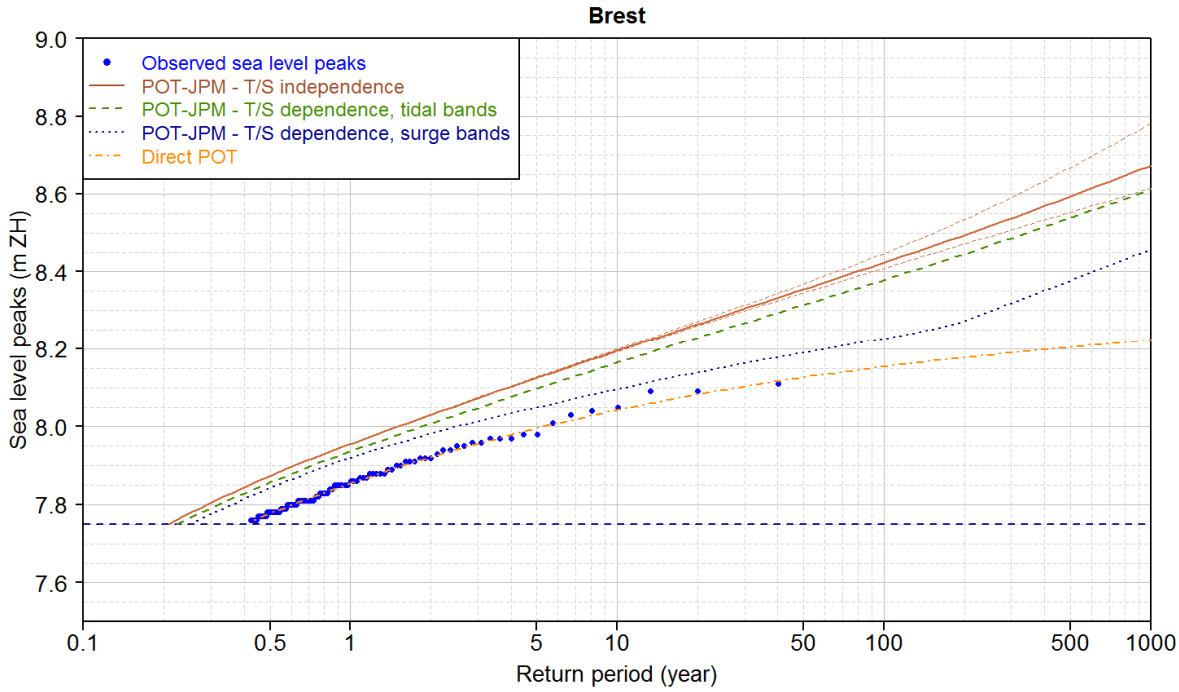

Figure 24. Return periods for sea level events. From the upper to the lower curve: indirect approach without tide–surge interaction, indirect approach with equi-probable tidal bands, indirect approach with equi-probable surge bands, direct approach.

Figure 24 displays the extrapolation plots for extreme sea level peaks with respect to return period for a direct POT extrapolation of sea level peaks and the new POT-JPM indirect approach. Tide-surge independence is either considered negligible (brown curve) or handled with two different methods. The observed peaks are also drawn.

Although an offset appears for small return periods (between 0.1 and 10 years), the interest of the indirect approach is obvious for large return periods. The shape of the curve from the direct extrapolation (a GPD with negative shape parameter, which yields a concave curve) is quite different from the straighter asymptotic behaviour of the fits from the indirect approach. The underestimation noticed by ***Haigh et al. (2010)*** is found again. Although we are more interested here in the methodology itself rather than in the quantitative results, they can be compared to the results of other studies (***Pirazzoli and Tomasin, 2007***, ***SHOM - CETMEF, 2008***, ***Haigh et al., 2010***) gathered in our overview paper published in 2011 in *La Houille Blanche* (**[M]**). In particular, the 100-year sea levels are very close to those found by ***SHOM - CETMEF (2008)*** after the method developed by ***Simon (1994)***, and a bit higher (0.10 to 0.15 m) than those found by ***Fortunato et al. (2016)***.

This offset can be observed on the many sites on both sides of the Channel analysed in ***Haigh et al. (2010)***, and has also been found when applying this methodology on high tide sea levels (skew surges and astronomical high water) at La Rochelle – La Pallice (Bay of Biscay), Brest (Atlantic Ocean), Cherbourg and Saint-Malo (Channel). However this offset is smaller when working on high water values than on hourly observations; for Brest the overestimation can be estimated at ~ 0.03 m in the former case versus ~ 0.08 m (accounting for tide-surge dependence) to ~ 0.10 m (independence assumption) in the latter. A significant difference between the high water values and the hourly values is that the extremal index is always very close to 1 in the former case: the transformations between the distributions of sequential and event-describing values are thus much easier and better modelled. Thus we can expect this offset to be caused, at least, by an extremal index model that is not accurate enough on the one hand, and by an incomplete handling of tide-surge interaction on the other hand.

We thus have a methodology for determining the univariate distribution of a phenomenon made of two components that can be dependent by combining their marginal distributions. The joint distribution of tide and surge is not required for the analysis because the information relating to joint occurrence is somehow included and reduced in the convolution operation. Thanks to the effort made above in regard to the terms, definitions and vocabulary (section 1.3), we can now say that we are considering the joint occurrence of two *physical quantities* that are *of the same kind*.

3.2. JOINT OCCURRENCE OF EXTREME WAVES AND SEA LEVELS: FROM BIVARIATE TO MULTIVARIATE

3.2.1. A new classification for multivariate analyses

The influence of sea states on the sea level at the shoreline through wave run-up (set-up and swash) or complex mechanisms described by ***Bertin et al. (2012)***; the combined effect of waves and sea level for coastal flooding, beach erosion, wave overtopping of coastal structures such as breakwaters, dykes and quays; the design of offshore structures submitted to concomitant loads from waves and wind: these are just some of the issues potentially called into play, and for which a methodology is required in order to estimate the extreme joint probabilities of two phenomena such as waves and sea levels.

Though the joint occurrence of wave height and sea level is a bivariate problem just like the joint occurrence of surge and tide, the difference obviously lies in the fact that we cannot eventually reduce the problem to the probability distribution of a single variable, such as sea level: the sum of significant wave height and sea level or, even more, the sum of significant wave height and wind speed or current velocity, is meaningless. We now consider the joint occurrence of two *physical quantities* that are *not of the same kind*.

More generally, and more importantly for this thesis, we can now see clear distinctions in bivariate analyses that are directly linked to what we call events.

In the *Coastal Engineering* paper (**[S]**), the following classification of multivariate analyses was proposed:

- **Type A**: a metocean process described by several parameters (e.g.: a sea state described by its significant wave height H_s, its peak period T_p, its peak direction θ_p, its directional spreading, etc.);
- **Type B**: a metocean process that can be broken down into several elementary processes (e.g. a sea state made of a swell system and a wind sea system; or the sea level made of a mean sea level, astronomical tide, meteorological surge, wave set-up, etc.);
- **Type C**: the joint occurrence of several distinct metocean processes (e.g. waves, sea level, wind, current).

We can now rewrite this classification, which is illustrated in Figure 24:

- **Type A**: a single *phenomenon* described by different *physical quantities* that are *possibly not of the same kind*;
- **Type B**: a *phenomenon* made of different *components*, described by *physical quantities of the same kind* between one component and another (e.g. H_s and T_p are not of the same kind, but $H_{s,wind\ wave}$ and $H_{s,swell}$ are);
- **Type C**: several *phenomena* described by *physical quantities* that are *possibly not of the same kind*.

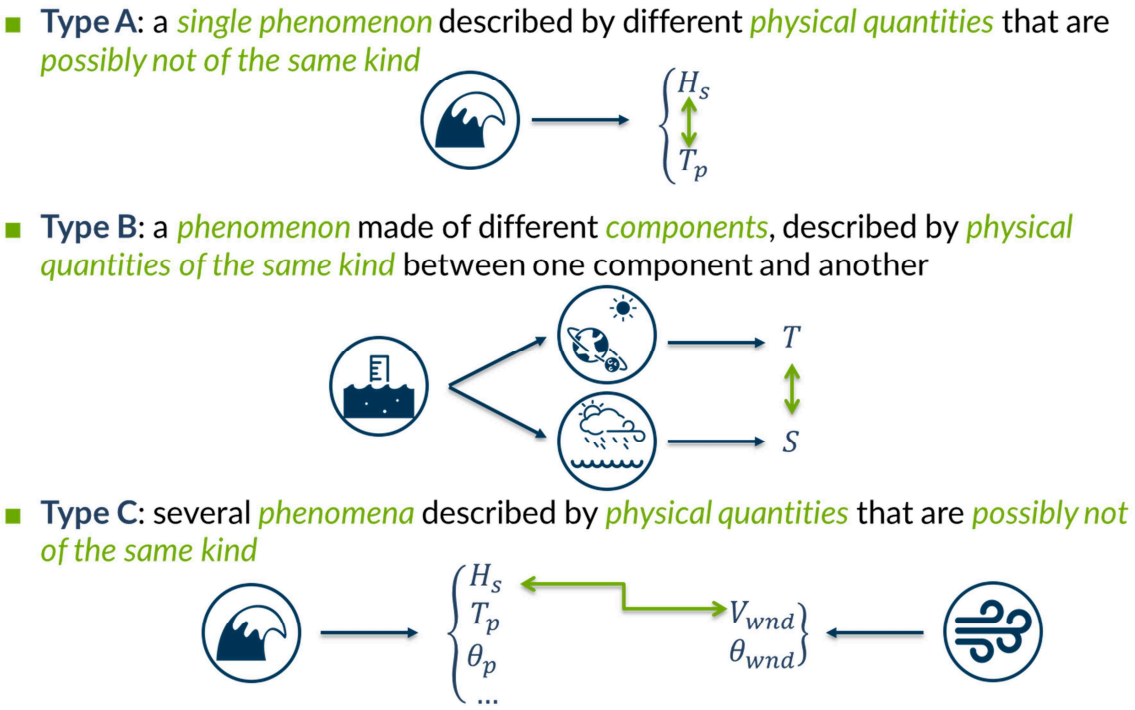

Figure 25. Illustration of the classification for multivariate analyses

Two main issues lie at the core of the analyses: **sampling** and **dependence**. They are intimately linked because dependence is assessed on the sample. **Sampling itself is nothing more than event definition, identification and description.**

3.2.2. Sampling: a description of events

In Type A analyses, sampling is generally straightforward. When dealing with directional extreme winds or H_s/T_p correlation, it is intuitive to define the event as an anomalous value of the quantity that measures the energy most directly, such as wave height, wind speed, current velocity, sea temperature, etc. The choice of **reference variable** or **event-defining variable** is easily made and the other variable is a mere covariate; the value concomitant to the peak of the reference variable is generally taken.

In Type B analyses, the components can be dealt with separately before being reduced to a single variable (the general phenomenon), and sampling can be performed on each component within the univariate framework described in section 2.2.

Things become much trickier in Type-C analyses. For instance, both wave height and wind speed are closely linked to energy and there is no obvious choice for setting a reference variable. In addition, it must be borne in mind that the joint occurrence of two (or more) phenomena is usually required because their combination causes a distinct phenomenon: beach erosion, wave overtopping, coastal flooding, dune breaching, etc. Furthermore, these **response phenomena** do not necessarily require the joint occurrence of large values of the **source phenomena**: the combination of a large value of one with a frequent value of the other may generate an event as far as the response phenomenon is concerned.

We thus begin to see that a point of view needs to be chosen. This also explains the variety of sampling methods found in the literature. Generally speaking, they differ in the answer they offer to the following question: should the method focus on the exceedance of one variable only, of both, or of one or the other? Figure 26 illustrates these possibilities: an event selection limited to domain 2 means a selection of simultaneous large values for both variables; domains 1+2 (resp. 2+3)

correspond to a selection for large values of a single variable; domains 1+2+3 correspond to a selection for large values of at least one of the two variables.

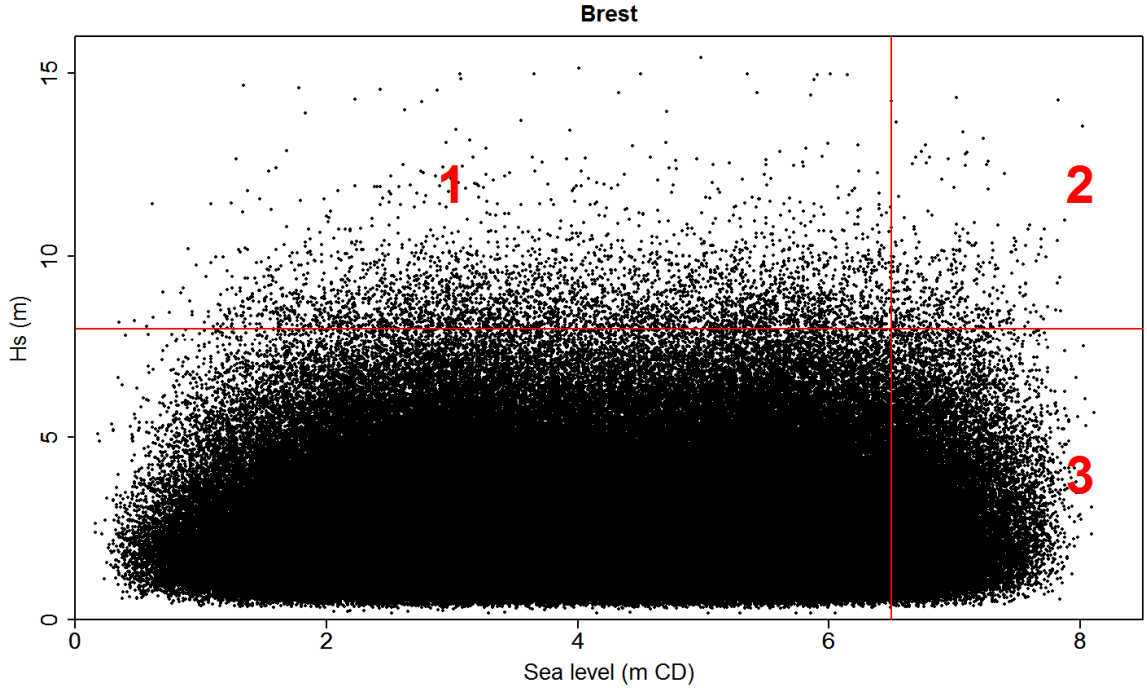

Figure 26. Possible domains for event selection on a Hs / sea level scatterplot

Source of data: cf. Figure 6

High tide sampling for wave / sea level analyses in a macrotidal environment, which consists in selecting the sea level at high water along with the concomitant H_s value (Figure 27), is based on the postulate that the analysis is performed for phenomena that can only occur at high tide (***Kergadallan, 2015***), which would not be the case with supplying seawater to cool nuclear power plants. The bivariate threshold proposed by ***Li et al. (2014)*** tries to consider both variables equivalently by selecting pairs that correspond to large values for both phenomena (Figure 28). Covariates such as wave period or direction may also be to be accounted for, notably to ensure that the selected events are identically distributed. This is why we proposed an additional sampling method, which consists first in using analytical formulae to calculate the nearshore wave height H_s' as a function of offshore wave height H_s, period T_p and direction θ_p, and then summing it to the sea level Z resulting in a "total water level" function[9] $Z + H_s'$ that is physically meaningless but unidimensional: a conventional univariate POT declustering process can be applied (Figure 29). In other words, we use a univariate response function whose desired properties are: (i) it selects large values of both source variables; (ii) it selects large combinations of the source variables; (iii) it accounts for covariates for the i.d. assumption (e.g. wave covariates so that waves that do not propagate nearshore are not selected); (iv) a response function that is "neutral", i.e. it is relevant for any true response function such as wave overtopping, wave load on a seawall, etc.

[9] An improvement to this function would be a normalization of the nearshore wave height H_s' and of the sea level so that both components have approximately the same weight in the response function. It has been checked that the resulting sampling is very close.

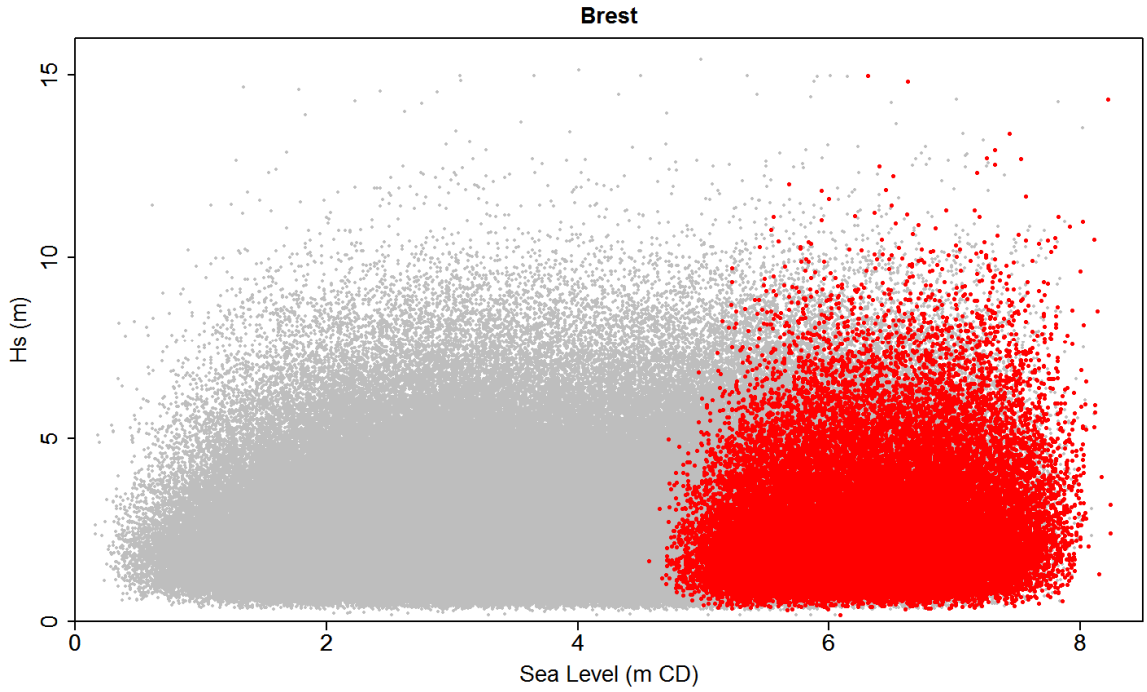

Figure 27. High tide sampling of sea level and wave height

Source of data: cf. Figure 6

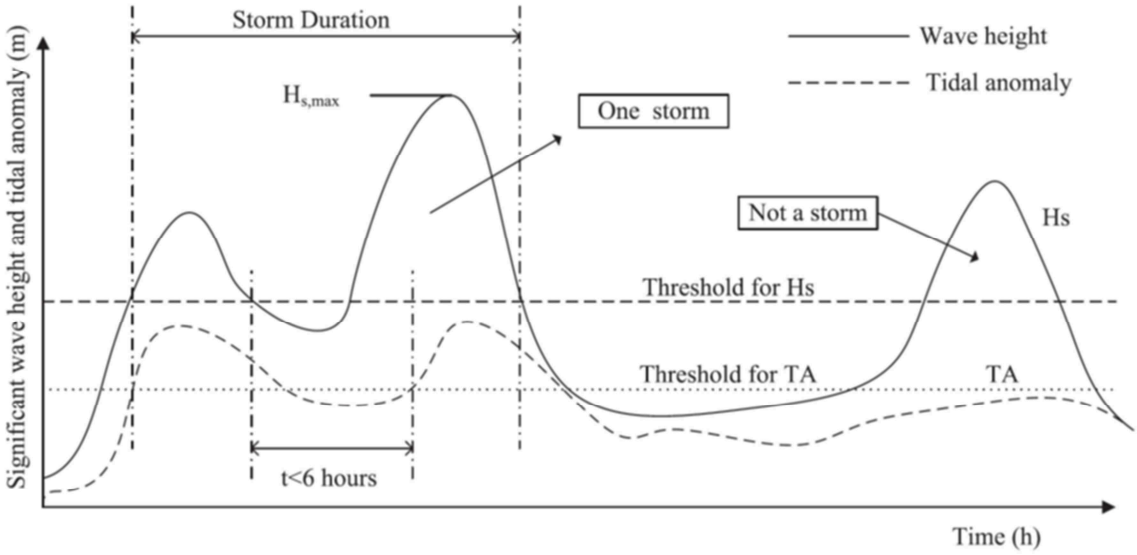

Figure 28. Definition of independent storm events with a bivariate threshold

source: Li et al. (2014)

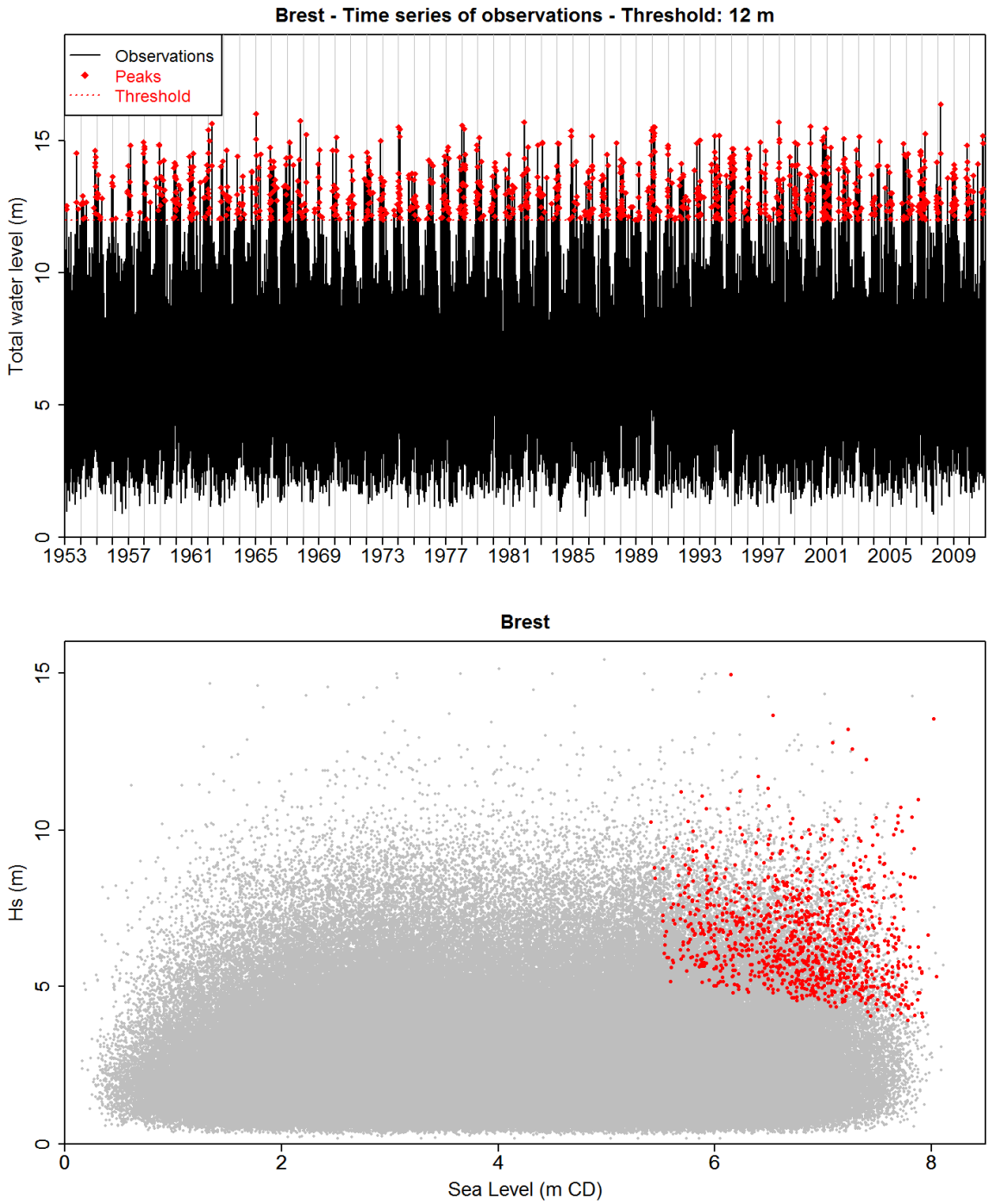

***Figure 29.** Top: time series of a univariate response function (sum of sea level and nearshore wave height) with threshold and peak of the selected events; bottom: sequential pairs (Hs, Z) of the time series (in grey) and selected event-describing pairs (in red)*

Source of data: cf. Figure 6

Of course, we should not expect to find the perfect function for this. Nonetheless, this function was introduced as an additional approach and in order to stress the main point of the sampling step: it all depends on the event definition that is chosen by the analyst.

An important change is that the event-describing variable of the univariate framework, most of the time a peak, is now an **event-describing tuple**, or pair in the bivariate case. Once again, it may be that none of its elements is a peak of its variable.

The three dimensions of sampling, namely event definition, identification and description, can thus be illustrated with one of the aforementioned methods: high tide sampling. It is clear that choosing this sampling necessarily *defines* an event as what happens at high tide; and vice-versa. *Identification* consists in finding the maxima of the variable "sea level". Last, the *description* of such events requires choices: should we consider the instantaneous surge at the time of the high tide or the skew surge, should we consider the wave height at the time of the high tide or its maximum value within a time window centred on the high tide, etc. The bivariate threshold method also illustrates these three sides: an event is *defined* as the joint occurrence of a large value for each variable; it is *identified* by the joint exceedance of a physical threshold for each one along with temporal criteria (minimum storm duration, maximal duration for fluctuations, etc.); and it is *described* by the peak values of each variable.

Once the choice of sampling method has been made, the i.i.d. bivariate (or multivariate) sample is set up and the rest of the methodology, though technical, is more straightforward.

3.2.3. Dependence: assessment and modelling

First the marginal distributions are determined. The mixture model introduced for the surges in section 3.1 is called upon: for the variable X (resp. Y), the i.i.d. sample of the event-describing tuple is reduced to a univariate sample made of the elements corresponding to X (resp. Y). The elements exceeding a statistical threshold are extrapolated by an extreme value distribution, then the parametric tail is connected to the empirical distribution of the elements below that threshold. Note that an extremal index is no longer needed: there is no temporal autocorrelation between the elements of the i.i.d. sample.

Next, dependence is assessed on the event-describing pairs (X_i, Y_i). To begin with, dependence may be analysed through various coefficients, statistics and plots, of which the scatterplot and the conventional correlation coefficients (Pearson's r, Spearman's ρ, Kendall's τ) are the simplest. Two tools are particularly interesting. The first one is the chi-plot introduced by ***Fisher and Switzer (1985)*** and further detailed in ***Fisher and Switzer (2001)***. Figure 30 displays an illustration of the chi-plot associated with a scatterplot showing a sample with positive dependence (correlation coefficient 0.5).

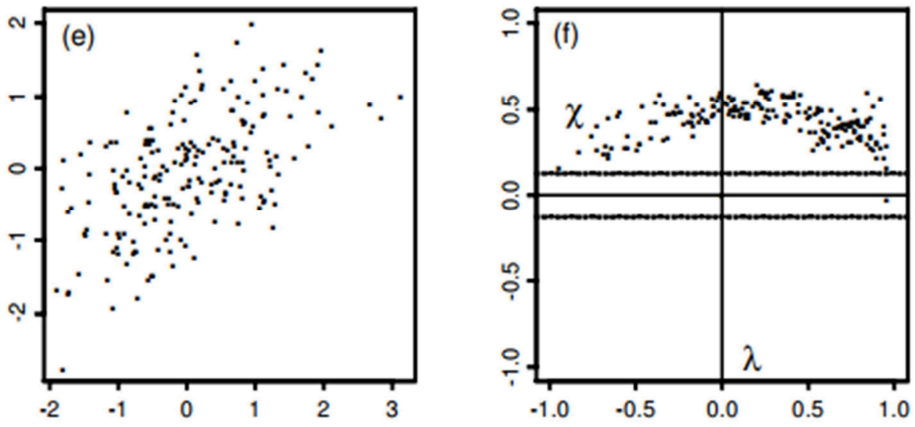

Figure 30. Chi-plot (right) for a sample with positive dependence

source: Fisher and Switzer (2001)

The second one is the upper tail dependence coefficient λ_U, which describes the limit probability that a variable exceeds a threshold when the other one does the same. Independence translates into λ_U tending to 0 whereas complete dependence translates into λ_U tending to 1.

To model dependence, we can use a mathematical object designed to describe the dependence structure between random variables: a copula. According to Sklar's theorem (***Sklar, 1959***), a multivariate distribution function H can be described by univariate marginal distributions F and a copula C:

$$H(x_1, \dots, x_d) = \mathbb{P}[X_1 \leq x_1, \dots, X_d \leq x_d] = C(F_1(x_1), \dots, F_d(x_d)) \quad (7)$$

Among all the existing copulas, the extreme value copula family is particularly suitable for our analyses, namely the copulas of Gumbel-Hougaard (logistic), Galambos (negative logistic) and Hüsler-Reiss. Goodness-of-fit tests such as the Cramér-Von Mises statistic allow the best-fitting copula to be selected.

3.2.4. Joint distribution: a first interpretation

Lastly, the estimated copula and the marginal distributions can be combined to get the joint probability distribution. The joint return period $\mathcal{T}$ associated with the event ($X > x$ and $Y > y$) is given by the following formula:

$$\mathcal{T}(x,y) = \frac{1}{\lambda_p \mathbb{P}[X > x, Y > y]} = \frac{1}{\lambda_p \left[1 + C_{\hat{\theta}}\left(F_X(x), F_Y(y)\right) - F_X(x) - F_Y(y)\right]} \quad (8)$$

The overall methodology is summarized in Figure 31 below.

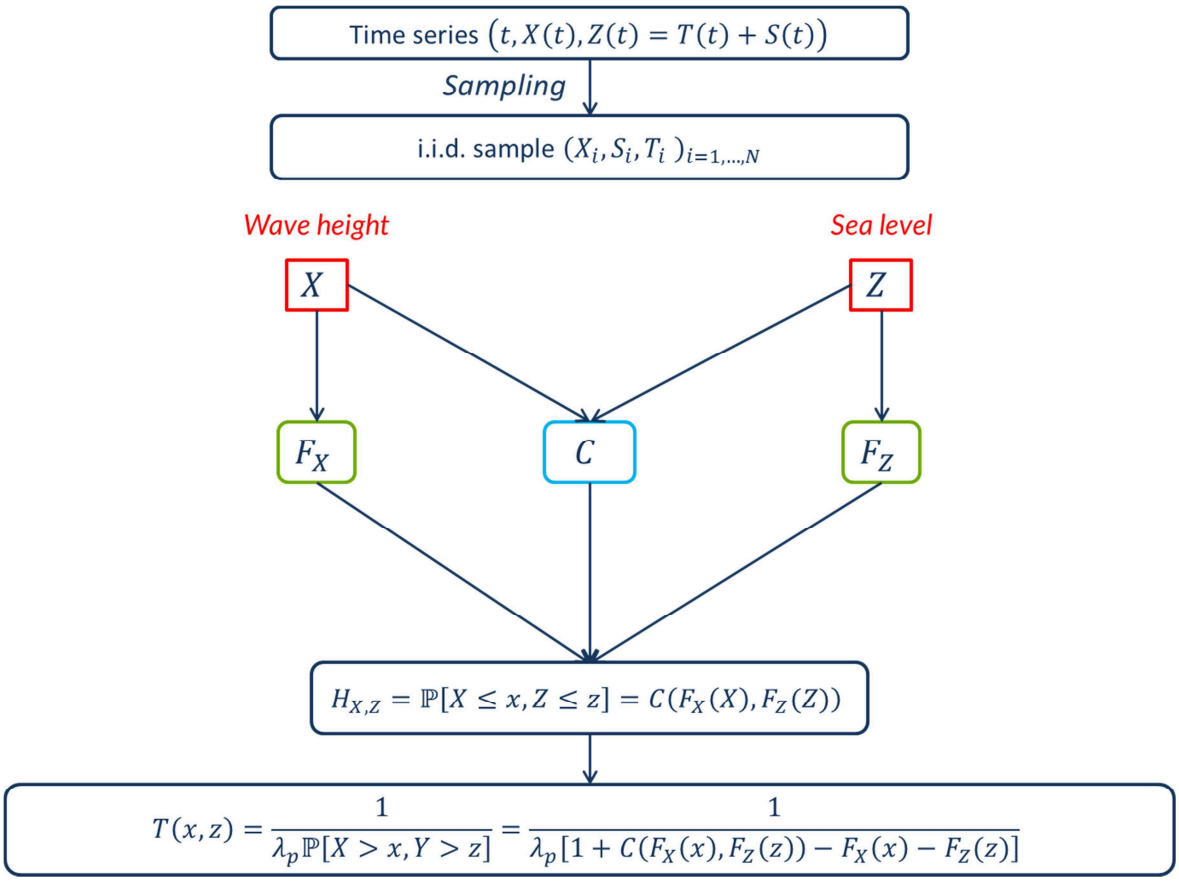

Figure 31. Sketch of the bivariate methodology for determining extreme joint probabilities of wave height and sea level

In the *Coastal Engineering* paper (**[S]**), this methodology was applied to a case study in Brest and compared with results from the JOIN-SEA software (the analysis being kindly performed and

provided by Dr P. Hawkes, HR Wallingford), which is widely considered as a technical reference in the domain (***Hawkes et al., 2002***, Figure 32).

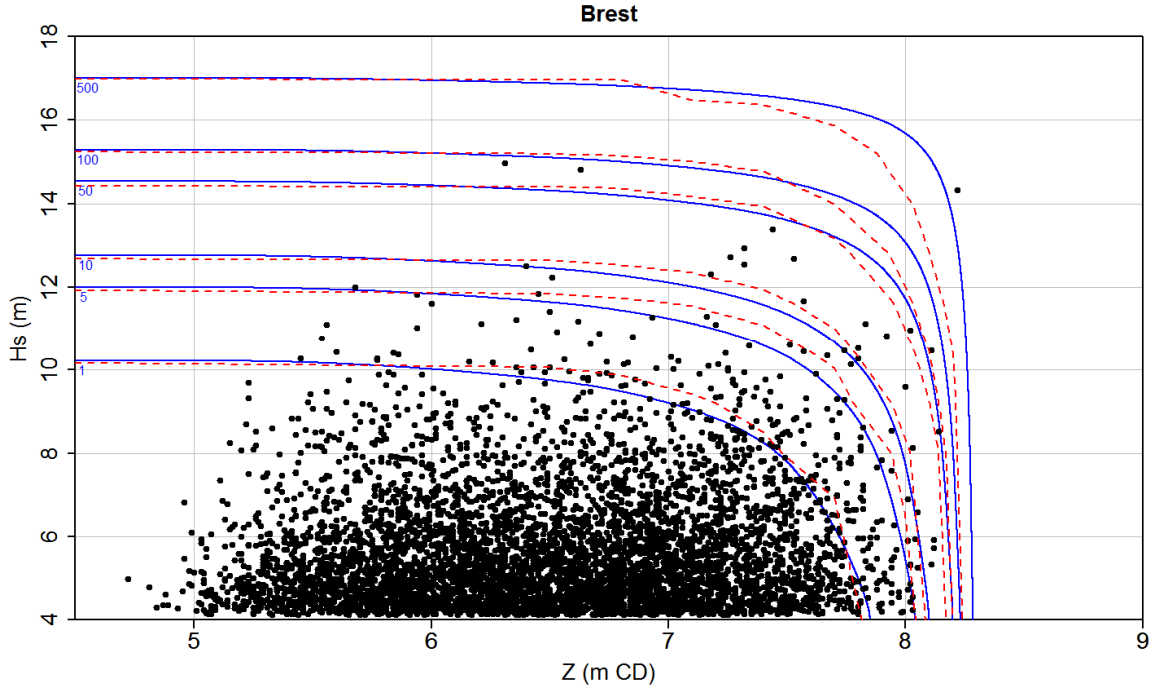

Figure 32. Comparison of joint return periods between the JOIN-SEA simulations (red dashed lines) and the bivariate methodology (blue plain lines)

Source of data: cf. Figure 6

In the paper, this methodology was refined in order to implement the POT-JPM approach for determining extreme sea levels presented in section 3.1 and to assess the dependence between wave heights and meteorological surge, instead of wave height and sea level. Indeed it may be expected, at least in environments such as the coasts of north-western Europe, that waves and surge are generated by the same kind of meteorological phenomenon, while the generation of the astronomical tide is totally independent of atmospheric processes[10]. The analysis is thus performed on wave height X, tide T and surge S, as depicted by the sketch in Figure 33. A convolution between the marginal distribution of S and T yields the distribution of sea level Z (possibly accounting for tide-surge interaction) while dependence is modelled between X and S, which provides the joint distribution of X and S. A 2D1D convolution between these two distribution (i.e. a convolution with the distribution of Z for each row of the joint distribution of X and S) then provides the joint distribution of X and $Z = S + T$.

[10] As seen and discussed in section 3.1, this is not true of its propagation in shallow areas.

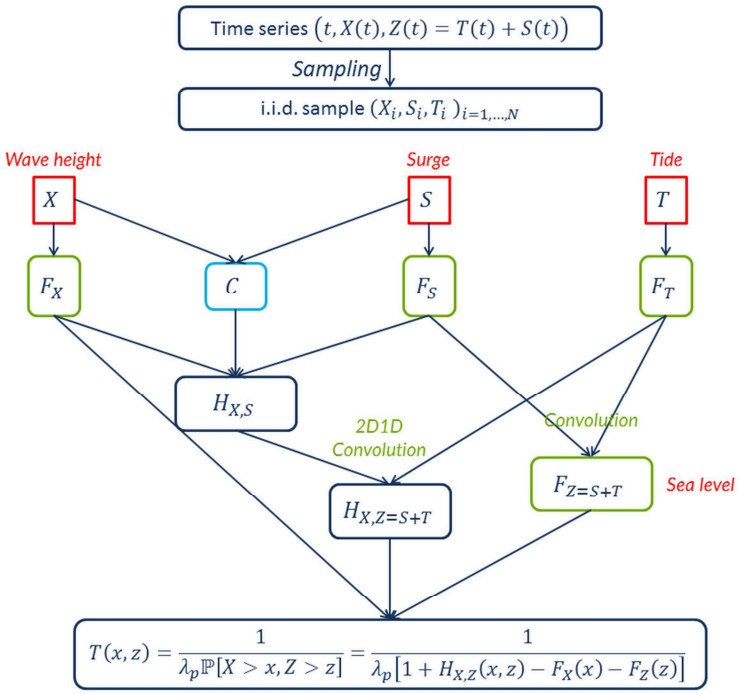

Figure 33. Sketch of the multivariate methodology for determining extreme joint probabilities of wave height and sea level

Figure 34 shows the scatterplots of the H_s/Z and H_s/S pairs selected from the time series by the sampling. Visually, H_s and S appear notably to be more correlated than H_s and Z.

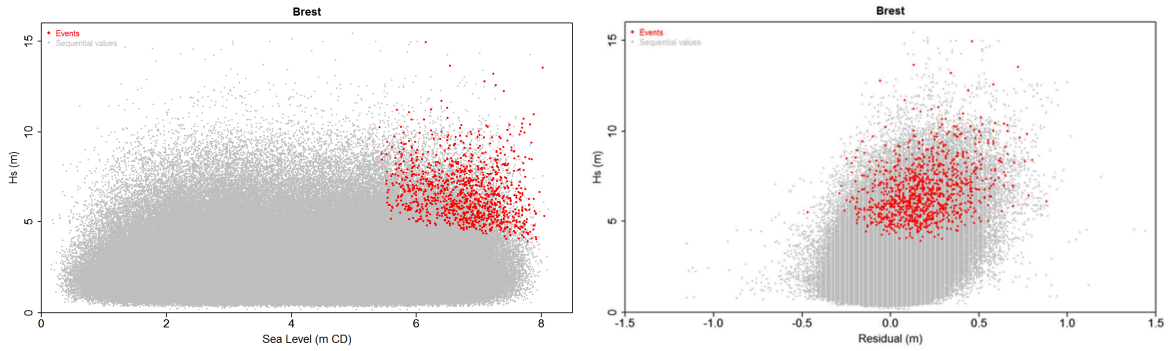

Figure 34. Scatterplot sequential values (grey) and event-defining pairs of Hs / sea level (left) and Hs / surge (right)

Source of data: cf. Figure 6

Indeed the chi-plot shows that this dependence between X and S is much stronger than between X and Zplot. We have also modified the appearance of the chi-plot in order to explain the two lobes that can be seen on the right-hand side of the plot. The upper one was found to correspond to (X, Y) pairs that are both larger than their respective medians, and the lower one to pairs that are both smaller. The plot display proposed in Figure 35 clearly shows that the positive dependence is stronger with the largest values.

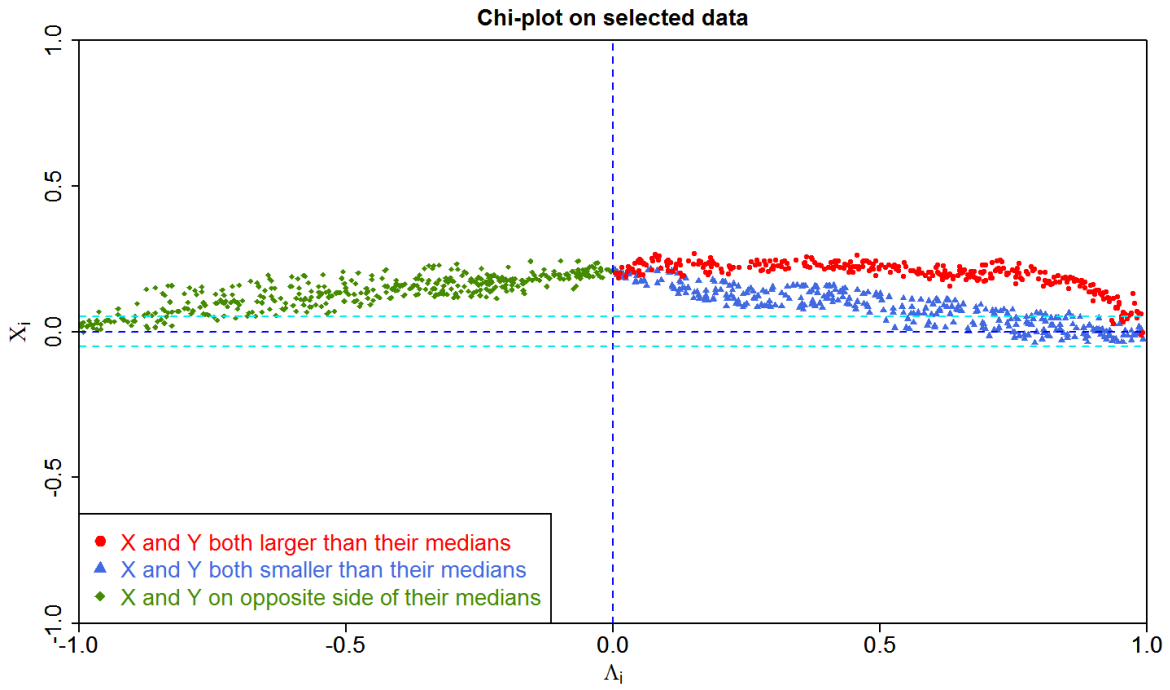

Figure 35. Modified Hs / surge chi-plot of the i.i.d. sample

The change in the results (Figure 36) can be observed on the extreme values of Z, which is now modelled using an indirect approach (JPM), and on the curvature of the contours, which directly depends on dependence.

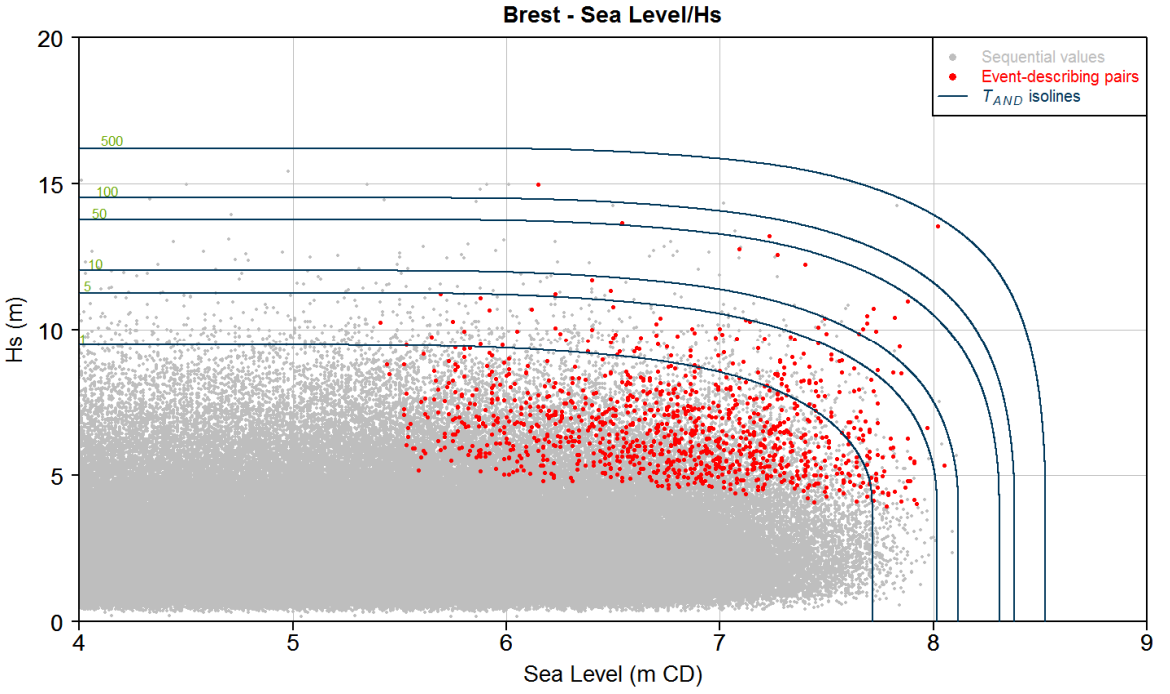

Figure 36. Contours of joint Hs / sea level return periods for the extended bivariate methodology

Source of data: cf. Figure 6

3.3. CONSIDERATIONS ON RETURN PERIODS

The two publications presented above, (**[P]** and **[S]**), describe methodologies for estimating the probability distributions of the sum of two components, or the joint occurrence of two quantities that are not of the same kind. Dependence is dealt with thanks to sophisticated statistical tools such as normalization by band, extremal index, extreme value copulas and so on. Different families of distributions and different estimators were compared.

But the main issue is still the correct interpretation of the analysis output. We present below different considerations and outputs that have not been published yet but that are the results of further work and engineering studies during 2017.

3.3.1. What is a multivariate return period?

The univariate return period can be given by Equation (2) because the definition of the event is generally unambiguous: it is an exceedance.

However, two variables offer a wider range of possible combinations to define the event A., i.e. the subset of all possible outcomes. ***Serinaldi (2015)*** lists the possibilities of probabilities to consider, which include the following:

$$p_{AND} = \mathbb{P}[U > u \cap V > v] = 1 - u - v + C(u,v) = \frac{\mu}{\mathcal{T}_{AND}} \tag{9}$$

$$p_{OR} = \mathbb{P}[U > u \cup V > v] = 1 - C(u,v) = \frac{\mu}{\mathcal{T}_{OR}} \tag{10}$$

$$p_{COND1} = \mathbb{P}[V > v \,|U > u] = \frac{1 - u - v + C(u,v)}{1 - u} = \frac{\mu}{\mathcal{T}_{COND1}} \tag{11}$$

$$p_{COND2} = \mathbb{P}[V > v \,|U \leq u] = 1 - \frac{C(u,v)}{u} = \frac{\mu}{\mathcal{T}_{COND2}} \tag{12}$$

$$p_{COND3} = \mathbb{P}[V > v \,|U = u] = 1 - \frac{\partial C(u,v)}{\partial u} = \frac{\mu}{\mathcal{T}_{COND3}} \tag{13}$$

where $\mu = 1/\lambda$ is the average inter-arrival time between two events, X and Y are two random variables, F_X and F_Y their marginal distributions, C their copula and $U = F(X)$ and $V = G(Y)$ are standard uniform random variables that allow us to work in the unit square $[0,1]^2$ where C is defined. In particular, we can reformulate the bivariate joint distribution function of X and Y as follows: $H(x,y) = C\big(F_X(x), F_Y(y)\big) = C(u,v)$.

These probabilities are illustrated in Figure 37. The bold black lines delimit the domains in which probability is computed (universe Ω), while the grey areas denote the critical regions fulfilling the condition related to each type of probability (event A).

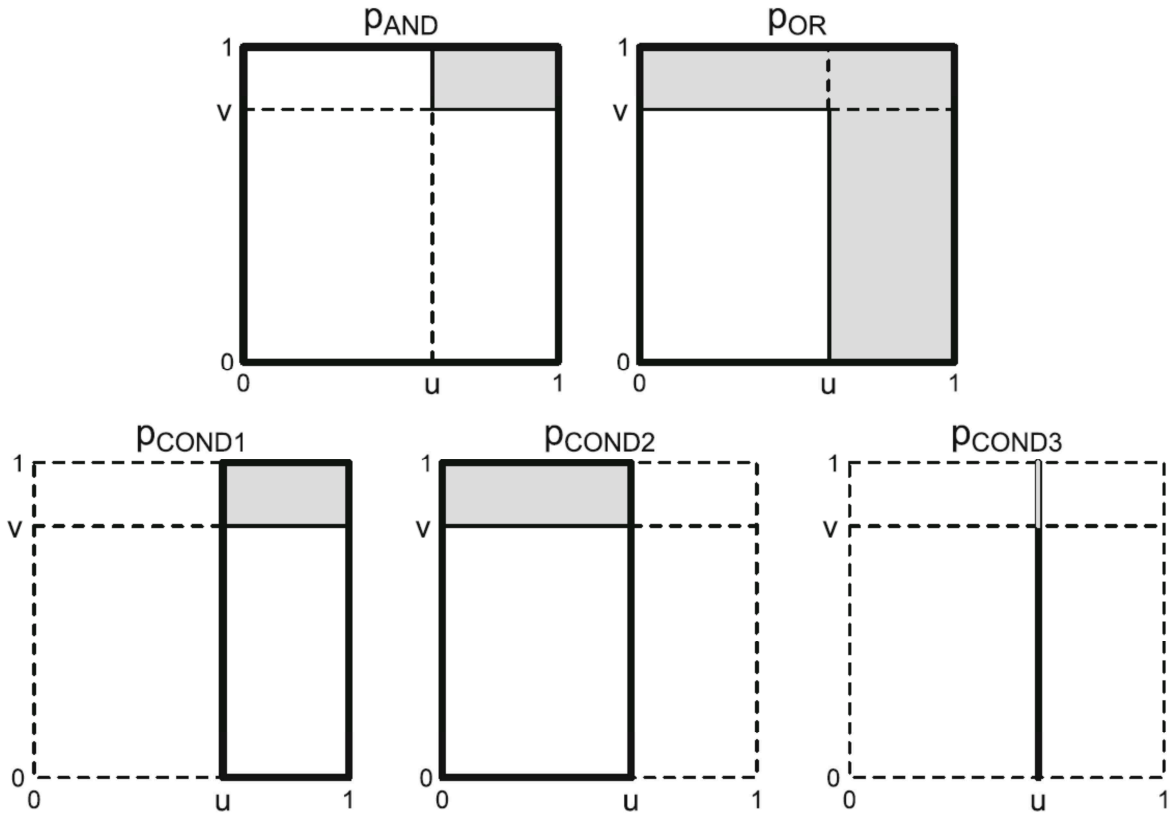

Figure 37. Domains and critical regions of several bivariate probabilities
source: Serinaldi (2015)

Considering the univariate return periods $\mathcal{T}_X$ and $\mathcal{T}_Y$ of X and Y and their associated probabilities of exceedance p_X and p_Y, the following relations can easily be derived (***Yue and Rasmussen, 2002***):

$$p_{OR} \geq \max(p_X, p_Y) \geq \min(p_X, p_Y) \geq p_{AND} \tag{14}$$

$$\mathcal{T}_{OR} \leq \min(\mathcal{T}_X, \mathcal{T}_Y) \leq \max(\mathcal{T}_X, \mathcal{T}_Y) \leq \mathcal{T}_{AND} \tag{15}$$

However, although the above relations are mathematically true, they are the results of the ordering and comparison of events whose ordering and comparison is somewhat meaningless. A choice thus has to be made between these different probabilities, depending on the aim of the study. The plots in Figure 32 and Figure 36 correspond to the $\mathcal{T}_{AND}$ definition.

In particular, this choice actually has to be made between the source variables and the response variables: the conditions of the probability (the definition of event A) may no longer depend on the source variables X and Y, but on a response variable $Z = g(X, Y)$. For instance, ***Volpi and Fiori (2014)*** defined a so-called "structure-based" return period:

$$p_S = \mathbb{P}[Z > z] = \mathbb{P}[g(U, V) > z] = 1 - F_Z(z) = \frac{\mu}{\mathcal{T}_S} \tag{16}$$

g is a function that links the source (environmental forcing) variables X and Y to one specific response (structural) variable Z. It can be seen that this is a reduction from the bivariate case back to the univariate definition of the probability of exceedance and the associated $\mathcal{T}$.

3.3.2. Bivariate return period of source variables vs. univariate return periods of response variables

The return period $\mathcal{T}_S$ associated with the (structural) response function is thus different from the joint return period associated with the source variables $\mathcal{T}_{AND}$. It is generally lower. We will illustrate this using two examples.

First let us consider the case of overtopping assessed on the basis of joint probabilities of wave height and sea level, as illustrated in Figure 38 from ***Hawkes et al. (2002)***. Let the design criteria be defined by the 10-year overtopping volume. Thanks to the methodology defined in this paper, we can accurately estimate the contour of the 10-year joint return period $\mathcal{T}_{AND}$ of H_s and Z. Along this contour, a single pair (H_s, Z) is associated with a worst case for overtopping (circled dot). We can then draw the contour of overtopping corresponding to this value[11], and we see a discrepancy between the hatched area of the joint exceedance probability of H_s/Z and the dotted area of overtopping probability. The latter is larger: it means the probability of the overtopping value being exceeded is higher and, as a consequence, its return period is lower.

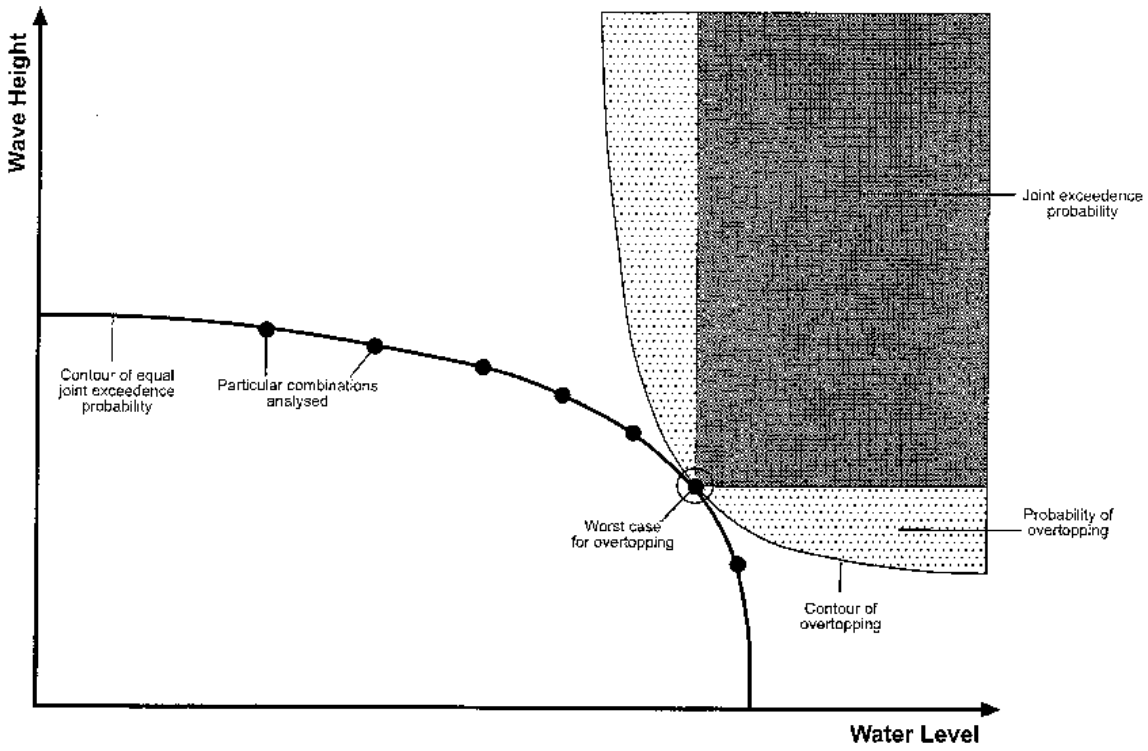

Figure 38. Joint exceedance probability and structure variable probabilities

source: Hawkes et al. (2002)

Another illustration is provided in Figure 39 with a very simple bivariate analysis H_s/Z, without accounting for covariates of any kind, and a very simple response function, $H_s + Z$. The extreme values can be extrapolated directly for this return function and drawn as contour lines in the (H_s, Z) plane. A colour code based on these extreme values is also applied to the event points. We find that the lines of equal univariate return period are almost tangential to the contours of equal joint

[11] For a given structure, overtopping depends on wave height and sea level, but also on wave period and obliquity. So in order to draw the contour of overtopping, assumptions must be made on the wave period and direction values to associate with each wave height value (or a full multivariate analysis). This is why a simpler example in which no covariates are involved is presented hereafter.

return period: when running along a bivariate return period contour, the univariate return period is lower, and equal (or close) on a single point.

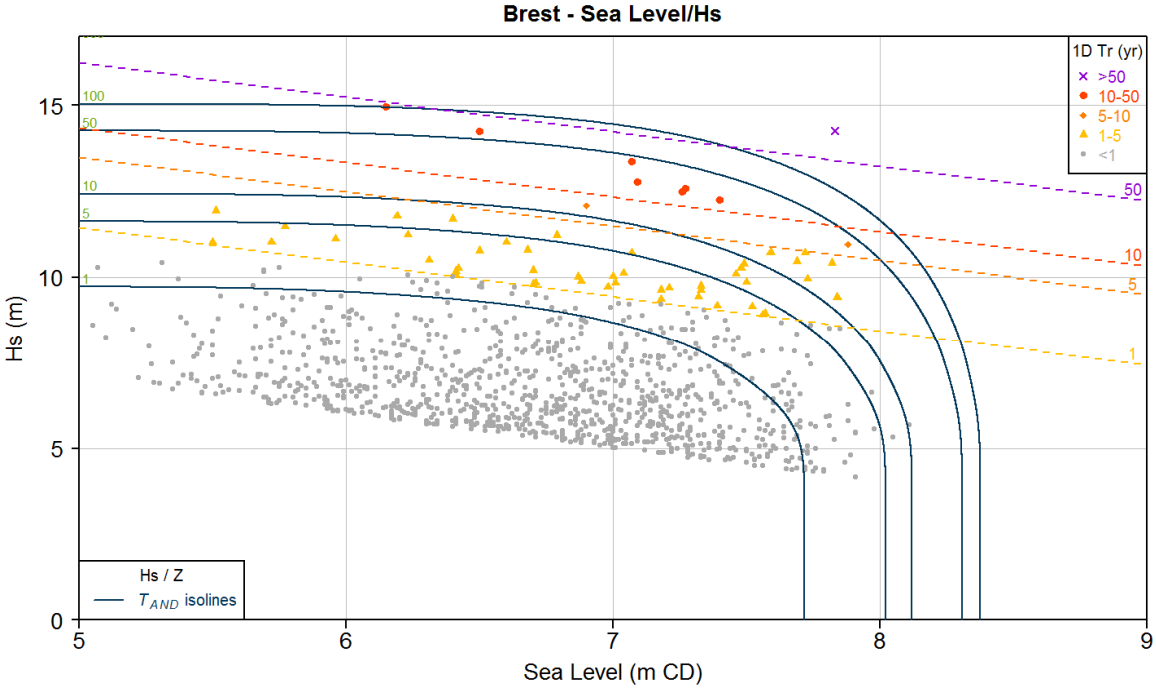

Figure 39. Contours of joint exceedance probabilities and lines of equal univariate return period of the response function (total water level)

Source of data: cf. Figure 6

3.3.3. Return periods and contours

3.3.3.1. Contours for event-describing values

It may be requested in some engineering studies to draw bivariate contours for return periods of 1, 10, 100 years or more. This is generally on account of a need to define a large number of load cases for the design of the structures. Because there may be different kind of loads to study, it is generally implicit that the return period refers to the source variables, and that it is $\mathcal{T}_{AND}$.

We defined return period hereinabove as the inverse probability of exceedance (univariate case) or as a joint or conditional exceedance (multivariate case), expressed in years using the Poisson parameter. A bivariate contour is related not to a notion of exceedance, but to one of inside / outside.

A proposed answer consists in drawing the contour associated with the quantile of the given return period of one of the source variables, named the **reference variable**. This kind of contour has been proposed by other authors (***Galiatsatou and Prinos, 2007***) but the need to choose the reference variable does not appear clearly. More specifically, for a return period of $\mathcal{T}$ years and with Y being the reference variable, the method is as follows. First, the $\mathcal{T}$-year quantile $y_{\mathcal{T}}$ of the reference variable is determined from its marginal distribution F_Y. Then the most probable value of X given $Y = y_{\mathcal{T}}$ is selected: this is the maximum of the joint density of X and Y along the line $Y = y_{\mathcal{T}}$. A contour of the density $p_{X,Y}(x_{max}, y_{\mathcal{T}})$ can then be drawn in the (X, Y) plane.

This approach is illustrated in Figure 40 for a bivariate analysis of storm duration / storm peak H_s off Cotonou, Benin. The top plot provides the $\mathcal{T}_{AND}$ isolines calculated on the event-describing pair H_s-peak / storm duration. The bottom plot displays bivariate contours, H_s being the reference variable. The horizontal dashed lines represent the marginal extreme values of H_s. The x-

coordinate of the maximum density values along these lines corresponds to the most probable storm duration for this H_s quantile and a contour of joint density can be drawn.

These two plots are drawn from the exact same joint distribution $H_{X,Y}$, just as a univariate distribution can be illustrated by its density, its cumulative distribution function or its survival function. We can understand from these figures that when contours of $\mathcal{T}_{AND}$ are horizontal or vertical, this means that the probability of joint exceedance p_{AND} does not vary along this part of the contour: this is because the joint density is nil in these areas. For instance, according to the plot below, the probability of observing a $(H_s = 3, Duration = 25)$ pair is negligible.

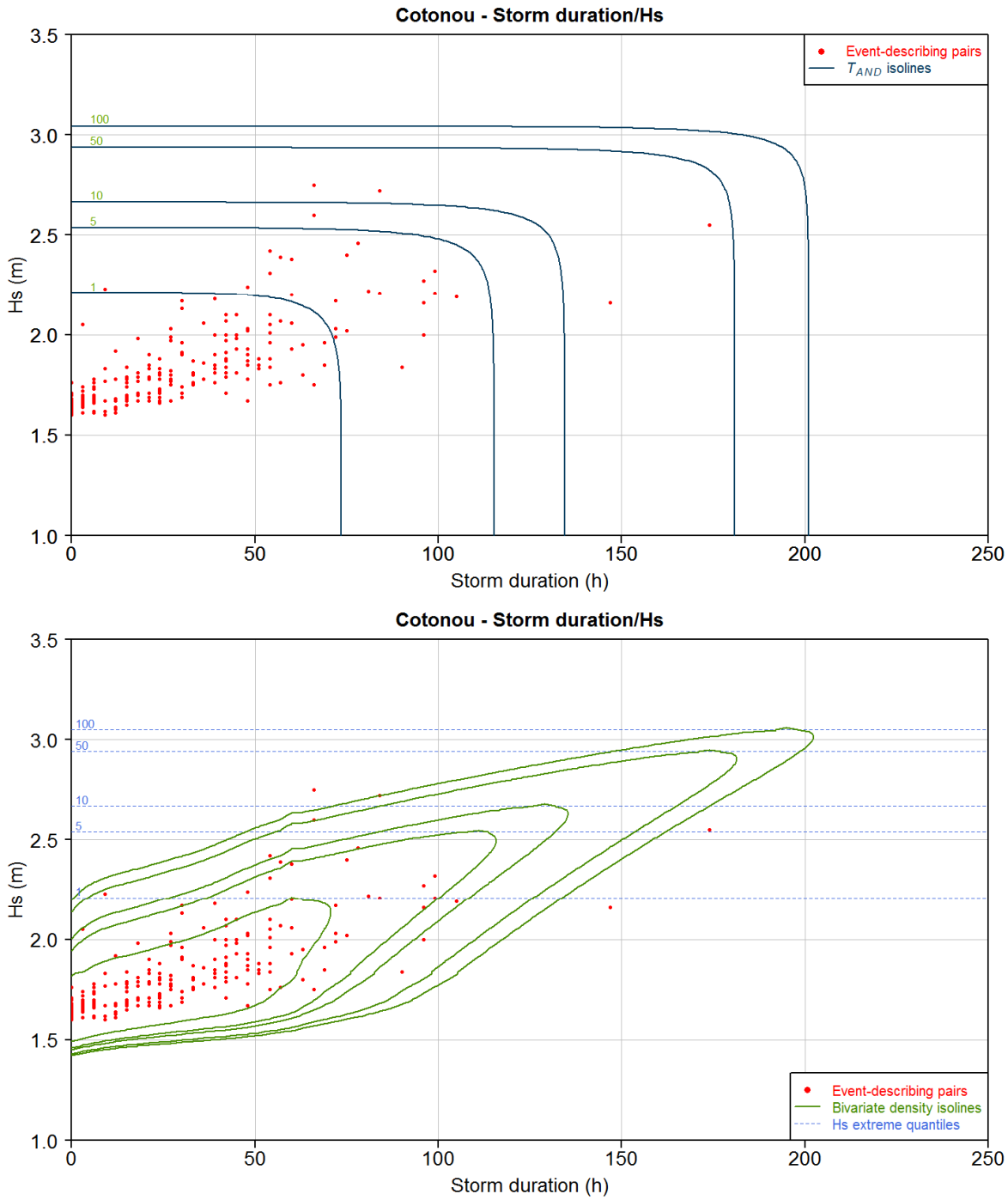

Figure 40. Top: isolines of joint return period; bottom: contours of density associated with extreme Hs for Hs / storm duration bivariate analysis

Source of data: 1992-2007 sea state hindcast by GlobOcean for ARTELIA (2008)

Interpreting the plots above is fairly straightforward (or at least not too confusing) because the contours are drawn from the joint density of the event-describing pairs. Whereas the peak value of H_s during a storm can be confused with the sequential values of H_s, storm duration is an event-describing variable that has no sequential equivalent. It is hence easy to understand that the analyses correspond to events only. Furthermore, this is a Type A analysis (one single

phenomenon described by several quantities) and the reference variable that defines the storm (here H_s) is obvious.

However, things become more complicated when the following contours are requested:

- in Type A analyses, two event-describing variables that directly derive from a sequential variable, such as H_s/T_p analyses;
- in Type C analyses, two variables describing different phenomena, each of which being legitimate to be a reference variable defining an event, such as wave height / wind speed analyses.

In the former case, the reference variable (event-defining variable) remains generally univocal; whereas in the latter case a choice is to be made, as will be illustrated further below.

3.3.3.2. Contours for sequential values

However, the contours may be requested no longer merely on events but for the full range of sequential observations. This is the bivariate equivalence of the univariate mixture model presented in section 3.1 and illustrated for the surge distribution in Figure 23. Of course, defining a bivariate tail is not straightforward and connecting the bulk of the distribution to this tail while handling the transformation from event-describing pairs to sequential values presents serious difficulties. Furthermore, we have seen that the concept of return period is linked to the concept of event: here it is used, and somehow confused, for the sequential values. This contradiction will be circumvented as follows: the bivariate density is derived for the sequential pairs, and the maximum value of the sequential density along the line associated to the extreme quantile of the event-describing reference variable is used for drawing the corresponding contour.

A preliminary method is exposed below, illustrated by a H_{m0}/T_p analysis off the coast of Groix island (French Atlantic coast). Hourly values of spectral sea state parameters are available for a 20-year duration (175,320 sequential pairs). A POT declustering is applied on H_{m0} and a sample of 297 peaks (event-describing pairs) is selected, as shown in Figure 41 below.

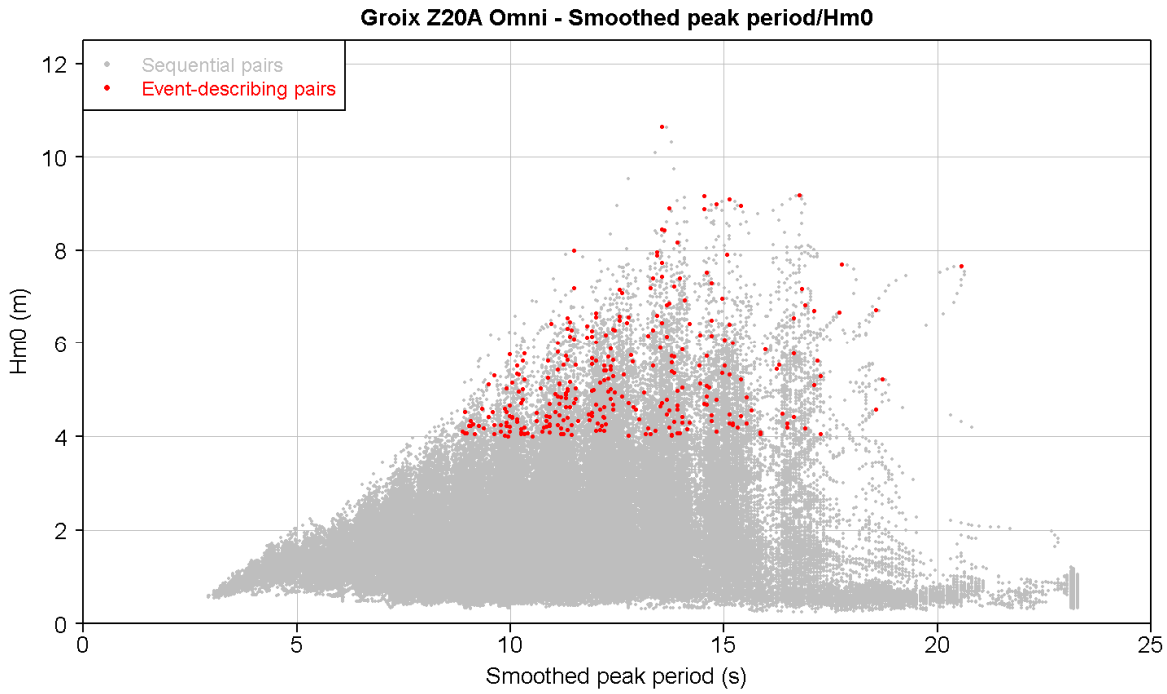

Figure 41. Sampling of the Hm0/Tp time series off Groix: scatterplot of the sequential pairs (grey dots) and event-describing pairs (red dots)

Source of data: 1996-2015 sea state hindcast by GlobOcean for ARTELIA (2016)

Figure 42 is the heat map (the highest values are transparent, and the lowest are in red) of the density of the parametric joint distribution of the event-describing pairs $(H_{m0}; T_p)$. From this bivariate density, it is easy to draw the isolines of density associated to the extreme H_{m0} peaks similarly to Figure 40, but we now aim at obtaining the joint distribution of the sequential pairs displayed in Figure 41.

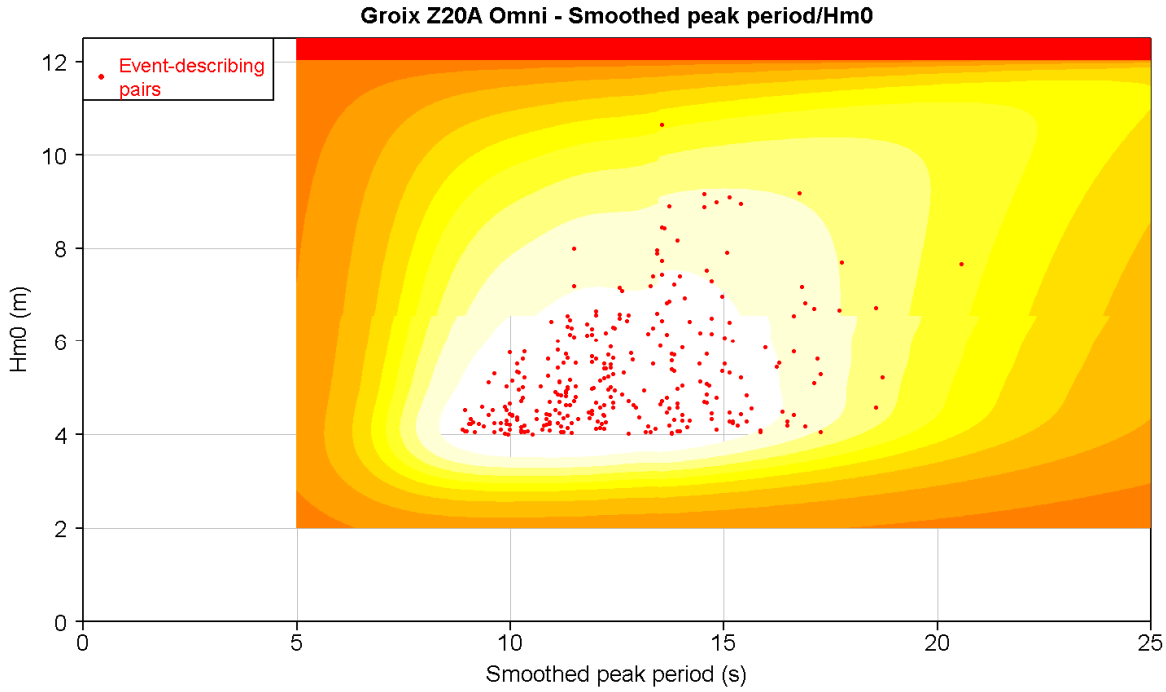

Figure 42. Joint density Hm0/Tp for event-describing pairs

Source of data: 1996-2015 sea state hindcast by GlobOcean for ARTELIA (2016)
Heat map: joint density values, from red (lowest) to transparent (highest)

The proposed approach consists in connecting the parametric distribution from the joint distribution of events to the empirical density estimated on the sequential (e.g. hourly) observations. It is thus necessary to define a parametric domain and an empirical domain. We propose a simple definition: the parametric domain is defined by the joint exceedance of a threshold in H_{m0}, noted $u_{H_{m0}}$, and a threshold in T_p, noted u_{T_p} (upper right-hand corner of the H_{m0}/T_p plane, similar to domain 2 in Figure 26). Within this domain, the joint density of the event-describing pairs needs to be transformed into the joint density of the sequential pairs belonging to this domain. This is done by multiplying the joint density of events by a factor equal to the ratio of the number of sequential pairs belonging to this domain over the total number of observations. In other words, if 5% of the pairs are located within the limits of the parametric domain, then the integration of the joint density within this domain is 0.05, while the integration within the empirical domain is 0.95.

It is thus necessary to set $u_{H_{m0}}$ and u_{T_p} in order to define this parametric domain. At first it seems natural to use the physical or statistical thresholds used for event identification or statistical extrapolation of H_{m0} and T_p, but numerous sensitivity tests have led us to set the following thresholds:

- $u_{H_{m0}}$: a value above which 10 storms/year on average are observed;
- u_{T_p}: the minimum period associated with storms that occur 5 times/year on average.

The density of the sequential pairs within the parametric domain is illustrated in Figure 43.

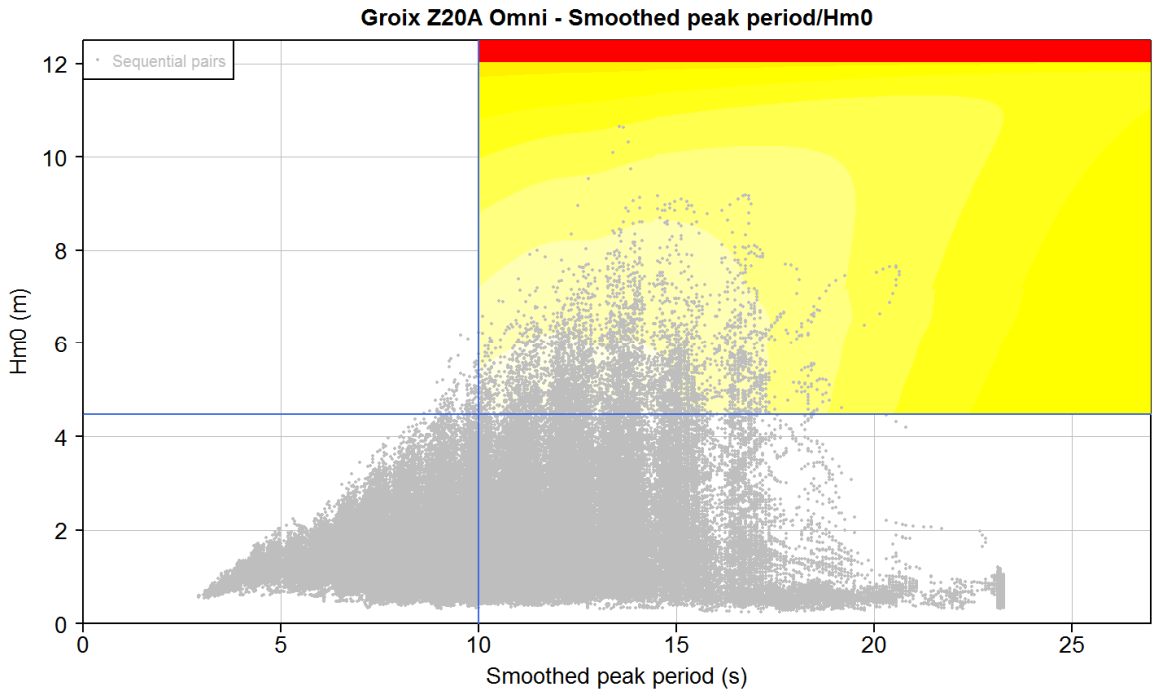

Figure 43. Parametric domain of the joint distribution for sequential pairs of Hm0 and Tp

Source of data: 1996-2015 sea state hindcast by GlobOcean for ARTELIA (2016)
Heat map: joint density values, from red (lowest) to transparent (highest)

The total joint distribution of hourly values of H_{m0} and T_p is thus the combination of the empirical bivariate density for $H_{m0} < u_{H_{m0}}$ or $T_p < u_{T_p}$ and the parametric joint density of the events for $H_{m0} \geq u_{H_{m0}}$ and $T_p \geq u_{T_p}$. Figure 44 shows:

- the sequential pairs (H_{m0}, T_p) in grey;
- the colour map of the bivariate density of (H_{m0}, T_p):
  - the parametric domain in the upper right-hand corner (i.e. the density displayed in Figure 43),
  - the empirical domain in the other three corners with the empirical bivariate density estimated by the kernel density estimator.

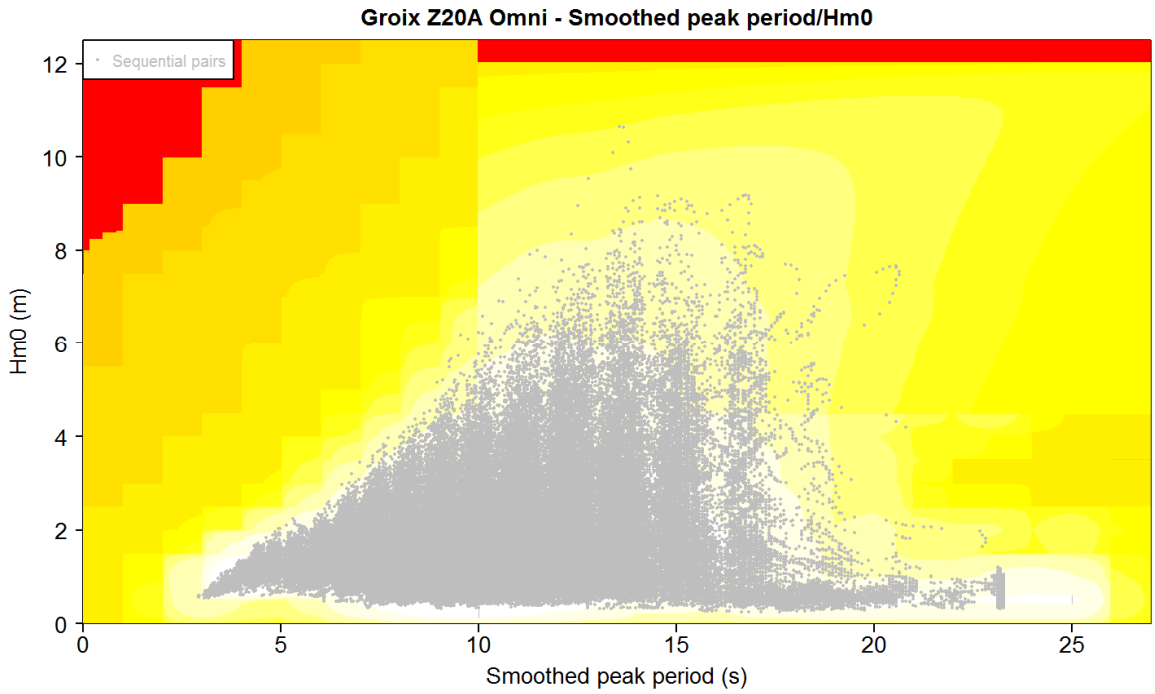

Figure 44. Joint distribution for sequential pairs of Hm0 and Tp

Source of data: 1996-2015 sea state hindcast by GlobOcean for ARTELIA (2016)
Heat map: joint density values, from red (lowest) to transparent (highest)

The isolines of bivariate density associated with the maximum of density along the lines of the extreme quantiles of the reference variable H_s can then be plotted, as previously described (Figure 45).

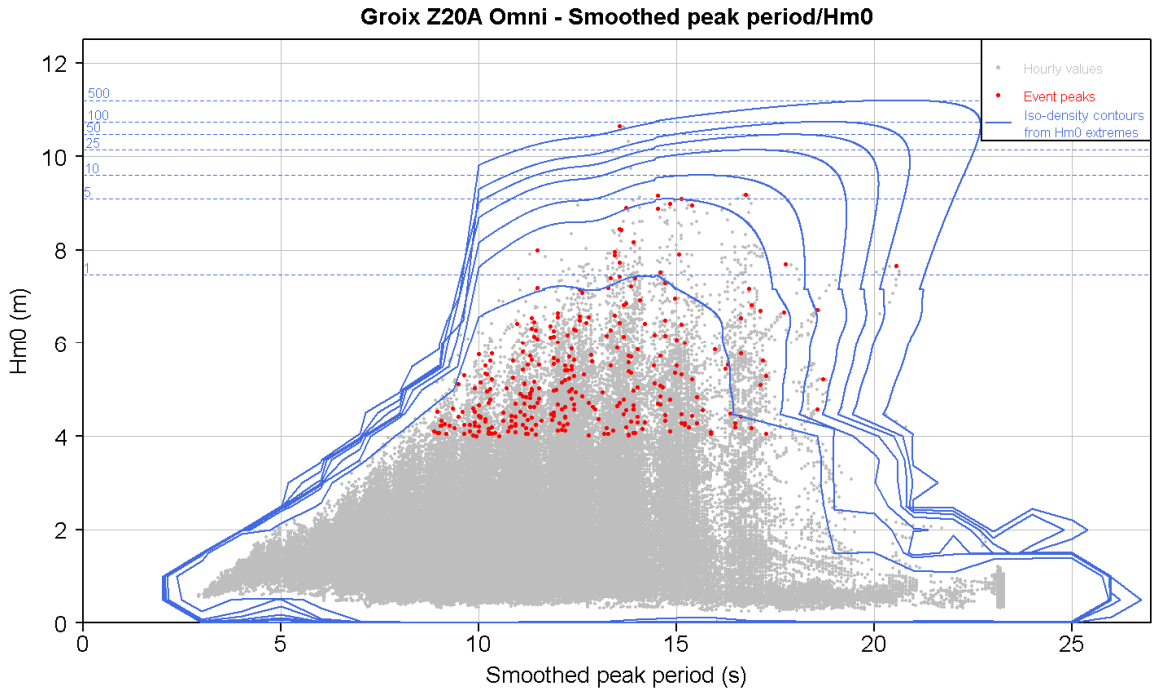

Figure 45. Contours of sequential Hm0/Tp density associated with extreme Hm0 peaks

Source of data: 1996-2015 sea state hindcast by GlobOcean for ARTELIA (2016)

However, it can be seen that the connections between the parametric and empirical domains are far from smooth. This is due partly to the clustering of the smoothed peak periods around the discrete values from the numerical modelling, partly to the mixture of populations that can easily be observed with the presence of both steep wind waves and long swells and lastly to the fact that the connecting approach is not quite accurate and requires improvements. Figure 46 shows, however, that the connection looks much better when the sample can be considered identically distributed, thanks for instance to a separation in fetch-based directional sectors.

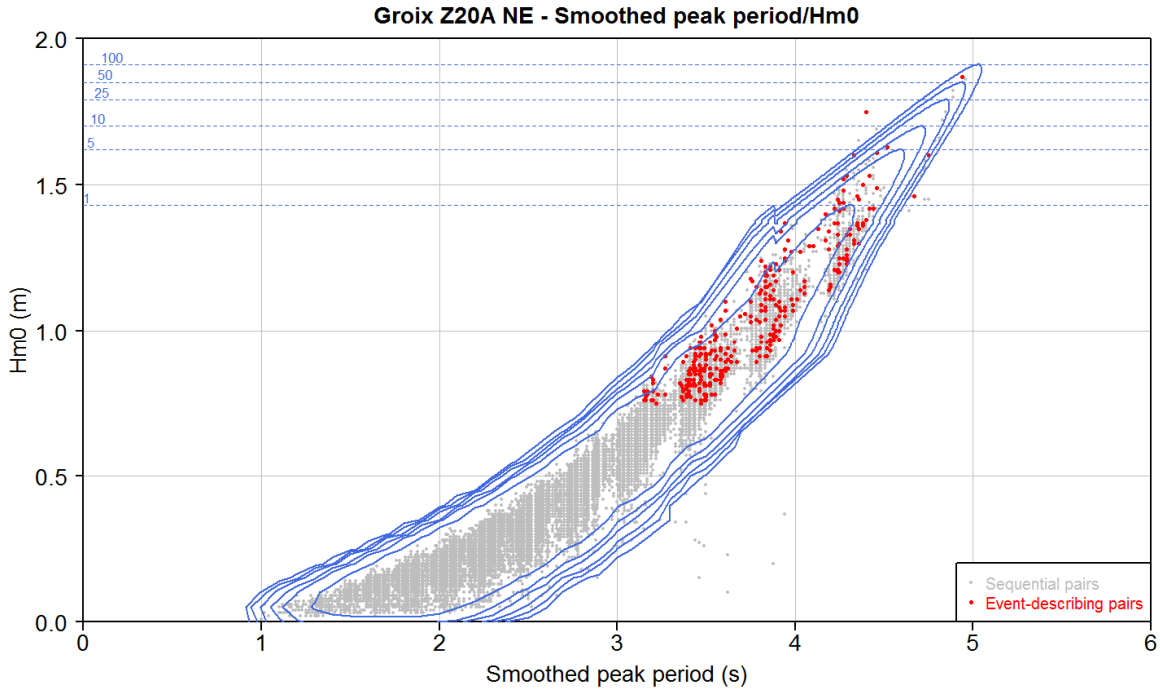

Figure 46. Contours of sequential Hm0/Tp density associated with extreme Hm0 peaks for a homogeneous directional sector

Source of data: 1996-2015 sea state hindcast by GlobOcean for ARTELIA (2016)

In the case of contours for Type-C analyses, we are faced once again with the issue of event definition and, hence, with the choice of an event-defining variable. This choice drives the sampling; thereafter, the methodology is similar to that described above. Two plots may hence be issued, as illustrated in Figure 47 for H_{m0} / wind speed W_s bivariate analyses.

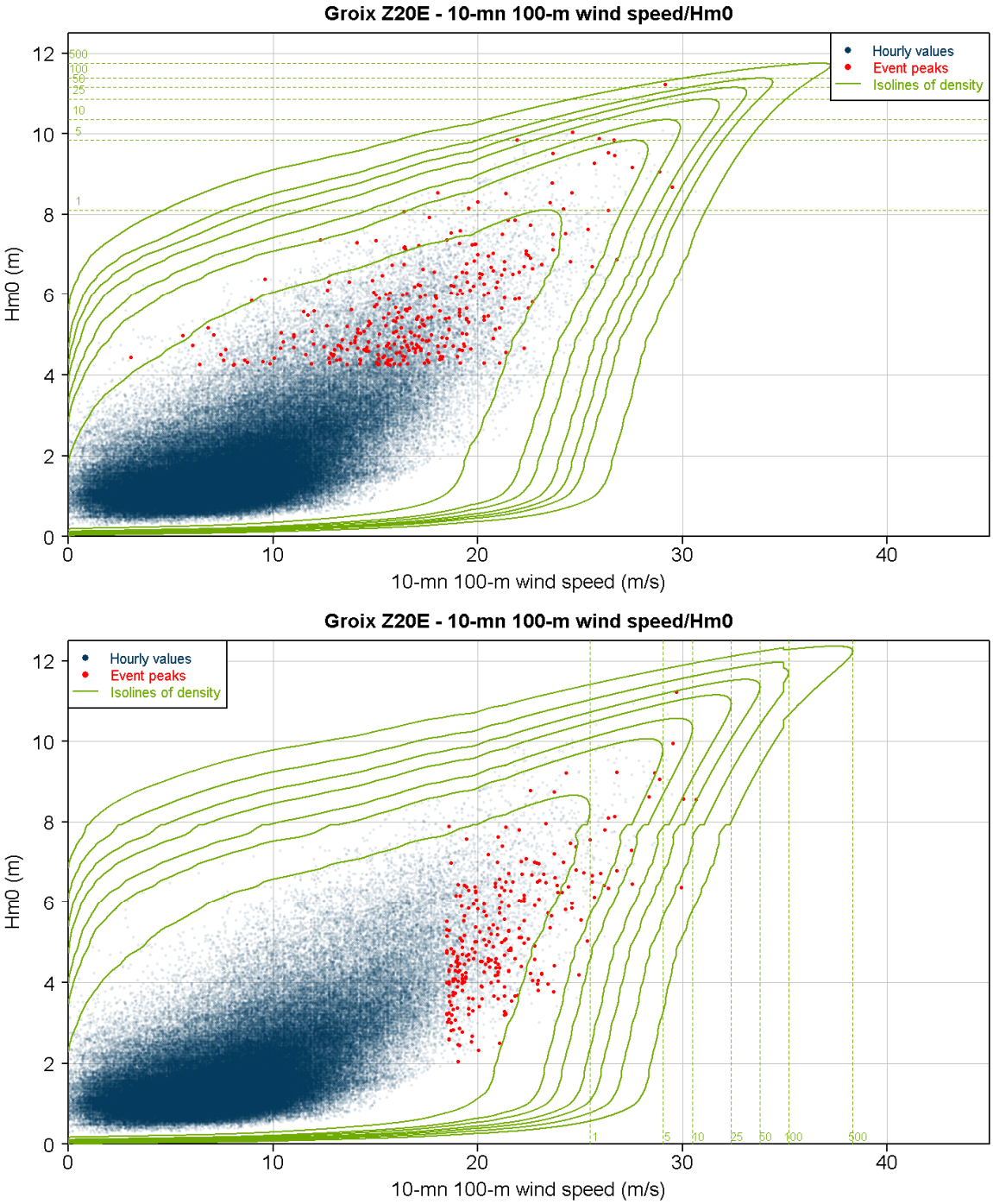

Figure 47. Contours of sequential Hm0/Ws density associated with extreme Hm0 peaks (top) and extreme wind speed peaks (bottom)

Source of data: 1996-2015 sea state hindcast and CFSR wind fields calibrated against satellite measurements by GlobOcean for ARTELIA (2016)

In the top plot, sampling was performed on the H_{m0} variable only, using the univariate declustering method presented in section 2.2. The concomitant wind speed is associated with the peak H_{m0} to constitute the event-describing pair. The marginal distributions, copula and joint distribution are then determined from this sample as presented above, and the contours of bivariate density are then drawn from the extreme quantiles of H_{m0}.

The method is exactly the same for the bottom plot, but the roles of H_{m0} and W_s have been exchanged. W_s is now the event-defining variable and the sampling is based on a declustering on wind speed peaks, with the associated H_{m0}. As a consequence, there is a difference not only in the reference variables between the two plots (from the marginal distribution of which we calculate the extreme quantiles and then find the maximum density and draw the contours), but also in the event-defining variables: this means that the joint distribution is not identical. In the former case, we are interested in wind speeds associated with wave events; in the latter we are interested in wave heights associated with wind events.

This is our final illustration of the importance of event definition, identification and description in extreme meteo-oceanic analyses (and in extreme environmental analyses more generally), and of the crucial fact that this choice to be made by the analyst depending on the aim of the study, takes place in the sampling (declustering) stage.

oOo

4. CONCLUSIONS AND PERSPECTIVES

4.1. MAIN RESULTS

The main results of the works carried out in the last decade are as follows:

- updating of the methodology for determining extreme wave heights or wind speeds:
  - development and justification of a two-step framework for extreme univariate over-threshold modelling introducing the concept of event and the separation of the physical and statistical thresholds,
  - proposal of practical tools for choosing the statistical threshold,
  - introduction of the parametric bootstrap approach for computing confidence intervals,
  - identification of a problematic issue in the behaviour of the Maximum Likelihood Estimator and proposal of a solution: use of 3-parameter distributions along with the L-moments estimator,
- application of the POT framework to the Joint Probability Method for determining extreme sea levels:
  - distinction between sequential values and event peaks through extremal indexes for surge and sea level,
  - construction of a mixture model for the surge distribution,
  - refinements for handling tide-surge dependence,
- application of the POT-JPM framework for the joint analysis of wave height and sea level:
  - proposal of an alternative sampling procedure,
  - separate analysis of tide and surge in order to model the dependence between wave height and surge to be incorporated in the joint distribution of wave height and sea level thanks to a 2D1D convolution operation,
  - use of extreme-value copulas,
  - improved presentation of the chi-plot,
- introduction of a new classification for multivariate analyses:
  - **Type A**: a single *phenomenon* described by different *physical quantities* that are *possibly not of the same kind*,
  - **Type B**: a *phenomenon* made of different *components*, described by *physical quantities of the same kind* between one component and another,
  - **Type C**: several *phenomena* described by *physical quantities* that are *possibly not of the same kind*,
- interpretation of the meaning of multivariate events:
  - link with the sampling procedure,
  - link with the different definitions of the return period,
  - in the bivariate case: transformation of the joint distribution of event-describing variables into the joint distribution of sequential pairs,
  - generation of alternative output plots such as contours of density for sequential pairs;
- a dedicated `R` package, `artextreme`, for implementing the methodologies presented above.

4.2. DISCUSSION

Even though it took years to appear clearly, the concept of "event" became central to our developments of probabilistic methodologies for determining extreme meteo-oceanic conditions. The main result of this centrality was a very practical and convenient approach in the univariate case. But when considering multivariate events, it was not merely useful; it required clear choices to be made in order to pursue the analysis, which made it essential.

Far beyond the important but *in fine* quite technical matters pertaining to choosing the distributions, determining the statistical threshold, the estimator, the computation of confidence intervals, the copula, etc. - all topics that can be discussed, improved and substituted - we believe that the originality of the work presented here consists of the intimate link between sampling, event definition and description, and the definition adopted for the return period, probability of exceedance or probability of encounter.

These steps cannot be automated in a universal method but must be tailored completely to the final objectives of the study in hand, which is precisely what is expected of an engineer. This is why the event definition was stated to be a matter of point of view.

Among many possible images, the following framework may be helpful to comprehend this: the Source-Pathway-Receptor (SPR) concept. It was forged originally to describe the flow of environmental pollutants from a source through different pathways to potential receptors (***Holdgate, 1979***) and later applied to coastal flooding by the UK Environment Agency (***HR Wallingford, 2002***) to describe the propagation of a flood from a source through flood defenses (pathways) to the floodplain beyond (receptors). We propose here to use this framework to visualize the different possibilities for defining the event. In our latest example, the *sources* may be the atmosphere (wind and pressure fields) and the astronomical forcing, the *pathway* would be the ocean (where the energy from the atmosphere and celestial bodies propagates through short and long waves) and the *receptor* would be the breakwater where overtopping occurs. Here, it is more intuitive that the return period associated with a *pathway event* (joint probabilities of waves and sea level) will differ from the one associated with a *receptor event* (overtopping) or *source event* (storm or exceptional syzygy[12]). Of course we may shift the terms of the SPR and consider the ocean as the source, the overtopping as the pathway and the buildings behind the levees as the receptors, but this does not change the idea.

Whatever the way to visualize this, it should always be made clear in engineering whether the design criteria are to be based upon the source variables or the response function. This will drive the choices of event, sampling and, possibly, methodology.

Lastly, the qualification of multivariate may be discussed. Indeed, the dimensions of the joint distributions and copulas have not exceeded two and it may be argued that only bivariate analyses have been presented up to now. This is true; however, we have seen how to select multiple covariates in addition to the two main source variables. They could be dealt with using n-dimensional copulas, of course, but also as covariates upon which the distribution parameters will depend. The latter approach has been widely explored by authors such as ***Jonathan et al. (2013)***. As regards the former, we do not have any knowledge of the use of genuine n-dimensional copulas in the field of meteo-oceanic extremes. Research is active on that topic, however, and a very good reference is ***Li et al. (2014)***.

[12] In astronomy, a syzygy (from the Ancient Greek σύζυγος "yoked together") is a straight-line configuration of three celestial bodies in a gravitational system.
Syzygy causes the bimonthly phenomena of spring and neap tides. At the new and full moon, the Sun and Moon are in syzygy. Their tidal forces act to reinforce each other, and the ocean both rises higher and falls lower than the average.
When a syzygy occurs when the Moon is close to its perigee and / or when it is crossing the ecliptic plane, the combination of the gravitational effects is further enhanced.

4.3. PERSPECTIVES

In the illustration above, it may be noted that the receptor (the breakwater) is local while the source event (storm) does not only extend in time but also in space. This is why the idea of going further into the spatial definition of event is quite appealing. A fair amount of work has already been done at the LHSV with the thesis of Jérôme Weiss (***Weiss et al., 2014***, ***Weiss, 2014***, chapter 3), notably regarding the spatio-temporal declustering of meteo-oceanic storms that allows for a broader view of the regional frequency analyses (***Hosking and Wallis, 1997***, ***Bardet et al., 2011***, ***Bernardara et al., 2011***) A logical and conceptual link with the concepts exposed here would certainly be most welcome.

Spatial extension naturally calls for temporal extension in addition. The interest of historical data has already been highlighted (see section 1.1.3). At the LHSV, a new thesis is ongoing with the efforts of Roberto Frau regarding the use of historical data in regional analyses (***Frau et al., 2017***). There again, the definition of event in the context of incomplete historical data (such as the number of events within a range of years, and the intensity of an event only known as a range or a minimal value) would probably require some clarifications.

Estimating the uncertainties is also a key part of extreme value analyses and they must be accounted for in engineering studies, particularly for design purposes. Introducing the parametric bootstrap method for estimating the confidence intervals was a first step forward, as it avoids obtaining symmetrical confidence intervals that are physically meaningless. However, the very framework of frequentist inference with the associated confidence intervals can be questioned and replaced with a Bayesian approach (see the discussion in section 1.3.2), which naturally integrates uncertainties in the inference process. The Bayesian framework is widely used in hydrology (***Gaume et al., 2010***) and is beginning to appear in maritime studies (***Egozcue et al., 2005***, ***Galiatsatou et al., 2008***, ***Bulteau et al., 2015***). In any case, it is always necessary to keep in mind the multiple sources of uncertainties:

- uncertainties relating to the data (accuracy of measurements or model outputs);
- uncertainties relating to the sampling (duration of the time series / sample size);
- uncertainties relating to sample representativeness (non-stationarity in particular);
- uncertainties relating to the choice of model.

In particular, the uncertainties relating to the choice of model were highlighted by an experiment carried out by the OSSË group at the initiative of Dr Kergadallan at a conference of the Société Hydrotechnique de France (SHF) in Lyons in 2013.

It was stated in section 0 that the mixture model for the joint distribution, particularly the connection between the parametric and empirical domains, was not very accurate. In particular, the dependence structure could be extended to the sequential pairs. This is definitely an interesting mathematical problem going forward.

Lastly, it has been said that even though the reasoning makes more than two variables intervene, the copulas and joint distributions used up to now do not go beyond the second dimension and are, strictly speaking, bivariate distributions.

The concepts and classifications presented here lay the groundwork for determining fully multivariate analyses that would include, in a single joint distribution, the different physical quantities describing two distinct phenomena: here again, a fine challenge for mathematicians who are aware of the beautiful complexities of the physics of extreme meteo-oceanographic events.

oOo

GLOSSARY

Symbol	Unit	Description
A	-	Probabilistic event (subset of Ω)
c	m.s^{-1}	Wave celerity
C	-	Copula function
d	-	Mean number of sequential values per year
D	yr	Lifetime duration
f	-	Univariate probability distribution function
F	-	Univariate cumulative distribution function
g	m.s^{-2}	Gravity acceleration
h	m	Water depth
H	-	Multivariate cumulative distribution function
H_s	m	Significant wave height
H_{m0}	m	Spectral significant wave height
k	-	Shape parameter
K	yr	Duration of a time series
N	-	Size of the sample of peek exceedances over u_s
N_p	-	Size of the sample of event-describing variables
p_X	-	Probability of exceedance of the random variable X: $p_X = 1 - F_X$
r	-	Pearson's correlation coefficient
S	m	Surge height (non-tidal residual)
t	s	Time
T	m CD	Astronomical tide level
T_p	s	Peak period of sea sate spectral or wave system density
u, v	-	Uniformly distributed marginal of the copula: $u = F_X(x), v = F_Y(y)$
u_p	-	Physical threshold for event identification
u_s	-	Statistical threshold
X, Y, Z	-	Event-describing variables
$Y = X - u_s$	-	Exceedance over the statistical threshold
Z	m CD	Sea level

θ	-	Vector of parameters of the statistical distribution
θ_p	°N	Peak direction of sea state or wave system spectral density
λ	yr^{-1}	Poisson rate parameter: mean number of threshold exceedances per year
λ_p	yr^{-1}	Mean number of events per year
λ_U	-	Upper tail dependence coefficient
μ	-	Location parameter
μ	yr	Inter-arrival time (reciprocal of Poisson parameter)
ρ	-	Spearman's correlation coefficient
σ	-	Scale parameter
τ	-	Kendall's correlation coefficient
ν	yr^{-1}	Number of sequential values per year
Ω	-	Probabilistic universe: set of possible outcomes
$\mathcal{T}$	yr	Return period
$\mathbb{P}$	-	Probability

REFERENCES

Published works

Publications in international peer-reviewed journals

[E] Mazas F., Hamm L., 2011. A multi-distribution approach to POT methods for determining extreme wave heights. *Coastal Engineering* **58**, 385–394.

Cited in 40 (Scopus) to 49 (Google Scholar) publications as of 20 September 2017

[F] Mazas F., Hamm L., 2012. Reply to Discussion by Z.J. You of "A multi-distribution approach to POT methods for determining extreme wave heights" by Mazas and Hamm. *Coastal Engineering* **65**, 16–18.

Cited in 2 (Scopus) to 2 (Google Scholar) publications as of 20 September 2017

[I] Bernardara P., Mazas F., Kergadallan X., Hamm L, 2014. A two-step framework for over-threshold modelling of environmental extremes. *Nat Hazards Earth Syst Sci* **14**, 635–647.

Cited in 17 (Scopus) to 25 (Google Scholar) publications as of 20 September 2017

[L] Mazas F., Garat P., Hamm L, 2014. Questioning MLE for the estimation of environmental extreme distributions. *Ocean Engineering* **92**, 44–54.

Cited in 3 (Scopus) to 3 (Google Scholar) publications as of 20 September 2017

[P] Mazas F., Kergadallan X., Garat P., Hamm L., 2014. Applying POT methods to the Revised Joint Probability Method for determining extreme sea levels. *Coastal Engineering* **91**, 140–150.

Cited in 6 (Scopus) to 7 (Google Scholar) publications as of 20 September 2017

[S] Mazas F., Hamm L., 2017. An event-based approach for extreme joint probabilities of waves and sea levels. *Coastal Engineering* **122**, 44–59.

Cited in 0 publication as of 20 September 2017

Publications in other journals

[B] Hamm L., Mazas F., Garcia N., Bailly B., 2010. Réconcilier théorie et pratique dans la détermination des houles extrêmes. *European Journal of Environmental and Civil Engineering* **14**, 127–148.

Cited in 1 (Scopus) to 1 (Google Scholar) publication as of 20 September 2017

[C1] Mazas F., Hamm L., 2010. Théorie statistique du renouvellement pour la détermination des houles extrêmes. Partie 1 : le point sur les méthodes disponibles. *La Houille Blanche* **4**, 96–102.

Cited in 1 (Scopus) to 1 (Google Scholar) publication as of 20 September 2017

[C2] Mazas F., Hamm L., 2010. Théorie statistique du renouvellement pour la détermination des houles extrêmes. Partie 2 : illustrations sur sites. *La Houille Blanche* **4**, 103–110.

Cited in 1 (Scopus) to 1 (Google Scholar) publication as of 20 September 2017

[M] Mazas F., Hamm L, 2011. Niveaux marins extrêmes en France : état des lieux. *La Houille Blanche* **5**, 5–11.

Cited in 0 (Scopus) to 2 (Google Scholar) publications as of 20 September 2017

International conferences with proceedings

[G] Bernardara P., Mazas F., Weiss J., Andreewsky M., Kergadallan X., Benoit M., Hamm L., 2012. On the two-step threshold selection for over-threshold modelling. *Coastal Engineering Proceedings* **1(33)**.

[K] Mazas F., Hamm L., Garat P., 2012. POT methods: a new insight into the estimation of extreme value distributions – Application to extreme wave heights. *Coastal Engineering Proceedings* **1(33)**.

[R] Mazas F., Hamm L., 2017. An event-based approach for extreme joint probabilities of waves and sea levels. *Coastal Engineering Proceedings* **1(35)**.

National conferences with proceedings

[A] Mazas F., Hamm L., 2008. Réconcilier théorie et pratique dans la détermination des houles extrêmes, in: Levacher, Daniel; Gaufrès, P.F. du L. (Ed.). Presented at the Xèmes Journées Nationales Génie Côtier - Génie Civil, Sophia-Antipolis, France.

Conferences without proceedings

[D] Mazas F., Hamm L., 2010. Estimation of extreme wave heights by POT methods revisited: sensitivity analysis to threshold and statistical distributions. Presented at the WISE 2010 - Waves In Shallow Environment, Brest, France.

[H] Bernardara P., Mazas F., Weiss J., Andreewsky M., Kergadallan X., Benoit M., Hamm L., 2013. A two-step framework for Over-Threshold Modeling of environmental extremes. EGU General Assembly 2013, Vienna, Austria.

[J] Mazas F., Garat P., 2011. A new insight into the estimation of extreme value distributions. Extreme Value Analysis (EVA) conference, Lyon, France, 2011.

[N] Mazas F., Hamm L., Kergadallan X., 2013. Extrapolation des niveaux d'eau extrêmes : incorporation de l'approche POT à la méthode des probabilités combinées. Journées REFMAR 2013, St-Mandé, France.

[O] Mazas F., Hamm L., Kergadallan X., 2013. Extrapolation of extreme sea levels: incorporation of Over-Threshold-Modelling to the Joint Probability Method. EGU General Assembly 2013, Vienna, Austria.

[Q] Mazas F., Hamm L., 2015. An event-based approach for extreme joint probabilities of waves and sea levels. EVAN 2015, Santander, Spain.

[T] Mazas F., Hamm L., 2017. An event-based approach for extreme joint probabilities of waves and sea levels. ICE 2017, Liverpool, United Kingdom.

[U] Mazas F., 2017. Multivariate events: definition, sampling and estimation of extreme values. EVAN 2017, Southampton, United Kingdom.

References from the scientific literature

[1] Aelbrecht D., Benoit M., Allilaire J., 2004. Renforcement de la protection contre l'inondation du front de Gironde sur le site du Blayais : apports conjoints des modélisations physique et numérique. *La Houille Blanche* **3**, 37-44.

[2] Bardet L., Duluc C.-M., Rebour V., L'Her J., 2011. Regional frequency analysis of extreme storm surges along the French coast. *Natural Hazards and Earth System Sciences* **11**, 1627–1639.

[3] Benoit M., Lafon F., Goasguen G., 2008. Constitution et exploitation d'une base de données d'états de mer le long des côtes françaises par simulation numérique sur 23 ans: Base ANEMOC en Atlantique-Manche-mer du Nord. *European Journal of Environmental and Civil Engineering*, **12**(1-2), 35-50.

[4] Bernardara P., Andreewsky M., Benoit M., 2011. Application of regional frequency analysis to the estimation of extreme storm surges. *Journal of Geophysical Research* **116**.

[5] Bertin X., Bruneau N., Breilh J.F., Fortunato A.B., Karpytchev M., 2012. Importance of wave age and resonance in storm surges: The case Xynthia, Bay of Biscay. *Ocean Engineering* **42**, 16-30.

[6] Bertin X., Prouteau E., Letetrel C., 2013. A significant increase in wave height in the North Atlantic Ocean over the 20th century. *Global Planetary Change* **106,** 77–83.

[7] Blitzstein J., Hwang J., 2014. *Introduction to Probability*. CRC Press.

[8] Boudière E., Maisondieu C., Ardhuin F., Accensi M., Pineau-Guillou L., Lepesqueur J., 2013. A suitable metocean hindcast database for the design of Marine energy converters. *International Journal of Marine Energy* **3**.

[9] Breysse D., 2009. UNIT: Project Cyber Ingénierie des Risques en Génie Civil. Chapter 2: Historique, vocabulaire, perception. Coordinator: H. Niandou.

[10] Bulteau T., Idier D., Lambert J., Garcin M., 2015. How historical information can improve estimation and prediction of extreme coastal water levels: application to the Xynthia event at La Rochelle (France). *Natural Hazards and Earth System Sci*ences **15**, 1135–1147.

[11] Charles E., Idier D., Thiébot J., Le Cozannet G., Pedreros R., Ardhuin F. and Planton S., 2012. Wave climate variability and trends in the Bay of Biscay from 1958 to 2001, *Journal of Climate* **25**, 2020-2039.

[12] Chelton D.B., Ries J.C., Haines B.J., Fu L. L., Callahan P.S., 2001. Satellite altimetry. *International Geophysics* **69**, 1-ii.

[13] Coles S., 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer.

[14] Cousin A., 2011. *Propositions pour une stratégie nationale de gestion du trait de côte, du recul stratégique et de la défense contre la mer, partagée entre l'Etat et les collectivités territoriale.* Rapport au gouvernement, 60.

[15] Cunnane C., 1973. A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. *Journal of Hydrology* **18**, 257–271.

[16] Dixon M.J., Tawn J.A., 1999. The effect of non-stationarity on extreme sea-level estimation. *Applied Statistics* **48**, 135–151.

[17] Dixon, M.J., Tawn, J.A., 1994. *Extreme sea-levels at the UK A-class sites: site-by-site analyses*. Lancaster University and The Proudman Oceanographic Laboratory.

[18] Flather R.A., 2001. Storm surges. In: Steele J., Thorpe S., Turekian K. (Eds.), *Encyclopaedia of Ocean Sciences*. Academic, San Diego, California, pp. 2882-2892.

[19] Fleming G., Frost L., Huntington S., Knight D., Law F., Rickard C., 2002. *Flood risk management: learning to live with rivers*. Institution of Civil Engineers, London.

[20] Fortunato A.B., Li K., Bertin X., Rodrigues M., Miguez B.M., 2016. Determination of extreme sea levels along the Iberian Atlantic coast. *Ocean Engineering* **111**, 471–482.

[21] Frau R., Andreewsky M., Bernardara P., 2017. Including historical data on the Regional Analysis of extreme storm surges. In *EGU General Assembly Conference Abstracts* **19**, p. 5101.

[22] Fuller W.E., 1914. Flood flows. *Transactions of the American Society of Civil Engineers* **77**.

[23] Galiatsatou P., Prinos P., 2007. Bivariate models for extreme of significant wave height and period. An application to the Dutch Coast, in: *Proceedings of the 2nd IMA Conference on Flood Risk Assessment*, 2007, Plymouth, Institute of Mathematics & its Applications, pp. 77–84.

[24] Galiatsatou P., Prinos P., Sanchez-Arcilla A. 2008. Estimation of extremes: Conventional versus Bayesian techniques. *Journal of Hydraulic Research* **46**, Extra Issue 2, pp. 211-223.

[25] Garnier E., Surville F., 2010. *La tempête Xynthia face à l'histoire; submersions et tsunamis sur les littoraux français du Moyen Âge à nos jours*. Saintes: Le Croît Vif.

[26] Gaume E., Gaál L., Viglione A., Szolgay J., Kohnová S., Blöschl G., 2010. Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites. *Journal of hydrology* **394**(1), pp. 101-117.

[27] Gōda Y., 1988. On the Methodology of Selecting Design Wave Height. in: *Proceedings of the 21st International Conference on Coastal Engineering*, Malaga, Spain.

[28] Gōda Y., Kobune K., 1990. Distribution Function Fitting for Storm Wave Data, in: *Proceedings of the 22nd International Conference on Coastal Engineering*, Delft, The Netherlands.

[29] Gōda Y., 2011. Plotting-position estimator for the L-moment method and quantile confidence interval for the GEV, GPA, and Weibull distributions applied for extreme wave analysis. *Coastal Engineering Journal* **53**, 111–149.

[30] Haigh I.D., Nicholls R., Wells N., 2010. A comparison of the main methods for estimating probabilities of extreme still water levels. *Coastal Engineering* **57**, 838–849.

[31] Hamdi Y., Bardet L., Duluc C.-M., Rebour V., 2015. Use of historical information in extreme-surge frequency estimation: the case of marine flooding on the La Rochelle site in France. *Natural Hazards and Earth System Sciences* **15**, 1515–1531.

[32] Hawkes P.J., Gouldby B.P., Tawn J.A., Owen, M.W., 2002. The joint probability of waves and water levels in coastal engineering design. *Journal of Hydraulic Research* **40**, 241–251.

[33] Holdgate M.W., 1979. *A Perspective of Environmental Pollution*. Cambridge University Press, Cambridge, UK.

[34] Hosking J.R.M., 1990. L-moments: Analysis and Estimation of Distributions using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society* **52**, 105–124.

[35] Hosking J.R.M., Wallis J.R., 1997. *Regional Frequency Analysis: an approach based on L-moments.* Cambridge University Press, Cambridge, UK.

[36] H.R. Wallingford, 2002. *Risk, Performance and Uncertainty in Flood and Coastal Defence – A review.* 2nd Edition, H.R. Wallingford, Wallingford, UK.

[37] Idier D., Dumas F., Muller H., 2012. Tide-surge interaction in the English Channel. *Natural Hazards and Earth System Sciences* **12**, 3709–3718.

[38] IPCC - Field C. B. (Ed.), 2012. *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press.

[39] JCGM, 2012. *Joint Committee for Guides in Metrology – The International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*, 3rd ed., JCGM 200:2012.

[40] Jonathan P., Ewans K., Randell D., 2013. Joint modelling of extreme ocean environments incorporating covariate effects. *Coastal Engineering* **79**, 22–31.

[41] Jones J.E., Davies A.M., 2008. On the modification of tides in shallow water regions by wind effects. *Journal of Geophysical Research* **113**, C05014.

[42] Kergadallan X., 2015. *Estimation des niveaux marins extrêmes avec et sans l'action des vagues le long du littoral métropolitain*. Université Paris-Est.

[43] Kinsman B., 1965. *Wind waves: their generation and propagation on the ocean surface*. Courier Corporation.

[44] Lang M., Ouarda T.B.M.J., Bobée B., 1999. Towards operational guidelines for over-threshold modeling. *Journal of Hydrology* **225**, 103–117.

[45] Laugel A., Tiberi-Wadier A. L., Benoit M., Mattarolo G., 2014. ANEMOC-2 Atlantique et Méditerranée: calibration et validation de deux nouvelles bases d'états de mer construites par simulations numériques rétrospectives sur 1979–2010. In *XIIIe Conférence Journées Nationales Génie Côtier Génie Civil*, Dunkirk, France (in French).

[46] Li F., van Gelder P.H.A.J.M., Ranasinghe R., Callaghan D.P., Jongejan R.B., 2014. Probabilistic modelling of extreme storms along the Dutch coast. *Coastal Engineering* **86**, 1–13.

[47] Masselink G., Castelle B., Scott T., Dodet G., Suanez S., Jackson D., Floc'h F., 2016. Extreme wave activity during 2013/2014 winter and morphological impacts along the Atlantic coast of Europe. *Geophysical Research Letters* **43(5)**, 2135-2143.

[48] Mathiesen M., Gōda Y., Hawkes P.J., Mansard E., Martin M.J., Peltier E., Thompson E.F., Van Vledder G., 1994. Recommended practice for extreme wave analysis. *Journal of Hydraulic Research* **32**, 803–814.

[49] Mazas F., Lorenzini J., Hamm L., Hajji H., Rowe S., Thouvenin V., 2015. Modélisation des états de mer dans le Golfe du Lion : apport des mesures satellitaires. *La Houille Blanche* 85–93.

[50] Payrastre, O., 2005. *Faisabilité et utilité du recueil de données historiques pour l'étude des crues extrêmes de petits cours d'eau - Etude du cas de quatre bassins versants affluents de l'Aude*. Ecole des Ponts ParisTech.

[51] Payrastre, O., Gaume, E., Andrieu, H., 2012. Information historique et étude statistique des crues extrêmes: quelles caractéristiques souhaitables pour les inventaires de crues historiques ?, in: *Houille Blanche (La) : Revue Internationale de L'eau 2013*. Presented at the Congrès SHF : «Evènements extrêmes fluviaux et maritimes», Paris.

[52] Péret J., Sauzeau T., 2014. *Xynthia ou la mémoire réveillée*, Gestes Editions, 296 pp.

[53] Pirazzoli P.A., Tomasin A., 2007. Estimation of return periods for extreme sea levels: a simplified empirical correction of the joint probabilities method with examples from the French Atlantic coast and three ports in the southwest of the UK. *Ocean Dynamics* **57**, 91-107.

[54] Pouvreau N., 2014. REFMAR: une coordination pour l'observation du niveau de la mer. *La Houille Blanche* **4**, 37-43.

[55] Pugh D.T., Vassie J.M., 1979. Extreme sea levels from tide and surge probability. *International Conference on Coastal Engineering* **1**.

[56] Salomon J.-N., 2002. L'inondation dans la basse vallée de la Garonne et l'estuaire de la Gironde lors de la "tempête du siècle" (27-28 décembre 1999) / Flooding in the Garonne valley and the Gironde estuary caused by the "storm of the century" (27-28 December 1999). *Géomorphologie : relief, processus, environnement* **8**, 127–134.

[57] Sauzeau T., 2011. Réinventer la culture du risque, in Acerra (Martine) et Mercier (Denis) *Xynthia, une tragédie prévisible*, Place Publique, hors-série, p.45-51.

[58] Sauzeau T., Acerra M., 2012. « Zones construites, zones désertes sur le littoral atlantique français : les leçons du passé », in D. Mercier, *Xynthia, regards de la géographie, de l'histoire et du droit*, Norois, environnement, aménagement, société, Rennes, n°222, 2012-1, p.123-143.

[59] Serinaldi F., 2015. Dismissing return periods! *Stochastic Environmental Research and Risk Assessment* **29**, 1179–1189.

[60] Shom, Cetmef, 2008. Statistique des niveaux marins extrêmes de pleine mer : Manche et Atlantique.

[61] Simon B., 1994. Les niveaux marins extrêmes le long des côtes de France et leur évolution. Rapport SHOM 001/94.

[62] Sklar A., 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* **8**, 229-231.

[63] Smith R.L., 1989. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science* **4**, 367-393.

[64] Stive M.J.F., 2012. Vers un nouveau plan delta pour garder les Pays-Bas à l'abri des inondations au cours du 21e siècle. *La Houille Blanche* **6**, 5–10.

[65] Tawn J., 1992. Estimating probabilities of extreme sea-levels. *Applied Statistics* **41**(1), 77-93.

[66] Tawn J.A., Vassie J.M., 1989. Extreme sea levels: the joint probabilities method revisited and revised, in: *Proceedings of the Institution of Civil Engineers*, Part 2., pp. 429–442.

[67] UNESCO-IOC, 2016. Manual on Sea-level Measurements and Interpretation, Volume V: Radar Gauges. Paris, Intergovernmental Oceanographic Commission of UNESCO. 104 pp. (IOC Manuals and Guides No.14, vol. V; JCOMM Technical Report No. 89).

[68] Van Vledder G., Goda Y., Hawkes P., Mansard E., Martin M.J., Mathiesen M., Peltier E., Thompson E., 1994. Case studies of extreme wave analysis: a comparative analysis. *Proceedings of the Second Symposium on Ocean Wave Measurement and Analysis*, New Orleans, Louisiana, USA, pp. 978–992.

[69] Volpi E., Fiori A., 2014. Hydraulic structures subject to bivariate hydrological loads: Return period, design, and risk assessment. *Water Resources Research* **50**(2):885–897.

[70] Weiss J., 2014. *Analyse régionale des aléas maritimes extrêmes*. Université Paris-Est.

[71] Yue S., Rasmussen P., 2002. Bivariate frequency analysis: discussion of some useful concepts in hydrological application. *Hydrologic Processes* **16**:2881–2898.

[72] Zhang W.-Z., Shi F., Hong H.-S., Shang S.-P., Kirby J. T., 2010. Tide-surge Interaction Intensified by the Taiwan Strait. *Journal of Geophysical Research* **115**, C06012.

PART 2
MAIN PUBLICATIONS

1. COPY OF MAIN PUBLICATIONS

1.1. COASTAL ENGINEERING 2011: A MULTI-DISTRIBUTION APPROACH TO POT METHODS FOR DETERMINING EXTREME WAVE HEIGHTS

Main publications citing this paper (source Google Scholar)

Asl S. J., Khorshiddoust A. M., Dinpashoh Y., Sarafrouzeh F., 2013. Frequency analysis of climate extreme events in Zanjan, Iran. *Stochastic environmental research and risk assessment*, **27(7)**, 1637-1650.

Bitner-Gregersen E. M., Bhattacharya S. K., Chatjigeorgiou I. K., Eames I., Ellermann K., Ewans K., ..., Nilva A., 2014. Recent developments of ocean environmental description with focus on uncertainties. *Ocean Engineering*, **86**, 26-46.

Bogner K., Pappenberger F., Cloke H. L., 2012. The normal quantile transformation and its application in a flood forecasting system. *Hydrology and Earth System Sciences*, **16(4)**, 1085-1094.

Bulteau T., Idier D., Lambert J., Garcin M., 2015. How historical information can improve estimation and prediction of extreme coastal water levels: application to the Xynthia event at La Rochelle (France). *Natural Hazards and Earth System Sciences*, **15(6)**, 1135-1147.

Breilh J. F., Bertin X., Chaumillon É., Giloy N., Sauzeau T., 2014. How frequent is storm-induced flooding in the central part of the Bay of Biscay?. *Global and Planetary Change*, **122**, 161-175.

Caballero-Megido C., Hillier J., Wyncoll D., Bosher L., Gouldby B., 2017. Comparison of methods for threshold selection for extreme sea levels. *Journal of Flood Risk Management*.

Calderon-Vega F., Vazquez-Hernandez A. O., Garcia-Soto A. D., 2013. Analysis of extreme waves with seasonal variation in the Gulf of Mexico using a time-dependent GEV model. *Ocean Engineering*, **73**, 68-82.

Chen X., Hofland B., Uijttewaal W., 2016. Maximum overtopping forces on a dike-mounted wall with a shallow foreshore. *Coastal Engineering*, **116**, 89-102.

Clancy C., O'Sullivan J., Sweeney C., Dias F., Parnell A. C., 2016. Spatial Bayesian hierarchical modelling of extreme sea states. *Ocean Modelling*, **107**, 1-13.

Durán-Rosal A. M., Fernández J. C., Gutiérrez P. A., Hervás-Martínez C., 2017. Detection and prediction of segments containing extreme significant wave heights. *Ocean Engineering*, **142**, 268-279.

Gaidai O., Ji C., Kalogeri C., Gao J., 2017. SEM-REV energy site extreme wave prediction. *Renewable Energy*, **101**, 894-899.

Godoi V. A., Bryan K. R., Stephens S. A., Gorman R. M., 2017. Extreme waves in New Zealand waters. *Ocean Modelling*.

Harley M., 2017. Coastal Storm Definition. *Coastal Storms: Processes and Impacts*, 1-21.

Lerma A. N., Bulteau T., Lecacheux S., Idier D., 2015. Spatial variability of extreme wave height along the Atlantic and channel French coast. *Ocean Engineering*, **97**, 175-185.

Li W., Isberg J., Engström J., Waters R., Leijon M., 2015. Study of the foundation design for a linear generator wave energy converter using stochastic methods. *Journal of Renewable and Sustainable Energy*, **7(6)**, 063112.

Li W., Isberg J., Waters R., Engström J., Svensson O., Leijon M., 2016. Statistical analysis of wave climate data using mixed distributions and extreme wave prediction. *Energies*, **9(6)**, 396.

Luo Y., Zhu L. S., 2014. A new model of peaks over threshold for multivariate extremes. *China Ocean Engineering*, **28(6)**, 761-776.

Mortlock T. R., Goodwin I. D., 2015. Directional wave climate and power variability along the Southeast Australian shelf. *Continental Shelf Research*, **98**, 36-53.

Sarafrouzeh F., 2013. Saeed Jahanbaksh Asl, Ali Mohammad Khorshiddoust, Yagob Dinpashoh &. *Stoch Environ Res Risk Assess*, **27**, 1637-1650.

Sartini L., Mentaschi L., Besio G., 2015. Comparing different extreme wave analysis models for wave climate assessment along the Italian coast. *Coastal Engineering*, **100**, 37-47.

Solari S., Losada M. Á., 2012. Unified distribution models for met-ocean variables: Application to series of significant wave height. *Coastal Engineering*, **68**, 67-77.

Solari S., Losada M. A., 2015. Statistical Methods for Risk Assessment of Harbor and Coastal Structures. In *Design of Coastal Structures and Sea Defenses* (pp. 215-272).

Teena N. V., Kumar V. S., Sudheesh K., Sajeev R., 2012. Statistical analysis on extreme wave height. *Natural hazards*, **64(1)**, 223-236.

Wang J., Zhai X., Liu C., Zhang Y., 2017. Determination of the Threshold for Extreme Load Extrapolation Based on Multi-Criteria Decision-Making Technology. *Strojniški vestnik-Journal of Mechanical Engineering*, **63(3)**, 201-211.

Wiesebron L. E., Horne J. K., Hendrix A. N., 2016. Characterizing biological impacts at marine renewable energy sites. *International Journal of Marine Energy*, **14**, 27-40.

You Z. J., 2011. Extrapolation of historical coastal storm wave data with best-fit distribution function. *Australian Journal of Civil Engineering*, **9(1)**, 73-82.

You Z. J., 2012. Discussion of “a Multi-distribution Approach to POT Methods for Determining Extreme Wave Heights” by Mazas and Hamm,[Coastal Engineering, 58: 385–394]. *Coastal Engineering*, **61**, 49-52.

You Z. J., Callaghan D., 2013. Discussion of “Modelling significant wave height distributions with quantile functions for estimation of extreme wave heights”[Ocean Eng. 54 (2012) 119–131]. *Ocean Engineering*, **70**, 208-210.

You Z. J., Nielsen P., 2013. Extreme Coastal Waves, Ocean Surges and Wave Runup. In *Coastal Hazards* (pp. 677-733). Springer Netherlands.

Coastal Engineering 58 (2011) 385–394

Contents lists available at ScienceDirect

Coastal Engineering

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

A multi-distribution approach to POT methods for determining extreme wave heights

Franck Mazas *, Luc Hamm

SOGREAH Consultants, 6 rue de Lorraine 38130 Echirolles, France

ARTICLE INFO

Article history:
Received 2 June 2010
Received in revised form 17 December 2010
Accepted 20 December 2010
Available online 14 January 2011

Keywords:
Extreme wave heights
POT
GPD
Multi-distribution
BIC/AIC
Climate variability

ABSTRACT

Determination of extreme wave heights using a Peaks-Over-Threshold (POT) approach is revisited. Firstly, the GPD-Poisson model is recalled. A double threshold is presented and justified, with objective tools for determining the high threshold. This model is then extended to other statistical distributions, namely the Weibull and Gamma distributions. Objective criteria (BIC and AIC) based upon likelihood are used to select the best-fitting distribution. This method is tested on two locations: the historical IAHR Haltenbanken dataset and a location at the entry of the Strait of Gibraltar. Finally, sensitivity analyses are carried out with respect to the high threshold and to the duration of the dataset to estimate the robustness of the approach presented.

1. Introduction

Statistical methods to determine extreme wave heights using the Peaks-Over-Threshold approach (POT) have been significantly improved for several years. The IAHR Working Group on Extreme Wave Analysis issued recommendations about the most appropriate way to proceed when determining extreme wave heights (Mathiesen et al., 1994). They recommended the use of the POT method along with a Weibull distribution estimated by maximum likelihood. A little later, several authors introduced the GPD-Poisson model (e.g. Coles, 2001), which is the most natural way to proceed when using the POT approach. While respecting the general guidelines of the IAHR Maritime Hydraulics-Working Group on Extreme Wave Analysis (Mathiesen et al., 1994), this model notably improves several key steps of the analysis, particularly by fitting a Generalized Pareto Distribution (GPD) to storm peaks while assuming that the number of storms in one year follows a Poisson distribution. It is now recommended (Hawkes et al., 2008) and widely used (e.g. Méndez et al., 2006; Thompson et al., 2009), although many authors still prefer other distributions, mainly the classical extreme distributions: GEV, Weibull, and Gumbel.

However, it should be recalled that the GPD-Poisson model is an asymptotic model. For this reason, other distributions might give better results.

We therefore propose to extend this model to a multi-distribution approach, using the Weibull and Gamma distributions in addition to the GPD. Objective criteria for choosing the most appropriate threshold and determining the best-fitting distribution are also presented.

This method is illustrated by case studies in the Northern Atlantic and in the Strait of Gibraltar.

2. POT method revisited

2.1. Brief justification of the GPD-Poisson model

Let us consider a sample of wave height data $(X_1,\ldots,X_n)$. These data follow an unknown continuous distribution, say F. Let u be a threshold and $y=x_{|x>u}-u$ the exceedance by x of the threshold u. So $Y=(Y_1,\ldots,Y_N)$ is the sample of the N threshold exceedances. The law of threshold exceedance is given by:

$$\mathbb{P}[Y<y]=\mathbb{P}[X<u+y|X>u]=\frac{F(u+y)-F(u)}{1-F(u)} \tag{1}$$

According to Pickands, 1975 (see also Embrechts et al., 1997), when u is large, this law is very nearly in the form of the Generalized Pareto Distribution defined as:

$$\begin{cases} G_{Y;k,\sigma}(y)=1-\left(1+k\frac{y}{\sigma}\right)^{-\frac{1}{k}} \text{ if } k\neq 0 \\ G_{Y;\sigma}(y)=1-\exp\left(-\frac{y}{\sigma}\right) \text{ if } k=0 \end{cases} \tag{2}$$

where k is the shape parameter and σ is a scale parameter. When $k>0$, the distribution has a heavy and unbounded tail and belongs to the

* Corresponding author. Tel.: +33 4 56 38 46 91; fax: +33 4 76 33 43 33.
E-mail addresses: franck.mazas@sogreah.fr (F. Mazas), luc.hamm@sogreah.fr (L. Hamm).

0378-3839/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.coastaleng.2010.12.003

Fréchet domain of attraction (a heavy tail is not exponentially bounded, and extreme values are more likely to occur than in distributions with exponential or lighter tails). When $k<0$, the distribution is bounded by $x_{max}=u+\sigma/k$ and belongs to the Weibull domain of attraction. Finally, when this parameter is zero, the GPD is the exponential distribution with scale parameter σ.

Still, it must be kept in mind that the GPD is an asymptotic law. This means we must be in its range of validity, i.e. u must be high enough. However, the higher the threshold, the greater the uncertainties because of the very small number of data left. It is the well known dilemma between bias and variance.

If we consider that the number of events (i.e. storms) in one year follows a Poisson distribution with parameter λ, we obtain the so-called GPD-Poisson model: the law of the exceedances is a Generalized Pareto Distribution and the storm occurrence is a Poisson process.

A Poisson distribution should thus be fitted to the data. However, the most common estimator for its unique parameter (e.g. the maximum likelihood estimator) is the empirical mean. We are thus able to link the number of storm occurrences with the return period T.

2.2. The multi-distribution POT model

2.2.1. Data homogenization

The first step in the analysis is to extract homogenous time series from the main continuous sea states time series (buoy measurement, hindcast data, etc.). If this step is omitted, storms from very different meteorological phenomena will be treated together, although it is most likely they are not identically distributed. Such homogenization can be carried out by separation in carefully chosen directional sectors, seasonal analysis (e.g. summer/winter monsoon) and separation of sea states into independent wave systems. Rare but very strong events such as hurricanes should also be checked if necessary. Actually, homogenization may be the most important step in the analysis (this point was stressed by Mathiesen et al., 1994), although it is often the least considered: the best statistical analysis cannot extract the "truth" out of wrongly prepared data ("Garbage In, Garbage Out").

2.2.2. Peak selection and double threshold

Once we have time series of homogenous sea states, we have to extract storm peaks. If we keep in mind that a rigorous statistical analysis requires independent and identically distributed (i.i.d.) data, we will pay special attention to obtaining independent storm peaks. Firstly, we should be careful concerning possible fluctuations in storms around the threshold. If the wave height falls below the threshold for a short period, say 3 h in a 24-hour storm, we should not cut the storm in two. Secondly, we should set a minimum period between two storms to ensure their peaks are independent. Finally, once the storm peaks are identified, outliers (i.e. values significantly larger than the other ones) must be checked carefully in order to be sure they really belong to the population and are not the result of some measurement error. If so, they could have a return period T much larger than the duration of the time series K. Thus they provide valuable information and we recommend keeping them in the sample.

The interest of a threshold is to consider that storm peaks above it have a statistically extreme behavior, i.e. they follow the same extreme distribution. However, we do not know the threshold value *a priori*. A simple way to proceed is therefore to use a double threshold (u_1,u_2). A low value u_1 is set to select both weak and strong storms. There is no need for precise criteria in the choice of u_1 because the procedure relies more heavily on u_2 (see below). Its aim is only to extract the storm peaks from the time series, reducing the sample size from 10,000 to 100,000 values to a few hundreds of peaks. u_1 shall be high enough to discriminate two consecutive storms and low enough to be below the "extreme area", i.e. the strong storms showing genuine statistically extreme behavior.

We obtain N_T peaks over a period of K years. Hence, the mean number of storms per year above u_1 is:

$$\lambda_T = \frac{N_T}{K} \tag{3}$$

Our experience in extra-tropical areas led us to set u_1 so as to have λ_T approximately between 5 and 10, although it is not an absolute constraint.

We have now to determine the high threshold u_2 above which storms have a statistically extreme behavior. As the GPD is the asymptotic law, it seems quite reasonable to use its properties to determine u_2. In particular, if a sample follows a GPD, the shape parameter k and the modified scale parameter $\sigma^*=\sigma-ku_2$ remain constant when u_2 increases. So if we fit a GPD to the exceedances of a threshold varying between u_1 and, for instance, a threshold corresponding to one storm per year, we can draw graphs of shape and modified scale parameters with respect to u_2 and search for "domains of stability" where they will remain roughly constant. As we want to be in the asymptotic domain, we are interested in the highest domain of stability. And as we want to have as much information as possible, we will choose the lowest threshold of this highest domain.

Thus we select N storm peaks over K years, namely $\lambda=N/K$ storms per year (as we have seen, this empirical mean is also the estimator of the Poisson parameter). After many tests, we believe it is appropriate for λ to stand approximately between 2 and 5. If K is low, a value around 5 is more advisable in order to ensure that N is large enough (with a minimum of 20–30). In contrast, if K is quite large (around 40–50), a value of around 2 is more appropriate.

2.2.3. Fit to multiple distributions

Storm peaks above u_2 are now to be fitted to a statistical distribution. As we have seen, the GPD is the asymptotic law, and thus a natural candidate. However, we do not know whether we are within the asymptotic domain. Thus, other distributions might fit the data better. We can try many of them and then determine the best-fitting one.

When looking for suitable distributions, it is useful to know their domain of attraction for maxima (Castillo and Sarabia, 1992). If they belong to the Fréchet domain (e.g. Pareto or beta laws), their tails are heavy and unbounded, which means they give too much weight to extreme events. Practice shows they are not appropriate for coastal engineering applications where the wave heights are physically bounded. If they belong to the Gumbel domain, their tails decrease exponentially. If they belong to the Weibull domain (e.g. GPD with negative shape parameter), their tails are bounded. We can thus limit our study to distributions belonging to Weibull or Gumbel domains of attraction for maxima.

Our tests have shown that along with the GPD, the Gamma distribution and 2-parameter Weibull distribution for minima often behave quite well. Although other distributions may be studied, we will work here with these two laws, whose cumulative distribution functions are respectively:

$$\text{Gamma}: F_{Y;k,\sigma}(y) = \mathbb{P}\{Y\le y\} = \frac{\gamma\left(k,\frac{y}{\sigma}\right)}{\Gamma(k)} \tag{4}$$

$$\text{Weibull}: F_{Y;k,\sigma}(y) = \mathbb{P}\{Y\le y\} = 1-\exp\left[-\left(\frac{y}{\sigma}\right)^k\right] \tag{5}$$

As for the GPD, we work with y, that is the threshold exceedance $x-u_2$, provided $x>u_2$. k and σ are respectively the shape and the scale parameters. Γ is the Gamma function, and γ is the lower incomplete gamma function.

The choice of the estimator is also quite important. Mathiesen et al. (1994) mention three estimators: least squares methods, the method of moments and the Maximum Likelihood Estimator (MLE). The statistical theory says that an estimator must be *robust*, i.e. it is not disturbed by an outlier and it must be *consistent*, i.e. the bias and the variance tend to zero when the sample size increases. Least squares methods, though easy to implement, are neither robust nor consistent. In particular, they are found to be sensitive to outliers (Mathiesen et al., 1994). They are therefore rejected. It is nevertheless noteworthy that Goda (2000) recommends this method with modified plotting position formulae. The method of moments may be used as first approximations but the small sample sizes hinder it. In particular, the method of moments gives too much bias for the typical sample sizes we are handling (Goda, 2000). To handle this difficulty, Hosking and Wallis (1987) have proposed an estimator based upon the Probability Weighted Moments. But it is known to be less efficient than the MLE. Finally, the most handy and appropriate method is to use the Maximum Likelihood Estimator (MLE). This estimator maximizes the likelihood function of the fit, which is defined by:

$$L(X_1,\ldots,X_N|\theta) = \Pi_{i=1}^{N} f_\theta(X_i;\theta) \tag{6}$$

where f_θ is the joint density function (with parameter vector θ) at the sample observations X_i. The log-likelihood function is usually used, since it is much easier to derive:

$$l(X_1,\ldots,X_N|\theta) = \Sigma_{i=1}^{N} \ln(f_\theta(X_i;\theta)) \tag{7}$$

Thus we have an optimization problem, as the likelihood function has as many variables as the distribution has parameters. In some cases, optimization algorithms may fail to maximize the likelihood.

However, the use of the MLE for two-parameter distributions such as the Weibull and Gamma distributions has a very disturbing drawback. These distributions are very sensitive to the distance between u_2 and the first peak. In other words, the estimated parameters will be quite different if the smallest value of the ordered sample of the threshold exceedances Y_1 is 0.1, 0.01 or 0.001. When we look at the 100-year wave height, the result varies between 14 and 16 m! The GPD is much less sensitive to this phenomenon. We think the explanation could be related to the shape of the density functions just above 0. A comparison with the method of moments estimator was carried out. From these tests, it appeared that the two-parameter distributions could be used with MLE, but only when u_2 meets a storm peak. As this peak is excluded, the first value of the exceedance sample is as far from u_2 as possible.

A solution would be to use the three-parameter Weibull and Gamma distributions (the latter being known as Pearson-III distribution) by adding a location parameter μ ($\mu<Y_1$, the first and smallest exceedance):

$$\text{Pearson}-\text{III (3}-\text{parameter Gamma)}: F_{Y;k,\sigma}(y) = \mathbb{P}\{Y\le y\} = \frac{\gamma\left(k,\frac{y-\mu}{\sigma}\right)}{\Gamma(k)} \tag{8}$$

$$\text{Weibull}: F_{Y;k,\sigma,\mu}(y) = P\{Y\le y\} = 1-\exp\left[-\left(\frac{y-\mu}{\sigma}\right)^k\right] \tag{9}$$

However, ML estimation of such distributions is very difficult, and the algorithms usually fit two-parameter distributions inside a discrete range of location parameters (Panchang and Gupta, 1989). Actually, it appears that quite often the maximum likelihood with respect to this location parameter μ is obtained for $\mu\to Y_1$ (with $\mu<Y_1$). Now, the Maximum Likelihood Estimator is known to provide poor results when the maximum is at the limit of the interval of validity of one of the parameters. In our applications, this is a major drawback of this estimator. We are currently carrying out further investigation on this subject and shall soon submit our results.

2.2.4. Best fit selection

Once several distributions are fitted to the data, we have to determine the best fit. For this purpose, we use objective Bayesian criteria. The first one is the Bayesian Information Criterion (BIC), also known as the Schwarz Criterion (Schwarz, 1978). It minimizes the bias between the fitted model and the unknown "true" model. Assuming asymptotic conditions (N large enough), BIC is given by:

$$\text{BIC} = -2\ln L + k_p \ln N \tag{10}$$

where L is the likelihood of the fit, N is the sample size (number of storm peaks above u_2) and k_p is the number of parameters of the distribution.

We can also use the closely related Akaike Information Criterion (AIC), which gives the model providing the best compromise between bias and variance (Akaike, 1973). It can be interpreted as the sum of two terms, the first one measuring bias and the second one measuring variance. Under the same assumptions as BIC, AIC is given by:

$$\text{AIC} = -2\ln L + 2k_p \tag{11}$$

For BIC as for AIC, the lower the criterion, the better the fit, so we will select the distribution providing the lowest criteria. Most of the time, both criteria give the same result. If they do not, we recommend keeping the distribution giving the most conservative return values.

2.2.5. Return values and confidence intervals

We now have only one distribution left, with MLE estimated parameters. We are interested in wave heights of return period T. It is actually a quantile of the estimated distribution, whose non-exceedance probability is $1-1/\lambda T$. These quantiles for GPD, Gamma and Weibull distributions are given by:

$$\text{GPD}: Hs_T = u_2 + \frac{\hat{\sigma}}{\hat{k}}\left[(\lambda T)^{\hat{k}}-1\right] \tag{12}$$

$$\text{Gamma}: Hs_T = u_2 + \Gamma^{-1}_{\hat{k},\hat{\sigma}}\left(1-\frac{1}{\lambda T}\right) \tag{13}$$

$$\text{Weibull}: Hs_T = u_2 + \hat{\sigma}\,[\ln(\lambda T)]^{\frac{1}{\hat{k}}} \tag{14}$$

Finally, confidence intervals are to be computed. Many authors (Coles, 2001) use the classical asymptotic method. Mathiesen et al. (1994) advocate the use of Monte-Carlo simulation techniques. A robust way is to use parametric bootstrap methods (Thompson et al., 2009). The principle is quite simple (see for instance Efron and Tibshirani, 1993a,b). From the estimated distribution with estimated parameter vector $\hat{\theta}_0$, a random sample of size N is generated and the same distribution is then fitted to this sample, leading to a slightly different estimated parameter vector $\hat{\theta}_1$, which will give a slightly different quantile $Hs_{T,1}$. After 100,000 iterations, a sample of 100,000 $Hs_{T,i}$ is obtained. The 90% confidence interval will be given by the percentiles $[Hs_{T,5\%}; Hs_{T,95\%}]$. It is advisable to correct the bias of the bootstrap. Bias is given by the difference between the empirical mean of $Hs_{T,i}$ and $Hs_{T,0}$, and it simply has to be removed from the percentiles previously obtained.

3. Case studies

3.1. Datasets

We shall study two different locations. The first one is the historical Haltenbanken dataset, provided by the IAHR Working Group on Extreme Wave Analysis (van Vledder et al., 1994). The Haltenbanken

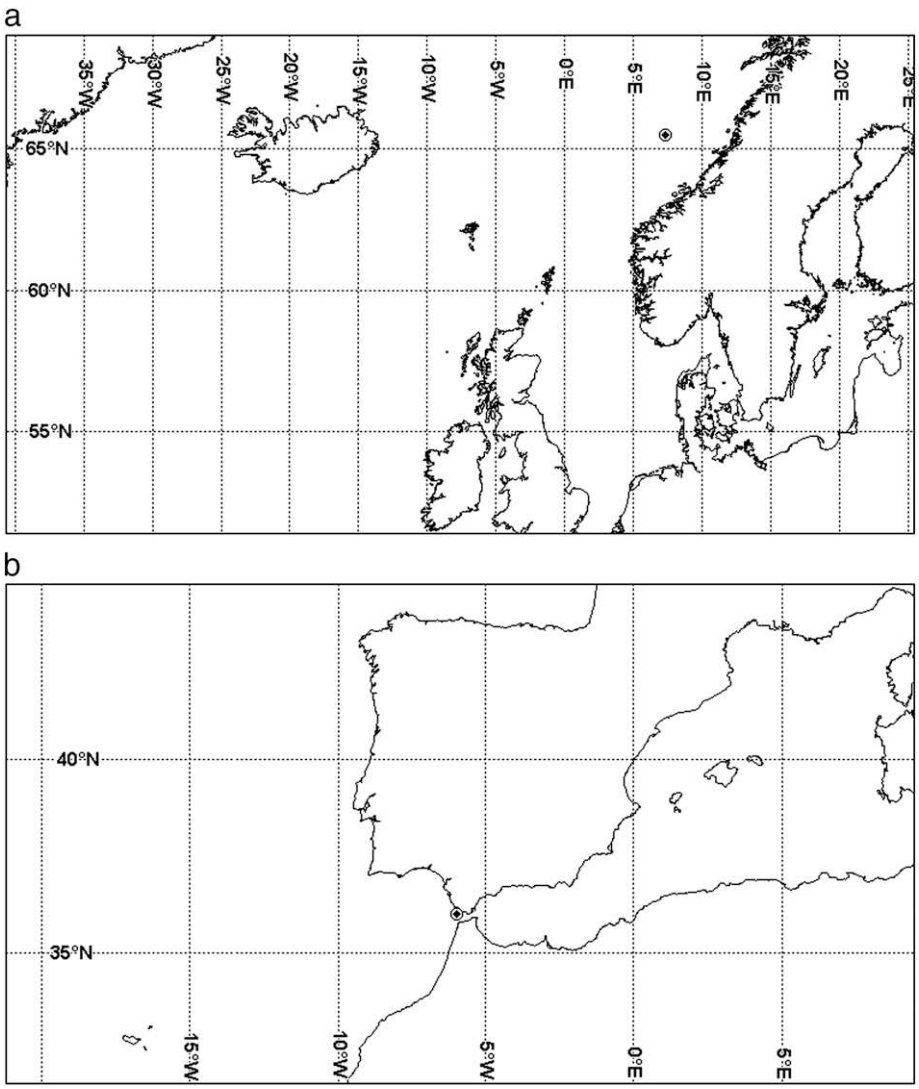

Fig. 1. a) Location of the Haltenbanken dataset. b) Location of the Gibraltar dataset.

buoy is located off the coast of Norway. Its coordinates are 65°5′N; 7°34′ E (Fig. 1a). The original dataset consists of 128 buoy-measured storm peaks above 7 m for a period of 9 years, so no pre-treatment on this sample was done and the peaks were considered to be independent and identically distributed. The shortness of the period must be stressed and will be discussed later.

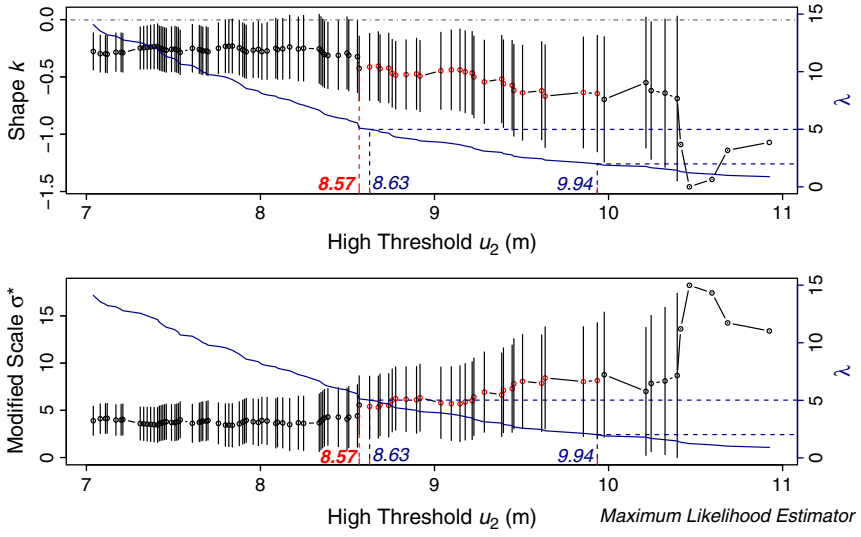

Fig. 2. Haltenbanken dataset: stability of shape and modified scale parameters for Generalized Pareto Distribution.

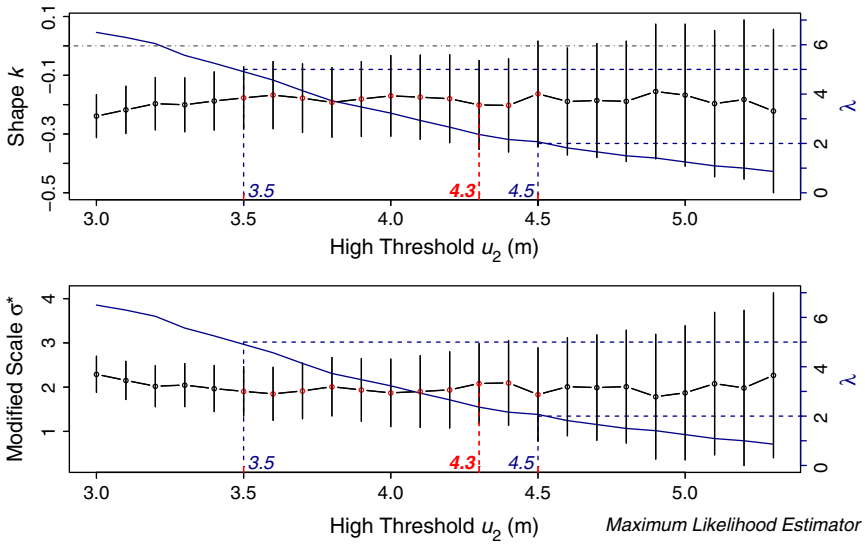

Fig. 3. Gibraltar dataset: stability of shape (above) and modified scale (below) parameters for Generalized Pareto Distribution.

The second dataset comes from the SIMAR-44 hindcast database provided by Puertos del Estado. We chose point 1056044, whose coordinates are 36°N; 6°W (Fig. 1b). It is located at the western entry of the Strait of Gibraltar. Wave and wind data are provided every 3 h from 1958 to 2001 for a total of 44 years. In contrast to Haltenbanken, the storm peaks have to be extracted here. As has been said above, the first and most important step is to homogenize the sample. A simple method, directional analysis, will be used. We will only consider the western sector, facing the Atlantic. From the model point, we can draw lines to Cape St-Vincent, Portugal north-westwards and towards the Moroccan coast near El-Jadida. Thus we obtain the following sector: [220°; 295°]. It is not a very wide sector, but unsurprisingly it is the dominant one, as 86% of the hindcasted waves come from this sector. We consider that all these waves are homogenous and, in particular, that all these western storms are generated by the same kind of Atlantic depressions and thus are identically distributed.

If we set a low threshold u_1 equal to 3 m, we select $N_T = 288$ storm peaks. As $K = 44$ years, we then have a mean number of total storms per year $\lambda_T = 6.55$, which seems quite reasonable.

3.2. Selection of high thresholds

We will now try to determine the best high threshold. For this purpose, as has been explained in Section 2.2.2, we will adjust a GPD to the data over a wide range of thresholds and look at the stability of the shape parameter k and of the modified scale parameter σ^* (Fig. 2). As has been previously discussed, the thresholds tested are those meeting the dataset values. We use tools available in the "ismev" package (Coles and Stephenson, 2006) developed for the R language (R Development Core Team, 2009). We modified these tools in order to take into account the remarks made previously. On the secondary axis, we draw the change in λ, so as to see easily the thresholds corresponding to a value of λ between 2 ($u_2 = 9.94$) and 5 ($u_2 = 8.63$). These limit thresholds are written in italics.

We can see two domains of stability where the parameters remain approximately constant. The lowest threshold of the highest domain, just below the value of 8.63, is 8.57 m. For this threshold, λ is 5.11, which is slightly higher than 5, but as K is very low (9 years), we can allow this small exceedance in order to have N large enough (46).

For Gibraltar, the choice is more difficult (see Fig. 3). The curves are rather flat when u_2 is higher than 3.5. (corresponding to $\lambda = 5$) This value could therefore be adopted. However, it is important to bear in mind that here K is large (44 years). A value closer to $\lambda = 2$ (corresponding to $u_2 = 4.5$ m) may therefore be more appropriate. Since there is small bump for 4.5 m, we will choose $u_2 = 4.3$ m. It is clear that choosing the right threshold is not always a straightforward matter. Thompson et al. (2009) presented methods for automated threshold selection, but these should be used rather when working with too many datasets for visual examination.

Table 1 recapitulates the characteristics of the samples.

3.3. Fit

For both datasets, the three distributions (GPD, Weibull, Gamma) are now fitted to the exceedances of the high threshold with the Maximum Likelihood Estimator. Table 2 provides BIC and AIC criteria for the two locations and the three distributions.

We can see that both criteria give the same result. For Haltenbanken, GPD is clearly selected, with Weibull then Gamma quite far off. In contrast, for Gibraltar, GPD gives poor results with respect to these criteria. The Gamma distribution is selected since it minimizes both BIC and AIC criteria.

Actually, as we use distributions with the same number of parameters (two), we could consider only the (log-)likelihood of the fit: this would give the same results as BIC and AIC. However, it is convenient to have a means of discriminating between fits for distributions with one, two or three parameters together.

Table 1
Characteristics of the samples.

	Haltenbanken	Gibraltar
K (years)	9	44
u_1 (m)	7	3
N_T (−)	128	288
λ_T (yr^{-1})	14.22	6.55
u_2 (m)	8.57	4.3
N (−)	46	104
λ (yr^{-1})	5.11	2.37

Table 2
BIC and AIC criteria for the fits of the three distributions for both datasets.

		GPD	Weibull	Gamma
Haltenbanken	BIC	120.0	122.0	122.4
	AIC	116.3	118.3	118.8
Gibraltar	BIC	216.6	212.4	211.4
	AIC	211.3	207.1	206.1

Table 3
Return values for the best-fitting distribution with 90% confidence intervals.

Return period (years)	Haltenbanken (GPD)	Gibraltar (Gamma)
100	12.7	8.8
	12.0–13.5	8.0–9.6
50	12.6	8.3
	11.9–13.3	7.6–8.9
20	12.4	7.6
	11.8–13.0	7.1–8.1
10	12.2	7.1
	11.7–12.7	6.7–7.5
5	11.9	6.5
	11.5–12.3	6.2–6.9
2	11.4	5.8
	11.0–11.8	5.6–6.0
1	10.8	5.3
	10.4–11.2	5.1–5.4

3.4. Return values and confidence intervals

The last step of the analysis is now to compute the return values for the return periods of interest using the quantile functions defined above for the best-fitting distribution. 90% confidence intervals are also computed using a parametric bootstrap approach with 100,000 iterations (bootstrap bias is corrected). Table 3 gives the results for 1, 2, 5, 10, 20, 50 and 100 years.

It can be seen that the Haltenbanken results are much lower than those given in van Vledder et al. (1994), where Hs_{100} varies between 14.2 and 15.8 m. The methods used in this paper were quite different, and nobody used a GPD at that time. The closest analysis was carried out by member D, who applied a 3-parameter Weibull distribution to the 46 storm peaks above 8.6 m. This member obtained a Hs_{100} of 14.7 m (12.6–16.9 90% CI). In our analysis, the 2-parameter Weibull distribution also gives 14.7 m for Hs_{100} (13.1–16.6 90% CI) but is rejected by the BIC/AIC criteria. It is also noteworthy that bootstrap confidence intervals for this GPD fit are much narrower than in the case of the Working Group analysis (1.5 m versus 2.5 to 5 m). Nonetheless, an explanation of such differences will be given later.

As for Gibraltar, the confidence intervals remain quite narrow; the fit seems reasonable.

4. Sensitivity analysis with respect to the high threshold

4.1. Purpose of the analysis

We have proposed an objective method for determining the high threshold. Nevertheless, the case of Gibraltar shows that a part of subjectivity may remain when choosing it. It follows that studying the

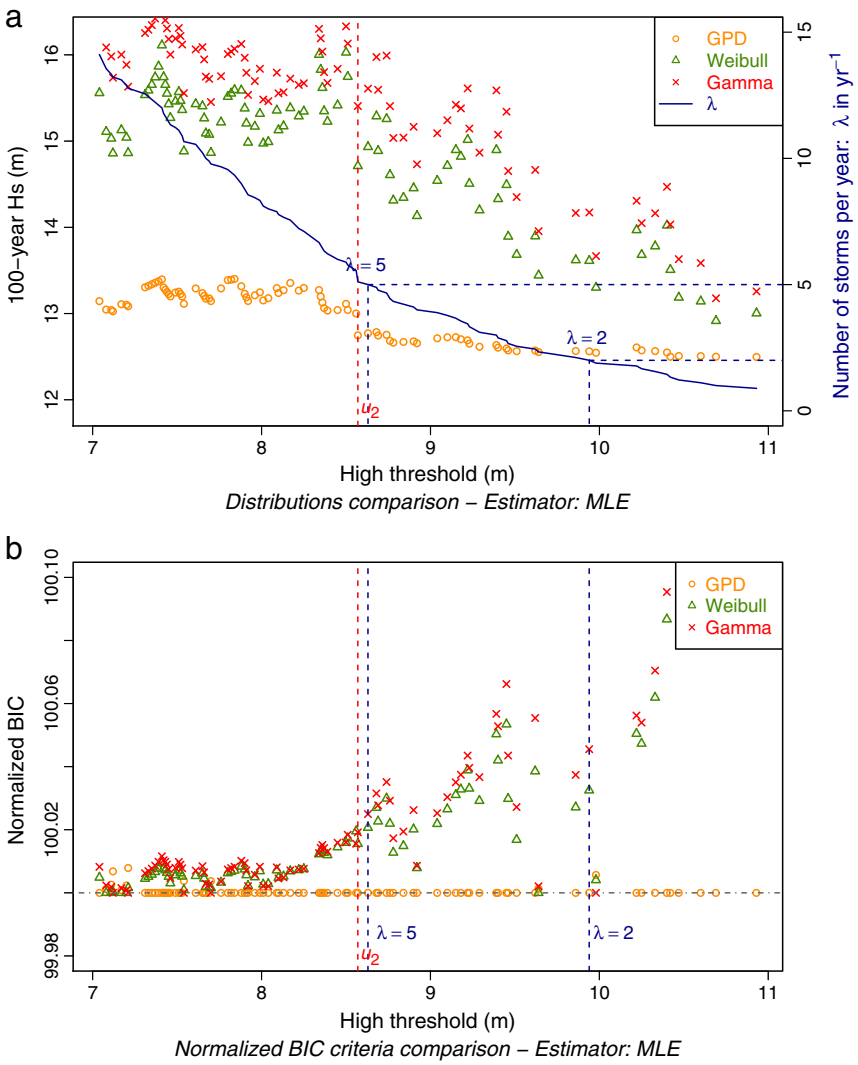

Fig. 4. Haltenbanken: a) stability of the 100-year significant wave height with respect to the high threshold. b) stability of the normalized BIC criterion with respect to the high threshold.

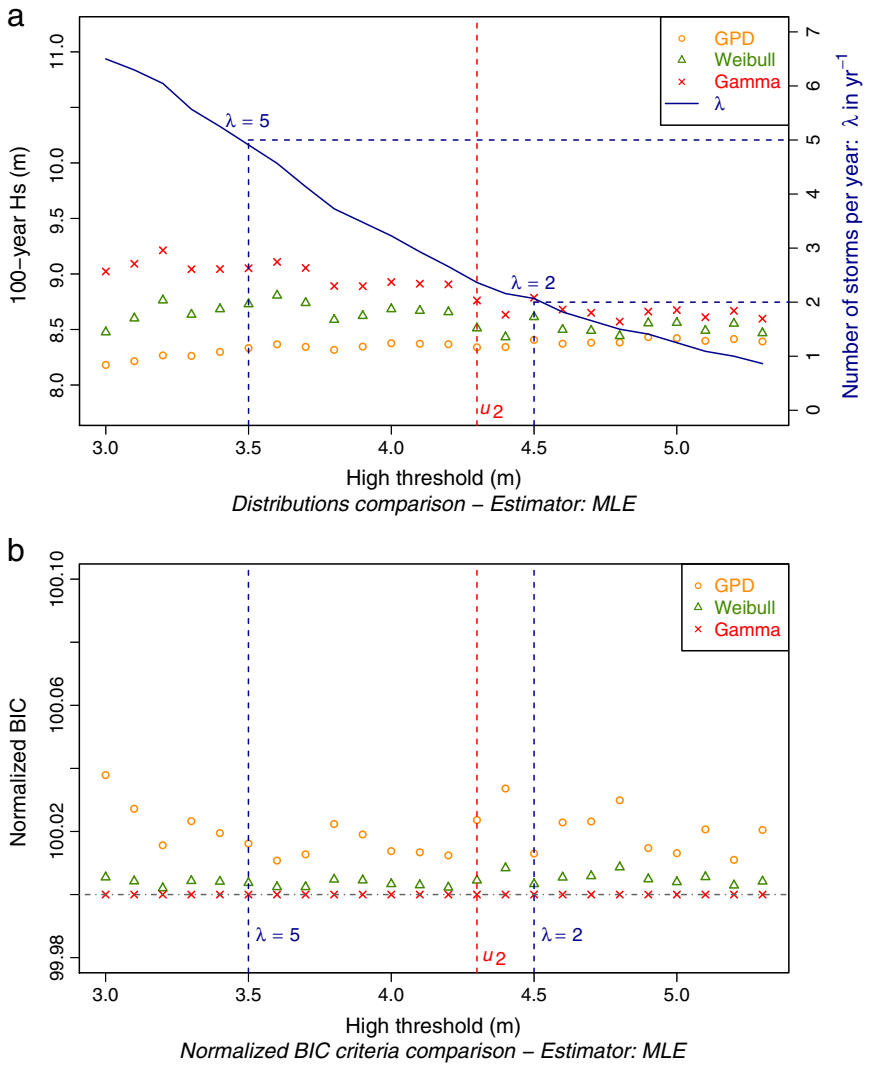

Fig. 5. Gibraltar: a) stability of the 100-year significant wave height with respect to the high threshold. b) stability of the normalized AIC criterion with respect to the high threshold.

change in return values and goodness of fit (i.e. in BIC/AIC criteria) can provide interesting information for validating the results or, conversely, reconsidering the choice of threshold.

Let us draw similar plots to those showing the stability of the GPD parameters, but this time with Hs_{100} and BIC/AIC criteria for each distribution. As for the criteria, it is actually necessary to normalize

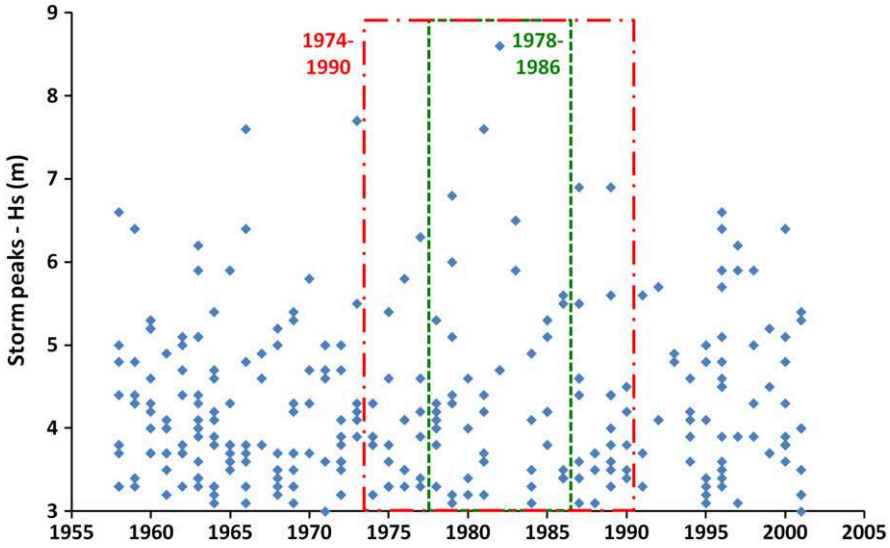

Fig. 6. Gibraltar: storm peaks above the low threshold u_1 with respect to the calendar years. Choice of the 9-year period 1978–1986 (dashed line) and of the 17-year period 1974–1990 (dashed-dotted line).

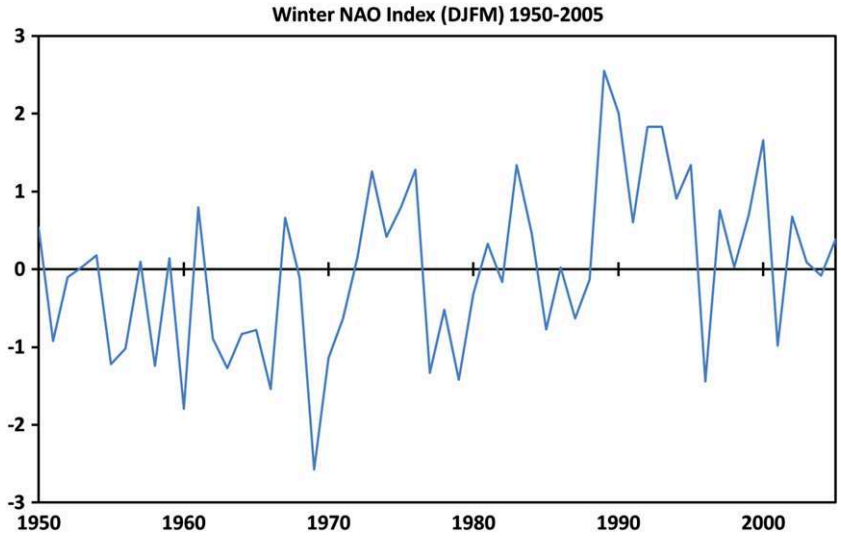

Fig. 7. PC-based winter NAO index from 1950 to 2005.

them in order to have clear graphs. Indeed, they will depend on the sample size and so direct comparison between all the tested thresholds is difficult. For each threshold, the criteria are computed for the three distributions studied, and the minimum criterion is set at 100. The relative difference of the other criteria with respect to the minimum are then added to 100.

5. Results

For the Haltenbanken dataset, we can see that from 8.57 m, that is the high threshold we chose previously, the GPD quantile is extremely stable, which is quite a remarkable result (Fig. 4a). The Weibull and Gamma distributions are much more unstable and seem to tend downwards towards the GPD value. The change in the normalized BIC criterion (Fig. 4b) shows that the GPD is almost always the best-fitting distribution in this case. Thus, the choice of u_2 on the basis of the stability of the GPD parameters appears to be particularly relevant here. It is even probable that the asymptotic domain starts at 8.57 m, thus giving further argument for choosing the GPD.

As in the case of the Gibraltar dataset, Fig. 5a shows that the three distributions converge towards a common value of Hs_{100}. Similarly to Haltenbanken, the GPD quantiles are below those of the other distributions. This pattern was observed in many tests: the GPD value is clearly not conservative compared to other distributions. As a matter of fact, the Weibull and Gamma distributions often behave well when no saturation (i.e. no "flattening" of the highest peaks) is observed in the data, whereas the GPD with a strongly negative shape parameter generally fits well when saturation occurs. Here, the (normalized) AIC criterion shows that the Gamma distribution is always the best-fitting one (see Fig. 5b). Once again, the choice of threshold appears to be relevant.

These stability plots for Hs_{100} and normalized criteria are a very helpful way of checking the results obtained previously. If it seems obvious that the return value is not at all representative, the choice of threshold will have to be reconsidered.

Table 4
Gibraltar: fit characteristics and 100-year significant wave height with 90% confidence intervals for the 9-year, 17-year and 44-year periods centered on 1982.

	9 years (1978–1986)	17 years (1974–1990)	44 years (1958–2001)
u_2 (m)	3.1	3.3	4.3
N (−)	44	79	104
λ (yr^{-1})	4.89	4.65	2.37
Min BIC/AIC distribution	Gamma	Gamma	Gamma
Hs_{100} GPD (m) 90% CI	9.3	9.1	8.3
	7.3–12.0	7.4–11.2	7.6–9.2
Hs_{100} Weibull (m) 90% CI	10.0	9.4	8.5
	8.1–12.3	8.1–11.0	7.8–9.3
Hs_{100} Gamma (m) 90% CI	10.5	9.6	8.8
	8.6–12.6	8.4–11.0	8.0–9.6

6. Sample duration

6.1. Purpose of the analysis

We have studied two datasets, one very short compared to the usual available duration (around 20 years) and the other quite long. It is likely that the very short duration of the Haltenbanken dataset (9 years) is the cause of the huge differences between the return values given by the three statistical distributions.

The Gibraltar dataset provides an opportunity to test the sensitivity of the return values with respect to the sample duration. Fig. 6 shows the westerly storm peaks above the low threshold u_1 with respect to the calendar years.

6.2. Link with the North Atlantic Oscillation

We also know that the Atlantic storm tracks are related to the North Atlantic Oscillation (NAO), i.e. the oscillation of the atmospheric pressure gradient between the Iceland low and the Azores high around a long-term mean (Hurrell, 1995). Bacon and Carter (1993) were among the first to suggest such a link. Recently, Dodet et al. (2010) used a 57-year hindcast (1953–2009) to quantify this link. They found correlation coefficients between the winter NAO index on the one hand, and the 90% *Hs* percentile, the mean wave direction winter-means and the peak period winter-means on the other hand. As for *Hs*, the Pearson correlation coefficients are close to 1 off the British Isles, close to zero off Galicia and become negative off the Moroccan coast. The coefficient is around −0.3 at the western entry of the Strait of Gibraltar. Taking account of the winter NAO index for this location thus provides valuable information.

If we look at the change in the (PC-based) winter NAO index from 1950 to 2005 (see Fig. 7), we see that from 1950 to approximately 1980 (except for 1973 to 1975), the index is mostly negative, which means the Atlantic storm tracks go preferentially southwards. After 1980, the index is mostly positive and storms sweep preferentially over northern Europe.

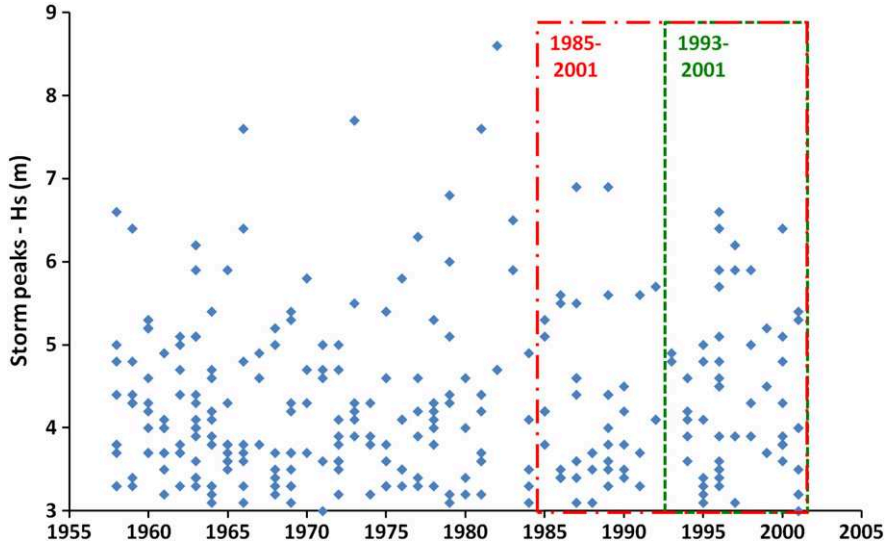

Fig. 8. Gibraltar: storm peaks above the low threshold u_1 with respect to the calendar years. Choice of the 9-year period 1993–2001 (dashed line) and of the 17-year period 1985–2001 (dashed-dotted line).

Méndez et al. (2006) use a non-stationary POT model to take into account the NAO index by allowing the GPD parameters to be time-dependent. However, it is not easy to determine the period of such an index, and the more parameters the distribution has, the more uncertainties there are in the final result. We thus kept on working with a stationary POT model.

7. Results

As the highest peak is reached in 1982 (not far from the switch between the negative and positive indices), this will be taken as a pivotal year. We will therefore study a 9-year period (1978–1986) and a 17-year (1974–1990) period, both centered on 1982, and the results will be compared with the 44-year dataset. The results are given in Table 4.

The main conclusion that may be drawn is that the return values are lower when the duration of the dataset increases. The extreme peak of 1982 (most probably very close to the 100-year wave height) clearly plays a role in this phenomenon, as its weight is greatly enhanced in the 9-year dataset compared to the 44-year one. The confidence intervals are therefore much wider in the short duration dataset. Another important fact is that the Gamma distribution is considered the best-fitting distribution by both the BIC and AIC criteria for the three datasets.

It is also noteworthy that the deviation between the 100-year wave heights for the three distributions is only 11% in the case of the 9-year sample and no more than 5–6% in that of the 17-year and the 44-year datasets.

The interest of a long dataset is clear, as the deviation between the best 100-year wave heights is around 9% between the 17-year and 44-year datasets. This interest is all the greater when the presence of an outlier is evident, as is the case here. A long period also allows cyclical regional climatic patterns with long (decadal or multidecadal) periods, such as the NAO, to be taken into account. However, engineers usually work with datasets whose duration is rather 15 to 20 years. Special attention should therefore be paid to such climatic patterns, as the dataset could cover periods with storm peaks that are lower or higher than a long-term mean.

If we choose now to study a 9-year and a 17-year period with weaker storms, the results are quite different. Let us carry out the analysis for the periods 1993–2001 (9 years) and 1985–2001 (17 years), when the NAO index is mostly positive and the storms rather weak (see Fig. 8). Results are given in Table 5.

BIC and AIC criteria now select the GPD for these two periods, though it gives far lower return values than the Weibull and Gamma distributions, whose return values remain by chance quite constant. Surprisingly, the GPD confidence intervals are narrower for the 9-year period whereas they are wider for the other two distributions.

Actually, it seems that when there is an outlier in a short-duration dataset, the GPD is much less sensitive to it than the Gamma and Weibull distributions. But when such a short duration corresponds to a calmer than usual period, the GPD returns wave heights that are too low, although it fits the data very well. This may well be the case for the Haltenbanken dataset, where no outlier appears. In spite of the likelihood-based criteria discrimination, a conservative choice for Haltenbanken would be to choose another distribution, say the Weibull one, with Hs_{100} around 15 m, which would be in accordance with observations (see for instance Magnusson et al., 2006).

It is clear that 9 years is definitely too short a period for a robust extreme wave heights analysis. When storms are thought to be stronger than usual in this period (or if the dataset contains an outlier), the GPD-Poisson model gives good results and seems quite stable. In contrast, if storms are thought to be rather weaker than usual, the GPD-Poisson model may produce return values that are too low, in spite of a very good statistical fit. In such a case, it is imperative to extend the duration of the dataset. If it is not possible, choosing the most conservative distribution could be safer than relying on the BIC/AIC criteria for design purposes.

Two conclusions may be drawn from this analysis when working with very short datasets (less than 10 years). Firstly, the interest of extending the GPD-Poisson model to other statistical distributions is manifest as we may obtain return values that are too low with this

Table 5
Gibraltar: fit characteristics and 100-year significant wave height with 90% confidence intervals for the calmer 9-year, 17-year and 44-year periods.

	9 years (1993–2001)	17 years (1985–2001)	44 years (1958–2001)
u_2 (m)	3.9	4.2	4.3
N (−)	32	43	104
λ (yr^{-1})	3.55	2.53	2.37
Min BIC/AIC distribution	GPD	GPD	Gamma
Hs_{100} GPD (m) 90% CI	6.6	7.0	8.3
	6.3–6.8	6.6–7.3	7.6–9.2
Hs_{100} Weibull (m) 90% CI	7.9	7.9	8.5
	6.9–9.2	7.1–8.9	7.8–9.3
Hs_{100} Gamma (m) 90% CI	8.7	8.4	8.8
	7.4–10.1	7.4–9.5	8.0–9.6

model. Secondly, in this case criteria based upon likelihood fail since the data are not fully representative of the local climate: engineers have to keep in mind that the ultimate goal is to provide safe design criteria and not the "purest" statistical fit.

To conclude with regard to the duration of the dataset, it may be said that a limit to the ratio between T and K, i.e. the return period and this duration, is necessary but not enough (this ratio is generally close to 5). This analysis has shown that K must be large enough with respect to the local climate in order to avoid covering only particularly weak or strong periods. We believe twenty years is a minimum period for a reasonably robust extreme wave analysis.

8. Conclusions

We carried out a complete review of a rigorous method for determining extreme wave heights using the GPD-Poisson model, in particular for choosing the high threshold. Although objective methods exist, it is clear that the choice may still be difficult. Parameters such as N and λ should be kept in mind when choosing u_2. Even so, several thresholds sometimes need to be tested.

It was also seen that although the GPD has the best theoretical justification for being selected as the asymptotic law, other distributions may give better results. Criteria for selecting the best-fitting distribution are presented. They are based upon the fit likelihood. However, analysts must always be very careful about the location of the high threshold with respect to the first exceedances, as instabilities can occur for both Gamma and Weibull distributions. We recommend using only high thresholds equal to the data values, but a better understanding of this purely mathematical phenomenon is necessary.

Sensitivity analyses for the return value and/or the criteria with respect to high thresholds are very helpful for post-checking the relevance of the choice made for u_2. However, these graphs should only be used as a verification tool, and not for decision-making.

This method was tested for two locations. As for the Haltenbanken dataset, the GPD-Poisson model had the best behavior and led to significantly lower 100-year wave heights than those calculated by the IAHR Working Group in 1993 (van Vledder et al., 1994), probably due to too short dataset duration. As for the Gibraltar location, the Gamma distribution was considered the best in relation to both the BIC and AIC criteria. Graphs illustrating the sensitivity analyses reinforced these estimations.

The interest of working on long duration datasets was also demonstrated. This interest is enhanced when the presence of outliers is suspected or when decadal or multidecadal climatic patterns may play a role. The multi-distribution approach appears to be necessary for very short datasets, although the means for discriminating the best-fitting distribution requires improvement. Indeed, we have shown that the GPD-Poisson model can lead to dangerously low return values when the analysis is carried out for a very short and rather calm period. In this particular case, if the period cannot be extended, BIC/AIC criteria may be put aside and the most conservative results may be chosen. Consequently, a dataset covering at least 20 years is strongly recommended.

We have thus a robust enlargement of the stationary GPD-Poisson model. It may be useful for engineers wishing to cover a wide range of situations. The choice of distributions proposed here is not exclusive, and others may be used. Engineers should remember that their aim is to determine safe design criteria rather than perfect statistical fits, so they must always be careful to ensure that the available data are fully representative of the local climate.

Acknowledgements

The authors wish to thank Dr. Philippe Garat from Pierre-Mendes-France University (Grenoble, France) for his help and advice.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory 267–281 Akademiai Kiado, Budapest.
Bacon, S., Carter, D.J.T., 1993. A connection between mean wave height and atmospheric pressure gradient in the North Atlantic. Int. J. Climatol. 13, 423–436.
Castillo, E., Sarabia, J.M., 1992. Engineering analysis of extreme value data: selection of models. J. Waterw. Port Coast. Ocean Eng. 118 (2), 129–146.
Coles, S., 2001. An Introduction to Statistical Modelling of Extreme Values. Springer, London.
Coles, S., Stephenson, A., 2006. ismev: An Introduction to Statistical Modelling of Extreme Values(Original S functions) (R port and R documentation files) http://www.maths.lancs.ac.uk/stephena/2006R package version 1.2.
Dodet, G., Bertin, X., Taborda, R., 2010. Wave climate variability in the North-East Atlantic Ocean over the last six decades. Ocean Model. 31, 120–131.
Efron, B., Tibshirani, R.J., 1993a. An Introduction to the Bootstrap. Chapman & Hall, London.
Efron, B., Tibshirani, R.J., 1993b. An Introduction to the Bootstrap, Monographs on Statistics and Applied Probabilities 57. Chapman & Hall, New York.
Embrechts, P., Kluppelberg, C., Mikosh, T., 1997. Modelling Extreme Events. Springer, New York.
Goda, Y., 2000. Random Seas and Design of Maritime Structures, 2nd Edition. Advanced Series on Ocean Engineering, Vol. 15. World Scientific.
Hawkes, P., Gonzalez-Marco, D., Sánchez-Arcilla, A., Prinos, P., 2008. Best practice for the estimation of extremes: a review. J. Hydraul. Res. Vol. 46 (Extra Issue 2), 324–332.
Hosking, J.R.M., Wallis, J.R., 1987. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics vol. 29, 339–349.
Hurrell, J.W., 1995. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitations. Science 269, 676–679.
Magnusson, A.K., Reistad, M., Breivik, O., Myklebust, R., Ash, E., 2006. Forecasting a 100-year wave event. Proceedings of the 9th International Workshop on Wave Hindcasting and Forecasting, Victoria, Canada, 24th to 30th September 2006.
Mathiesen, M., Goda, Y., Hawkes, P.J., Mansard, E., Martín, M.J., Peltier, E., Thompson, E.F., Van Vledder, G., 1994. Recommended practice for extreme wave analysis. J. Hydraul. Res. 32 (6), 803–814.
Méndez, F.J., Menéndez, M., Luceño, A., Losada, I.J., 2006. Estimation of long-term variability of extreme significant wave height using a time-dependent Peaks Over Threshold (POT) model. J. Geophys. Res. 111, C07024. doi:10.1029/2005JC003344.
Panchang, V.G., Gupta, R.C., 1989. On the determination of three-parameter Weibull MLE's. Commun. Statist. Simula. 18 (3), 1037–1057.
Pickands, J., 1975. Statistical inference using extreme order statistics. Ann. Statist. 3, 119–131.
R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. 3-900051-07-0. http://www.R-project.org, Vienna, Austria
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.
Thompson, P., Cai, Y., Reeve, D., Stander, J., 2009. Automated threshold selection methods for extreme wave analysis. Coastal Eng. 56, 1013–1021.
Van Vledder, G., Goda, Y., Hawkes, P., Mansard, E., Martin, M.J., Mathiesen, M., Peltier, E., Thompson, E., 1994. Case studies of extreme wave analysis: a comparative analysis. Proceedings of the Second Symposium on Ocean Wave Measurement and Analysis, New Orleans, Louisiana, USA, pp. 978–992.

1.2. NATURAL HAZARDS AND EARTH SYSTEM SCIENCES 2014: A TWO-STEP FRAMEWORK FOR OVER-THRESHOLD MODELLING OF ENVIRONMENTAL EXTREMES

Main publications citing this paper (source Google Scholar)

Caballero-Megido C., Hillier J., Wyncoll D., Bosher L., Gouldby B., 2017. Comparison of methods for threshold selection for extreme sea levels. *Journal of Flood Risk Management*.

Chen C., Beltrame G., 2017. An Adaptive Markov Model for the Timing Analysis of Probabilistic Caches. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, **23**(1), 12.

Durán-Rosal A. M., Fernández J. C., Gutiérrez P. A., Hervás-Martínez C., 2017. Detection and prediction of segments containing extreme significant wave heights. *Ocean Engineering*, **142**, 268-279.

Hamdi Y., Bardet L., Duluc C. M., Rebour V., 2015. Use of historical information in extreme-surge frequency estimation: the case of marine flooding on the La Rochelle site in France. *Natural Hazards and Earth System Sciences*, **15(7)**, 1515-1531.

Jane R., Dalla Valle L., Simmonds D., Raby A., 2016. A copula-based approach for the estimation of wave height records through spatial correlation. *Coastal Engineering*, **117**, 1-18.

Lerma A. N., Bulteau T., Lecacheux S., Idier D., 2015. Spatial variability of extreme wave height along the Atlantic and channel French coast. *Ocean Engineering*, **97**, 175-185.

Lin-Ye J., Garcia-Leon M., Gracia V., Sanchez-Arcilla A., 2016. A multivariate statistical model of extreme events: An application to the Catalan coast. *Coastal Engineering*, **117**, 138-156.

Mondal A., Mujumdar P. P., 2015. Return levels of hydrologic droughts under climate change. *Advances in Water Resources*, **75**, 67-79.

Mortlock T. R., Goodwin I. D., 2015. Directional wave climate and power variability along the Southeast Australian shelf. *Continental Shelf Research*, **98**, 36-53.

Naghettini M., de Andrade Pinto E. J., 2017. At-Site Frequency Analysis of Hydrologic Variables. In *Fundamentals of Statistical Hydrology* (pp. 311-389). Springer International Publishing.

Rutkowska A., Willems P., Niedzielski T., 2016. Relation between design floods based on daily maxima and daily means: use of the Peak Over Threshold approach in the Upper Nysa Kłodzka Basin (SW Poland). *Geomatics, Natural Hazards and Risk*, 1-22.

Silva A. T., Naghettini M., Portela M. M., 2016. On some aspects of peaks-over-threshold modeling of floods under nonstationarity using climate covariates. *Stochastic environmental research and risk assessment*, **30(1)**, 207-224.

Silva A. T., Portela M. M., Naghettini M., Fernandes W., 2017. A Bayesian peaks-over-threshold analysis of floods in the Itajaí-açu River under stationarity and nonstationarity. *Stochastic Environmental Research and Risk Assessment*, **31(1)**, 185-204.

Solari S., Egüen M., Polo M. J., Losada M. A., 2017. Peaks Over Threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. *Water Resources Research*, **53(4)**, 2833-2849.

Vukmirović V., Vukmirović N., 2017. Stochastic analysis of flood series. Hy*drological Sciences Journal*, 1-15.

Weiss J., Bernardara P., Benoit M., 2014. Formation of homogeneous regions for regional frequency analysis of extreme significant wave heights. *Journal of Geophysical Research: Oceans*, **119(5)**, 2906-2922.

Weiss J., Bernardara P., Benoit M., 2014. Modeling intersite dependence for regional frequency analysis of extreme marine events. *Water Resources Research*, **50(7)**, 5926-5940.

Nat. Hazards Earth Syst. Sci., 14, 635–647, 2014
www.nat-hazards-earth-syst-sci.net/14/635/2014/
doi:10.5194/nhess-14-635-2014

Natural Hazards and Earth System Sciences
Open Access

A two-step framework for over-threshold modelling of environmental extremes

P. Bernardara[1,2,5], **F. Mazas**[3], **X. Kergadallan**[2,4], **and L. Hamm**[3]

[1]LNHE, EDF R&D, Chatou, France
[2]Université Paris Est, Saint Venant Laboratory for Hydraulics, ENPC, EDF R&D, CETMEF, Chatou, France
[3]ARTELIA, Grenoble, France
[4]CEREMA, Brest, France
[5]EDF Energy R&D UK Centre, London, UK

Correspondence to: P. Bernardara (pietro.bernardara@edf.fr)

Received: 1 August 2012 – Published in Nat. Hazards Earth Syst. Sci. Discuss.: –
Revised: 1 November 2013 – Accepted: 21 November 2013 – Published: 20 March 2014

Abstract. The evaluation of the probability of occurrence of extreme natural events is important for the protection of urban areas, industrial facilities and others. Traditionally, the extreme value theory (EVT) offers a valid theoretical framework on this topic. In an over-threshold modelling (OTM) approach, Pickands' theorem, (Pickands, 1975) states that, for a sample composed by independent and identically distributed (i.i.d.) values, the distribution of the data exceeding a given threshold converges through a generalized Pareto distribution (GPD). Following this theoretical result, the analysis of realizations of environmental variables exceeding a threshold spread widely in the literature. However, applying this theorem to an auto-correlated time series logically involves two successive and complementary steps: the first one is required to build a sample of i.i.d. values from the available information, as required by the EVT; the second to set the threshold for the optimal convergence toward the GPD. In the past, the same threshold was often employed both for sampling observations and for meeting the hypothesis of extreme value convergence. This confusion can lead to an erroneous understanding of methodologies and tools available in the literature. This paper aims at clarifying the conceptual framework involved in threshold selection, reviewing the available methods for the application of both steps and illustrating it with a double threshold approach.

1 Introduction

A reliable estimation of extreme natural hazard is important for the protection of remarkable natural sites, urban areas, industrial facilities, etc. In particular, extreme natural events include floods, heavy rainfalls, high and low temperatures, strong winds, high sea levels or sea surges, oceanic waves, among many others.

Traditionally, the estimation of the probability of occurrence of such extreme events is performed by fitting a probability distribution to a sample of historical observations for a given phenomenon observed at a given site, usually recorded as a time series of observations. In this framework, the extreme value theory (EVT) (Fréchet, 1928; Gnedenko, 1943; Gumbel, 1958; Pickands, 1975) offers a sound theoretical framework.

In particular, Pickands' theorem (that can be seen as a central limit theorem for extreme values) states that, in a sample composed by independent and identically distributed (i.i.d.) values, the distribution of the data exceeding a given threshold converges towards a generalized Pareto distribution (GPD) (Pickands, 1975). Following this theoretical result, the over-threshold modelling (OTM) approach widely spread in extreme value analyses, together with the application of the GPD, (Davison and Smith, 1990; Simiu and Heckert, 1995; Embrechts et al., 1997; Palutikof et al., 1999; Coles, 2001; Mackay et al., 2001; Pandey et al., 2001; Rosbjerg and Madsen, 2004; Ribatet et al., 2007). It is widely recognized that the choice of the threshold is a critical point

in this approach and the final estimation could significantly depend on its value (Onoz and Bayazit, 2001; Li et al., 2012).

In this theoretical framework, the choice of the appropriate threshold should be a statistical optimization procedure. Given an empirical sample of i.i.d. observations, the selected threshold must be high enough to meet the hypothesis of convergence on the GPD but it should be low enough to limit the variability of the GPD parameter calibration on the subsample of observations over the threshold (Beirlant et al., 1996). This is the well-known dilemma between bias and variance.

However, environmental variables are often handled as time series, i.e. discrete realizations of this variable, coming from either observation or modelling, far from being i.i.d. More precisely, environmental time series are often composed by dependent values because of the strong temporal autocorrelation (e.g. Zawadzky, 1987; Smith, 1988; Colombo et al., 1999; Walton, 2000; Marani, 2003; Bernardara et al., 2006). The autocorrelation is indeed explained by the dynamical behaviour of the subjacent physical system and its momentum. Since the pioneering works of Hurst (1951), some studies even suggest that the autocorrelation of some environmental time series could be infinite (Schertzer and Lovejoy, 1997; Elek and Markus, 2004; Koscielny-Bunde et al., 2006).

An attempt to cope at the same time with EVT and autocorrelated data is the introduction of the extremal index, (Leadbetter et al., 1983; Smith and Weissman, 1994; Embrechts et al., 1997; Ancona-Navarrete and Tawn, 2000; Coles, 2001; Beirlant et al., 2004). The extremal index is an extra parameter that allows taking into account data auto correlation on the extreme value theorem. The extremal index represents the reciprocal of the mean size of event clusters. Estimating this index allow to apply the EVT theorem results directly on a series of auto correlated observations.

However, in general, the EVT cannot be applied directly to the observed data and a data pre-processing is needed in order to build the i.i.d. sample required by its hypothesis. This data pre-processing is often called physical declustering, because it tends to extract independent observations from the time series, which are naturally (physically) clustered.

Moreover, environmental time series can be composed by nonidentically distributed (nonhomogeneous) values. Indeed, natural phenomena can have very different physical genesis, they can exhibit strong seasonality of the observed phenomena or they can depend on other covariates. Among others, Adamowski (2000), Garavaglia et al. (2010) and Allamano et al. (2011) show that mixing heterogeneous samples can lead to biased estimation of extreme value probability of occurrence. Garavaglia et al. (2010) introduced a compound distribution for extreme rainfalls taking into account seasonality and different physical genesis. For wave heights, probability distribution and even time series autocorrelation may depend on direction, fetch, water depth and other covariates (Mathiesen et al., 1994; Jonathan and Ewans, 2007; Taylor et al., 2009; MacKay et al., 2010; Mazas and Hamm, 2011). Many possibilities exist for getting time series of homogeneous physical phenomena (clipping, decomposition, cleaning, etc.) but they are beyond the scope if this paper. In the following, if not stated differently, it will be considered that the time series are identically distributed, but they are still not composed of independent values.

Accordingly to the previous considerations and within the framework of the OTM, the constitution of the sample for the statistical inference of the extreme return levels of the environmental variable logically requires two successive and complementary steps: the physical declustering and the statistical optimization.

The need for both declustering and statistical optimization was generally recognized in the past. However they were often confused or merged together, arising methodological questions and confusing the meaning of the two operations. For instance, Lang et al. (1999) stated that "two different approaches can be adopted for threshold selection: the first one is based on physical criteria [...] and the second one is based on purely mathematical and physical considerations", but they neither separate both steps, nor recognized their complementarity. It is indeed important to clarify which, among the numerous parameters to be defined for an OTM analysis, often largely arbitrary (Takvor and Panagiota, 2001), are involved in the physical declustering and which ones are involved in the optimization procedure. Even when in the past the two steps were performed separately (Dupuis, 1998; Egozcue et al., 2005; Bernardara et al., 2008, 2011; Garavaglia et al., 2010; Bardet et al., 2011; Mazas and Hamm, 2011; Wahl et al., 2011), or when approaching the two steps at the same time (i.e. the extremal index approach), the underlying concepts were not clearly exposed.

With the previous considerations in mind, this paper aims at clarifying the general conceptual framework of threshold selection for over-threshold modelling, distinguishing in particular the physical declustering procedure from the statistical optimization. A large literature review of the existing methods for both steps is given.

The main improvement of this effort of review and clarification is that, distinguishing both steps, the existing methods can be used in the right context, namely the physical declustering can be done based on physical arguments and the statistical optimization is performed later with purely statistical methods.

Note that this theoretical discussion is relevant in a multidisciplinary context, including different environmental applications (e.g. hydrology, meteorology, ocean sciences). This is an important point toward sharing of the knowledge of OTM techniques between different domains and different scientific communities. As a consequence, in this paper literature review and examples are based on environmental phenomena as different as floods, heavy rainfalls, extreme winds, high sea levels, extreme sea surges, and oceanic waves.

The paper is organized as follows. In Sect. 2, the two steps are depicted and the methodological framework is clarified. In Sect. 3 a review of the current methodology for physical declustering and statistical optimization of the statistical threshold is given. It is also shown that often in the past the two steps were merged together and sometimes confused. In Sect. 4 this general framework is applied to two different case studies for the estimation of hydrological and maritime extreme observations. In particular a double threshold is introduced. In Sect. 5 some final conclusions are drawn.

2 Distinguishing two steps for OTM

Let $Z(t)$ be a time series (or, more generally, a spatial field) of discrete realizations of a given environmental variable, Z (e.g. a river discharge, a significant wave height, a sea level, a surge, a temperature, a wind speed, etc.) at a given resolution, Δt (i.e. time step, either regular or irregular). This can be the result of observation or modelling. The time series is assumed to be identically distributed.

For extreme value estimation applications, times series lasting several years are generally needed. For this reason, their size can be very important, depending on its duration, K, and time step, Δt. For example, a daily series lasting 20 yr contains around 7000 values, while an hourly series lasting 5 yr contains more than 40 000 values. Note also that environmental measures are often submitted to failure of measurement device, thus the time series can be incomplete.

2.1 Step 1: physical declustering

The physical declustering aims at extracting a sample of i.i.d. values, X_i, from the time series, $Z(t)$. This step can be viewed as an identification procedure, through purely physical consideration, of independent events. We claim that the notion of event and its correct understanding is a fundamental concept of EVT analysis applied to a time series. An event is defined here as a continuous physical phenomenon of the environmental variable, notably out of its mean regime, as can be instinctively comprehended by anyone: a storm for wind speed or wave height, a flood for river discharge, a heat or cold wave for temperature, a drought for rainfall, etc.

The events have a given duration that is often longer that the resolution Δt of the time series. In this case an event is composed of a set of consecutive discrete realizations of the variable, called cluster. The analyst should then define a random variable X describing the events. Very often this event-describing variable X represents the maximum value of $Z(t)$ within the event, or cluster, and is often called the "peak" of the event. However, X may also be the result of any mathematical transformation of the cluster values: this is the case for the volume of a flood, for example. Generally speaking, X can be any characteristic of the event. The actual definition of X will depend mainly 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, of the time step of the series and on the physic and dynamic characteristics of the observed process. For instance, the declustering of a daily temperature time series will require different techniques than the declustering of hourly rainfall observations. In general, the knowledge of the physics of the studied phenomenon drives the physical declustering choices that should guarantee: first the independence of the selected events and second that no event is omitted in the process.

It is quite important to stress that Z and X are per se different random variables. It is quite clear in the case of daily river discharge vs. flood volume, but it is also true in the case of three-hourly wind speed vs. the peak wind speed of a storm, for instance. In particular, even if each event can be associated to a particular instant of occurrence on the time line, X does not depend on time any more.

A sample of N_{T} i.i.d. values X_i is thus obtained. Its size is much lower than the size of the time series $Z(t)$, generally in the order of few hundreds instead of several thousands.

In Sect. 3 a review of practical methods for physical declustering of environmental time series is given.

2.2 Step 2: statistical optimization

The physical declustering allowed the setting up of an i.i.d. sample: the EVT hypothesis is now met and the well-known statistical models can be applied.

Let us introduce u_{s}, which stands for statistical threshold and let us define the random variable $Y = X - u_{\mathrm{s}}$, given $X > u_{\mathrm{s}}$. Y is the exceedance of X above the threshold u_{s}. Thus a sample Y_i of size N can be defined from the sample $X_i : Y_i = X_i - u_{\mathrm{s}}$, given $X_i > u_{\mathrm{s}}$. Note the sample size reduction ($N \leq N_{\mathrm{T}}$) as the X_i values falling below (or equal to) u_{s} are excluded from the analysis.

Within the theoretical framework of OTM, and in particular according to Pickands' theorem, when u_{s} increases, the probability distribution of the sample Y_i converges toward the generalized Pareto distribution (GPD) whose cumulative distribution function of the GPD, in its three parameter formulation, is given by

$$F(y) = 1 - \left[1 + k\left(\frac{y-\mu}{\sigma}\right)\right]^{-\frac{1}{k}}, \quad (1)$$

where k, $k \neq 0$, is the shape parameter, also indicated as ξ (or $-\xi$) in statistic literature, σ is the scale parameter and μ is the location parameter, with $y > \mu$ for $k > 0$ and $\mu < y < \mu - \sigma/k$ for $k < 0$. Note that following Pickand's theorem, the location parameter is generally set equal to zero. Note also that the modified scale parameter, $\sigma^* = \sigma - k u_{\mathrm{s}}$, is often used as scale indicator. u_{s} is thus the optimal threshold providing the best compromise between the convergence of the Y_i through a GPD (bias minimization) and the necessity

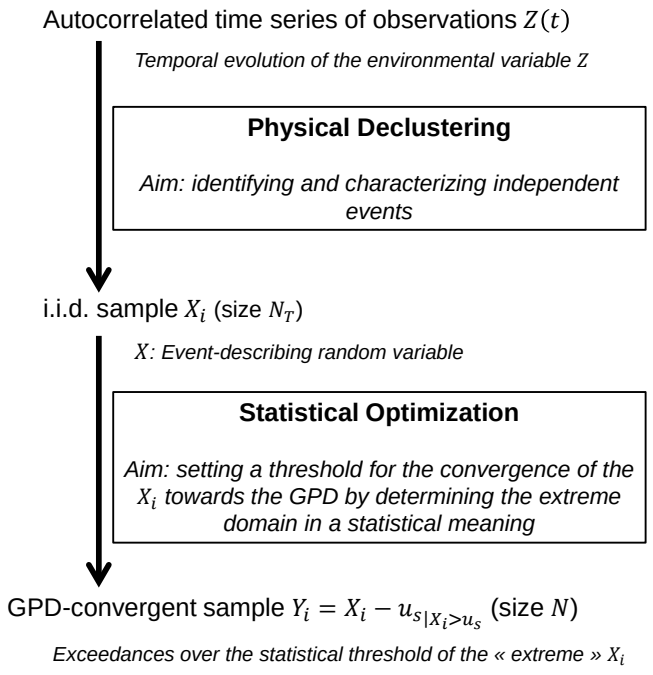

Fig. 1. General framework for identifying extreme data for over-threshold modelling (OTM).

to keep enough data Y_i for the estimation of its parameters (variance minimization). The extrapolation of the estimated GPD will yield the estimated return levels (or extreme quantiles). 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; see Sect. 3 for a general review. It does not depend on the particular random variable (environmental or not) and it is general for every extreme value application.

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

3 Review of methods for physical declustering and for statistical optimization

In this section a literature review of the physical declustering and the statistical optimization techniques is given. Its aim is not to particularly recommend any of these to the detriment of the others, but rather to catalogue the different practices. At the end of this section a discussion is proposed to understand why and how these two steps were often merged and confused in the past and we point out that this knowledge is important in order to perform correctly both steps.

3.1 Methods for physical declustering (Step 1)

3.1.1 General principles

As stated in the previous sections, the physical declustering procedure aims at building the i.i.d. sample X_i, on which all the statistical analyses are based, by identifying and characterizing the events.

Generally speaking, this procedure does not require the "over-threshold" concepts per se. One could imagine, for instance, manually extracting a sample of extreme and independent events from a historical record of observations. Garavaglia et al. (2010) introduced the concept of central rainfall, defined as the rainfall observation $Z(t)$ for which $z(t-1) < z(t) > z(t+1)$ for declustering the rainfall observations series.

However, an overview of the literature (see in particular Lang et al., 1999) shows that the techniques based on the definition of a threshold to be exceeded spread widely. Adamowski (2000) claims, moreover, that the choice of the threshold for declustering is also a critical step in order to select homogenous events and to avoid merging different populations of observations (this is the concept of identical distribution mentioned above).

Following such an approach, the clusters (events) are usually defined as the series of consecutive values of $Z(t)$ above a given threshold, which is called here u_p, for physical threshold. Note that this approach was often called partial duration series sampling in the past (Cunnane, 1973; Rosbjerg, 1985; Rosbjerg et al., 1992; Madsen and Rosbjerg, 1997).

As mentioned in Sect. 2.1, the event-describing variable X could be any transformation of the values of Z within the cluster. For example, a temporal integration of the consecutive value of river discharge over the physical threshold is sometimes used in hydrology to characterize the volume of a flood. However, in most environmental applications, the maximum value of $Z(t)$ observed during the cluster i, or event peak X_i, is retained to describe the event. For this reason the name of peak over threshold (POT) spread widely in the literature for this sampling technique.

It has been stressed above that the physical declustering must not only identify the events, but also guarantee their independence. As a consequence, the actual definition of a cluster (event) generally relies upon two different families of parameters: on the one hand, the physical threshold u_p, whose value is expressed in the same units as $Z(t)$ (but not necessarily as X_i, e.g. the volume of a flood); on the other hand, one or several parameters needed for ensuring the independence of the different clusters.

Several physically based criteria are available in the literature for the definition of the physical threshold and for the characterization of the independence of the clusters.

3.1.2 Physical threshold

The choice of the threshold u_p is a key step for the identification of events. Several approaches were proposed in the past for the definition of the physical threshold.

First, and quite simply, the choice of the physical threshold can be defined by an expert prior knowledge. Expert prior knowledge can be based, in a heuristic approach, on the practical consequences for the phenomenon to cross a given threshold value. For example, Bonazzi et al. (2012) define a wind velocity threshold corresponding to the level at which damage to buildings is likely to occur. Perception threshold, defined by expert estimation, is used to perform OTM for historical and non-systematic observation, (Ouarda et al., 1998; Barriendos et al., 2003; Payrastre et al., 2005, 2011; Hamdi et al., 2013).

For threshold selection, Lang et al. (1999) defined the mean number of events per year, $\lambda_T = N_T/K$, where K is the total duration of the time series, in years, and N_T is the number of physical events to be selected (and also the sample size of the X_i). Then they discussed the evolution of λ_T when the threshold increases and they identified four domains. First, the threshold is below the minimum of the time series: thus the entire time series is considered as an event, though it has no physical sense. Second, as the threshold value raises between the minimum and (roughly) the mean value of the series λ_T increases: the higher the threshold, the more events there are. This actually means that just shortfalls below a low threshold are identified. Third, λ_T reaches a maximum (when the threshold is close to the mean of the series) and begins to decrease. Fourth, when the threshold is larger than the maximum of the series, no more exceedances can be extracted and $\lambda_T = 0$. The authors require that λ_T be in the third domain, though far from both the lower and upper limits of the domain. In the proposed conceptual framework, it can be stated that events (as defined above) are identified in the third domain. However, though this recommendation is relevant, it is easily fulfilled and is not specific enough to be really useful in practical applications.

Another widely used approach relies on the idea that u_p can be tuned in order to obtain a physically reasonable value of λ_T. More generally, the choice of the actual number λ_T is based on expert knowledge of the physics and the dynamics of the process. For example, in hydrological applications it is suggested to choose $\lambda_T < 5$, which is a large number of floods to observe, for a given site, on average, per year. Obviously, this number should depend on an on-site hydrological regime. Working on skew surges, in a regional analysis framework, (Bernardara et al., 2011) suggested to set u_p so that $\lambda_T = 1$, while for local analysis Walton (2000) fixed $\lambda_T = 3$. Analyzing significant wave heights, (Mazas and Hamm, 2011) suggested that the value of λ_T should be roughly between 5 and 10, also depending on the value of the time series duration (closer to 5 for long time series, closer to 10 for short ones). Tawn and Vassie (1989) suggested a value of λ_T around 5 for extreme sea surge. Floris et al. (2010) analyzed λ_T in a framework of extreme rainfall analysis.

Some authors set the threshold using a given quantile of the time series $Z(t)$. Ruggiero et al. (2010) set it to the 99.5th percentile of the data, working with wave heights. Rosbjerg et al. (1992) suggest calculating the physical threshold as the mean value of the observed series plus three standard deviations. Notice that, though it looks like a statistical approach, there is no optimization process in it.

These criteria for the selection of the relevant physical threshold were compared and simultaneously used in the past, for instance Ntegeka and Willems (2008) stated that "an extreme event can be selected based on frequency, intensity, threshold exceedances or physical expected impacts".

3.1.3 Parameters for ensuring the independence of the events

In order to ensure the independence of the selected events, many physically based criteria have been developed. Several of these criteria are recurrent in the literature and will be presented here.

The most common techniques consist in setting temporal parameters, most of the time based on the minimal time lag between two events. The idea is quite simple: after a given period of time, the autocorrelation between the observations becomes negligible and two events can be safely considered independent. The definition of this time lag is directly derived from the physics of the natural phenomenon: it should be longer than the typical duration of the physical processes (usually meteorological ones) generating the events. Thus it can be set by an expert prior knowledge. However, the time lag should not be too long in order to avoid discarding independent events and thus missing valuable information. For instance, in north-eastern Europe, extreme wave heights may be generated by successive storms moving along the storm track every 24 h or so; therefore setting the time lag to 48–72 h could lead to miss information. Many applications of this approach are available in the literature: Egozcue et al. (2005) studied wave height hazards along the Mediterranean coast of Spain and set the time lag to 4 days; Haigh et al. (2010) studied the extreme sea levels along the English Channel and required the surge peaks to be separated by 30 h at least; USWRC (1976), Cunnane (1979) and Lang et al. (1999) imposes that successive river flood events be separated by at least as many days as five plus the natural logarithm of square miles of the basin area. Willems (2000) required that two rainfall events are separated by at least a 12 h lag.

The time lag can also be defined using the autocorrelation function of the time series $Z(t)$. For extreme wave heights, Mathiesen et al. (1994), propose requiring that it cannot be longer than the time interval for which the autocorrelation function of the series drops under 0.3–0.5. In a similar way, while studying storm surge extremes along the US East

Coast, Walton (2000) established the typical duration of an event via the autocorrelation of the surge series and found that the drop off of the autocorrelation function to a noise level value close to zero was on the order of 24–72 h. Note that some authors define the time lag between the end of an event and the beginning of another while others define it between two peaks.

In general, both physical and statistical methods aim at the definition of the correlation length, thus they should naturally converge.

Other temporal parameters can be used; for instance, in their atlas of waves along the Italian coasts, (Franco et al., 2004) allowed short fluctuations of the time series below the threshold u_p up to a maximal duration (6 h) and also set a minimal storm duration (12 h). This last parameter may be useful for some applications; for instance, waves generated by a very short storm will not cause damage to a breakwater. In the sea wave analysis field, (Takvor and Panagiota, 2001) extracted independent the sea state by looking at wave energy reductions between consecutive time steps. In contrast, (Smith, 1988) examined the typical duration of extreme wave conditions and did not see any rationale for using such a parameter.

Another technique consists in using a secondary threshold: in this approach, two events are considered independent when the signal $Z(t)$ falls below this value. In particular, this secondary threshold may be defined as a fraction of the physical threshold u_p (in this case the value of this fraction can indeed be considered as the parameter to set) or as a fraction of the peak value of one of these events. For instance, (USWRC, 1976) requires (among other criteria) that the intermediate flows between two consecutive flood peaks must drop below 75 % of the lowest of these two peaks, while (Cunnane, 1979) imposes that the flow must drop below $2/3$ of the first peak value.

The independence of the selected events (or more generally the independence of the selected clusters) has been checked in the past via the analysis of the probability distribution of the occurrences of the events for a given time interval. In fact, the Poisson distribution (Haight, 1967) is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time if these events occur with a known average rate and independently. Thus, if the number of occurrences of events follows a Poisson distribution, the events are supposed to be independent (Cunnane, 1979; Rosbjerg et al., 1992; Lang et al., 1999). Some authors, (e.g. Ashkar and Ouarda, 1996, Silva et al., 2011) selected the physical threshold corresponding to the best adaptation of the number of occurrences to the Poisson distribution and checked the independence hypothesis looking at the uniform distribution of the arrival time of events depending on the observation period support.

Note that for several of these parameters, their value should be somehow dependent on the value of u_p. For instance, if the time lag between two events is defined between the end of the first and the beginning of the second one, or if a fraction of u_p is used to ensure the independency, the values to be considered could be different if u_p is rather low or rather high. This is particularly true for the minimal event duration or the maximal duration of fluctuations below u_p, even though these parameters are quite scarcely used.

3.2 Methods for statistical optimization (Step 2)

Once the sample of the i.i.d. data X_i is built, the statistical optimization consists in choosing the relevant value of the statistical threshold u_s to take into account for the estimation of the GPD model on the observations exceeding the threshold, toward which the sample is supposed to converge (Beirlant et al., 1996).

The threshold selection criteria here are statistically based and they aim at meeting the EVT hypothesis and the best compromise between bias and variance. In this step, the question is which ones of these events are extreme from a statistical point of view?

A first class of such methods is based on the maximization of the goodness of fit between the probability distribution and the data or the minimization of the asymptotic mean square error of the estimators. Several authors suggested choosing the value of u_s providing the best GPD adaptation to the empirical data. That can be done through the optimization of χ^2 or the Kolmogorov–Smirnov test. For instance, (Bernardara et al., 2011) used this approach for fitting a regional surge probability distribution. Anderson–Darling (AD) goodness-of-fit test is suggested and employed by Choulakian and Stephens (2001) and Haylock (2011). The adaptation of GPD to empirical data above the threshold could also be checked via the L-moments. In particular, for GPD the relation between L-moments of order 3 and 4 is known and L-moments plot technique can be used (Hosking and Wallis, 1997).

A similar class of methods are based on the minimization of the variance estimation of the Hill semi-parametric estimator of the tail index, (Hill, 1975; Hall, 1982; De Haan and Peng, 1998). (Beirlant et al., 1996) and (Willems 1998) introduced a systematic methodology based on these principles. In (Willems, 2000) an application to rainfall observations is given. (Neves and Fraga Alves, 2004) give a short review on these methods and they propose an automatic selection procedure.

A well-known property of the GPD is that the shape and modified scale parameters will remain constant when the threshold increases. Following this property, (Davison and Smith, 1990; Lang et al., 1999; Egozcue et al., 2005) suggested choosing the threshold so that the mean of the exceedances above the threshold, $E(X - u_s)$, is a linear function of the threshold value, indicating a range where the GPD parameters are not depending on the threshold selection. This technique is also known as MRL (Mean Residual Life) plot, (Coles, 2001). In this framework, (Beguéria, 2005) chose the

value of u_s in maximizing the fit of a linear function on the mean excess function.

(Coles, 2001) also proposed the STM (stability method) consisting in defining an optimum where the shape parameter of the GPD distribution is approximately constant for small threshold changes. (Mazas and Hamm, 2011) performed a sensitivity analysis of shape and modified scale parameters with respect to the (statistical) threshold in order to identify domains of stability. Then they selected the lowest threshold (minimization of variance) of the highest domain of stability (minimization of bias). (Thompson et al., 2009) used the property of stability of the GPD modified scale parameter to introduce a procedure for automatizing the threshold selection. They define 100 equally spaced threshold values between the median of the time series $Z(t)$ and its 98th percentile. For each value, the stability of the modified scale parameter is tested by the Pearson normality tests. The first (i.e. lowest) value satisfying the test is retained as the statistical threshold.

The stability of some relevant quantiles of the GPD distribution has been used in the past for the selection of the optimal u_s value, for instance by Rosbjerg et al. (1992), among many others.

3.3 Review discussion

The concept of the exceedances over a given threshold was used both for physical declustering and statistical optimization (Smith, 1984; Lang et al., 1999; Parent and Bernier, 2003).

This is explained by the fact that this concept is, on the one hand, useful for defining events as independent clusters of observations and, on the other hand, is also consistent with the EVT concepts of GPD convergence.

However, the use of the same concept of "exceedances over a threshold" for the two different steps of the analyses led to some incoherencies and confusions.

First of all, as pointed out in the previous section, it should be highlighted that the domains of application of the two steps are different. The declustering procedure applies to highly correlated data, such as the environmental time series of observations and it was studied mainly by earth scientists in the past. The statistical threshold optimization applies to a large number of statistical problems as it was treated in the past mainly by the statistical community. That leads to some incoherencies on the vocabulary used and on the concepts definition. For example, (Takvor and Panagiota, 2001), in a review of declustering techniques, called the physical declustering "statistical pre-processing", a definition that can confuse the reader. In a similar way, the concept of "POT method" often includes the whole methodology for the determination of the probability of occurrence of the extremes values, including the GPD model, while it should be restricted to the declustering step. The same comment holds for the "partial duration series" definition which refers to the first part of physical declustering but it was used in the past to indicate the whole analysis. Also, note that following the EVT vocabulary, the word "extreme" is restricted to the values exceeding the threshold u_s while the full sample of the X_i represent the whole i.i.d. population describing the different events. However, in practice, the X_i are often arbitrarily considered as an "extreme" population, which may be confusing.

Another example of incoherencies can be found in the number of data to be selected. As explained in section 3.1, (Lang et al., 1999) pointed out that the number of peaks over the threshold, X_i can decrease but also increase when the threshold u_p increases. This is logical in a physical point of view, because the sample of the X_i completely changes depending on the value of u_p. However, for the statistical optimization threshold u_s the number of exceedances over the threshold Y_i must decrease when the threshold u_s increases. Introducing the concept of physical event to be identified in the time series is thus much relevant for understanding such a behaviour.

Note also that, following the statistical theory (Pickands, 1975), the EVT requires choosing all the values over a given threshold, u_s, and not only some of them, as in the case of physical declustering using the peaks over u_p.

Moreover, the EVT states that the Y_i sample converges towards a GPD distribution, while the X_i sample, representing a whole population of i.i.d. events could be described by any statistical model. Indeed, several authors (Goda, 1988, 2010, 2011; Mathiesen et al., 1994; Goda et al., 2010) considered other distributions than the GPD (or GPD family).

In practice, another strong rationale arises for separating both steps. As shown by the literature review in Sect. 3.2, most of the methods for determining the statistical threshold require testing many values of u_s. If the physical declustering has not been performed prior to this, it will have to be done as many times as there are tested values of u_s, instead of just once. If one keeps in mind that the declustering of time series, whose size can be up to several hundreds of thousands of data, can be quite computer intensive, the interest of running this step once for all instead of 10 to 100 times is obvious. Furthermore, it has been shown in Sect. 3.1 that during the declustering process, the parameters ensuring the independence of the events may depend on the u_p threshold value, or be relevant for a small range of u_p values only. Then there is much interest in setting them accurately with regard to the value of u_p once and for all, instead of repeating this many times. This will be illustrated in the second case study in Sect. 4.3.

Note that the clusters population, namely the structure of the events, described by the clusters of $Z(t)$ over u_p, present an interest in itself. It is important, for some application to calculate some statistics such as the probability distribution of the event size or duration, the internal correlation or the shape characteristics of a general event. This shows again the interest in separating the two steps.

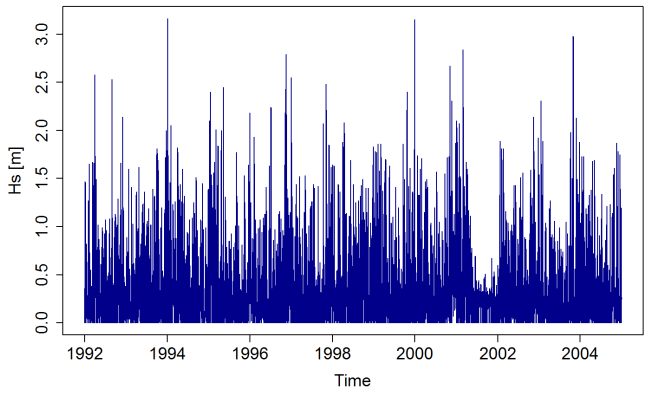

Fig. 2. Swell significant wave height time series off Marseille.

Even though the methodological framework was not yet clarified, in the past few authors highlighted some points which direct attention to the different meaning and application of the two steps. A very good illustration is provided by Cruise and Arora (1990) who noticed that the threshold level for physical declustering often had to be raised significantly to meet exponentially based tests on the POT distribution (Lang et al., 1999).

Some other authors in the past proposed extreme value analyses in which the two steps are well distinguished. For example (Dupuis, 1998; Egozcue et al., 2005; Bardet et al., 2011; Bernardara et al., 2011; Mazas and Hamm, 2011) applied two different thresholds, one for the physical declustering, the other for the statistical optimization. (Garavaglia et al., 2010) used the central rainfall concept for physical declustering and arbitrarily fixed the u_s at the 70 % quantile of the distribution, without any optimization. (Bernardara et al., 2008) used classical declustering criteria for daily discharge series and a specific optimization for the shape parameter of the GPD distribution.

However, the theoretical framework distinguishing two steps was not clearly defined. Hence, as a conclusion of this review effort, we deem that the lack of the concept of event is a major cause of the confusion observed in the past. It is useful for the understanding of EVT analyses of auto-correlated time series and it has a sound physical basis.

4 Applications to environmental variables

4.1 The double threshold approach

In order to illustrate the proposed framework, and in coherence with the literature, we propose here to provide both physical declustering and statistical optimization through a threshold approach.

This approach is applied to two case studies, a wave height study and an extreme discharge study, in order to illustrate that the methodological framework is valid in different fields of natural hazard estimation.

It is shown that this approach allows selecting the correct techniques and carrying out a complete analysis, extracting all the relevant information.

Table 1. Summary of parameters for the two case studies.

	Wave heights, Marseille	Discharge/flood volume, Rieutord
K [yr]	13	19.33
Δt	3 h	1 day
n	38 005	7062
u_p	1.4 m	10 $m^3 s^{-1}$
λ [events yr^{-1}]	10	3.6
N_T	130	70
u_s	1.87 m	6.48 Mm^3
N	43	25

4.2 Wave heights

We consider in this first illustrative example a time series of simulated three-hourly significant wave heights H_s offshore Marseille, France (5.3104° W, 43.3460° N; water depth: 34 m). The duration of the data is $K = 13$ yr, and the size of the time series is $n = 38,005$ data. In order to ensure the homogeneity of the data, a decomposition of the sea states have been performed and only the swell component have been retained. The H_s time series is plotted in Fig. 2.

In this case study, Z is a three-hourly significant wave height of the swell component (in metres), the events to be identified are swell storms and they are classically described by the random variable X "storm peak" (in metres), that is, the maximum value within the cluster. A physical threshold has been set in order to obtain a sample of $\lambda_T = 10$ 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. The declustering has been performed by using a minimal duration of 24 h between two storms to ensure their independence. Furthermore, a minimal storm duration of 6 h has been set (because very short events do not cause important damage to coastal structures) and fluctuations below the threshold within a same storm have been allowed for less than 12 h. These parameters can be considered relevant for the chosen physical threshold but it would not be the case for a higher or lower threshold. It is important to notice that this first step allowed defining a population of events which can be analyzed to extract relevant statistics. In particular, the extraction yields $N_T = 130$ events, the mean storm duration of the events is around 15 h. In all, 96 % of the events last less than 36 h, which is consistent with the physics. Unsurprisingly, a strong seasonality is observed: 28 storms in fall, 38 in winter, 23 in spring and 4 in summer.

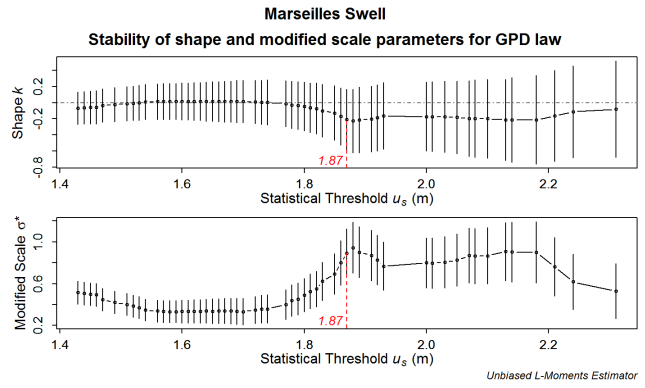

Fig. 3. Stability of shape parameter and modified scale parameter for Marseille series of swell waves.

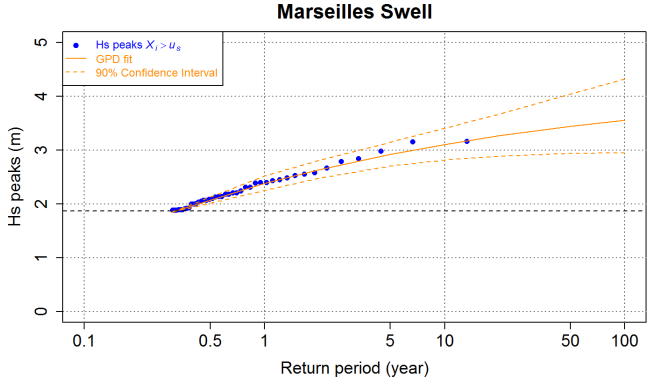

Fig. 4. GPD fit for the swell off Marseille.

The statistical optimization is then performed on the sample X_i. Note that in this case, Z and X have the same dimension; since $X_i \geq u_\mathrm{p}$ it can thus be written that $u_\mathrm{p} \leq u_\mathrm{s}$ (meaning that testing values of $u_\mathrm{s} < u_\mathrm{p}$ would be useless). The stability of the GPD shape parameter k and modified scale parameter $\sigma^* = \sigma - ku$ with respect to the statistical threshold u_s is illustrated in Fig. 3, along with the associated 95 % confidence intervals computed by the asymptotic method. 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 $u_\mathrm{s} = 1.87$ m, yielding $N = 43$. Note that, here $u_\mathrm{s} = u_\mathrm{p} + 0.47$ m and the number of observation has been reduced from $n = 38,005$ to $N = 43$.

The GPD parameters are estimated by the L-moments estimator (Hosking and Wallis, 1997). The fit is illustrated in Fig. 4. A summary of the parameters estimate for this case study is given in Table 1.

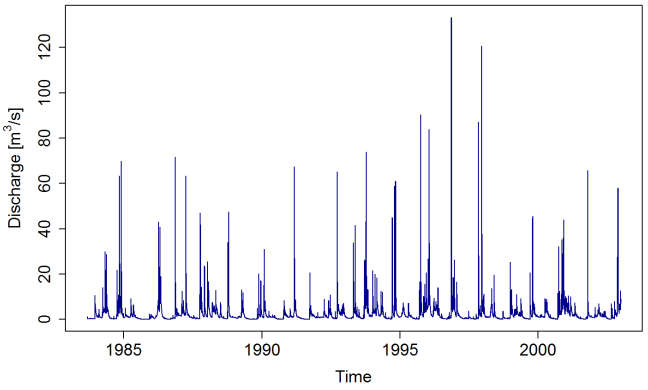

Fig. 5. Discharge series of the Loire river at Rieutord.

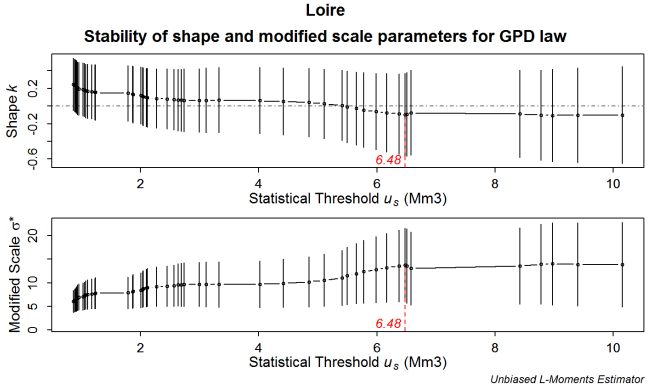

Fig. 6. Stability of shape and modified scale parameters of Loire river sample.

4.3 Discharge and flood volumes

The selected time series is the Loire river discharge daily series at Rieutord, from 01/09/1983 to 31/12/2002. The duration of the data is $K = 19.33$ yr, and the size of the time series is $n = 7062$ data. The discharge series is plotted in Fig. 5. The local modulus of the river is estimated around 2.7 $\mathrm{m^3\,s^{-1}}$. A physical threshold of $u_\mathrm{p} = 10$ $\mathrm{m^3/s}$ based on an expert judgment has been applied. The inter-event duration has been set to 10 days. That leads to an average number of flood events per year $\lambda_\mathrm{T} = 3.6$, which is physically acceptable and coherent with expert prior knowledge based on physical characteristics of the discharge phenomenon. $N_\mathrm{T} = 70$ flood events are thus retained. Their mean duration is around 103 h. 50 % of the flood peaks last only one day, while 9 % last two days and 14 % last three days. A strong seasonality can be observed, with 27 flood peaks in fall, 23 in winter, 19 in spring and only one in summer.

It is decided to describe these flood events by their volume: X is the temporal integration of the discharge over the flood duration. In particular, the physical threshold is assumed to define completely the flood, allowing the computation of its volume. In this case study, Z is a daily discharge in $\mathrm{m^3\,s^{-1}}$,

Table 2. Evolution of the size of the upper part of the X_i sample with respect to u_p.

	Number of events whose peak is above				
u_p	$5\,m^3\,s^{-1}$	$10\,m^3\,s^{-1}$	$15\,m^3\,s^{-1}$	$20\,m^3\,s^{-1}$	$25\,m^3\,s^{-1}$
$5\,m^3\,s^{-1}$	75	52	37	33	26
$10\,m^3\,s^{-1}$	–	70	47	40	32
$15\,m^3\,s^{-1}$	–	–	49	41	33
$20\,m^3\,s^{-1}$	–	–	–	42	34
$25\,m^3\,s^{-1}$	–	–	–	–	34

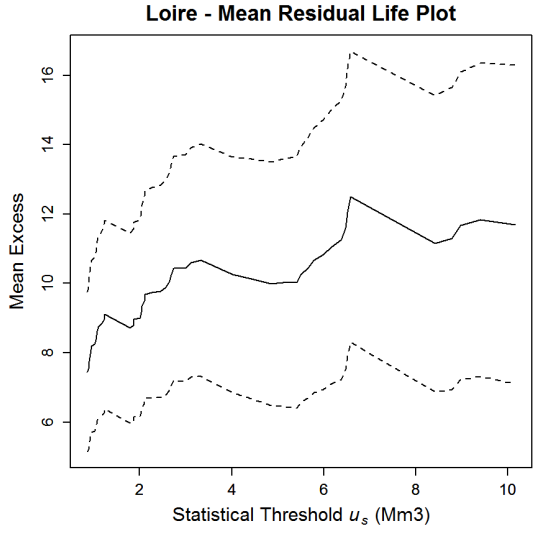

Fig. 7. Mean residual life plot for the flood volumes of Loire river sample.

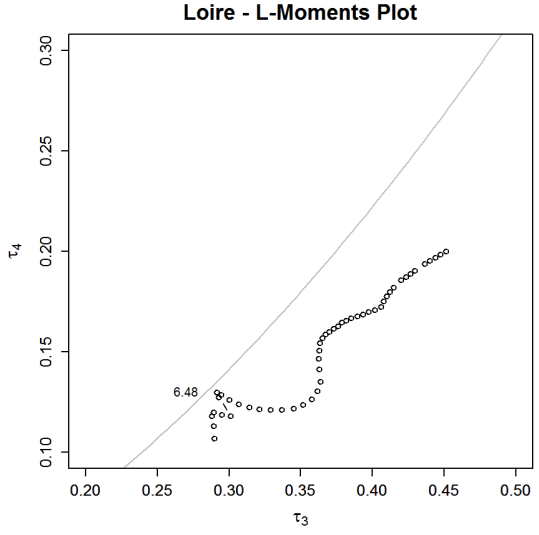

Fig. 8. L-moments plot for the flood volumes of the Loire river sample.

an event is a flood and X is the volume of the flood, in Mm^3. This is of course a very rough estimate of the flood volume: in a genuine hydrological study, it should be computed based on the hydrogram of each flood. This application is for illustrative purpose.

The stability of GPD parameters with respect to the statistical threshold u_s is given in Fig. 6. The shape (resp. modified scale) parameter slowly decreases (resp. increases) from about $1\,Mm^3$ to $6.48\,Mm^3$, then remains remarkably constant. This result is confirmed by MRL Plot, Fig. 7. L moment analysis depending on the statistical threshold u_s is shown in Fig. 8. Here it is found that the threshold value of $6.48\,Mm^3$ is the only one for which the L moments are almost lying on the theoretical GPD curve. Thus the statistical threshold is set to $u_s = 6.48\,Mm^3$, yielding a sample of $N = 25$ flood volume exceedances Y_i. In Fig. 9, the corresponding GPD calibration is shown, along with the 90 % confidence interval. A summary of the parameters estimate for this case study is given in Table 1.

In this case study, the inter-event duration is defined between the end of an event and the beginning of another. As has been stressed in Sect. 3.1.3, a consequence is that not only the lower part but also the upper part of the X_i sample, and thus ultimately of the Y_i sample to be fit, may vary when a broad range of u_p values is tested while the independence criteria remain constant. This is the case for this declustering procedure on this sample. Different values of u_p have been tested: 5, 10, 15, 20 and $25\,m^3\,s^{-1}$. For each value, a number of events are identified. Then, for each of these X_i samples (the events are here described by the discharge peaks), the number of events exceeding a higher value than the threshold is counted. The results are given in the Table 2. For instance, if the physical threshold is set to $5\,m^3\,s^{-1}$, the number of peaks exceeding $25\,m^3\,s^{-1}$ is 26, while, if the physical threshold is set to $25\,m^3\,s^{-1}$, their number increase to 34. This illustrates that one should be careful when choosing or tuning the independence criteria and this is an additional reason for separating the physical declustering step from the statistical optimization step, all the more since the latter one often requires investigating a wide range of threshold values.

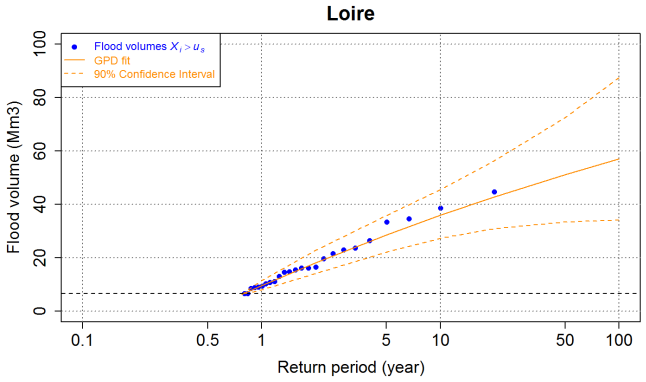

Fig. 9. GPD fit for the flood volumes of the Loire river at Rieutord.

5 Conclusions

This paper clarifies the general framework for "over the threshold" exceedance modelling, distinguishing in particular the physical declustering procedure to the statistical optimization. The two steps have a very different meaning and very different techniques have been proposed in the past for the application of these two steps. The literature is wide and an effort of review of the existing methods for both steps is done. It allows choosing the most appropriate methods for each step. In particular, we highlighted that declustering techniques are mostly based on the analysis and on the characterization of the physics of the phenomenon, while statistical optimization is a purely statistical problem. A consequence is the importance of the notion of physical event. It was often underlying in the literature, but we deem it most important to make it explicit. From our point of view, the distinction between the auto-correlated observations, at a regular time step, of the time series and the independent and self-consistent physical events should become central in the extreme value analysis of environmental variables. We also claim that fully apprehending both the difference and complementarity of these two steps allows a clearer understanding of the meaning of the different available tools and of the parameters needed for the rest of the analysis (i.e. the independence criteria parameters) and thus, ultimately, the appropriate application of the existing threshold selection methods. Through two simple practical examples, we show that each step has a clearly distinct role. It is also worthy to note that this discussion is relevant for several domains of natural hazard estimation, and even more generally to any EVT analysis of auto-correlated time series.

Edited by: T. Glade
Reviewed by: P. Jonathan and one anonymous referee

References

Adamowski, K.: Regional analysis of maximum and partial duration flood data by nonparametric and L-moments methods, J. Hydrol., 229, 216–223, 2000.

Allamano, P., Laio, F., and Claps, P.: Effects of disregarding seasonality on the distribution of hydrological extremes, Hydrol. Earth Syst. Sci., 15, 3207–3215, 2011.

Ancona-Navarrete, M. A. and Tawn, J. A.: A comparison of methods for estimating the extremal index, Extremes, 3, 5–38, 2000.

Ashkar, F. and Ouarda, T. B. M. J.: On some methods of fitting the generalized Pareto distribution, J. Hydrol., 177, 117–141, 1996.

Bardet, L., Duluc, C. M., Rebour, V. and L'Her, J.: Regional frequency analysis of extreme storm surges along the French coast, Nat. Hazards Earth Syst. Sci., 11, 1627–1639, 2011.

Barriendos, M., Coeur, D., Lang, G., Llasat, M. C., Naulet, R., Lemaitre, F., and Barrera, A.: Stationarity analysis of historical flood series in France and Spain (14th–20th centuries), Nat. Hazards Earth Syst. Sci., 3, 583–592, 2003.

Beguéria, S.: Uncertainties in partial duration series modelling of extremes related to the choice of the threshold value, J. Hydrol., 303, 215–230, 2005.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L.: Statistics of Extremes, Theory and Applications, 2004

Beirlant, J., Teugels, J. L., and Vynckiee, P.: Practical analysis of Extreme Values, Leuven University Press, 1996

Bernardara, P., Andreewsky, M., and Benoit, M.: Application of the Regional Frequency Analysis to the estimation of extreme storm surges, J. Geophys. Res., 116, C02008, doi:10.1029/2010JC006229, 2011.

Bernardara, P., De Michele, C., and Rosso, R.: A simple model of rain in time: An alternating renewal process of wet and dry states with a fractional (non-gaussian) rain intensity, Atm. Res., 2006.

Bernardara, P., Schertzer, D., Sauquet, E., Tchiguirinskaia, I., and Lang, M.: The Flood probability distribution tail: how heavy is it?, Stoch. Environ. Res. Risk Assess., 22, 107–122, 2008.

Bonazzi, A., Cusack, S., Mitas, C., and Jewson, S.: The spatial structure of European wind storms as characterized by bivariate extreme-value Copulas, Nat. Hazards Earth Syst. Sci., 12, 1769–1782, 2012.

Choulakian, V. and Stephens, M. A.: Goodness-of-fit tests for the Generalized Pareto Distribution, Technometrics, 43, 478–484, 2001.

Coles, S.: An introduction to statistical modeling of extreme values, London, Springer-Verlag, 2001

Colombo, A. F., Etkin, D., and Karney, B. W.: Climate Variability and the Frequency of Extreme Temperature Events for Nine Sites across Canada: Implications for Power Usage, J. Climate, 12, 2490–2502, 1999.

Cruise, J. F. and Arora, K.: A hydroclimatic application strategy for the Poisson partial duration model, Water Resour. Bull., 26, 431–442, 1990.

Cunnane, C.: A particular comparison of annual maxima and partial duration series methods of flood frequency prediction, J. Hydrol., 18, 257–271, 1973.

Cunnane, C.: A note on the Poisson assumption in partial duration series models, Water Resour. Res., 15, 489–494, 1979.

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

De Haan, L. and Peng, L.: Comparison of tail index estimators, Statist. Neerlandica, 52, 60–70, 1998.

Dupuis, D. J.: Exceedances over High Thresholds: A Guide to Threshold Selection, Extremes, 1, 251–261, 1998.

Egozcue, J. J., Pawlowsky-Glahn, V., and Ortego, M. I.: Wave-height hazard analysis in Eastern Coast of Spain – Bayesian approach using generalized Pareto distribution, Adv. Geosci., 2, 25–30, 2005.

Elek, P. and Markus, L.: A long range dependent model with nonlinear innovations for simulating daily river flows., Nat. Hazards Earth Syst. Sci., 4, 277–283, 2004.

Embrechts, P., Klüppelberg, C., and Mikosh, T.: Modelling extremal events, Stochastic Modelling and Applied Probability, Springer-Verlag, 33, p 648, 1997.

Floris, M., D'Alpaos, A., Squarzoni, C., Genevois, R., and Marani, M.: Recent changes in rainfall characteristics and their influence on thresholds for debris flow triggering in the Dolomitic area of Cortina d'Ampezzo, north-eastern Italian Alps, Nat. Hazards Earth Syst. Sci., 10, 571–280, 2010.

Franco, L., Piscopia, R., Corsini, S., and Inghilesi, R.: L'Atlante delle onde nei mari italiani - Italian Wave Atlas, Full Final Report by APAT-University of Roma Tre, sponsored by AIPCN Italian Section and Italia Navigando, 2004.

Fréchet, M.: Sur la lois de probabilité de l'écart maximum, Annales de la société polonaise de mathématique (in French), 6, 93–122, 1928.

Garavaglia, F., Gailhard, J., Paquet, E., Lang, M., Garçon, R., and Bernardara, P.: Introducing a rainfall compound distribution model based on weather pattern sub- sampling, Hydrol. Earth Syst. Sci., 14, 951–964, 2010.

Gnedenko, B.: Sur la distribution limite du terme maximum d'une série aléatoire (in French), Ann. Math., 44, 423–453, 1943.

Goda, Y.: On the Methodology of Selecting Design Wave Height, Proc. 21st Int. Conf. on Coast. Eng., Malaga, Spain, 1988.

Goda, Y.: Random seas and design of maritime structures, Singapore, World Scientific Publishing, 2010.

Goda, Y.: Plotting-position estimator for the L-moment method and quantile confidence interval for the GEV, GPA, and Weibull distributions applied for extreme wave analysis, Coast. Eng. J., 53, 111–149, 2011.

Goda, Y., Kudaka, M., and Kawai, H.: Incorporating of Weibull distribution in L-moments method for Regional Frequency Analysis of Peak Over Threshold wave heights, ICCE, Shangai, 2010.

Gumbel, E. J.: Statistics of Extremes, New York, Columbia University Press, Columbia, 1958.

Haigh, I. D., Nicholls, R., and Wells, N.: A comparison of the main methods for estimating probabilities of extreme still water levels, Coast. Eng., 57, 838–849, 2010.

Haight, F. A.: Handbook of the Poisson Distribution, New York, John Wiley & Sons, 1967.

Hall, P.: On some simple estimate of an exponent of regular variation, J. R. Stat. Soc., 44, 37–42, 1982.

Hamdi, Y., Duluc, C.-M., Deville, Y., Bardet, L., and Rebour, V.: Use of historical information in extreme storm surges frequency analysis, EGU General Assembly, Vienna, Geophysical Research Abstracts, 2013.

Haylock, M. R.: European extra-tropical storm damage risk from a multi-model ensemble of dynamically-downscaled global climate models, Nat. Hazards Earth Syst. Sci., 11, 2847–2857, 2011.

Hill, B. M.: A simple General Approach to Inference About the Tail of a Distribution, The Annals of Statistics, 3, 1163–1174, 1975.

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

Hurst, H. E.: Long-Term storage capacity of reservoir, Trans. Am. Soc. Civil Eng., 116, 770–799, 1951.

Jonathan, P. and Ewans, K.: The effect of directionality on extreme wave design criteria, Ocean Eng., 34, 1977–1994, 2007.

Koscielny-Bunde, E., Kantelhardt, J. W., Braun, P., Bunde, A. and Havlin, S.: Long-term persistence and multifractality of river runoff records: Detrended fluctuation studies, J. Hydrol. Hydrofractals '03, 322, 120–137, 2006.

Lang, M., Ouarda, T. B. M. J., and Bobée, B.: Towards operational guidelines for over-threshold modeling, J. Hydrol., 225, 103–117, 1999.

Leadbetter, M. R., Lindgren, G., and Rootzen, H.: Extremes and Related Properties of Random Sequences and Processes, New York, Springer-Verlag, 1983.

Li, F., Bicknell, C., Lowry, R., and Li, Y.: A comparison of extreme wave sample analysis methods with 1994-2010 offshore Perth dataset., Coast. Eng., 69, 1–11, 2012.

Mackay, E. B. L., Challenor, P. G., and Bahaj, A. S.: A comparison of estimators for the generalised Pareto distribution, Ocean Eng., 38, 1338–1346, 2001.

MacKay, E. B. L., Challenor, P. G., and Bahaj, A. S.: On the use of discrete seasonal and directional models for the estimation of extreme wave conditions, Ocean Eng., 37, 425–442, 2010.

Madsen, H. and Rosbjerg, D.: The partial duration series method in regional index-flood modeling, Water Resour. Res., 33, 737–746, 1997.

Marani, M.: On the correlation structure of continuous and discrete point rainfall, Water Resour. Res., 39, SWC21–SWC28, 2003.

Mathiesen, M., Goda, Y., Hawkes, P. J., Mansard, E., Martin, M. J., Peltier, E., Thompson, E. F., and Van Vledder, G.: Recommended practice for extreme wave analysis, J. Hydraul. Res., 32, 803–814, 1994.

Mazas, F. and Hamm, L.: A multi-distribution approach to POT methods for determining extreme wave heights, Coast. Eng., 58, 385–394, 2011.

Neves, C. and Fraga Alves, M. I.: Reiss and Thomas' automatic selection of the number of extremes, Computational Statistics & Data Analysis, 47, 689–704, 2004.

Ntegeka, V. and Willems, P.: Trends and multidecadal oscillation in rainfall extremes, based on a more than 100 years time series of 10 minutes rainfall intensities at Uccle, Belgium, Water Resour. Res., 44, W07402, doi:10.1029/2007WR006471, 2008.

Onoz, B. and Bayazit, M.: Effect of the occurrence process of the peaks over threshold on the flood estimates, J. Hydrol., 244, 86–96, 2001.

Ouarda, T. B. M. J., Rasmussen, P. F., Bobée, B., and Bernier, J.: Use of historical information in hydrology frequency analysis, J. Water Sci. 11, 41–49, 1998.

Palutikof, J. P., Brabson, B. B., Lister, D. H., and Adcock, S. T.: A review of methods to calculate extreme wind speeds, Meteorol. Appl., 6, 119–132, 1999.

Pandey, M. D., Van Gelder, P. H. A. J. M., and Vrijling, J. K.: The estimation of extreme quantiles of wind velocity using L-moments in the peaks-over-threshold approach, Struc. Saf., 23, 179–192, 2001.

Parent, E. and Bernier, J.: Encoding prior experts judgments to improve risk analysis of extreme hydrological events via POT modeling, J. Hydrol., 283, 1–18, 2003.

Payrastre, O., Gaume, E., and Andrieu, H.: Use of historical data to assess the occurrence of floods in small watershed in the French Mediterranean area., Adv. Geosci., 2, 313–320, 2005.

Payrastre, O., Gaume, E., and Andrieu, H.: Usefulness of historical information for flood frequency analyses: Developments based on a case study, Water Resour. Res., 47, W08511, doi:10.1029/2010WR009812, 2011.

Pickands, J.: Statistical Inference Using Extreme Order Statistics, The Annals of Statistics, 3, 119–131, 1975.

Ribatet, M., Sauquet, E., Grésillon, J.-M., and Ouarda, T. B. M. J.: A regional Bayesian POT model for flood frequency analysis, Stoch. Environ. Res. Risk Assess., 21, 327–339, 2007.

Rosbjerg, D.: Estimation in partial duration series with independent and dependent peak values, J. Hydrol., 76, 183–195, 1985.

Rosbjerg, D. and Madsen, H.: Advanced approaches in PDS/POT modelling of extreme hydrological events, Hydrology – Science and Practise for the 21st Century, B. H. Society, B. Webb., 217–220, 2004

Rosbjerg, D., Madsen, H., and Rasmussen, P. F.: Prediction in partial duration series with generalized pareto-distributed exceedances, Water Resour. Res., 28, 3001–3010, 1992.

Ruggiero, P., Komar, P. D., and Allan, J. C.: Increasing wave heights and extreme value projections: The wave climate of the US Pacific Northwest, Coast. Eng., 57, 539–552, 2010.

Schertzer, D. and Lovejoy, S.: Universal multifractals do exist!: Comments, J. of Appl. Meteorol., 36, 1296–1303, 1997.

Silva, A. T., Portela, M. M., and Naghettini, M.: Nonstationarities in the occurrence rates of flood events in Portuguese watersheds, Hydrol. Earth Syst. Sci., 8, 8609–8639, 2011.

Simiu, E. and Heckert, A.: Extreme Wind Distribution Tails: A 'Peaks Over Threshold' Approach, National Science Foundation, 1995.

Smith, O. P.: Duration of extreme wave conditions, J. Waterw. Port C.-ASCE, 114, 1–17, 1988.

Smith, R. L.: Threshold methods for sampling extremes, Statistical extremes and applications, T. d. O. J. Dordrecht, teh Netherlands, 621–638, 1984

Smith, R. L. and Weissman, I.: Estimating the Extremal Index, J. R. Stat. Soc., 56, 515–528, 1994.

Takvor, S. H. and Panagiota, A. M.: The Effect of Declustering in the r-Largest Maxima Model for the Estimation of Hs Design Values, The Open Ocean Eng. J., 4, 34–43, 2001.

Tawn, J. A. and Vassie, J. M.: Extreme sea level: the joint probability method revisited and revised, Proc. Instn. Civ. Engrs., 429–442, 1989.

Taylor, P. H., Barker, V. E., Bishop, D., and Eatock Taylor, R.: 100-year waves, teleconnections and wave climate variability, 11th International Workshop on Wave Hindcasting and Forecasting, Halifax, Canada, 2009.

Thompson, P., Cai, Y., Reeve, D., and Stander, J.: Automated threshold selection methods for extreme wave analysis, Coast. Eng., 56, 1013–1021, 2009.

USWRC: Guidelines for determining flood flow frequency H. C. Bull.17, Washington DC, United States Water Resources Council, p. 73, 1976.

Wahl, T., Mudersbach, C., and Jensen, J.: Assessing the hydrodynamic boundary conditions for risk analyses in coastal areas: a stochastic storm surge model, Nat. Hazards Earth Syst. Sci., 11, 2925–2939, 2011.

Walton, T. L.: Distribution for storm surge extremes, Ocean Eng., 27, 1979–1293, 2000.

Willems, P.: Hydrological applications of extreme value analysis, Hydrology in a changing environment, H. a. C. K. Wheater. Chichester, John Wiley & Sons, 15–25, 1998

Willems, P.: Compound intensity/duration/frequency-relationships of extreme precipitation for two seasons and two storm types, J. Hydrol., 233, 189–205, 2000.

Zawadzky, I.: Fractal structure and exponential decorrelation in rain, J. Geophys. Res., 92, 1816–1818, 1987.

1.3. OCEAN ENGINEERING 2014: QUESTIONING MLE FOR THE ESTIMATION OF ENVIRONMENTAL EXTREME DISTRIBUTIONS

Main publications citing this paper (source Google Scholar)

Solari S., Egüen M., Polo M. J., Losada M. A., 2017. Peaks Over Threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. *Water Resources Research*, **53(4)**, 2833-2849.

Barker A., Murphy J., Pakrashi V., 2015. Reliability of extreme wave prediction methods. 12th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP12 Vancouver, Canada, July 12-15, 2015.

Ocean Engineering 92 (2014) 44–54

Contents lists available at ScienceDirect

Ocean Engineering

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

Questioning MLE for the estimation of environmental extreme distributions

Franck Mazas [a,*], Philippe Garat [b], Luc Hamm [a]

[a] ARTELIA Group, 6 rue de Lorraine, 38130 Echirolles, France
[b] Université Pierre Mendes-France – Laboratoire Jean Kuntzmann – UMR 5224, BSHM, 1251 avenue centrale BP 47, 38040 Grenoble Cedex 09, France

ARTICLE INFO

Article history:
Received 10 April 2013
Accepted 27 September 2014

Keywords:
Extreme wave heights
POT
MLE
L-moments
Threshold
Location parameter

ABSTRACT

In determining extreme environmental variables, such as wave heights, with the Peaks-Over-Threshold (POT) method, it has become common practice in the metocean community to use the GPD–Poisson model fitted by the Maximum Likelihood Estimator (MLE). However, Mazas and Hamm (2011) pointed out some difficulties in getting stable estimations of extreme quantiles with this method. Further investigation reported in the present paper enable to understand that this problem is linked to the behavior of the likelihood function and to solve it by introducing a location parameter and replacing maximum likelihood estimated two-parameter distributions by L-moments estimated three-parameter distributions. Applications on real and simulated data highlight the distinction between the location parameter of a statistical distribution and the statistical threshold chosen in the POT context. With three-parameter distributions, MLE is no more suitable and it is found that the L-moments estimator can be a valid alternative. With these two improvements, stable quantiles are obtained not only with the GPD but also with other distributions such as Weibull and Gamma (Pearson-III).

1. Introduction

Determining extreme univariate environmental variables, such as wave heights, wind speeds or river discharges, is a complex methodology requiring a good knowledge of the physics involved and the use of sound and reliable statistical methods.

In the coastal engineering field where we are operating, an international joint effort was carried out twenty years ago to provide practical recommendations for performing such an analysis (Mathiesen et al., 1994, see also Goda et al., 2010). In this state-of-the-art paper, the following five steps were advised:

- selection of the sample: the Peaks-Over-Threshold (POT) method was recommended. It was highlighted that the storm peaks above the chosen threshold represent a censored sample of all storm peaks;
- choice of model distributions: a three-parameter Weibull-min distribution was recommended in general while advising to choose a Gumbel distribution when the shape parameter of the Weibull is found to be near 1;
- choice of fitting methods: the maximum likelihood estimator (MLE) was favored without excluding the least squares method with suitable plotting position formula;
- goodness of fit tests: visual inspection of quantile-quantile plots together with various numerical tests were recommended;
- return value and encounter probability: a Poisson distribution was recommended to estimate the encounter probability; and
- computation of confidence limits: a Monte-Carlo approach was suggested.

Since that work, new statistical approaches for modeling of extreme values (see Coles, 2001 for a broad overview) disseminated widely in our community including in particular the Generalized Pareto distribution (GPD). As a result, the methodology combining the POT declustering technique with a GPD–Poisson model fit by the maximum likelihood estimator (MLE) is now the most widely used approach in present day research in the field of oceanographic and coastal engineering (see for instance, among many other examples, Tancredi et al., 2006; Thompson et al., 2009; Mackay et al., 2010; Solari and Losada, 2012).

Much appreciated for its sound theoretical justifications, this model is now used as the basis for numerous developments concerning in particular the use of covariates to take into account phenomena such as temporal non-stationarity at multiple time scales (see for example Méndez et al. (2006, 2008), Nogaj et al. (2007)) or the

* Corresponding author. Tel.: +33 4 56 38 46 91; fax: +33 4 76 33 43 33.
E-mail addresses: franck.mazas@arteliagroup.com (F. Mazas), philippe.garat@iut2.upmf-grenoble.fr (P. Garat), luc.hamm@arteliagroup.com (L. Hamm).

http://dx.doi.org/10.1016/j.oceaneng.2014.09.034
0029-8018/

influence of the direction of sea states (see for example Jonathan and Ewans, 2007).

However, intensive practice of this model over a wide range of meteo-oceanographic conditions encountered in our international engineering practice as well as the diversity of available data led us to unexpected difficulties not reported in the literature that are shortly summarized hereafter.

First, the unique choice of the GPD as a candidate for the choice of a model distribution appeared to be too restrictive. It can provide underestimations of design conditions as observed in several cases. Our analysis of this difficulty led us to suggest an extension of the model to a multi-distribution approach including objective criteria of choice (Mazas and Hamm, 2011, hereinafter referred to as MH2011). Such an extension was also favored by other authors, see for instance Goda and Kudaka (2009) and Van Vledder et al. (1994).

Secondly, a clarification was needed regarding the identification and extraction of storms from the sequential time series for setting up the i.i.d. sample required for meeting the hypotheses of the Extreme Value Theory (EVT). A first step was the introduction of a double threshold approach as explained in MH2011. The discussion by You (2012) of this point and the preparation of our reply (Mazas and Hamm, 2012) was the starting point of a joint research work where we elaborated in details the reasons for a two-step framework for threshold selection in Over-Threshold Modeling (OTM) (Bernardara et al., 2014, hereinafter referred to as BM2014).

Thirdly, accurate sensitivity tests with respect to the threshold revealed an instability of the quantiles linked to the behavior of the MLE itself for which the help of a statistician was needed. The results of this joint research questioning MLE as the best estimator for fitting extreme distributions in the OTM framework is the main subject of this paper. In order to understand this difficulty and, ultimately, to solve it, it appeared that we have to question three choices that are actually intimately linked:

- setting the threshold;
- accounting for a location parameter;
- choosing a reliable estimator.

We shall consider one specific variable, namely the wave heights, but the following is also applicable to all environmental variables: wind speed, surge height, river discharge, temperature.

To smoothly report the results of this work, the paper is organized as follows. First, the multi-distribution model proposed in Mazas and Hamm (2011) is briefly recalled in Section 2. The limits of the MLE are then illustrated with reference to the case studies of Haltenbanken (Northern Atlantic) and simulated GPD datasets in Section 3. Alternative estimators based on probability weighted moments and L-moments are briefly introduced in Section 4 (and detailed in Appendix). Lastly, a comparison between the fit of 2-parameter distributions estimated by MLE and 3-parameter distributions estimated by L-moments is presented in Section 5. Results are discussed in Section 6. This leads to the proposal of an improved, more reliable methodology.

2. The multi-distribution stationary model

The present section is a brief abstract of the model presented in MH2011, complemented by the considerations and clarifications about the declustering methodology described in BM2014. For more details, justification and illustrations please refer to these papers.

2.1. Data processing and physical declustering

Oceanographic or coastal engineers work with time series of sea states. These may be buoy measurements, numerically hindcast time series, etc. The first task is therefore to identify the various populations composing the sea states on site and then separate them in order to work on homogeneous – that is to say identically distributed – data. For instance, ocean swell, local wind seas or cyclonic events, among other possible populations, should be analyzed separately.

Though identically distributed, these time series of sequential values (e.g. every 1 h) still show strong temporal autocorrelation. In particular extreme values are likely to occur in clusters, corresponding to storms, or more generally physical events. This is true when the time step of the series is shorter than the typical duration of the event. Consequently, a step of physical declustering is required to grant the independence of extreme data. A popular approach in various engineering fields is the Peaks-Over-Threshold (POT) approach. Physical events (storms) are defined by exceeding a value termed physical threshold and denoted u_p, following the nomenclature of BM2014. Additionally, temporal parameters should be used to make sure that the selected events are independent, for instance by requiring a minimal duration between two successive events (typically 24–48 h for sea states). This threshold must be physically meaningful and linked to the underlying physics of the phenomenon. BM2014 present an extensive review of different possibilities for its determination; in this paper we will set u_p so as to get between 5 and 10 events per year in average, which is consistent with the physics of significant storms in western Europe. Then, the storm is characterized by its maximum value, or peak, denoted X. The result of this step of physical declustering consists of a sample of N_T independent and identically distributed (i.i.d.) storm peak data X_i larger than u_p.

2.2. Optimization of the statistical threshold

Having reduced the working sample from a few tens of thousands of data to a few tens or hundreds, it is then necessary to determine the optimal threshold of the i.i.d. sample, i.e. the threshold that has a statistical meaning and above which it can be assumed that the peaks have an “extreme behavior”. Still following BM2014, it may be termed the statistical threshold, denoted u_s. In particular, and in contrast with the physical threshold, determining u_s requires a statistical optimization to be performed on an i.i.d. sample and not on an auto-correlated time series.

As explained in Coles (2001) and many other papers, the GPD–Poisson model is the natural candidate for the extreme analysis of an i.i.d. sample issued from a POT declustering. Indeed, the Generalized Pareto Distribution is the law that asymptotically (i.e. for a fairly high threshold) approximates the exceedance law of a threshold. In addition, it is considered that the number of events during a year follows a Poisson process. Consequently, the determination of u_s is usually based on theoretical properties of the GPD: see BM2014 for a review of different methods.

In this paper, the following property will be used: if a sample is derived from a GPD, the shape parameter k remains constant when the threshold rises. Hence a GPD is fit to the peak exceedances for a broad range of u_s values and “domains of stability” where k is roughly constant are looked for. Since the GPD law is asymptotic, the bias should be lowest within the domain of stability corresponding to the highest thresholds. However, to limit the increase in variance due to the reduction in sample size and thus loss of information, the aim is to select the lowest threshold of this highest domain of stability; or in other words, to find the best compromise between bias and variance.

Finally, N storm peaks exceed the statistical threshold u_s, from which can be derived a sample of N threshold exceedances $(Y_1, \ldots, Y_N)$, where $Y = X - u_s$, given $X > u_s$. MH2011 recommended studying only values of u_s equal to the storm peaks (explanations will be provided in Section 3.2).

2.3. Choice of statistical distributions

The GPD is particularly suitable for OTM analyses due to its asymptotic properties. However, given the very large number of cases that may occur, other distributions may be more suitable. For instance, Goda and Kudaka (2009) pointed out the risk of underestimation of return levels due to finite upper bounds and favor a multi-distribution approach. Van Vledder et al. (1994) also encouraged the analysts to try several candidate distributions, a recommendation that was expressed by the 1993 IAHR Working Group (Mathiesen et al., 1994). Furthermore, it should be kept in mind that the GPD is an approximation of the law of threshold exceedances (see Eq. (1) in MH2011), only valid within the asymptotic domain (Pickands, 1975). Consequently, it cannot be guaranteed a priori that this domain will be reached and thus that the GPD will be appropriate. For many practical case studies, the difficulty in determining the statistical threshold u_s as explained above suggests such a situation. This is another argument for a multi-distribution approach.

Through our practical work we have adopted the Weibull distribution (traditionally used by many authors, see for example Mathiesen et al., 1994 or Goda, 2000) and the Gamma distribution (Dorsch et al., 2008). However, in a multi-distribution approach, analysts should feel free to test other distributions that could give good results for their particular datasets. The cumulative distribution functions of these three laws are as follows:

$$2-\text{parameter GPD}: \begin{cases} F_{Y;k,\sigma}(y) = 1-\left(1+k\frac{y}{\sigma}\right)^{-1/k} \text{if } k \neq 0 \\ F(y) = 1-\exp\left(-\frac{y}{\sigma}\right) \text{if } k = 0 \end{cases} \quad (1)$$

$$2-\text{parameter Weibull}: F_{Y;k,\sigma}(y) = 1-\exp\left[-\left(\frac{y}{\sigma}\right)^k\right] \quad (2)$$

$$2-\text{parameter Gamma}: F_{Y;k,\sigma}(y) = \frac{\gamma\left(k,\frac{y}{\sigma}\right)}{\Gamma(k)} \quad (3)$$

In the three cases, k is a shape parameter and σ a scale parameter. If $k < 0$, the GPD law has a finite terminal point $\omega = -\sigma/k$, that is to say that the probability of $y > \omega$ is nil. It should be noted that if y is replaced by $x - u_s$, it can be seen immediately that u_s plays a similar role to that of a location parameter. However, it is set before the estimation; this means they are indeed two-parameter laws.

2.4. Fitting the data to the distributions

There are various estimation methods, the best known being the least squares, moments and maximum likelihood (MLE) estimators. The last of these is very widespread. It is in particular asymptotically unbiased, robust and consistent. For these reasons, in spite of the reservations already expressed in MH2011, it is this estimator that was chosen in this previous study. The choice of the estimator was also discussed in You (2012) and our reply (Mazas and Hamm, 2012).

2.5. Calculation of quantiles and confidence intervals

Once the fits have been performed, it is easy to determine the wave heights for a given return period T by means of the quantile function applied to the probability of non-exceedance $1-1/(\lambda T)$, where λ is the average number of storm peaks larger than u_s per year.

The confidence intervals are calculated by using the parametric bootstrap technique, based on numerous resampling operations (see for example Efron and Tibshirani, 1993).

Because these three distributions have the same number of parameters (2), the model with the largest likelihood can be selected as the best-fitting distribution.

3. Limitations of the Maximum Likelihood Estimator

3.1. Order statistics and formalism

Here, we examine the influence of the careful choice of the statistical threshold u_s on the fit of a distribution and more specifically when u_s varies between two consecutive values of the i.i.d. sample of the storm peaks. The following formalism is introduced for this purpose.

The sample $(X_1, \ldots, X_{N_T})$ of the N_T storm peaks above the physical threshold u_p may be sorted in ascending order so as to work on *order statistics* and denoted $(X_{1:N_T}, \ldots, X_{N_T:N_T})$, so that $u_p < X_{1:N_T} \leq \ldots \leq X_{i:N_T} \leq \ldots \leq X_{N_T:N_T}$. The sample of the N extreme peaks above the statistical threshold u_s is then $(X_{(N_T-N+1):N_T}, \ldots, X_{N_T:N_T})$ with $X_{(N_T-N):N_T} \leq u_s < X_{(N_T-N+1):N_T} \leq \ldots \leq X_{N_T:N_T}$. It is important to note that this sample is strictly identical for any threshold u_s within the interval $\mathcal{I}_u = [X_{(N_T-N):N_T}; X_{(N_T-N+1):N_T}[$. (It should be recalled that when u_s is strictly equal to a data value, this data is excluded from the sample of the threshold excesses in order to have only strictly positive exceedance values.)

Similarly, we can sort the sample $(Y_1, \ldots, Y_N)$ of the N peak exceedances above the threshold u_s in ascending order: $0 < Y_{1:N} \leq \ldots \leq Y_{N:N}$, with $Y_{i:N} = X_{i:N} - u_s$. But in contrast with the extreme peaks above, this sample $(Y_{1:N}, \ldots, Y_{N:N})$ will vary and be translated when u_s varies in $\mathcal{I}_u$. In particular, the first value $Y_{1:N}$ can take any value within $]0; X_{(N_T-N+1):N_T} - X_{(N_T-N):N_T}]$ (i.e. $\mathcal{I}_u - X_{(N_T-N):N_T}$).

Keeping in mind that the statistical fit is performed on the $Y_{i:N}$ data, the consequence is that selecting the same storm peaks (i.e. the same physical events) can yield infinity of samples to be fitted. Because the physics of the phenomenon is invariant it should therefore be expected that this infinity of possible fits when u_s runs through $\mathcal{I}_u$ results in the same results concerning the quantiles (wave heights for any return period); the following section will show it that this is not the case with the MLE.

3.2. Illustrations on site

First, these limits will be illustrated with one of the case studies performed by MH2011: the classical Haltenbanken dataset provided by the IAHR Maritime Hydraulics – Working Group on Extreme Wave Analysis (Van Vledder et al., 1994), which consists of 128 storm peaks higher than $u_p = 7$ m, measured by a buoy over a 9-year period.

The statistical threshold was determined by a sensitivity analysis on the GPD shape parameter and set at $u_s = 8.57$ m (meeting four storm peaks at the same value). The following data peak equals 8.63 m, so with this choice of u_s, $Y_{1:N} = 0.06$ m and $\mathcal{I}_u = [8.57; 8.63[$. At this stage, it can be assigned to each ordered data $X_{i:N_T}$ above u_p an empirical probability of non-exceedance (or plotting position formula), for instance the Weibull formula $p_i = i/(N_T+1)$, which can be transformed into an empirical return period: $T_i = 1/[\lambda_T(1-p_i)]$. This allows a visualization of the sample, as illustrated in Fig. 1.

Let us note $\overline{u_s}$ the optimal value of the statistical threshold as chosen with the classical methods (8.57 m). However, this value was determined here with a fine step of 0.01 m, corresponding to the number of digits of the available data. One could imagine that

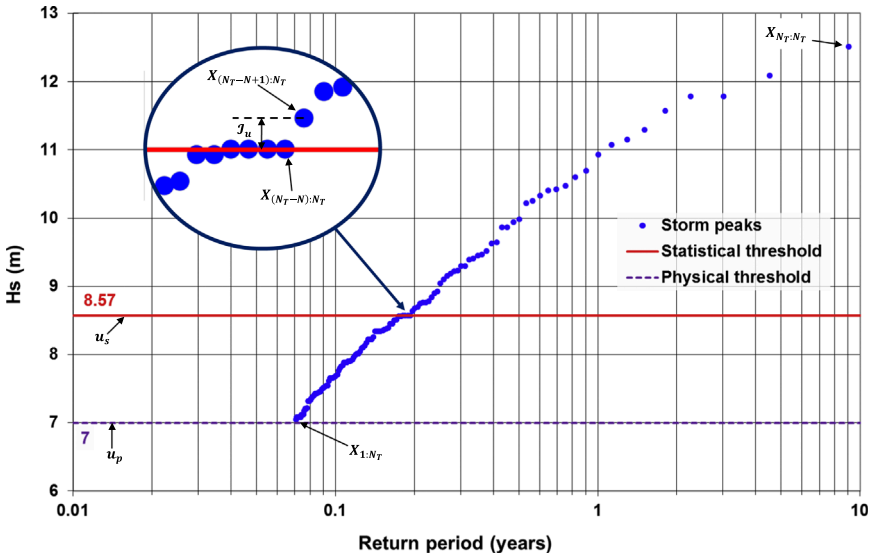

Fig. 1. Haltenbanken: storm peaks as a function of their empirical return period

with a coarser step of 0.05 or 0.1 m, for instance, a value of 8.60 m would have been chosen. Because the peak following $\overline{u_s}$ is higher than 8.60 m, strictly the same N peaks X_i would have been selected with this choice, but the sample of peak exceedances Y_i would have been translated of 0.03 m. So a question arises: what is the influence of this shift on the fit and eventually on the results?

If u_s runs through $\mathcal{I}_u$ by a very fine step, say 0.0001 m (i.e. between $\overline{u_s}$ and $X_{(N_T-N+1):N_T}-0.0001$), the fit can be performed for the corresponding sample $(Y_{1:N}, \ldots, Y_{N:N})$ at each step and the 1-in-100-year H_s and log-likelihood can be calculated for each distribution and each step. The results are presented in Figs. 2 and 3.

Fig. 2a clearly shows that the GPD is not quite sensitive to the position of u_s in $\mathcal{I}_u$ as far as the quantile is concerned, with a slight (but not nil) increase in Hs_{100}, whereas the Weibull and Gamma distributions may increase considerably on approaching the upper bound. In the log-likelihood profiles presented in Fig. 2b, it can be seen that the maximum is reached on the open upper bound with a strictly positive derivative, except in the case of the Weibull distribution for which a global maximum exists. It may be added that when examining the second dataset of MH2011 (see also Section 5.3), the quantiles and log-likelihood profiles of all three distributions, Weibull included, show this behavior of monotonous increase.

Focusing on the GPD, this behavior can be observed between any values of the peak sample. The above plot of Fig. 3 shows the evolution of the 1-in-100-year *Hs* (curve below, primary axis) and of the shape parameter (curve above, secondary axis). Data values are represented by points. It can clearly be seen that both the shape parameter and the quantile can be quite unstable when the threshold varies between two peaks, with very distinct sub-curves. These typical sub-curves may be called "*virgae*", referring to a *virga*: a cloud whose shape is similar to a comma. These *virgae* also appear for the log-likelihood: for the sake of clarity, the below plot of Fig. 3 is limited to a zoom around 8.57 m.

In MH2011 these instabilities were considered to be a good reason for restricting the possible values of u_s to those taken by the sample data, but no further investigations were made. The results of those are presented below.

3.3. Introduction of a location parameter

Letting u_s vary within $\mathcal{I}_u$ is equivalent to the following procedure: first, setting the value of u_s equal to $X_{(N_T-N):N_T}$, and second, introducing a location parameter μ that can take any value between 0 (included) and $Y_{1:N}=X_{(N_T-N+1):N_T}-X_{(N_T-N):N_T}$

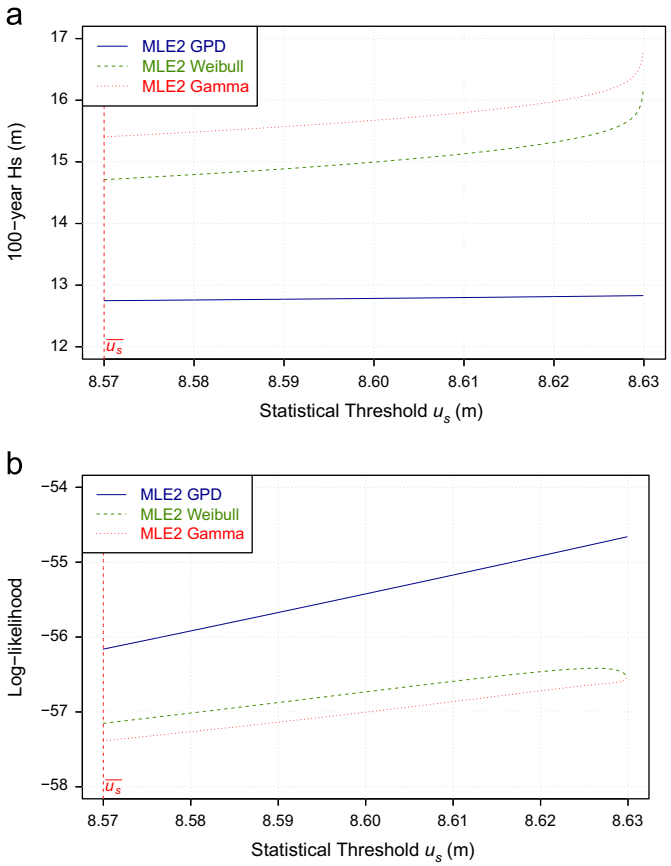

Fig. 2. Haltenbanken: change in Hs_{100} (a) and in the log-likelihood (b) when u_s runs through $\mathcal{I}_u$

(excluded). Actually, μ can also be negative and varies within $\mathcal{I}_\mu=]-\infty; X_{(N_T-N+1):N_T}-X_{(N_T-N):N_T}[$. The cumulative distribution functions of the 3-parameter distributions are then written as follows:

$$3-\text{parameter GPD}: \begin{cases} F_{Y;k,\sigma,\mu}(y)=1-\left(1+k\frac{y-\mu}{\sigma}\right)^{-1/k} & \text{if } k\neq 0 \\ F_{Y;\sigma,\mu}(y)=1-\exp\left(-\frac{y-\mu}{\sigma}\right) & \text{if } k=0 \end{cases} \quad (4)$$

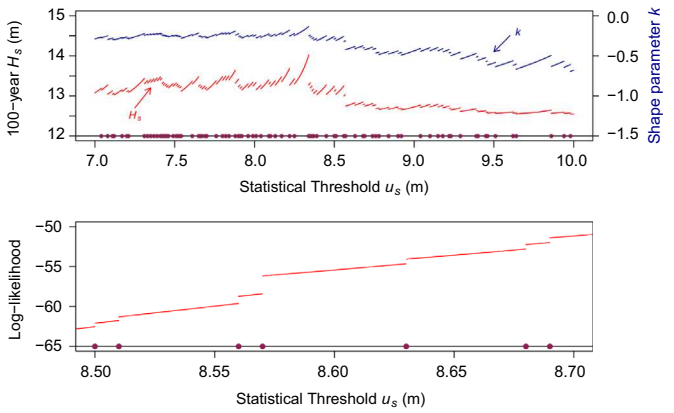

Fig. 3. Haltenbanken – change in the ML-estimated GPD shape parameter (top plot, above curve), in Hs_{100} (top plot, below curve) and in the log-likelihood (down plot, zoom) with respect to u_s. Points represent the data values $X_{i:N_T}$.

$$3-\text{parameter Weibull}\ : F_{Y;k,\sigma,\mu}(y)=1-\exp\left[-\left(\frac{y-\mu}{\sigma}\right)^{k}\right] \tag{5}$$

$$3-\text{parameter Gamma}\ : F_{Y;k,\sigma,\mu}(y)=\frac{\gamma\left(k,\frac{y-\mu}{\sigma}\right)}{\Gamma(k)} \tag{6}$$

with $\mu < Y_{1:N}$. It should be noted that the 3-parameter Gamma distribution is more widely known as the Pearson-III distribution. In the following, we will focus on the 3-parameter GPD.

3.4. Behavior of log-likelihood maxima on simulated GPD datasets

Let us randomly generate a sample of size $N=100$ from a theoretical 3-parameter GPD with the following set of parameters: $k=-0.25$, $\sigma=1.5$, $\mu=0.05$, and let us estimate the parameters from this sample using the Maximum Likelihood Estimator. Let us recall that the domains of validity of the parameters are as follows: $\mu < Y_{1:N}$, $\sigma>0$ and $k>-\sigma/(Y_{N:N}-\mu)$. Note that we examine the case $k<0$, corresponding to a finite upper endpoint, because it is the most frequent case in environmental applications (for instance, the wave height or wind speed is expected to be physically limited).

The log-likelihood is computed for a large number of discretized parameters set: $(k_i;\sigma_j;\mu_l)$. In particular, μ takes a finite number of discrete values between a lower bound, say -0.5, and $Y_{1:N}-\epsilon$, where ϵ is the discretization step, say 0.002. This operation results in a large number of discrete log-likelihood values within the 3-dimensional parameter space. This exercise aims at analyzing and understanding the behavior of the log-likelihood within this space, in particular regarding the value of μ. For this purpose, we will first examine its local maximum when μ is constant (i.e. for each hyperplane $(k;\sigma)$ of the parameter space $(k;\sigma;\mu)$) then its global maximum. Conclusions will be drawn in Section 3.5.

Besides, it is noteworthy that standard algorithms for 3-parameter laws make the location parameter vary within its discretized validity interval (Panchang and Gupta, 1989) and then, for each scanned value of μ, solve the classical 2-parameter set of equations.

Fig. 4 displays the contours of the log-likelihood values for two hyperplanes of the parameter space: $\mu=0$ and $\mu=Y_{1:N}-\epsilon$. The line $\sigma=-k(Y_{N:N}-\mu)$ is drawn as it is a limit of the domain of validity. The local maximum within this hyperplane is marked by a cross. It can be seen that for each hyperplane, the local (2-dimensional) maximum is reached within the domain of validity and not at one of its bound. It implies that the derivatives at this local maximum are nil.

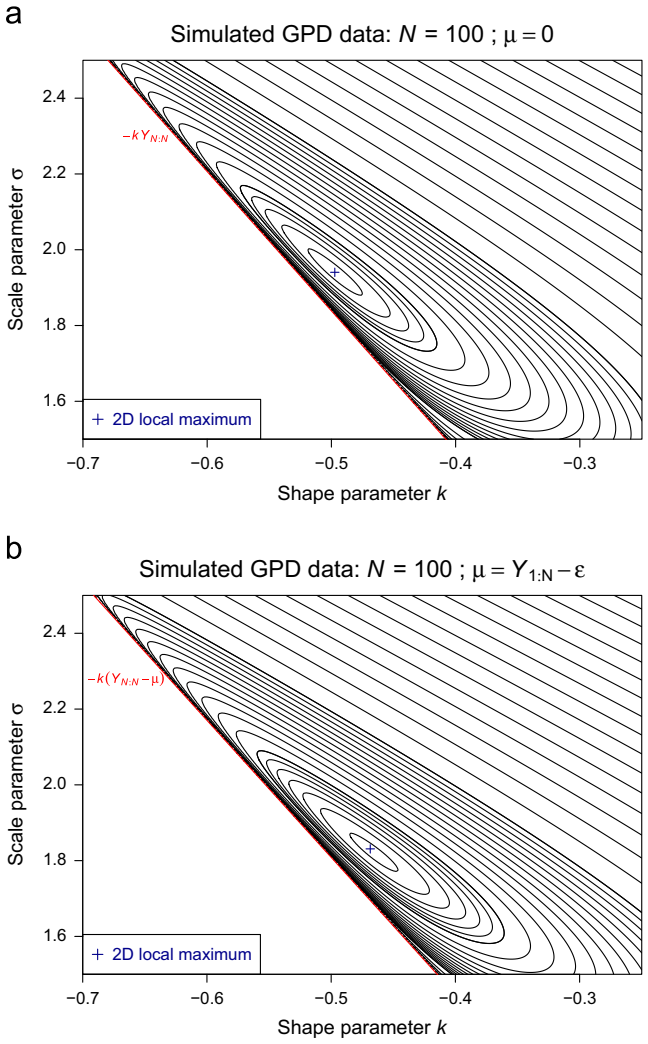

Fig. 4. Simulated GPD data: log-likelihood for the hyperplanes (a) $\mu=0$ and (b) $\mu=Y_{1:N}-\epsilon$.

Fig. 5 shows the variation of this local maximum log-likelihood for each hyperplane $\mu=$ constant of the parameter space with respect to the value of μ. The global maximum is reached for $\mu=Y_{1:N}-\epsilon$, i.e. when μ tends to its open upper bound. It can also be observed that the derivative $\partial l/\partial\mu$ is strictly positive, not nil. This is fully consistent with what can be observed for real environmental datasets (see Fig. 2b).

Lastly, Fig. 6 displays the evolution of the pair $(k;\sigma)$ corresponding to the local maximum of the log-likelihood for each hyperplane. It can be observed that when $\mu\to Y_{1:N}$, the pair $(k;\sigma)$ does not converge towards any stable value. It should be noted that this exercise has been repeated with a positive value of k (0.05 and 0.15) as well as with different sample sizes of $N=25$ and $N=50$: the same results can be observed.

3.5. Conclusions on the validity of the Maximum Likelihood Estimator

In the previous sections, it has been shown both for environmental and simulated datasets of different sample sizes that the global maximum likelihood is reached at an open upper bound of the parameter space, with a non-nil local derivative and without any convergence of the values of the estimated set of parameters $(\hat{k};\hat{\sigma};\hat{\mu})$. It is actually not surprising for the GPD because the maximum of its probability density function is for $y=0$. Yet

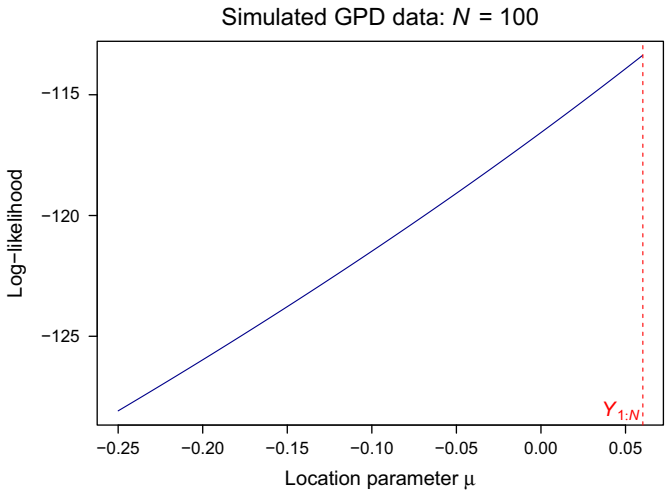

Fig. 5. Simulated GPD data: change in the local maximum log-likelihood for each hyperplane $\mu = cste$.

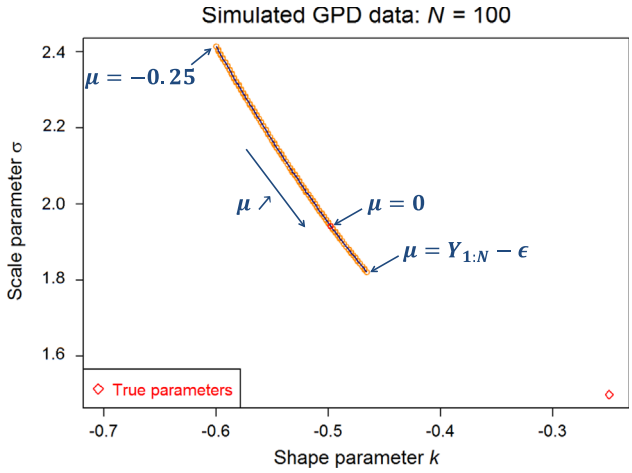

Fig. 6. Simulated GPD data: evolution of the pair $(k;\sigma)$ corresponding to the local maximum log-likelihood for each hyperplane $\mu = cste$.

proving the asymptotic properties (consistency, efficiency) of the Maximum Likelihood Estimator requires that the true (and unknown) vector of parameters be an interior point of an open set (Lehmann, 1983, chapter 6), and that the likelihood function converges towards a global (and finite) maximum on this point with locally nil derivatives.

In other words, the conclusion can be drawn that using distributions with no location parameter but fixing the threshold arbitrarily between two data values is in fact implicitly equivalent to fixing the location parameter and that, in this case, the conditions for the Maximum Likelihood Estimator to be valid are not met, which can produce instabilities in the quantiles being investigated.

As a consequence, and this is a key result of our study, it is essential to dissociate two clearly distinct concepts: that of selecting data via the statistical threshold u_s and that of carefully determining the origin of the distribution, which is the role of the location parameter μ. Note also that if the estimate of μ is negative, the origin of the distribution may be lower than some peaks below u_s: however, these peaks are not to be included in the fit, which highlights the difference between u_s and μ. Practically speaking, during the step of statistical optimization of the threshold (see Section 2.2), this is a justification that limiting the values of u_s to be tested to the values of the peaks $X_{i:N_T}$ is necessary and sufficient: for each of these values, the same exceeding peaks will be fit, including an estimation of the location parameter. Thus this step should be considered as a sensitivity study to the dataset, in contrast with a sensitivity study to a parameter.

It is therefore recommended to work with distributions that include this parameter. Because the MLE does not produce proper parameter estimation, alternative estimators will be examined.

4. Alternative estimators for 3-parameter distributions

4.1. Quick review

In addition to the Maximum Likelihood Estimator, the other classic estimator is that of the method of moments. This involves relating the theoretical moments of the distribution, which depend on its parameters, with the empirical moments of the sample, namely its empirical mean, variance, etc. This very simple method is however unsuited to small samples (typically less than 50–80 data for environmental applications). In particular, it has a severe bias, poor variance and sometimes slow convergence. It is therefore unwise to use it for applications such as determining extreme waves. Generalizations have been made, see for example Bera and Bilias (2002) for a history of the estimation and a summary of the estimators connected with the method of moments and those connected with the likelihood function.

There are numerous other estimators, such as for example the likelihood moment estimator (Zhang, 2007), the generalized moments method (Hansen, 1982) or the generalized empirical likelihood method (Owen, 2001). de Zea Bermudez and Kotz (2010a, 2010b) have reviewed many estimators for the GPD, with a particular focus on the 2-parameter form. However, this study suggests the use of a relatively simple method: the Probability Weighted Moments and L-moments estimators. These were used by Van Gelder et al. (2000) in the Dutch North Sea, by Goda and Kudaka (2009) and Goda et al. (2010) in the Japan Sea and by Pandey et al. (2001) for determining extreme wind velocities. In the latter paper, the authors studied these estimators with and without a location parameter and they favored its inclusion.

4.2. Probability Weighted Moments and L-moments

An alternative to the method of moments was proposed by Greenwood et al. (1979). They defined the Probability Weighted Moments (PWM) as moments weighted by non-exceedance probabilities. The disadvantage of probability weighted moments is that they have no immediate statistical meaning. To remedy the situation, Hosking (1990) defined L-moments. The idea is to combine the PWMs linearly (hence the *L* in the name L-moments) so that they can be interpreted more easily. The L-moments λ_r are linear functions of the expected order statistics of a random sample derived from a distribution and hence can be related quite simply to its parameters.

A consequence is that estimating the L-moments allows an estimation of the parameters just like the classical method of moments, including the location parameter. Thus the L-moments estimator is used for estimating 3-parameter distributions in the following. The definition and estimation of PWMs and L-moments are extensively presented in Appendix.

5. Comparison of 2- and 3-parameter distributions

5.1. Sensitivity study to the statistical threshold

When estimating 3-parameter distributions by the L-moments and letting the statistical threshold u_s vary in $\mathcal{I}_u$ in the same way

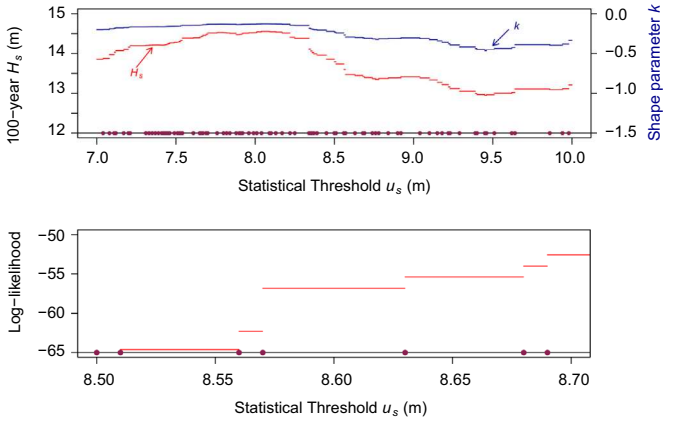

Fig. 7. Haltenbanken – change in the L-moments estimated 3-parameter GPD shape parameter and in Hs_{100} (above) and in the log-likelihood (below, zoom) with respect to u_s. Points represent the data values $X_{i:N_T}$.

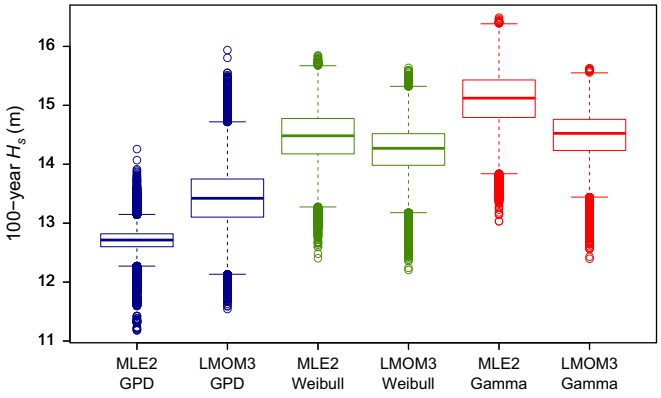

Fig. 8. Haltenbanken: Boxplots of Hs_{100} for 100,000 sub-samplings.

as in Section 3.2, the shift in the threshold is strictly compensated by an identical shift in the location parameter, meaning that all the different samples $(Y_i - \mu)$ remain constant, in contrast with the different samples of the Y_i. As a consequence, the shape and scale parameters, quantiles (Hs_{100}) and log-likelihood remain strictly constant in $\mathcal{J}_u$. The estimator of the L-moments coupled with the introduction of a location parameter is therefore insensitive to any "artificial", or non-physical, translation of the sample. This stability is illustrated for the GPD fit of the Haltenbanken dataset in Fig. 7, to be linked with Fig. 3.

In Section 5.2–5.4, further comparison of the performance of 2-parameter MLE (hereinafter referred to as MLE2) and 3-parameter L-moments estimator (hereinafter referred to as LMOM3) is presented for the GPD, Weibull and Gamma laws, based both on simulated and real-world environmental datasets.

5.2. Sub-sampling of environmental datasets

From the N peaks above u_s of the Haltenbanken dataset, n storm peaks are randomly selected ($N = 46$, $n = 35$, i.e. a rate of 76% and more than 13 billion possible combinations), then the resulting sub-sample is fitted to the GPD, Weibull and Gamma laws by MLE2 and LMOM3. The sub-sampling is performed 100,000 times. This technique is used for studying the variability of the estimators using boxplots, which allow visualization of the median and variance of the estimated parameters, the quantiles, the likelihood of the fit, etc., or more specifically the position of the quartiles and hence their relative spacing. The boxplots of quantile Hs_{100} for each distribution are plotted in Fig. 8. From a physical point of view, sub-sampling can be seen as the equivalent of random breakdowns of the buoy or any measurement devices.

It can be seen that for the same law (in its 2- or 3-parameter formulation), the differences with regard to the medians of Hs_{100} may be significant (up to 6–7%) from one estimator to another (compare boxplot 1 with 2, 3 with 4, 5 with 6). In particular, the MLE2–GPD yields Hs_{100} that are significantly lower than the LMOM3-GPD, and therefore non-conservative from an engineering point of view. For this case study, using the L-moments estimator greatly reduces the difference between the Hs_{100} from one distribution to another: it falls from 20% between the GPD law and Gamma law with the MLE to 9% between the 3-parameter GPD law and the Pearson-III law with the L-moments (compare boxplots 1, 3 and 5 on the one hand with 2, 4 and 6 on the other). The variance or scatter of the MLE2 estimates is significantly less than with LMOM3 for the GPD (see first two boxplots) while it is roughly the same for Weibull and Gamma.

To compare the three fits using MLE2 in order to select the best one, BIC and AIC were used in MH2011. However, these criteria involve maximizing the likelihood (Schwarz, 1978; Akaike, 1973) and this is no longer the case with the L-moments. Focusing on likelihood is thus no longer the most pertinent way to get a descriptive measure of goodness-of-fit. Other measures are therefore necessary. Among many authors, Mathiesen et al. (1994) provide a review of the most commonly used goodness of fit tests: Kolmogorov–Smirnov or Chi-square tests, Cramér–von Mises or Anderson–Darling statistics, etc. We choose the Kolmogorov–Smirnov (KS) distance, that is the maximum distance between the empirical distribution function curve F_N of the sample and the theoretical cumulative distribution function F of the law estimated using the same sample. It is written as follows:

$$D_N = \sup_y |F_N(y) - F(y)| \tag{7}$$

The empirical distribution function of the sample of the Y_i is written as follows:

$$F_N(y) = \frac{1}{N}\sum_{i=1}^{N} I_{Y_i \le y} \tag{8}$$

I is the indicator function, which has a value of 0 or 1. On the basis of D_N, a p-value can be derived bijectively. This is more significant as it relates this distance to the size N of the sample. The closer it is to 1, the more the hypothesis that "the sample is derived from the distribution studied" is likely to be true. It is therefore a good way of discriminating the distributions fitted to a sample. In this exercise, it is used for comparing the goodness of fit of a statistical law with and without a location parameter. The boxplots of the KS p-values computed for the fits of the 100,000 sub-samplings are presented in Fig. 9. It can be seen that for each law, LMOM3 obtains much better p-values than MLE2: the boxplots are narrower and closer to 1. This is particularly obvious for the GPD. Hence, as far as the GPD, Weibull and Gamma laws are concerned, introducing and estimating the location parameter improved the fits.

5.3. Further justification for the location parameter by a hybrid KS–ML estimator

The analysis carried out in Section 3.2 (Fig. 2) can be extended, this time studying the change in this KS p-value when u_s varies. At each step, 2-parameter distributions are fitted by MLE. u_s is now allowed to be smaller than the lower bound $X_{(N_T-N):N_T}$, but without including the corresponding peaks. It can also be seen as setting discrete values of μ in $\mathcal{J}_\mu$, possibly negative, and fitting 2-parameter distributions for each discrete value.

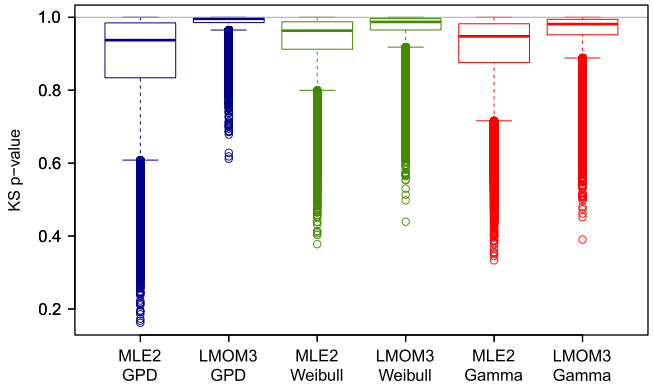

Fig. 9. Haltenbanken: Boxplots of the Kolmogorov–Smirnov p-value for 100,000 sub-samplings.

The case study of Haltenbanken is supplemented by the second case study of MH2011, consisting of 44 years of hindcast data at a point located at the western entrance to the Strait of Gibraltar, extracted from the SIMAR-44 database from Puertos del Estado, Madrid, Spain. There are $N_T = 288$ i.i.d. peaks $X_{i:N_T}$ above the physical threshold $u_p = 3$ m and the statistical threshold is set at $u_s = 4.3$ m. The first peak above u_s is at 4.4 m ($Y_{1:N} = 0.1$ m). The results are presented in Fig. 10 (once again, $\overline{u}_s$ is the original choice of the statistical threshold determined as explained in Section 2.2). In addition, the estimate of μ by LMOM3 associated the KS p-value of the fit are plotted (dots).

While there was no maximum log-likelihood in two cases out of three (Fig. 2b), the KS p-value here exhibits global maxima. Hence one may estimate μ on the basis of this criterion: this is actually a hybrid "ML–KS estimator", with ML-estimated shape and scale parameters and KS-estimated location parameter. This estimation can be compared with the estimates by LMOM3, summed up in Table 1.

It is noteworthy that the estimates are consistent between the two estimators: in particular, all estimates are negative for the Haltenbanken dataset and positive (except for LMOM3 Gamma) for the Gibraltar dataset. Even though the choice of the KS test may not be the best one, it appears that the introduction of a location parameter improves the fit. The Haltenbanken case also shows that μ can be negative, making it necessary to distinguish it from the statistical threshold u_s whose role should be limited to extreme data selection or censorship.

5.4. Monte-Carlo simulations

The performance of MLE2 and LMOM3 is further examined by using simulated datasets in a Monte-Carlo approach. For each law (GPD, Weibull, Gamma), a set of shape and scale parameters is chosen so as to be typical of environmental applications: $(k_0 = -0.25; \sigma_0 = 1.5)$ for the GPD, $(k_0 = 1.25; \sigma_0 = 1.5)$ for Weibull and $(k_0 = 2; \sigma_0 = 0.75)$ for Gamma. A sensitivity study is performed on the location parameter μ_0 which can take five values: -0.01, 0, 0.01, 0.025 and 0.05. Considering that these samples may correspond to datasets of H_s peaks with duration of $K = 20$ years and mean number of events per year of $\lambda = 5$, the 100-year H_s quantile can be computed for each law. Depending on the value of μ_0, its value is around 9.75 m for the GPD, 11.5 m for Weibull and 11.35 m for Gamma. For each law and set of parameters, 10,000 samples of size $N = K\lambda = 100$ are randomly generated and fit by MLE2 and LMOM3. Last, the bias and root mean square error (RMSE) are computed for the 100-year H_s quantile (Table 2).

For the GPD and Gamma laws, there is a significant decrease of the absolute value of the bias when using the 3-parameter L-moments (typically 2–4 times smaller). For the Weibull law, such a decrease is less obvious when μ_0 is close to 0. It is quite noteworthy that 2-parameter MLE yields negatively biased quantiles, that is under-estimated 100-year H_s, while 3-parameter L-moments tend to yield positively biased quantiles for the GPD and Weibull laws, which is conservative for engineering purposes. This is consistent with the results of the sub-sampling on environmental datasets (see Fig. 8, Section 5.2). Last, RMSE is typically 20–30% larger with 3-parameter L-moments.

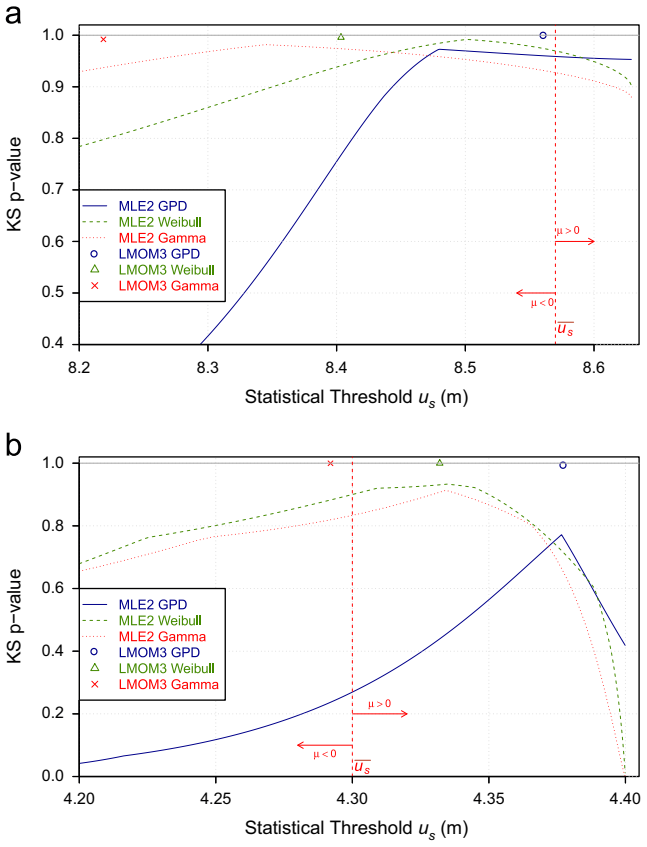

Fig. 10. Change in Kolmogorov–Smirnov p-value when μ runs through $\mathcal{I}_\mu$ – (a) Haltenbanken and (b) Gibraltar.

Table 1
Location parameter estimates and KS p-value using hybrid ML–KS estimator and L-moments estimator.

	ML–KS		L-moments	
	$\hat{\mu}$	KS p-value	$\hat{\mu}$	KS p-value
Haltenbanken				
GPD	−0.090	0.9728	−0.010	0.9999
Weibull	−0.069	0.9918	−0.167	0.9957
Gamma	−0.225	0.9817	−0.351	0.9920
Gibraltar				
GPD	0.077	0.7717	0.077	0.9936
Weibull	0.035	0.9331	0.032	0.9996
Gamma	0.034	0.9133	−0.008	0.9996

6. Discussion

This study aims at highlighting instabilities of parameter estimation by the Maximum Likelihood Estimator while following the classical methodology of extreme wave heights (or other environmental variables) extrapolation, combining POT declustering and GPD–Poisson model; but also with other distributions widely spread in the

Table 2
Simulated data: bias and RMSE of the estimated 100-year H_s.

	2-parameter MLE		3-parameter L-moments	
	Bias	RMSE	Bias	RMSE
$\mu_0=-0.01$				
GPD	−0.177	0.593	0.066	0.724
Weibull	−0.056	0.765	0.028	0.933
Gamma	−0.032	0.625	0.016	0.829
$\mu_0=0$				
GPD	−0.167	0.596	0.077	0.739
Weibull	−0.048	0.768	0.042	0.924
Gamma	−0.039	0.619	0.009	0.819
$\mu_0=0.01$				
GPD	−0.191	0.593	0.079	0.737
Weibull	−0.047	0.764	0.050	0.932
Gamma	−0.060	0.617	−0.009	0.820
$\mu_0=0.025$				
GPD	−0.217	0.588	0.074	0.741
Weibull	−0.067	0.757	0.024	0.926
Gamma	−0.068	0.625	−0.026	0.844
$\mu_0=0.05$				
GPD	−0.270	0.584	0.061	0.727
Weibull	−0.097	0.773	−0.011	0.916
Gamma	−0.101	0.630	−0.045	0.816

literature. These instabilities on the parameters logically affect the quantile computation: in other words, different threshold values falling between two storm peaks yield different extreme wave heights, though it seems counterintuitive.

Based both on real environmental and simulated datasets, this study shows that these instabilities are linked with the likelihood of the statistical distributions which tends to increase as the threshold tends to the lowest storm peak exceedance $Y_{1:N}$. As for the GPD, this can be explained by the shape of the probability density function whose maximum is reached at $y=0$, whatever the sign of k (positive or negative). However, as a consequence, the hypotheses necessary for proving the MLE asymptotic properties, in particular its consistency, are not met. We introduce the use of a location parameter so as to make a clear distinction between selecting extreme storm peaks to be extrapolated (the role of u_s) and estimating the origin of the distribution of the extreme sample (the role of μ). Actually, one could argue that letting u_s vary between two peaks is exactly the same as setting the threshold and introducing μ. Thus this could seem a superfluous sophistication of the vocabulary: however, it has been shown that the optimal value of μ could be negative and fall below a storm peak that is not included in the extreme sample. Hence introducing μ not only clarifies the concepts but is also necessary for determining the optimal set of parameters.

Using 3-parameter distributions require an alternative estimator to the MLE. Estimators based on PWMs or L-moments appear to be highly suitable and have been used in this study for comparative purposes. Parameter estimates are easily computed and any shift, or translation, in the sample of the extreme peak excesses $Y_{i:N}$ is strictly compensated by an inverse shift in the location parameter. Hence, the shape and scale parameters and, as a consequence, the quantiles (wave heights for any return period) are constant as far as the same storm peaks are fitted. Accounting for a location parameter also leads to a notable improvement in the goodness of fit for real-world environmental datasets, as evaluated by the Kolmogorov–Smirnov distance (or p-value), as well as for simulated datasets, as evaluated by the bias and RMSE of the 100-year H_s quantile. Future work could include an extension of this comparison with other goodness-of-fit criteria. Comparing the behavior of the log-likelihood and of the KS p-value (Figs. 2b and 10) is an extra clue that the likelihood is not appropriate for estimating the location parameter. Future work could also include a more in-depth comparison of the performance of 2-parameter MLE vs. 3-parameter L-moments with extended sets of parameters and with different sample sizes.

It is also necessary to question the practical usefulness and relevance of this work. Does the application of our recommendations improve the results? If there is such an improvement, is it really worth bothering with both u_s and μ (not to mention u_p)? Or, in other words, how severe is the problem? The answer depends on the distribution (GPD, Weibull, Gamma or others) and on how the threshold is set. In particular, the real-world examples above (Figs. 2 and 3) show that if u_s is reasonably well chosen, the quantile of the GPD, though not constant, is rather stable, the amplitude of the *virgae* being limited. Regarding Weibull and Gamma, the divergence appears for threshold values just below the first peak. Thus it can be considered that this point has generally little effect on the final results. Nonetheless, we consider that the analysts should be aware of this issue, especially when the threshold is set using techniques implying a sensitivity study with respect to u_s: in such a case, it implicitly plays the role of a location parameter. Attention should also be paid to the step of this sensitivity study, in relation with the number of significant digits of the data. Moreover, the severity of the problem has been found to be more acute on other environmental datasets, in particular for the Gibraltar dataset with 100-year H_s variations when u_s varies between two data values up to 0.2 m for the GPD and up to 1 m for Weibull and Gamma.

In any case, it has been shown that 3-parameter distributions could yield significantly different extreme quantiles while performing better in goodness-of-fit tests. In addition to the mere knowledge of the causes of this *virga* phenomenon, this should favor introducing μ.

7. Conclusions

The presence of *virgae* in the estimation of parameters and quantiles is linked to the non-existence of a global maximum of the likelihood function within the domain of validity of the parameters of the distribution. This study shows that strictly speaking, the location parameter should not be confused with the statistical threshold. Their respective roles: selecting the data to be fitted for the statistical threshold; accurately setting the origin of the distribution for the location parameter; should be clearly differentiated in order to avoid producing unstable estimations. Such a conclusion is valid for any POT data or OTM analysis.

The methodology for determining extreme wave heights presented in MH2011 can thus be improved by replacing ML-estimated 2-parameter distributions by L-moments-estimated 3-parameter distributions using the KS p-value instead of BIC/AIC. This methodology can then be summarized as follows:

(1) Homogenization of time series;
(2) physical declustering and selection of i.i.d. storm peaks by POT approach using a physical threshold u_p;
(3) determination of an optimal statistical threshold u_s by a stability analysis of the GPD shape and modified scale parameters;
(3) Simulation and experimental results are provided to demonstrate the efficiency of the method.
(4) fit by L-moments of 3-parameter GPD, Weibull and Gamma distributions;
(5) selection of the best fit using the Kolmogorov–Smirnov p-value; and

(6) computation of return levels (quantiles) and confidence intervals (by parametric bootstrap).

Future works could include further investigations for determining the best estimator for 3-parameter distributions, as well as the best goodness-of-fit criterion. In particular, it could be explored whether hybrid estimators similar to the one presented in Section 5.3 could perform better than the L-moments.

Acknowledgments

The authors wish to thank J.R.M. Hosking for his advice and assistance in understanding the L-moments method, as well as the reviewers for their help and relevant suggestions which truly helped improving this paper.

Appendix. Definition and estimation of L-moments

Definition of PWMs and L-moments

Greenwood et al. (1979) defined the Probability Weighted Moments (PWM) as follows:

$$M_{p,r,s} \equiv E[X^p \{F(X)\}^r \{1-F(X)\}^s] = \int x^p \{F(x)\}^r \{1-F(x)\}^s dF(x) \quad (9)$$

where p, r and s are real numbers. The moments are therefore weighted by exceedance probabilities $1-F$ and/or non-exceedance probabilities F. More particularly, we shall consider the following PWMs:

$$\alpha_r \equiv M_{1,0,r} = E[X\{1-F(X)\}^r], \quad r \in \mathbb{N} \quad (10)$$

$$\beta_r \equiv M_{1,r,0} = E[X\{F(X)\}^r], \quad r \in \mathbb{N} \quad (11)$$

Each of these two forms of PWMs is sufficient to characterize the distribution F completely and they are interchangeable:

$$\alpha_r = \sum_{k=0}^{r} (-1)^k \binom{r}{k} \beta_k \; ; \; \beta_r = \sum_{k=0}^{r} (-1)^k \binom{r}{k} \alpha_k \quad (12)$$

The disadvantage of probability weighted moments is that they have no immediate statistical meaning. To remedy the situation, Hosking (1990) defined L-moments. The idea is to combine the PWMs linearly (hence the L in the name L-moments) so that they can be interpreted more easily. Considering an ordered sample, the L-moments are then defined as follows:

$$\lambda_r \equiv \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E[X_{r-k:r}] \; ; r \in \mathbb{N}^* \quad (13)$$

where $X_{k:N}$ is the kth order statistic of a random sample of size N derived from the distribution F. The λ_r are therefore linear functions of the expected order statistics. In particular, λ_1 is the expected value of the distribution, and hence a measurement of the location, and λ_2 is a measurement of the range or scatter of the data, in a similar way to the conventional second moment, variance. Because of this, the higher-order moments can be normalized by λ_2; the ratios of the L-moments are thus defined in the following manner:

$$\tau_r \equiv \frac{\lambda_r}{\lambda_2} \; ; r = 3, 4, \ldots \quad (14)$$

τ_3 may thus be interpreted as a measurement of asymmetry and τ_4 as a measurement of flatness (or, conversely, of "peakedness"). λ_1, λ_2, τ_3 and τ_4 are therefore referred to respectively as the *L-location*, *L-scale*, *L-skewness* and *L-kurtosis*.

L-moments are connected to PWMs as follows:

$$\lambda_r = (-1)^{r-1} \sum_{k=0}^{r-1} p^*_{r-1,k} \alpha_k = \sum_{k=0}^{r-1} p^*_{r-1,k} \beta_k \; ; \; r \in \mathbb{N}^* \quad (15)$$

with

$$p^*_{r,k} = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} \quad (16)$$

This means that working with α_r, β_r or λ_r is strictly equivalent. Therefore, for each distribution studied, it is a case of choosing the formulation that expresses the PWMs or L-moments as simply as possible depending on the parameters. As an example, there is the following relation for the 3-parameter Weibull distribution:

$$\alpha_r = \frac{\mu}{1+r} + \sigma \frac{\Gamma(1+\frac{1}{k})}{(1+r)^{1+\frac{1}{k}}}, \quad r \in \mathbb{N} \quad (17)$$

In the same way, the relation between the L-moments and the three parameters of the GPD is written in the following way:

$$\begin{cases} \lambda_1 = \mu + \frac{\sigma}{1-k} \\ \lambda_r = \sigma \frac{\Gamma(1-k)\Gamma(r-1+k)}{\Gamma(1+k)\Gamma(r+1-k)}, \; r = 2, \; 3, \ldots \end{cases} \quad (18)$$

Conversely, the simplest formulation will be used to express the parameters as a function of the PWMs or L-moments. Explicit expressions may be established when the cumulative distribution function of the distribution is invertible, as is the case with the GPD and Weibull law

$$3-\text{parameter GPD} : \begin{cases} k = \frac{3\tau_3 - 1}{1+\tau_3} \\ \sigma = \lambda_2 (1-k)(2-k) \\ \mu = \lambda_1 - \frac{\sigma}{1-k} \end{cases} \quad (19)$$

$$3-\text{parameter Weibull} : \begin{cases} \mu = 4\frac{\alpha_0\alpha_3 - \alpha_1^2}{4\alpha_3 + \alpha_0 - 4\alpha_1} \\ \sigma = \frac{\alpha_0 - \mu}{\Gamma\left(\frac{\ln\left(\frac{\alpha_0 - 2\alpha_1}{\alpha_1 - 2\alpha_3}\right)}{\ln 2}\right)} \\ k = \frac{\ln 2}{\ln\left(\frac{\alpha_0 - 2\alpha_1}{2(\alpha_1 - 2\alpha_3)}\right)} \end{cases} \quad (20)$$

It is therefore simply a question of estimating α_0, α_1 and α_3 in order to calculate the three parameters μ, k and σ of the Weibull distribution, and to do the same with λ_1, λ_2 and τ_3 for the GPD.

In contrast, other distributions have a non-reversible cumulative distribution function. This is the case for example with the Gamma distribution with two or three parameters (Pearson-III). In such cases, it is not possible to express the parameters as a function of the PWMs and/or L-moments, and the equations must be solved numerically by iterative methods.

Estimation of PWMs and L-moments

Given a sample sorted in ascending order and of size N, the PWMs are estimated in the following form (Landwehr et al., 1979):

$$a_r \equiv \frac{1}{N} \sum_{i=1}^{N} \frac{\binom{N-i}{r}}{\binom{N-1}{r}} X_{i:N} \quad (21)$$

$$b_r \equiv \frac{1}{N} \sum_{i=1}^{N} \frac{\binom{i-1}{r}}{\binom{N-1}{r}} X_{i:N} \quad (22)$$

a_r and b_r are unbiased estimators of α_r and β_r, which also verify the identities (12). By analogy with (15), an unbiased estimator of the L-moments is immediately deduced from this

$$l_r \equiv (-1)^{r-1} \sum_{k=0}^{r-1} p^*_{r-1,k} a_k = \sum_{k=0}^{r-1} p^*_{r-1,k} b_k \tag{23}$$

Another option is to use other estimators of PWMs defined by using empirical plotting-position formulae $p_{i:N}$

$$\tilde{a}_r \equiv \frac{1}{N} \sum_{i=1}^{N} (1-p_{i:N})^r X_{i:N} \tag{24}$$

$$\tilde{b}_r \equiv \frac{1}{N} \sum_{i=1}^{N} p^r_{i:N} X_{i:N} \tag{25}$$

$$\tilde{l}_r \equiv \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{k=0}^{r-1} p^*_{r-1,k} (p_{i:N})^k \right] X_{i:N} \tag{26}$$

$p_{i:N}$ is usually of the form

$$p_{i:N} = \frac{i+a}{N+b} \tag{27}$$

Various formulations exist (see for instance Seckin et al., 2010, for a detailed comparison), such as the Weibull or Cunnane (1978) formulae: $(a,b)=(0,1)$ for the former and $(-0.4, 0.2)$ for the latter. Hosking and Wallis (1987) suggest $(a,b)=(-0.35, 0)$.

The parameters are then estimated as in the case of the conventional moments method, i.e. by equating the theoretical PWMs or L-moments of the distribution, which depend on these parameters, with the estimators a_r, b_r, l_r, or $\tilde{a}_r$, $\tilde{b}_r$, $\tilde{l}_r$.

If the unbiased formulation is chosen, it is then strictly equivalent to estimate the parameters of a distribution by a_r, b_r or l_r and if an empirical probability formula is chosen, the same is true with $\tilde{a}_r$, $\tilde{b}_r$ or $\tilde{l}_r$ (in both cases, the identities (12) and (15) may be used).

Theoretically, there is no reason to prefer formula (27) and the associated estimators (24)–(26) to the unbiased formulations . In certain ranges of shape parameter, Hosking and Wallis (1987) showed that this formula could give better results for the GPD law. Our tests showed, however, that the differences between the two formulations in the following comparisons are not significant. For greater clarity we shall therefore restrict ourselves to the unbiased estimators a_r, b_r or l_r.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Proceedings of the Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Bera, A.K., Bilias, Y., 2002. The MM, ME, ML, EL, EF and GMM approaches to estimation: a synthesis. J. Econom. 107 (1–2), 51–86.

Bernardara, P., Mazas, F., Weiss, J., Andreewsky, M., Kergadallan, X., Benoît, M., Hamm, L., 2014. On the two-step threshold selection for over-threshold modeling. Coast. Eng. Proc. 1 (33), http://dx.doi.org/10.9753/ice.v33.management.42.

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

Cunnane, C., 1978. Unbiased plotting position formula. A review. J. Hydrol. 37, 205–222.

de Zea Bermudez, P., Kotz, S., 2010a. Parameter estimation of the generalized Pareto distribution – Part I. J. Stat. Plan. Inference 140, 1353–1373.

de Zea Bermudez, P., Kotz, S., 2010b. Parameter estimation of the generalized Pareto distribution – Part II. J. Stat. Plan. Inference 140, 1374–1388.

Dorsch, W., Newland, T., Tassone, D., Tymons, S., Walker, D., 2008. A statistical approach to modeling the temporal patterns of ocean storms. J. Coast. Res. 24 (6), 1430–1438.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, London.

Goda, Y., 2000. Random seas and design of maritime structures, second ed. Advanced Series on Ocean Engineering, 15. World Scientific.

Goda, Y., Kudaka, M., 2009. Bounded and unbounded distribution functions for extreme wave analysis. In: Proceedings of the Asian Pacific Coasts, World Scientific, pp. 8–14.

Goda, Y., Kudaka, M., Kawai, H., 2010. Incorporation of Weibull distribution in L-moments method for regional frequency analysis of peaks-over-threshold wave heights. In: Proceedings of the 32nd Conference on Coastal Engineering, Shanghai, China, 2010, World Scientific. Available online: ⟨https://journals.tdl.org/ICCE/issue/current⟩.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C., Wallis, J.R., 1979. Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. Water Resour. Res. 15 (5), 1055–1064.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Hosking, J.R.M., 1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. J. R. Stat. Soc. 52 (1), 105–124.

Hosking, J.R.M., Wallis, J.R., 1987. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics 29, 339–349.

Jonathan, P., Ewans, K., 2007. The effect of directionality on extreme wave design criteria. Ocean Eng. 34, 1977–1994.

Landwehr, J.M., Matalas, N.C., Wallis, J.R., 1979. Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. Water Resour. Res. 15 (5), 1055–1064.

Lehmann, E.L., 1983. Theory of Point Estimation. Wiley & Sons, Australia.

Mackay, E., Challenor, P., Bahaj, A., 2010. On the use of discrete seasonal and directional models for the estimation of extreme wave conditions. Ocean Eng. 37, 425–442.

Mathiesen, M., Goda, Y., Hawkes, P.J., Mansard, E., Martín, M.J., Peltier, E., Thompson, E.F., van Vledder, G, 1994. Recommended practice for extreme wave analysis. J. Hydraul. Res. 32 (6), 803–814.

Mazas, F., Hamm, L., 2011. A multi-distribution approach of POT methods for determining extreme wave heights. Coast. Eng. 58, 385–394.

Mazas, F., Hamm, L., 2012. Reply to discussion by Z.J. You of "A multi-distribution approach to POT methods for determining extreme wave heights" by Mazas and Hamm. Coast. Eng. 65, 16–18.

Méndez, F.J., Menéndez, M., Luceño, A., Losada, I.J., 2006. Estimation of long-term variability of extreme significant wave height using a time-dependent Peaks Over Threshold (POT) model. J. Geophys. Res. 111, C07024. http://dx.doi.org/10.1029/2005JC003344.

Méndez, F.J., Menéndez, M., Luceño, A., Medina, R., Graham, N.E., 2008. Seasonality and duration in extreme value distributions of significant wave height. Ocean Eng. 35, 131–138.

Nogaj, M., Parey, S., Dacunha-Castelle, D., 2007. Non-stationary extreme models and a climatic application. Nonlinear Process. Geophys. 14, 305–316.

Owen, A.B., 2001. Empirical Likelihood. CRC Press.

Panchang, V.G., Gupta, R.C., 1989. On the determination of three-parameter Weibull MLE's. Commun. Statist. Simul. 18 (3), 1037–1057.

Pandey, M.D., van Gelder, P.H.A.J.M., Vrijling, J.K., 2001. The estimation of extreme quantiles of wind velocity using L-moments in the peaks-over-threshold approach. Struct. Saf. 23, 179–192.

Pickands, J., 1975. Statistical inference using extreme order statistics. Ann. Stat. 3, 119–131.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.

Seckin, N., Yurtal, R., Haktanir, T., Dogan, A., 2010. Comparison of probability weighted moments and maximum likelihood methods used in flood frequency analysis for Ceyhan river basin. Arab. J. Sci. Eng. 35, 49–69.

Solari, S., Losada, M., 2012. Unified distribution models for met-ocean variables: application to series of significant wave height. Coast. Eng. 68, 67–77.

Tancredi, A., Anderson, C., O'Hagan, A., 2006. Accounting for threshold uncertainty in extreme value estimation. Extremes 9 (2), 87–106.

Thompson, P., Cai, Y., Reeve, D., Stander, J., 2009. Automated threshold selection methods for extreme wave analysis. Coast. Eng. 56, 1013–1021.

Van Gelder, P.H.A.J.M., De Ronde, J.G., Neykov, N.M., Neytchev, P., 2000. Regional frequency analysis of extreme wave heights: trading space for time. Coast. Eng., 1099–1112 (April 2001).

Van Vledder, G., Goda, Y., Hawkes, P., Mansard, E., Martin, M.J., Mathiesen, M., Peltier, E., Thompson, E., 1994. Case studies of extreme wave analysis: a comparative analysis. In: Proceedings of the Second Symposium on Ocean Wave Measurement and Analysis. ASCE, New Orleans, Louisiana, USA, pp. 978–992.

You, Z.J., 2012. Discussion of "a multi-distribution approach to POT methods for determining extreme wave heights" by Mazas and Hamm [Coast. Eng., 58, pp. 385–394]. Coast. Eng. 61, 49–52.

Zhang, J., 2007. Likelihood moment estimation for the generalized Pareto distribution. Aust. N.Z. J. Stat. 49 (1), 69–77.

1.4. COASTAL ENGINEERING 2014: APPLYING POT METHODS TO THE REVISED JOINT PROBABILITY METHOD FOR DETERMINING EXTREME SEA LEVELS

Main publications citing this paper (source Google Scholar)

Fortunato A. B., Li K., Bertin X., Rodrigues M., Miguez B. M., 2016. Determination of extreme sea levels along the Iberian Atlantic coast. *Ocean Engineering*, **111**, 471-482.

Bulteau T., Idier, D., Lambert J., Garcin M., 2015. How historical information can improve estimation and prediction of extreme coastal water levels: application to the Xynthia event at La Rochelle (France). *Natural Hazards and Earth System Sciences*, **15(6)**, 1135-1147.

Lin-Ye J., Garcia-Leon M., Gracia V., Sanchez-Arcilla A., 2016. A multivariate statistical model of extreme events: An application to the Catalan coast. *Coastal Engineering*, **117**, 138-156.

Caballero-Megido C., Hillier J., Wyncoll D., Bosher L., Gouldby B., 2017. comparison of methods for threshold selection for extreme sea levels. *Journal of Flood Risk Management*.

Antonini A., Archetti R., Lamberti A., 2017. Wave simulation for the design of an innovative quay wall: the case of Vlorë Harbour. *Natural Hazards and Earth System Sciences*, **17(1)**, 127.

Coastal Engineering 91 (2014) 140–150

Contents lists available at ScienceDirect

Coastal Engineering

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

Applying POT methods to the Revised Joint Probability Method for determining extreme sea levels

Franck Mazas [a,*], Xavier Kergadallan [b], Philippe Garat [c], Luc Hamm [a]

[a] ARTELIA Eau & Environnement, 6 rue de Lorraine, 38130 Echirolles, France
[b] CEREMA/DTecEMF, 155 rue Pierre Bouguer, 29280 Plouzané, France
[c] Université Pierre Mendes-France, Laboratoire Jean Kuntzmann, UMR 5224, BSHM, 1251 avenue centrale, BP 47, 38040 Grenoble Cedex 09, France

ARTICLE INFO

Article history:
Received 14 November 2012
Received in revised form 19 May 2014
Accepted 20 May 2014
Available online xxxx

Keywords:
POT
Mixture model
Sea level
Storm surges
JPM
Tide–surge interaction

ABSTRACT

Newly exposed concepts of POT declustering (Bernardara et al., 2014) within the GPD-Poisson model are applied to the joint probability of tide and surge for determining extreme sea levels, as a variation of the Revised Joint Probability Method (RJPM, Tawn and Vassie, 1989). A mixture model is proposed for the meteorological residual (surge) component with a non-parametric (empirical) density for the bulk values and parametric models for both the lower and upper tails. In particular, a distinction is made between values observed at regular time steps, called *sequential values*, and the clusters of extreme values, or *events*, on which the statistical extrapolations are performed. The sea level distribution is obtained by convolution of the tide and surge density functions. Confidence intervals are also proposed. This model is applied to the case study of Brest, France using both hourly and high water values. Two methods for handling tide–surge interaction are presented and discussed and a comparison with a direct approach is made.

1. Introduction

Accurate estimation of extreme sea levels is crucial for assessing risk in coastal areas and particularly for ensuring the safety of the people living there. In France, Atlantic storm Xynthia of February 27–28th, 2010 caused extensive flooding because of the joint occurrence of a large storm surge with the high water of a spring tide, highlighting the need to predict such events accurately by properly taking into account both the tidal and surge components of the sea level.

There are two traditional ways of determining extreme sea levels. The first includes what are called direct methods: they deal with the sea level variable by directly extrapolating observations. In contrast, the second family includes indirect methods: considering that the astronomical tide is a fully deterministic variable while the surge is a stochastic one, they are dealt with separately then recombined by convolution.

Pugh and Vassie (1979) proposed the first indirect approach with the Joint Probability Method (JPM). The empirical probability density functions of the hourly observations of both the tidal and surge components were computed then convoluted. One of the main drawbacks of the method was that a null probability was assumed for surges larger than the highest observation: thus the highest predicted sea level was simply the addition of the maximum surge observation and of the Highest Astronomical Tide (HAT). It is clear that such a procedure can yield dangerous underestimations, particularly in the case of short datasets. Furthermore, hourly surge values are strongly correlated and thus dependent, which can affect the estimation of extreme value return periods.

Tawn and Vassie (1989) proposed an extension of the JPM, called the Revised Joint Probability Method (RJPM), to correct these deficiencies. Tawn (1992) gives a detailed presentation of the RJPM. In particular the extreme value theory is called on to fit a parametric model to the upper tail of the surge distribution by extrapolating the highest observations. These extreme values can be extracted either by identifying the highest data peaks of the time series, with approximately 5 peaks per year on average, or by an *r*-largest approach. Therefore the annual maxima of hourly surges follow a Generalized Extreme Value (GEV) distribution, or one of its special cases: the Gumbel distribution. The clustering of surge and sea level extremes is then accounted for by an extremal index for converting the distribution of hourly values into return periods. As a result, the return level is parameterized as the GEV parameters and two extremal indices representative of the mean duration of surge and sea level events.

Tide–surge interaction has also been included in the RJPM (Tawn, 1992; Dixon and Tawn, 1994, hereinafter referred to as DT94; Dixon and Tawn, 1999), improving in particular the reliability of the results

* Corresponding author. Tel.: +33 4 56 38 46 91; fax: +33 4 76 33 43 33.
E-mail addresses: franck.mazas@arteliagroup.com (F. Mazas), xavier.kergadallan@developpement-durable.gouv.fr (X. Kergadallan), philippe.garat@iut2.upmf-grenoble.fr (P. Garat), luc.hamm@arteliagroup.com (L. Hamm).

http://dx.doi.org/10.1016/j.coastaleng.2014.05.006
0378-3839/

for shallow water sites. Spatial extensions of these different methods have also arisen, such as the Spatial Revised Joint Probability Method (SRJPM) (Dixon and Tawn, 1995, 1997), but this paper will remained focused on site-by-site analyses (DT94). A good presentation of these methods along with a comparison of their results for eighteen sites around the English Channel can be found in Haigh et al. (2010).

One of the key points of an indirect method is thus the statistical model for the tail(s) of the surge distribution. In other words, it is a problem of extrapolating the extreme surge values and then transforming them into extreme sea level values. Another difficulty lies in the necessary distinction between *events*, such as storms, that can be considered as independent and are fit for a proper statistical extrapolation, and observations or realizations of the variable (sea level or surge) at a regular time step δt, called *sequential values*, that are usually highly auto-correlated and occur in clusters (the events). This distinction between events and sequential values has been thoroughly discussed by Bernardara et al. (2014).

Determining the extremes of environmental variables is a widely studied topic. Over the past decade, a broad consensus has emerged in favor of the joint use of the POT (Peaks-Over-Threshold) approach for declustering extreme events and of the Poisson-GPD model for fitting their peaks (see for instance Coles, 2001, for an overview of the extreme value theory), especially when the observation period is short compared with the return periods to be studied. Two of the present authors have also presented an extension of this model in the form of a multi-distribution approach for determining extreme wave heights (Mazas and Hamm, 2011, hereinafter referred to as MH2011).

This paper aims to apply the approach of POT declustering based on events as described by Bernardara et al. (2014) and the GPD-Poisson model to the family of JPM-type approaches, and in particular the RJPM, for determining extreme sea level values. Particular attention is paid to the mixture model for the surge component, which allows probabilities to be quantified for the entire range of sea level values, not just for the extreme ones. Methods for estimating the confidence intervals are also proposed. Indeed the proposed methodology can be viewed as a variation of the classical RJPM.

The sea level components and processing of the time series of observations are discussed in Section 2. The mixture model for the surge component is described in Section 3. The distribution of sea levels along with confidence intervals is given in Section 4. The issue of tide–surge interaction is briefly presented in Section 5. The proposed methodology is applied to Brest in Section 6 for hourly sea levels and high tide levels. The results are discussed in Section 7.

2. Sea level observations

2.1. Components of the sea level

This paper focuses on the two main components of the sea level: the astronomical tide and the meteorological surge. However, in practice, the analyst deals with a time series of sea level measurements and it is necessary to be fully aware of all the possible components. Indeed, the sea level is influenced by many physical processes and thus can be considered as the sum of the following:

- mean sea level: can be defined over different periods of time and is subject to fluctuations at these same timescales: long-term trend (e.g. climate change or isostatic rebound), decadal oscillations (e.g. due to geostrophic currents), seasonal oscillations, etc.;
- astronomical tide: a purely deterministic phenomenon that can be predicted very accurately at local level;
- meteorologically induced surges: caused by wind and pressure (both static and dynamic) effects on the free surface of the sea;
- coastal seiches: low-frequency waves whose periods typically range from 10 to 50 min and are generated by abrupt meteorological changes;
- wave set-up (or set-down): caused by the action of waves, whose high-frequency fluctuations are filtered by tide gauges but whose breaking can generate low-frequency waves (periods of 1 to 5 min);
- exceptional phenomena such as tidal waves caused by earthquakes, volcanic eruptions, and submarine landslides, which can easily be removed from the analysis.

Note that the first component is a level relative to a datum whereas the following are extra elevations, which may be either positive or negative.

Current practice is to filter the signal of sea level observations to calculate the mean value and center the signal on it. If the period is long enough, a spectral or harmonic analysis is performed to extract the different tidal constituents and re-build the time series of the astronomical tide, then remove it from the centered signal to get the non-tidal residual.

For the sake of simplicity, it will be assumed in this paper that the non-tidal residual is uniquely the meteorologically induced elevation, which will simply be called the surge component and can be either negative or positive. The mean sea level is included in the astronomical tidal component, which is relative to a datum that may be a local chart datum or the mean sea level. We can then write:

$$Z(t) = T(t) + S(t) \tag{1}$$

where Z is the sea level (in m relative to the datum), T the tidal level (in m relative to the datum), S the surge component (non-tidal residual, in m) and t the time.

2.2. Processing of the dataset of sea level observations

Some authors (e.g. Bardet et al., 2011; Bernardara et al., 2011) only use the maxima of sea levels over one tide, i.e. the predicted astronomical high tide and the skew surge (defined as the highest sea level observed around the predicted time of high tide, minus the predicted astronomical high tide), in order both to reduce the dependency of the two components and to improve the reliability and quality of the data. The present method can easily be adapted to such an approach by using the high water values of Z, T and S instead of their observations at a regular time step: it will be illustrated with the case study of Brest, with a comparison of analyses based on hourly sea levels on the one hand, and high water levels on the other hand. In the following presentation of the methodology (Sections 2 to 4), the full range of sea level data will be considered, i.e. sequential values at a regular time step δt, or for the sake of simplicity "hourly values" whatever the value of δt.

Let K be the duration of the time series of observations, in years, and ν be the number of observations per year. In the case of hourly values, $\nu = 365.25 \times 24 = 8766$. Computing the predicted astronomical tide over the duration of the dataset eventually yields three time series of νK data: $Z(t)$, $T(t)$ and $S(t)$.

2.3. Distribution of astronomical tidal levels

The first step is to compute the deterministic variable, i.e. the astronomical tide. In order to (almost) fully determine its density function, it is necessary and (almost) sufficient to get the tidal level values over a continuous period of 18.6 years, corresponding to the lunar nodal cycle. If a different period is chosen, the density function will be biased, unless it is a multiple of 18.6 years. However, it can be considered that any period lasting more than a few years will give a very good approximation.

Let us define the interval $X_T = [x_{T,\min}; x_{T,\max}]$ covering the entire range of the tidal level values $T(t)$. $x_{T,\min}$ is equal to or slightly lower than the Lowest Astronomical Tide (LAT) and $x_{T,\max}$ is equal to or slightly higher than the Highest Astronomical Tide (HAT). For computation

purposes, X_T is discretized at a regular step, denoted δx. Typically, δx can be 0.01, 0.05 or 0.1 m. X_T is called the support of the tidal level distribution.

The empirical density function p_T is then computed over X_T by a non-parametric kernel density estimator. It is recommended that the estimated density should be assumed to be nil for values outside the range [LAT; HAT] (bounded support).

2.4. Distribution of frequently observed surges (bulk distribution)

The same processing as for the tidal signal is applied to the surge observations $S(t)$, but this time over the entire length of the dataset. In a similar way to above, and again for practical computation purposes, let us define $X_S = [x_{S,\min}; x_{S,\max}]$ the support of the surge distributions. $x_{S,\max}$ (resp. $x_{S,\min}$) is a positive (resp. negative) value that must be higher (resp. lower) than the extreme values that will be extrapolated (it could be, for instance, the expected 1000-year or 10,000-year surge value), including the confidence intervals if computed. When discretizing X_S, δx is taken to be the same as for discretizing X_T in order to make discrete convolution of the two distributions easier. The empirical density function of the observed surges over the support X_S is denoted p_S.

3. Mixture model for the distribution of surges

3.1. Overview of the mixture model

With regard to the surge component, a mixture model is built:

- modeling of the bulk (frequently observed values) by an empirical (non-parametric) density function as described in Section 2.4;
- extrapolation of the upper and lower tails (high and low extreme values) by a parametric model based on Peaks-Over-Threshold (POT) declustering and Extreme Value Theory (EVT);
- connection of the bulk and tail distributions.

Many mixture models have been presented in the literature for both the central and extreme domains: Mendes and Lopes (2004) propose a model where the central domain is assumed to be normal while the tails are fitted by two GPDs; Behrens et al. (2004) also combine a parametric distribution (either Gamma, Weibull or Normal) for the bulk values with a GPD for the tail above a certain threshold; Tancredi et al. (2006) present a semi-parametric mixture model with piecewise uniform distributions from a low threshold up to a high threshold above which a Point Process model is fitted to the tail; MacDonald et al. (2011) propose a flexible extreme value mixture model combining a non-parametric kernel density estimator for the bulk domain with a tail model based on Bayesian inference.

However, these models do not directly account for the distinction between sequential (e.g. hourly) and auto-correlated observations and independent events. The present mixture model performs the statistical extrapolation on i.i.d. events. A transformation between events and sequential values is thus necessary.

The extrapolation is applied to both tails of the distribution, i.e. both highest and lowest values of the non-tidal residual. The methodology for the lower tail can easily be derived from the upper tail; hence, for the sake of simplicity, only the methodology for extreme positive surges will be presented.

Furthermore, Bernardara et al. (2014) presented a justification of the double-threshold approach of MH2011, establishing a clear distinction between the physical declustering (see Section 3.2) and the statistical optimization (see Section 3.3), while slightly changing the nomenclature of both thresholds. The notations of Bernardara et al. (2014) will be used here.

3.2. Physical declustering of extreme surge events

The extreme values of environmental variables generally occur in clusters, particularly when the time step of the time series is significantly smaller than the typical duration of the physical process that generates this variable. In the case of sea levels, and under temperate latitudes (for instance in the North-East Atlantic), hourly measured surges are generated by low-pressure areas, or storms, whose local effect typically lasts around 12–36 h. A consequence is that the successive sequential values are strongly auto-correlated and cannot be considered as independent. It is noteworthy that the clustering is accounted for in the RJPM (see Section 3.4).

Still, an extreme statistical analysis (within the simple univariate model) requires a sample of independent and identically distributed (i.i.d.) data. First, it will be assumed that the extreme surges are all generated by the same type of meteorological processes so that the data can be considered as identically distributed, or homogeneous. This assumption can quite easily be considered as valid in Western Europe, for instance, but a classical example of heterogeneity in coastal areas such as the East Coast of the US could be the presence of hurricane-generated surges in the time series. In such cases, heterogeneous populations should be considered separately.

Second, the independence condition will be fulfilled thanks to physical declustering. The most widely used method is the Peaks-Over-Threshold (POT) method. Following the vocabulary of Bernardara et al. (2014), clusters of sequential values exceeding a threshold, denoted u_p for *physical threshold*, are defined as *events* (i.e. storms). Temporal criteria are applied to ensure their independence: two events are considered independent if there is a time lag of at least 24 h between the end of an exceedance and the beginning of another. Short fluctuations below the threshold during a storm are also allowed. The events are then described by their maximum value, or peak.

Eventually, a sample of N_T i.i.d. peaks higher than u_p is built: $\left(P_1^S, \ldots, P_i^S, \ldots, P_{N_T}^S\right)$ where P^S is the random variable "surge peak". A simple guideline derived from our practical experience is that the physical threshold should be set so that the mean number of storms per year $\lambda_T = N_T/K$ is roughly between 5 and 10 (see MH2011).

3.3. Statistical extrapolation of extreme surge peaks

Once this i.i.d. sample is extracted from the time series, the extreme value theory (EVT) can be called on for the extrapolation. The distribution of the P_i^S is unknown. Let u_s be a *statistical threshold*, in contrast with the physical one. u_s is applied to the sample of surge peaks, so $u_s \geq u_p$. This defines the sample of the N peak exceedances over u_s: $Y_i^{P^S} = \left(P_i^S - u_s\right)_{|P_i^S > u_s}$. Thus Y^{P^S} is the random variable "surge peak excess over the (statistical) threshold".

Under the i.i.d. assumption, the EVT states (Pickands, 1975) that the asymptotic form (i.e. when u_s is high enough) of the distribution of the $Y_i^{P^S}$ tends toward the Generalized Pareto Distribution (GPD), whose cumulative distribution function is:

$$G_{Y^{P^S};k,\sigma}(y) = 1 - \left(1 + k\frac{y}{\sigma}\right)^{-\frac{1}{k}} \tag{2}$$

where k and σ are respectively the shape and scale parameters, with $y > 0$ and $y \leq -\sigma/k$ if $k < 0$.

The statistical threshold must now be determined. There are many methods: a first class is based on maximization of the goodness of fit between the probability distribution and the data or minimization of the asymptotic mean square error of the estimators (Bernardara et al., 2011); some are based on graphical visualization of certain theoretical properties of the GPD (see in particular Coles, 2001) while others aim at automatic threshold selection (Thompson et al., 2009). Each analyst

should feel free to use their favorite methodology; in this paper u_s will be determined by using the stability plots of the GPD shape and modified scale parameters.

As presented by Bernardara et al. (2014), this statistical threshold u_s should not be confused with the physical one u_p. The latter is applied to a highly auto-correlated time series of sequential values and requires physical considerations in order to identify independent events and build an i.i.d. sample of the event-describing variable, such as the storm peak. In particular, during this step, particular attention should be paid to the independence criteria: time lag between two events, maximum duration of fluctuations below the threshold during the event, minimum storm duration, etc. In contrast, the choice of the former aims at identifying the extreme domain of the peaks in a statistical sense and consequently requires genuine statistical optimization based on an i.i.d. sample. Furthermore, this optimization step is generally performed by testing many threshold values over a broad range; thus merging these two thresholds would have two consequences: first it would be highly inefficient from a computing standpoint (declustering of the entire time series would be performed at each threshold tested) and second it could yield different samples for the same final statistical threshold because the independence criteria may depend on the threshold value.

Once this extreme sample is built, a GPD can be fitted to the N exceedances $Y_i^{P^s}$ over u_s. In this study, the L-moments estimator defined by Hosking (1990) is adopted. It has been used to estimate GPD parameters for environmental variables such as wave heights (Goda, 2011) and wind velocities (Pandey et al., 2001).

The confidence intervals for the parameters and the quantiles of the estimated distribution are computed by the parametric bootstrap method (see for instance Efron and Tibshirani, 1993).

MH2011 proposed an extension of the GPD model to other distributions (namely the Weibull and Gamma distributions), arguing that it cannot be granted that the asymptotic domain of the EVT is reached. Such multi-distribution approaches are adopted by different authors (e.g. Goda, 2011; Hosking and Wallis, 1997). However, although the use of other distributions can very easily be included into this methodology, this paper will be restricted to the single use of the GPD.

Eventually, the distributions of the peak excesses $Y_i^{P^s}$ and, hence, of the extreme surge peaks over u_s, P_i^S, are obtained. They are characterized by the probability density functions $G_{Y^{P^s};\hat{k},\hat{\sigma}}$ or $G_{P^s;\hat{k},\hat{\sigma}}$; and the cumulative distribution functions $G_{Y^{P^s};\hat{k},\hat{\sigma}}$ or $G_{P^s;\hat{k},\hat{\sigma}}$, where $\hat{k}$ and $\hat{\sigma}$ are the parameter estimates. It should be noted that the density is null for surge peak values falling below the origin of the distribution u_s, while its sum for values above it is equal to 1. The limit between the empirical domain for the bulk values and the parametric domain for extreme values is thus defined by u_s.

3.4. Distribution of extreme surges

The next step is to transform the distribution of the extreme surge peaks P^S into the distribution of the extreme hourly surges S. In other words, the parametric law of all the hourly exceedances over u_s, is looked for. Reference may be made to Leadbetter (1983) and the use of the extremal index for local clustering of extreme values (see Eq. (4) in Tawn and Vassie, 1989):

$$\Pr\{\max(Y_1,\ldots,Y_N)<x\} = F^{n\theta_n(x)}(x) \tag{3}$$

with the finite version of θ_n introduced by Tawn and Vassie (1989). These authors highlighted that in an intuitive and simple approach, the extremal index θ may be interpreted as being the reciprocal of $d_S(s)$, the mean number of hourly surges per peak, i.e. the mean duration of surge events in hours. $d_S(s)$ can be estimated directly from the original time series of hourly surges (see below).

Let (S_i) be a sample of data whose unknown distribution is F_S, and $Y_i^S = S_i - u_s$, given $S_i > u_s$. Y^S is the random variable "exceedances of S over u_s, given S larger than u_s". The law of the exceedances of S over the threshold u_s, i.e. the probability of having $S > s = u_s + y$, is:

$$\Pr\{S>s|S>u_s\} = \Pr\{S>u_s+y|S>u_s\} = \Pr\left\{Y^S>y\right\} = \frac{1-F_S(u_s+y)}{1-F_S(u_s)}. \tag{4}$$

Knowing that $1-\Pr\left\{Y^S>y\right\} = \Pr\left\{Y^S\le y\right\} = G_{Y^S}(y)$, it is possible to write:

$$G_{Y^S}(y) = \frac{F_S(u_s+y)-F_S(u_s)}{1-F_S(u_s)}. \tag{5}$$

Thus, the first step consists in determining G_{Y^S}, the conditional law of hourly surge excesses above u. It can be derived from the conditional law of extreme surge peak excesses, $G_{Y^{P^S}}$, and from the mean number of hourly observations per peak, or per event, d_S.

For the sake of simplicity, we will work with the conditional variables S and P^S, and their conditional distributions G_S and G_{P^S}, given $S > u_s$ and $P^S > u_s$. Let $n_S(s)$ (resp. $n_{P^S}(s)$) be the number of hourly surges (resp. surge peaks) above s: $n_S(s) = n_{P^S}(s) \times d(s)$. Dividing both sides of this equation by: $n_S(u_s) = n_{P^S}(u_s) \times d(u_s)$, we have:

$$\frac{n_S(s)}{n_S(u_s)} = \frac{n_{P^S}(s)}{n_{P^S}(u_s)} \times \frac{d_S(s)}{d_S(u_s)}. \tag{6}$$

The ratio $n_S(s)/n_S(u_s)$ is taken as an estimate of the probability of exceedances of hourly surge values above s, given $s > u_s$, and thus as an estimate of $1 - G_S(s)$, while $n_{P^S}(s)/n_{P^S}(u_s)$ is taken as an estimate of $1-G_{P^S}(s)$. Furthermore, $d_S(u_s) = n_S(u_s)/n_{P^S}(u_s) = n_S(u_s)/N$. Thus Eq. (6) yields:

$$G_{S;\hat{k},\hat{\sigma}}(s) = 1-\frac{N}{n_S(u_s)}d_S(s)\left[1-G_{P^S;\hat{k},\hat{\sigma}}(s)\right]. \tag{7}$$

Now let us fit a parametric distribution to the observations of $d(s)$. Three distinct regimes can be identified. First, between $s = u_s$ and $s = s_1$, the values of $d_S(s)$ steadily decrease from $d_S(u_s)$ (typically between 5 and 15) to 1. Second, between $s = s_1$ and $s = \max(P^S)$, the number of observed hourly surges is the same as the number of observed peaks and $d_S(s) = 1$. Third, no more values are observed and $d_S(s)$ cannot be computed. When $\log[d_S(s)]$ is plotted against $\log(s)$, a linear trend appears for the first regime. Hence we propose to fit a linear regression model for the first regime in log–log scale while for the second and third regimes $d_S(s) = 1$.

We now seek to determine $F_S(s)$ for $s > u_s$, which is nothing else than $F_S(u_s + y)$. From Eq. (5):

$$F_S(u_s+y) = [1-F_S(u_s)] \times G_{Y^S}(y) + F_S(u_s). \tag{8}$$

$1 - F_S(u_s)$ is the probability that the surge observations exceed u_s. Intuitively, it can be seen that a very good approximation of this probability will be given by the ratio of the number of observations above u_s, $n_S(u_s)$, to the total number of observations νK:

$$1-F_S(u_s)\simeq\frac{n_S(u_s)}{\nu K}. \tag{9}$$

Thus, Eqs. (8) and (9) yield:

$$F_S(u_s+y) = 1+\frac{n_S(u_s)}{\nu K}\left[G_{Y^S}(y)-1\right]. \tag{10}$$

With respect to s, this gives:

$$F_S(s) = 1+\frac{n_S(u_s)}{\nu K}\left[G_S(s)-1\right], \quad \text{for } s>u_s. \tag{11}$$

Finally, Eqs. (7) and (11) yield:

$$F_S(s) = 1 + \frac{N}{\nu K} d(s)\left[G_{P^s}(s) - 1\right], \quad \text{for } s > u_s. \tag{12}$$

This eventually yields the cumulative distribution function F_S of the hourly surges of the extreme regime, that can easily be transformed into the corresponding probability density function f_S.

The approach used here is very similar to that of the RJPM, except that the index varies with s above the threshold. Also, it has been adapted to the law of threshold exceedance (Eq. (4)).

3.5. Connection of surge distributions of central and extreme regimes

The last step in determining the probability distribution of the surges is to connect the distributions for the bulk, or central regime (empirical distribution p_S of observed surges) and the extreme regime (estimated distribution f_S of extreme surges). The probability density function $\tilde{p}_S$ can thus be defined over the support X_S as follows:

$$\tilde{p}_S = \begin{cases} p_S \text{ for } s \le u_s \\ f_S \text{ for } s > u_s \end{cases}. \tag{13}$$

It is to be noted that if a parametric distribution is also fitted to the lower tail, it will also be connected to the bulk regime distribution below an appropriate threshold, quite similarly to the upper tail.

4. Distribution of sea levels

4.1. Distribution of sequential sea levels

The probability density function f_Z of the hourly sea level Z can now be computed by convoluting the empirical pdf p_T of the tidal levels and the final pdf $\tilde{p}_S$ of the hourly surges.

The cumulative distribution function F_Z can easily be derived from f_Z, as well as the return periods of hourly sea levels. Let T_r be any return period, in years. The relationships between T_r and the corresponding quantile $z\ (T_r)$, i.e. the hourly sea level of return period T_r, are the following:

$$\begin{cases} T_r(z) = \dfrac{1}{\nu\left[1 - F_Z(z)\right]} \\ z(T_r) = F_Z^{-1}\left(1 - \dfrac{1}{\nu T_r}\right) \end{cases}. \tag{14}$$

4.2. Distribution of sea level events

This last equation provides return periods associated with a given sea level value. However, extreme sea levels also occur within clusters, i.e. events, that may be storms or simply high tides. Confusion over the return period associated with an event should again be avoided, as required for coastal engineering applications.

The approach for obtaining return periods for extreme sea level peaks is simply the inverse of the transformation presented in Section 3.4., where the distribution of extreme surge peaks was to be converted into a distribution of extreme hourly surge values. Let P^Z be the random variable "sea level peak", and G_{P^Z} its (conditional) cumulative distribution function. The sea level peaks P_i^Z are defined with respect to a threshold u^Z, which is a physical threshold identifying extreme sea level events, similarly to u_p for storm surge events (see Section 3.2). Thus its value does not come from an optimization procedure and does not need to be accurate, but it must be physically meaningful. We propose using a level whose return period is about 0.1 to 1 year (i.e. 1 to 10 "sea level events" per year). We also define N^Z, the number of events exceeding u^Z, $\lambda^Z = N^Z/K$, the mean number of sea level events per year, $n_{Z(u^Z)}$, the number of hourly observations exceeding u^Z, and $d_Z(z)$, the mean number of hourly sea level observations per event above z, that is the reciprocal of the extremal index for the sea level (Leadbetter, 1983; Tawn and Vassie, 1989). Adapting Eq. (12) yields:

$$G_{P^Z}(z) = 1 + \frac{\nu K}{N^Z} \frac{1}{d_Z(z)} \left[F_Z(z) - 1\right], \quad \text{for } z > u^Z. \tag{15}$$

The fit of $d_Z(z)$ is similar to the fit of $d_S(s)$. However, it may happen that no clear trend can be seen in the first regime. In that case, a simple mean value could be substituted for the linear regression.

The return period of these events is:

$$T_r(z) = \frac{1}{\lambda^Z\left[1 - G_{P^Z}(z)\right]}. \tag{16}$$

Combining Eqs. (15) and (16) eventually yields:

$$\begin{cases} T_r(z)_{|\text{events}} = \dfrac{1}{\dfrac{\nu}{d_Z(z)}\left[1 - F_Z(z)\right]} \\ z(T_r)_{|\text{events}} = F_Z^{-1}\left(1 - \dfrac{1}{\dfrac{\nu}{d_Z(z)} T_r}\right) \end{cases}. \tag{17}$$

The extreme sea level peak associated with an event can then be computed for any return period T_r.

4.3. Confidence intervals for return levels

It would be useful to be able account for the uncertainty associated with the estimation of extreme surges when determining the probability distribution of sea levels.

The confidence interval for the quantiles of the extreme surge peak distribution is computed by parametric bootstrapping. The typical number of iterations is 10,000–100,000. In this method, a new set of parameters (k, σ, μ) is estimated at each iteration. A rigorous method would be to follow the methodology from Sections 3.4 to 4.2 for, say, 10,000 iterations. The result would be 10,000 slightly different sea level pdfs and, for each return period, a sample of 10,000 slightly different quantiles whose 5% and 95% percentiles would provide the 90% confidence interval. However, discrete convolution is very computer-intensive, and so is parametric bootstrapping. This solution, though quite rigorous, would be way too computer-intensive.

A good and much less demanding approximation consists in defining "equivalent GP distributions" for both the lower and upper bounds of the confidence interval for the quantiles. The three parameters of the equivalent GPDs are estimated by least square methods by simple interpolation between three pairs (quantile; return period). These equivalent distributions for extreme surge peaks can then be transformed into equivalent distributions for the extreme hourly surge values, then eventually convoluted with the tidal levels to provide a confidence interval for the extreme quantiles of the sea level. In this way only two additional convolution operations are performed instead of 10,000–100,000 for each tail.

5. Tide–surge interaction

5.1. Characterization of tide–surge dependence

Both the tide and the surge components behave as long waves during their propagation to the shore. So their celerity and amplitude at a definite location, say a tide gauge, depend on the total water depth: as a consequence, the value of one of the components influences the other and both interact. Generally speaking, larger surges are expected at low tide because they tend to rise in shallower waters. This interaction is in

contradiction with the convolution hypotheses, where independence is assumed. Furthermore, storm surges tend to occur during winter, while the largest tide amplitudes occur around the equinoxes. This is a seasonal non-stationarity.

A first step is to characterize this interaction so as to determine whether it has an influence at the study site. Different diagnostics have been proposed, among which the test proposed by DT94. First, the tidal range is split into five bands of equal probability. The surges are then separated into five samples of approximately the same size, depending on the band within which the concomitant tidal level falls. If there is no interaction, the number of surges per tidal band exceeding a high value η, say the 99.75% empirical quantile of the total surge distribution, should be close to the expected value $e = \nu K \times 0.0025/5$, with νK being the total number of surge observations. In reality, $n_{S,i}(\eta)$ surge values exceed η for each tidal band i, $i = 1\ldots5$. A test for interaction using a standard χ^2 test statistic can thus be proposed:

$$\chi^2 = \sum_{i=1}^{5} \frac{\left(n_{S,i}(\eta) - e\right)^2}{e}. \tag{18}$$

If this statistic exceeds the 95% significance level of the test $\chi^2_{4,0.95} = 9.49$, then it can be considered that tide and surge interact.

5.2. Accounting for tide–surge interaction

There are several possibilities for handling tide–surge interaction. They all have their advantages and drawbacks. An extensive study is beyond the scope of this paper, nonetheless some of them will be briefly presented here.

A first possibility is described in DT94 (see also Dixon and Tawn, 1999). The parametric upper tail of the surge distribution is modeled conditional on the tide. An appropriate transformation of the surge variable with the tidal level as a covariate yields a normalized surge series that is stationary with respect to the tide. Like the independence test presented above, this transformation implies splitting the tidal range into several equi-probable tidal bands, within which the empirical distribution of the associated surges is characterized in order to determine the coefficients for the normalization to be performed:

$$S_t^* = \frac{S_t - a(T_t)}{b(T_t)} \tag{19}$$

where $a(T_t)$ and $b(T_t)$ are linear functions of the empirical 98% and 99% quantiles for each tidal band. This methodology has been tested at many sites on the UK coast and is said to work adequately by the authors. Here we propose a different normalization: $a(T_t) = 0$ and $b(T_t)$ is the mean of the surge values exceeding the empirical 95% quantile of the corresponding tidal band. This normalization has been found to perform better at several sites on the French Atlantic/Channel coast, according to the χ^2 test statistic proposed by DT94.

An alternative approach consists in reversing the reasoning: instead of defining equi-probable tidal bands yielding sub-samples of the surge component, let us define bands of surge observations (e.g. based on the 0.05, 0.5, 0.9 and 0.99 surge quantiles) and yield sub-samples of the observed tidal levels. The tide distribution is then modeled by a non-parametric density estimator for each band. During the convolution operation, the tidal distribution to be associated with the surge mixture model will depend on the surge value. This method has its advantages, in particular for the mixture model presented in this paper, which is based upon the distinction event vs. sequential values. During an event, the tidal level typically varies between a low tide and a high tide; hence the sequential surge values during the same event are scattered in the different sub-samples with the approach of the tidal bands. However this approach also has its drawback. When the entire range of tidal levels is considered, the bounds of the support of the variable are known, i.e. LAT and HAT. Furthermore, when a complete lunar nodal cycle is available, the full distribution is known. So it can be expected that the empirical density estimation (e.g. by kernel density estimators) will be very accurate. In contrast, splitting the tidal levels into several sub-samples implies that the distribution and its bounds are not completely known (for instance it is expected that the tidal levels associated with the highest surges will not reach HAT level). Thus the estimation of the tide distribution is likely to be less accurate.

6. Case study: Brest

6.1. Presentation of the dataset

The model is illustrated with a dataset of sea level observations at Brest, France. Two approaches are tested: first with hourly sea levels, second with high tides and skew surges. At first tide and surge are assumed to be independent; then the tide–surge interaction is examined. The location of Brest is presented in Fig. 1. The exact coordinates of the station are 4.49503994°W; 48.38290024°N. The local datum is the *Zéro Hydrographique* (ZH), or Hydrographic Datum, defined in 1996 (0.5 m higher than the previous ZH).

A continuous time series of hourly sea level observations from 1953/01/01 to today is available from REFMAR (*Réseaux de référence des observations marégraphiques*, http://refmar.shom.fr/). In the present study, the time series lasts until 1992/12/31; i.e. 40 years. The precision of the data is 0.01 m. The eustatic trend is analyzed following the methodology presented by Bernardara et al. (2011). A positive trend of +1.48 mm/y is identified and removed from the analysis so as to get a stationary time series with a mean sea level (MSL) of +4.14 m ZH.

The hourly tidal levels have been computed with the SHOMAR software developed by the *Service Hydrographique de la Marine* (SHOM) over a saros period of 18.6 years (from 1963/01/01 to 1981/08/08) for the determination of the tide probability density function. High tides and skew surges were extracted following the methodology described in Bernardara et al. (2011).

6.2. Hourly sea levels — Tide–surge independence assumption

6.2.1. Empirical probability density function of the observations

The official Lowest and Highest Astronomical Tides at Brest are +0.25 m ZH and +7.93 m ZH. Over the 18.6-year long period considered, the minimum and maximum values predicted are +0.31 m ZH and +7.90 m ZH. $x_{T,\min}$ and $x_{T,\max}$ are set respectively at LAT and HAT, with $\delta x = 0.01$ m.

The minimum and maximum values of the surge time series are respectively −1.50 and 1.42 m. The latter value corresponds to the storm of October 15th–16th, 1987, which is known to have produced a much higher storm surge than anything ever recorded there. The second largest peak is 1.14 m. It can thus be considered that the highest possible surge value will not be much higher than 1.5 m. In order to account for the confidence interval to be computed later, $x_{S,\min}$ and $x_{S\max}$ are set respectively at −3 and 3 m. It is recalled that these bounded and discretized supports are used for computing the convolution of the density functions.

6.2.2. Mixture model for the surge distribution

The time series of hourly observed surges is physically declustered by using a physical threshold $u_p = 0.4$ m. A minimum event duration is set at 3 h and there must be at least 24 h between two consecutive events in order to assume their independence. The result is $N_T = 308$ storm peaks, i.e. $\lambda_T = N_T/K = 7.7$ events per year on average.

The statistical threshold u_s is determined by studying the stability of the GPD shape and modified scale parameters with respect to u_s (see Coles, 2001) with the guidelines of MH2011: first, domains of stability (ranges of u_s within which these parameters remain roughly constant) are identified; second, the lowest value (minimization of

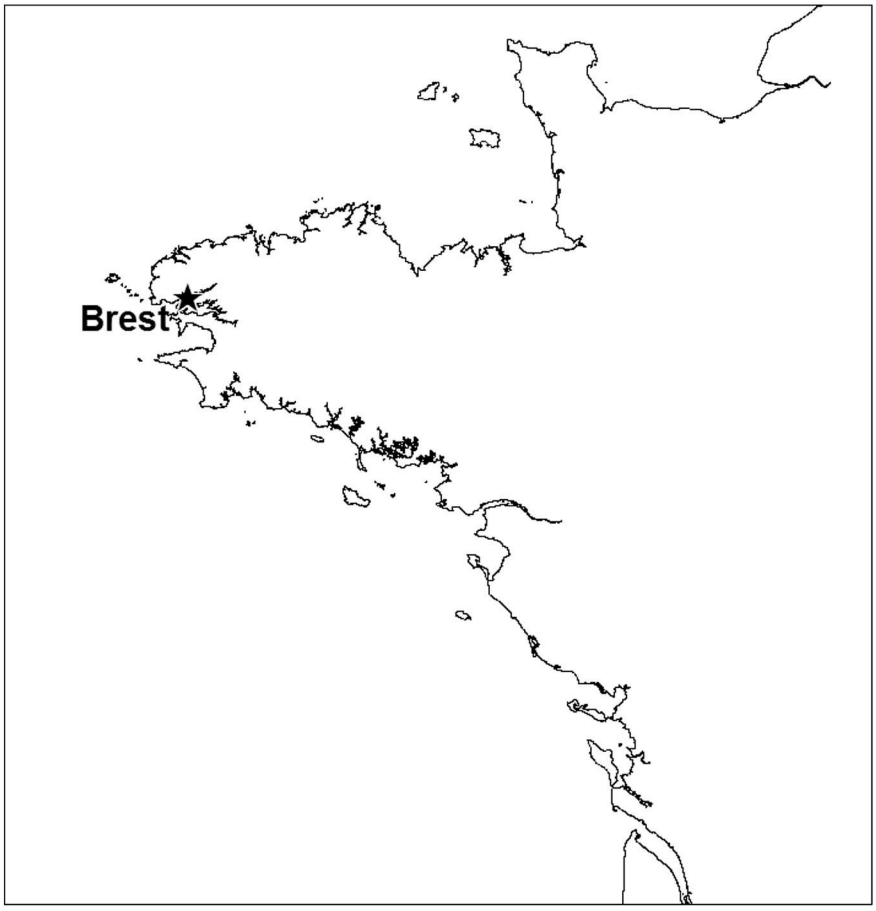

Fig. 1. Location of Brest.

variance) of the highest domain of stability (minimization of bias) is set at u_s. Because of the influence of the event of 1987 (which obviously is an outlier), no clear domains of stability can be identified; so this event is removed from the sample for this analysis. Two regions can be distinguished (Fig. 2): the first one from 0.4 to 0.6 m or so, and the second one above 0.67 m. In order to minimize the bias, the highest region is chosen, and in order to minimize the variance its lower value of 0.67 m is taken as the statistical threshold. Since the domains of stability are not particularly clear for this case, this choice was cross-checked with other diagnostic tools such as the mean excess life plot (Coles, 2001) and the L-moments plot (Hosking, 1990), which support this choice. Such a threshold yields $N = 45$ events, corresponding to a mean number of storms per year of $\lambda = N/K = 1.13$. This is slightly below the range 2 to 5 recommended by MH2011, but the period (40 years) is long enough to accept it.

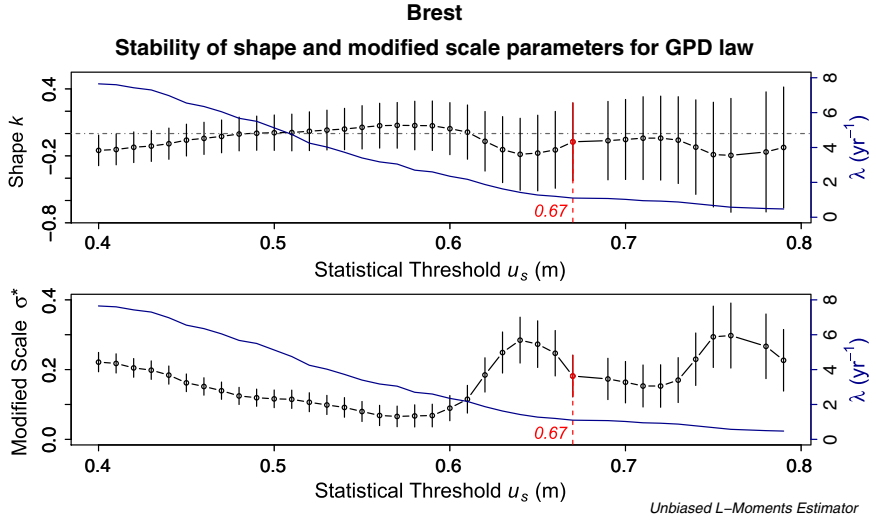

Fig. 2. Stability of the GPD shape and modified scale parameters with respect to the statistical threshold u_s. Secondary axis: mean number of surge peaks per year λ.

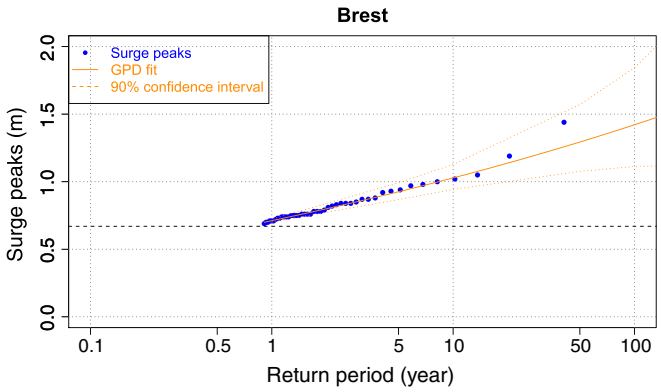

Fig. 3. Extrapolation of extreme surge peaks by a GPD.

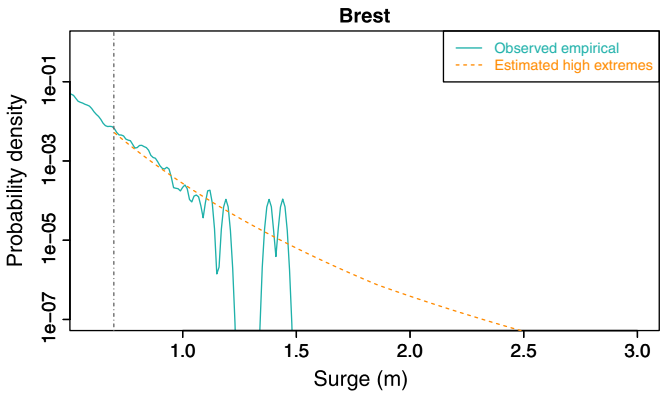

Fig. 5. Upper tail of hourly surge values pdf.

The GPD parameters are estimated by the unbiased L-moments estimator (Hosking, 1990). The fit with the 90% confidence intervals is illustrated in Fig. 3.

The extreme storm surge peak distribution is transformed into the extreme surge distribution as described in Section 3.4. Fig. 4 illustrates the modeling of the mean number of hourly surge values per event d_S while Fig. 5 illustrates the modeling of the upper tail of the hourly surge values by extrapolation in contrast with the empirical density.

The advantage of extrapolation is clearly visible in this figure: a non–nil probability is assigned to values above 1.5 m. In the empirical pdf, there is also a gap around 1.25 m: this is because the four highest values (whose corresponding "bumps", which were smoothed by the kernel density estimator, are quite visible in the figure) are 1.44, 1.38 (second highest hourly surge value of the October 1987 storm), 1.19 (November 10, 1963) and 1.12 m (October 1987 again).

The method of "equivalent distributions" described in Section 4.3 is used. The plot of the quantile function (with respect to the return period) for the 90% confidence interval is presented in Fig. 6. It shows good agreement with the discrete values computed for five return periods with the parametric bootstrap method for return periods up to 100 years.

6.2.3. Distribution of sea levels

The probability density functions of the hourly surges (both observed and extreme regimes), astronomical tide and sea levels are plotted in Fig. 7, with the associated 90% confidence intervals.

The effect of extrapolating both tails of the surge distribution is clearly visible on the lower and upper tails of the sea level distribution. The distribution of hourly sea level is converted into the distribution of sea level peaks, defined by the exceedance of the sea level threshold $u^Z = 7.7$ m. The change in mean number of observed hourly sea levels per event and its fit is given in Fig. 8. The sea level density is then to be transformed into return levels for event peaks with respect to return periods, as shown in Fig. 9.

The extreme sea level events for return periods of 1, 5, 10, 50, 100, 500 and 1000 years can then be derived, with their confidence intervals. They are summed up in Table 1.

Because the astronomical tide is predominant at Brest, the confidence intervals are reduced to a few centimeters for return periods of less than 50 years. It would be otherwise in locations where the stochastic component (surge) would be of greater relative importance.

6.3. High water levels — Tide–surge independence assumption

The analysis is performed with skew surges and high tide levels. Results are given in Table 2. Both approaches show good agreement (maximum difference of ± 0.03 m). The confidence interval width is very similar for return periods up to 500 years.

6.4. Tide–surge interaction

The dependence between tidal levels and meteorological surges can be assessed as described in Section 5.1. The value of the χ^2 statistic is found to be 42.39 for the hourly values dataset and 2.78 for the high water levels dataset. Hence the former dataset exhibits significant dependence while the latter can be considered as independent. This is not surprising: actually, the entire high water levels dataset is not very different from the sub-sample built from the upper tidal band for the hourly values.

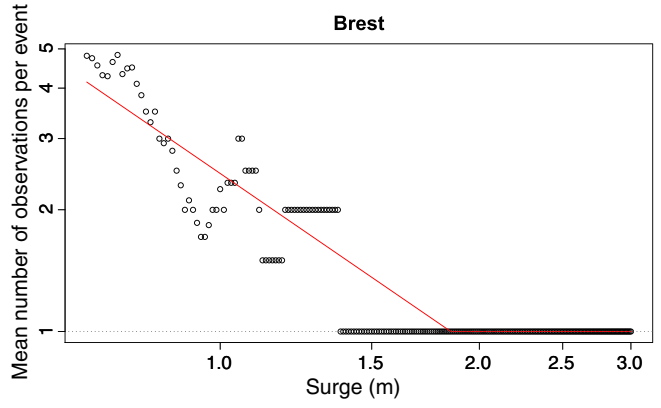

Fig. 4. Mean number of hourly values per event with respect to surge height: observations (circles), model (lines).

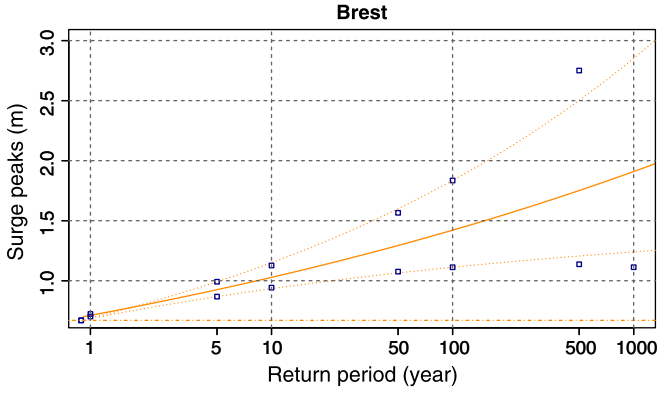

Fig. 6. Equivalent distributions for the 90% confidence interval of the surge peaks distribution. The squares represent discrete values as computed by the parametric bootstrap method.

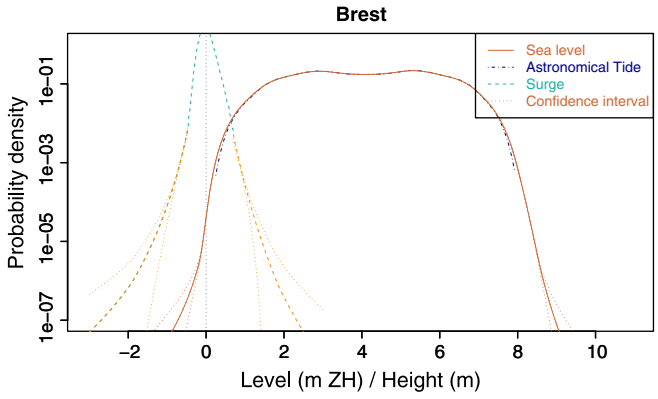

Fig. 7. Probability density functions of hourly surge, astronomical tide and sea level.

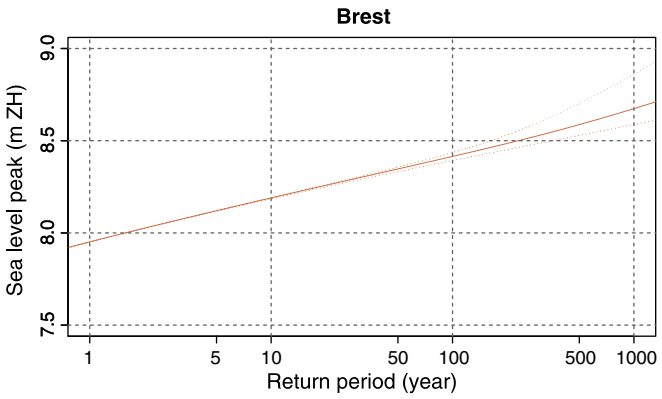

Fig. 9. Return levels for sea level events.

If the extrapolation of high water levels is considered to be sound and reliable, this would suggest that the tide–surge interaction has no significant effect on the estimation of extreme hourly sea levels, since both approaches converge. However, high water levels are somewhat higher than the maxima of hourly sea levels over a tidal cycle because they are the maximum value of a spline function interpolated between these hourly values. As a consequence, extreme hourly sea levels should be slightly lower than extreme high water levels. This means that the assumption of independence yields an overestimation of extreme sequential sea levels, which is a conservative approach for engineering projects aiming to protect coastal areas.

Nonetheless, the influence of tide–surge interactions was explored for the case of hourly values. A first approach involved splitting the dataset into five equi-probable tidal bands with the new surge normalization procedure described in Section 5.2. This procedure is found to perform better than the one proposed by DT94: the χ^2 statistic is reduced from 42.39 to 2.66, well below the 95% significance level of the test $\chi^2_{4,0.95} = 9.49$, while it is reduced to 5.35 with the method of DT94. This greater reduction of the statistic was also observed at other sites on the French coast of the Channel. Furthermore, the bounds of the stationary surge dataset are much more reduced: $-3.18 \leq S^* \leq 4.09$ vs. $-24.23 \leq S^* \leq 17$, which is much easier to handle from a practical point of view.

Statistical extrapolation for both tails is applied to the stationary dataset of normalized surges S^* (Sections 3.2 to 3.4). Then for each band (i.e. for each surge sub-dataset), the resulting tail distributions are transformed back to the original surge value S and connected to the empirical bulk distribution of the corresponding band. This yields five surge distributions: the surge distribution to be used during convolution with the tide distribution depends on the value of the tidal level.

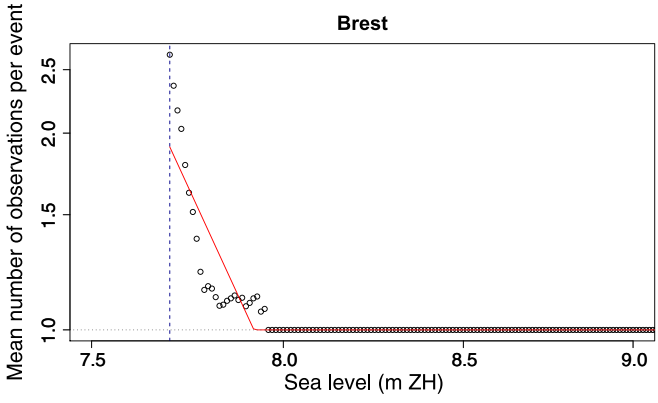

Fig. 8. Mean number of hourly values per event with respect to sea level: observations (circles), model (lines). Dashed line: sea level threshold u^Z.

A second approach involved splitting the dataset of surge observations into five bands defined by the 0.05, 0.5, 0.9 and 0.99 surge quantiles. The χ^2 statistic is now 12.96, which is far lower than 42.39 but still above $\chi^2_{4,0.95}$ (though just below the 99% confidence level $\chi^2_{4,0.99} = 13.28$). The empirical distribution of the tidal levels is estimated for each of these bands over the full period of observations (not just over a saros), with a nil density for values below the minimum or above the maximum tidal level of the corresponding band. This yields five tidal level distributions: the tide distribution to be used during the convolution procedure depends on the value of the surge.

Results for both approaches are given in Table 3 (confidence intervals were not computed). A slight decrease can be observed with the first approach (tidal bands): −0.07 m for a return period of 1000 years, −0.05 m for 100 years and −0.04 m for 10 years. However, it is likely that the dependence is not completely accounted for and that this estimation is still somewhat conservative. In contrast, the decrease is much more noticeable with the second approach. One could argue that the support of the tidal level distributions is too narrow; hence the probability of high tidal levels is too low and the sea level peaks would be under-estimated. It is also noteworthy that the estimates with both approaches fall below the lower bound of the 90% confidence interval of the results yielded by the tide–surge independence assumption: this is another strong argument in favor of modeling the tide–surge interaction for this case study.

6.5. Comparison with a direct approach

The extreme hourly sea level peaks were also estimated using a classical direct approach (POT declustering along with the Poisson-GPD model, as described in Sections 3.2 and 3.3 for surge peaks) for comparison purposes, though this approach is known to yield underestimated results (Haigh et al., 2010). Results are given in Table 4 and the extrapolation curve is compared with the different indirect approaches (with and without tide–surge interaction) in Fig. 10.

Table 1
Extreme sea level events at Brest as computed by the POT-JPM with hourly values assuming tide–surge independence.

Return period (years)	Hourly values — Independence assumption	
	Sea level peak (m ZH)	90% confidence interval
1000	+8.67	+8.59; +8.86
500	+8.59	+8.53; +8.70
100	+8.41	+8.39; +8.43
50	+8.35	+8.33; +8.35
10	+8.19	+8.18; +8.19
5	+8.12	+8.12; +8.12
1	+7.95	+7.95; +7.95

Table 2
Extreme sea level events at Brest as computed by the POT-JPM with high water values assuming tide–surge independence.

Return period (years)	High water — Independence assumption	
	Sea level peak (m ZH)	90% confidence interval
1000	+8.67	+8.58; +8.78
500	+8.58	+8.52; +8.68
100	+8.40	+8.37; +8.46
50	+8.32	+8.30; +8.37
10	+8.16	+8.15; +8.18
5	+8.09	+8.09; +8.10
1	+7.93	+7.93; +7.94

As expected, extreme sea level peaks are much lower in the case of return periods. However, the offset is already considerable with the 1-year sea level peak: around −0.1 m in comparison with the indirect approach. This is far from negligible, in particular for a 40-year long time series. It is likely that a kind of tide–surge interaction has not been properly accounted for. It may be that the two proposed treatments of tide–surge dependency do not deal correctly with the temporal dependence of seasonality. Future works should attempt to identify and characterize this hidden mechanism. Meanwhile, a somewhat rough and arbitrary correction, or calibration, could be made by reducing the results of the indirect approaches (with hourly values) by the value of this offset: −0.1 m for Table 1 and −0.08 m for Table 3.

7. Conclusions

This model aims to include the latest developments of POT declustering and of the Poisson-GPD model into the general framework of the RJPM. In particular, the mixture model for the surge distribution makes a clear distinction between auto-correlated values observed at a regular time step, and events, i.e. clusters of extreme values (Bernardara et al., 2014). Such a mixture model can also be applied to other variables, such as wave heights (Solari and Losada, 2012), when it is necessary to assign a probability density to the entire range of possible realizations of the variable. Also, this model can be used to determine the density function over the entire range of sea levels, which may be very useful for probabilistic design methods which require probabilities for the extreme

Table 3
Extreme sea level events at Brest as computed by the POT-JPM when accounting for tide–surge interaction.

Return period (years)	Equi-probable tidal bands	Surge bands
	Sea level peak (m ZH)	Sea level peak (m ZH)
1000	+8.60	+8.47
500	+8.52	+8.38
100	+8.36	+8.23
50	+8.30	+8.19
10	+8.15	+8.10
5	+8.09	+8.05
1	+7.93	+7.92

Table 4
Extreme sea level events at Brest as computed by the direct POT approach with hourly values.

Return period (years)	Sea level peak (m ZH)	90% confidence interval
1000	+8.19	+8.07; +8.37
500	+8.18	+8.07; +8.33
100	+8.14	+8.06; +8.23
50	+8.12	+8.05; +8.19
10	+8.04	+8.00; +8.08
5	+8.00	+7.96; +8.03
1	+7.85	+7.84; +7.86

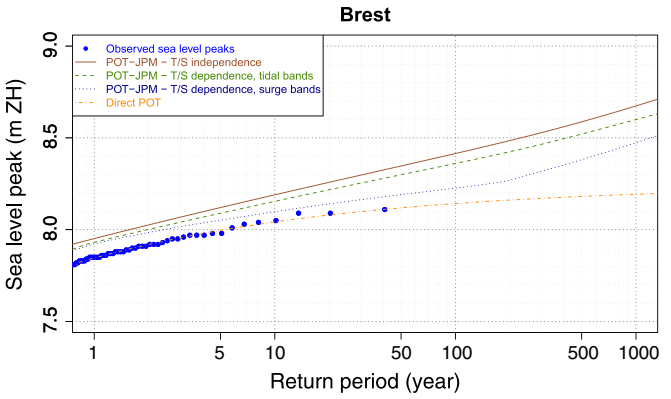

Fig. 10. Return levels for sea level events. From the upper to the lower curve: indirect approach without tide–surge interaction, indirect approach with equi-probable tidal bands, indirect approach with equi-probable surge bands, direct approach.

domain but also for frequent values (Solari and Losada, 2012). Furthermore, a good approximation of confidence intervals can be provided. This model was applied to sea level measurements from Brest, France. The 1-in-100-year hourly sea level peak is estimated to be +8.41 m ZH with a 90% confidence interval of [+8.39; +8.43]. An analysis based on high water levels yields very similar results with a 1-in-100-year level estimation of +8.40 m ZH with a 90% confidence interval of [+8.37; +8.46]. However, an analysis of different possibilities for handling tide–surge interaction suggests that these results are over-estimated in the case of hourly values. Furthermore, it appears from comparison with a direct analysis that this interaction is not fully accounted for: a notable offset remains within the interpolation domain, not to speak of the extrapolation domain. Future work will therefore focus on understanding this offset and the associated mechanism, which may be related to seasonality effects. Meanwhile, a simple correction can be considered.

Acknowledgments

The authors wish to thank Dr Pietro Bernardara, Jérôme Weiss, Michel Benoit and Marc Andreewsky from EDF/LNHE for their help and advice, as well as the reviewers for their helpful comments and suggestions that greatly contributed to the improvement of this paper.

References

Bardet, L., Duluc, C.M., Rebour, V., L'Her, J., 2011. Regional frequency analysis of extreme storm surges along the French coast. Nat. Hazards Earth Syst. Sci. 11, 1627–1639.
Behrens, C.N., Lopes, H.F., Gamerman, D., 2004. Bayesian analysis of extreme events with threshold estimation. Stat. Model. 4, 227–244.
Bernardara, P., Andreewsky, M., Benoit, M., 2011. Application of regional frequency analysis to the estimation of extreme storm surges. J. Geophys. Res. 116.
Bernardara, P., Mazas, F., Kergadallan, X., Hamm, L., 2014. A two-step framework for over-threshold modelling of environmental extremes. Nat. Hazards Earth Syst. Sci. 14, 635–647.
Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer, London.
Dixon, M.J., Tawn, J.A., 1994. Extreme sea-levels at the UK A-class sites: site-by-site analyses. Lancaster University and the Proudman Oceanographic Laboratory, Tech. Rep., March 1994.
Dixon, M.J., Tawn, J.A., 1995. Extreme sea-levels at the UK A-class sites: optimal site-by-site analyses and spatial analyses for the east coast. Lancaster University and the Proudman Oceanographic Laboratory, Tech. Rep., August 1995.
Dixon, M.J., Tawn, J.A., 1997. Spatial analyses for the UK coast. Lancaster University and the Proudman Oceanographic Laboratory, Tech. Rep., June 1997.
Dixon, M.J., Tawn, J.A., 1999. The effect of non-stationarity on extreme sea-level estimation. Appl. Stat. 48 (2), 135–151.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC.
Goda, Y., 2011. Plotting-position estimator for the L-moment method and quantile confidence interval for the GEV, GPA, and Weibull distributions applied for extreme wave analysis. Coast. Eng. J. 53 (2), 111–149.
Haigh, I., Nicholls, R., Wells, N., 2010. A comparison of the main methods for estimating probabilities of extreme still water levels. Coast. Eng. 57, 838–849.
Hosking, J.R.M., 1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. J. R. Stat. Soc. 52 (1), 105–124.

Hosking, J.R.M., Wallis, J.R., 1997. Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press, Cambridge, UK.
Leadbetter, M.R., 1983. Extremes and local dependence in stationary sequences. Probab. Theory Relat. 65, 291–306.
MacDonald, A., Scarrott, C.J., Lee, D., Darlow, B., Reale, M., Russell, G., 2011. A flexible extreme value mixture model. Comput. Stat. Data Anal. 55, 2137–2157.
Mazas, F., Hamm, L., 2011. A multi-distribution approach to pot methods for determining extreme wave heights. Coast. Eng. 58, 385–394.
Mendes, B., Lopes, H.F., 2004. Data driven estimates for mixtures. Comput. Stat. Data Anal. 47, 583–598.
Pandey, M., van Gelder, P., Vrijling, J., 2001. The estimation of extreme quantiles of wind velocity using L-moments in the Peaks-Over-Threshold approach. Struct. Saf. 23, 179–192.
Pickands, J., 1975. Statistical inference using extreme order statistics. Ann. Stat. 3 (1), 119–131.
Pugh, D.T., Vassie, J.M., 1979. Extreme sea levels from tide and surge probability. Proc. 16th Coast. Engng Conf., Hamburg, 1978, vol. 1. American Society of Civil Engineers, New York, pp. 911–930.
Solari, S., Losada, M.A., 2012. Unified distribution models for met-ocean variables: application to series of significant wave height. Coast. Eng. 68, 67–77.
Tancredi, A., Anderson, C.W., O'Hagan, A., 2006. Accounting for threshold uncertainty in extreme value estimation. Extremes 9, 87–106.
Tawn, J., 1992. Estimating probabilities of extreme sea-levels. Appl. Stat. 41 (1), 77–93.
Tawn, J.A., Vassie, J.M., 1989. Extreme sea levels: the joint probabilities method revisited and revised. Proc Instn Civ Engrs, Part 2, 87, pp. 429–442.
Thompson, P., Cai, Y., Reeve, D., Stander, J., 2009. Automated threshold selection methods for extreme wave analysis. Coast. Eng. 56, 1013–1021.

1.5. COASTAL ENGINEERING 2017: AN EVENT-BASED APPROACH FOR EXTREME JOINT PROBABILITIES OF WAVES AND SEA LEVELS

Coastal Engineering 122 (2017) 44–59

Contents lists available at ScienceDirect

Coastal Engineering

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

An event-based approach for extreme joint probabilities of waves and sea levels

Franck Mazas*, Luc Hamm

ARTELIA, 6 rue de Lorraine, 38130 Echirolles, France

ARTICLE INFO

Keywords:
Joint probabilities
Event
POT-JPM
Extreme-value copula
Upper tail dependence coefficient
Wave height
Sea level
JOIN-SEA
Chi-plot

ABSTRACT

A methodology for determining extreme joint probabilities of two metocean variables, in particular wave height and sea level, is presented in the paper. This methodology focuses in particular on the sampling of the time series, which should be based on the notion of event: either the event generating the variables whose joint probabilities are wanted (such as a storm generating waves and surges) or the event that is a result of the combination of these variables (such as a beach erosion event generated by waves at high sea level). A classification is proposed for multivariate analyses in order to help the choice of the sampling method. The dependence between the variables is analysed using tools such as the chi-plot, of which an enhanced presentation is proposed, then is modelled by extreme-value copulas (Gumbel-Hougaard, Galambos and Hüsler-Reiss) estimated by Canonical Maximum Likelihood or by the upper tail dependence coefficient. Joint return periods are then computed. A comparison is made with a simulation from the JOIN-SEA software on a dataset of wave height and sea levels offshore Brest, France. Then the bivariate methodology is extended to a multivariate framework. The distribution of sea level is determined by an indirect approach (extrapolation of extreme surges then convolution with the astronomical tide) and the dependence is analysed between the wave height and the surge component only. A bidimensional convolution between the joint distribution of wave height and surge and the distribution of the astronomical tide yields the joint distribution of wave height and sea level. The application of this method to the dataset of Brest and its comparison with the bivariate approach are finally discussed.

1. Introduction

The estimation of extreme environmental variables has been widely studied in the univariate cases. In the field of coastal engineering, the methodologies have been progressively enhanced and now provide reliable estimates of extreme wave heights, storm surges, wind speed..., provided the quantity of data is sufficient. Looking back, a gradual convergence towards the so-called “GPD-Poisson” model appears. Fitting a Fisher-Tippett distribution (now known as GEV distribution) to a sample of annual maxima was among the first popular methodologies. In the mid-1990s, the IAHR Working Group on Extreme Wave Analysis recommended using Peaks-Over-Threshold (POT) declustering along with a Weibull distribution estimated by maximum likelihood [50]. A few years later, the logics of the POT declustering was pushed a step further by fitting a Generalized Pareto Distribution (GPD) to the peak excesses, since the law of the exceedances over a threshold u asymptotically tends to a GPD when u is high enough [55]. The GPD-Poisson model is now widely recommended [36] for univariate extremes and a detailed description can be found in Coles [9].

Along with fellow researchers, the authors proposed several improvements within this general framework, based upon their daily practice of determining design waves and sea levels for coastal engineering projects. First, Mazas and Hamm [52] advocated a multi-distribution approach in order to deal with their experience that the GPD often has a tendency to under-estimate the quantiles such as 100-year H_s if u is not high enough. This approach considers additional distributions to the GPD (e.g. the Weibull, Exponential or Gamma distributions), following other authors (e.g. Gōda and Kudaka [27] or Van Vledder et al. [72]). Second, Mazas et al. [51] investigated the behaviour of the Maximum Likelihood Estimator (MLE) when u varies, concluding that the conditions for the MLE to be valid ([47], chapter 6) are not met and proposing to use the L-moments estimator [39] instead along with a location parameter μ (3-parameter GPD). Last, Bernardara et al. [2] provided a deeper justification of the recommendation by Mazas and Hamm [52] for a two-step framework for over-threshold modelling. The first step consists of event identification and characterization, based on physical considerations, in order to set up an i.i.d. (independent and identically distributed) sample from the time

* Corresponding author.

http://dx.doi.org/10.1016/j.coastaleng.2017.02.003
Received 24 May 2016; Received in revised form 30 January 2017; Accepted 5 February 2017

series. The POT declustering that selects peaks over a physical threshold u_p is the most frequent method. The second step is a statistical optimization aiming at determining the optimal threshold u_s above which the peak excesses should be fitted by a proper statistical distribution such as the GPD.

However, many coastal phenomena, such as overtopping, beach erosion, load on a structure, coastal flooding or others are the result of the combined actions of two physical processes or more such as sea level, waves, currents or winds. The most common cases are the combination of high sea levels and large wave heights and the joint occurrence of a storm surge along with a high astronomical tide level. Thus the univariate extreme models must be extended to a bivariate and more generally multivariate dimension.

However multivariate analyses may be quite different as regards sampling, dependence modelling, output... Generally speaking, three types of multivariate cases may be distinguished and we propose the following classification:

- Type A: a metocean process described by several parameters (e.g.: a sea state described by its significant wave height H_s, its peak period T_p, its peak direction θ_p, its directional spreading...);
- Type B: a metocean process that can be broken down in several elementary processes (e.g. a sea state made of a swell system and a wind sea system; or the sea level made of a mean sea level, the astronomical tide, the meteorological surge, the wave set-up...);
- Type C: the joint occurrence of several distinct metocean processes (e.g. waves, sea level, wind, current).

In the literature, the joint probabilities of tide and surge are among the soonest to have drawn deep attention, which is not surprising considering the stakes in safety associated with coastal flooding in macrotidal environments: this requires a Type B analysis. A methodology was built and progressively enhanced for determining extreme sea levels by what is called an indirect approach [31], i.e. a separate analysis of the deterministic astronomical tide component T and of the stochastic meteorological surge component S. Pugh and Vassie [57] introduced the Joint Probability Method (JPM), extended to the Revised Joint Probability Method (RJPM) by Tawn and Vassie [70] who introduced a GEV for fitting extreme surges and the extremal index. Tawn [69], Dixon and Tawn [16] then refined the RJPM by accounting for tide-surge dependence.

Indeed, the dependence that may exist between the different variables is a key issue of multivariate analysis. This dependence may be of different types: it may exist because large values of metocean processes such as waves, surges and winds are often generated together by a larger scale process (a storm), or because of interactions between two metocean processes: e.g. onshore wave height or surge magnitude depend on the water depth and thus on the water level.

The modelling of the dependence is at the core of the models built for Type C analyses: estimating the joint occurrence of two (or more) distinct metocean processes such as waves and sea levels. Tawn [67] was amongst the first to apply the parametric models for the dependence function discovered by Gumbel [30], particularly the logistic model, in the field of coastal engineering. The word “copula” was not present yet, though it was used by Sklar [62] as soon as the late 1950s, but here we have a mathematical object that links the distributions of the variables *via* their dependence. The logistic model is now widely known as the Gumbel-Hougaard copula.

In 1994, the UK Ministry of Agriculture, Fisheries and Food (MAFF, now DEFRA) started the funding of joint research projects at HR Wallingford and Lancaster University. A complete methodology for determining the joint probability of waves and sea levels was established ([10,32–34,54], among other references) and the software JOIN-SEA was developed and spread. Joe et al. [41], Zachary et al. [78] and De Haan and De Ronde [13] also applied these “dependence functions” for multivariate environmental extremes. Last, Heffernan and Tawn [37] proposed a semi-parametric approach which applies whether the variables are asymptotically dependent or asymptotically independent and is suitable for highly multidimensional analyses.

During the last decade or so, the copula functions spread into the field of coastal engineering and were used for analysing the combination of two or more variables: wave heights and periods [14]; storm surges and wind waves [15,73]; wave height, wave period and storm duration [12]; wave height, wave period, storm duration and storm surge [48].

Sampling and dependence are closely linked and here it is assumed that a sampling approach based on the notion of event is relevant for a thorough understanding of the physics. Callaghan et al. [5] highlight the interest of working with meteorological events, defined by the peak wave and sea level conditions but also by their duration and the spacing between two successive events, in order to understand the occurrence of beach erosion events. Li et al. [48] also consider events (storms) for the assessment of coastal flooding hazard in the Netherlands. Mazas et al. [53] proposed a “POT-JPM” approach for determining extreme sea levels (Type B analysis), distinguishing the distribution of events (surge events and sea level events) from the distribution of sequential values (e.g. hourly surges or high tide sea levels). The POT-JPM approach also provides confidence intervals for extreme sea levels (the uncertainty coming from the extrapolation of extreme surges) and can account for tide-surge dependence.

On February 2010, 28th, the storm Xynthia wreaked havoc on the French Atlantic coastline, causing more than 30 casualties because of coastal flooding in inhabited low areas and reminding coastal engineers the terrible effect of the combination of a storm surge occurring at the high water of a spring tide. But an in-depth analysis also shown the role played by the waves, that increased the ocean roughness and whose breaking added a set-up component [3]. As a consequence, an accurate estimation of such an event requires considering waves and sea levels on the one hand, and a separate analysis of storm surge and astronomical tide on the other hand.

The present paper is the result of the attempts by the authors to extend their event-based framework to the joint occurrence of two distinct metocean processes (Type C analysis), while applying an indirect approach for determining extreme sea levels. It should then be understood as a continuity of the work presented in the aforementioned references.

In Section 2, a bivariate methodology is presented, which can be applied for determining extreme joint probabilities of waves and sea levels (Type C), storm duration and peak H_s, storm peak H_s and T_p (Type A)... While following a classical four-step framework, the different stages are made consistent with previous publications and a particular focus is made on data sampling. Different methods are discussed, among which high tide sampling, bivariate threshold and the use of a univariate response function that allows accounting for covariates.

In Section 3, this methodology is extended to incorporate an indirect approach for determining extreme sea levels (separate analysis of tide and surge, then recombined by convolution). The main interest consists in the dependence that is modelled between the waves and the meteorological surge, instead of the total sea level.

In Section 4, the dataset of the case study of Brest, France is presented and applied first to the bivariate methodology with comparison with JOIN-SEA results, kindly provided by Dr Peter Hawkes (HR Wallingford), then to the multivariate methodology. The differences between the methodologies and the extension to other applications are discussed in Section 5.

2. Bivariate methodology

In the U.K., current advanced engineering practice largely relies on the approach implemented in the JOIN-SEA software [34]. Despite the practical difficulties in routinely using this software, which thus is not

to be regarded as a universal reference, the principles of this approach have largely spread in neighbouring countries. This methodology involves the following four steps:

1. Data selection/Sampling of the joint time series for extreme analysis.
2. Modelling of marginal distributions of Hs and Still Water Level (SWL).
3. Analysis and modelling of the dependence structure (Hs, SWL).
4. Computation of joint probabilities and curves of joint return periods.

The present study addresses alternatives to original choices implemented in JOIN-SEA.

2.1. Sampling

The sampling of the time series is a critical point of any extreme analysis, be it univariate or multivariate. The analyst usually deals with time series of the variables, observed (i.e. measured or modelled) at regular time steps: they can be called sequential values. Because the physical phenomena have a non-nil duration and a momentum, the sequential values are usually auto-correlated (see, among others, , Smith [63] or Walton [74]). The finer the time step and the longer the duration of the physical phenomenon, the more auto-correlated the sequential data. Still, the extreme value theory requires (at least in a first step) working with an i.i.d. sample. Homogeneous subsets of the time series must thus be identified (in order to consider i.d. data) then declustering selects independent data.

Bernardara et al. [2] detail a two-step framework in the univariate case. In a first step, considering the underlying physical (e.g. metocean) processes, the aim is to identify physical *events* such as a storm, a flood, a heatwave... within the time series of the sequential observations. Using exceedances over a physical threshold u_p is a practical way to perform this identification. The independence of the identified events is checked using criteria such as minimal duration between events or others. An event has a certain duration that is typically longer than the time step of the series: this is why threshold exceedances in a time series occur in clusters. The events are then characterized by an *event-describing random variable*, most frequently their peak value. This step results in the extraction of an i.i.d. sample from the auto-correlated time series. Note that the event-describing variable and the sequential observations are two different random variables (e.g. peak H_s vs. hourly H_s). In a second step, the Extreme Value Theory (EVT) is called upon for determining the statistical threshold u_s above which the exceedances of the event-describing variable may be modelled by a proper statistical distribution such as the GPD.

In the multivariate case, setting up an i.i.d. multivariate sample from the time series may be the first source for headaches to the analyst. A first trick is that it is not necessary that all the variables be extreme: average wave heights occurring at an extreme sea level may cause coastal flooding, for instance. As a consequence, the frequently observed values (the bulk of the distribution) will have to be included in the analysis which seems contradictory with the event-based, peaks-over-threshold approach.

In order to overcome this contradiction, it is first necessary to be aware that bivariate or multivariate analyses may be relevant for quite different cases.

Sampling for Type A cases can be reduced to the univariate case when a single parameter is enough for identifying the event. For instance, it can be interesting for designing breakwaters to complement the knowledge of extreme wave heights by the joint probabilities H_s/T_p or H_s/storm duration. In that case, the storm is defined by the exceedances of H_s over the physical threshold u_p and the i.i.d. sample is made of the peak H_s, the associated T_p and the duration of the exceedances of H_s over u_p. Note that unlike H_s and T_p, storm duration is an event-describing variable that has no sequential equivalent: this highlights the fact that sequential observations and event-describing variables are indeed different random variables. When such a simple sampling is irrelevant, Type A cases become similar to Type C cases (see below).

When considering the most common of Type B analyses, i.e. joint occurrence of tide and surge, it appears that, because the variables are considered separately before being recombined by convolution, there again sampling can be reduced to the univariate case and a surge event is easily distinguished from a sea level event. The "POT-JPM" approach proposed by Mazas et al. [53] clearly distinguishes the distribution of events (surge events and sea level events) from the distribution of sequential values (e.g. hourly surges or high tide sea levels). The link between the distribution of peak and sequential values is provided by the extremal index θ (interpreted as the reciprocal of the mean duration of an event), introduced by Leadbetter [45] and incorporated by Tawn and Vassie [70] into the RJPM.

Though an analysis of the physics of the phenomena and a Peaks-Over-Threshold approach provides a very practical framework in the univariate case, it is much more difficult in the multivariate case of Type C. A combination of high level and moderate waves may cause overtopping, for instance: it is thus necessary to select data for both variables representative of both usual and rare conditions, while granting the independence of data. In some cases, the analysis may be even more delicate: Callaghan et al. [5] show the importance of event duration and duration between events when considering beach erosion. Hence the conceptual approach of threshold exceedances must be adapted. This is illustrated by the scatterplot H_s/Z (sea level) offshore Brest in Fig. 1. Should we select independent pairs (H_s,Z) above a threshold relative to a single variable (domains 1+2 or 2+3), or when both variables exceed their own threshold (domain 2)? Or should we be a little more sophisticated?

A first approach in the literature is specific to the joint occurrence of waves and sea levels, especially in tide-dominant areas. It is based on the postulate that for coastal structures, flooding or overtopping will only occur at high water. Consequently, the high-tide sampling is quite straightforward: the sea level at high tide (or closest to) is selected along with the associated H_s. Thus only one record per tidal cycle is extracted (i.e. nearly 2 per day). It is thus assumed that the lag between two successive high waters is long enough to consider that two successive records are independent. It seems acceptable for areas where the tide is diurnal (24 h 50 mn between two high waters); it is probably a rather strong assumption for semi-diurnal areas (12 h 25 mn): Walton [74] found that the autocorrelation function of a hourly surge series measured at Sandy Hook, NJ, dropped to a noise level close to zero after 24–72 h. Nonetheless, this method is used in JOIN-SEA [54] (along with a recommendation to separate the data in two or three separate populations if needed for dealing with different sectors, seasons, wave systems...) and is also recommended in France: CETMEF [8], among other examples. However, the selection of H_s is not necessarily straightforward with this sampling: [44] highlights the difficulties to examine and proposes practical solutions. If sea states are available offshore and sea levels onshore (e.g. at a tide gauge), the wave propagation time should be accounted for. Waves may also be spatially variable, and / or subject to uncertainty if the data is the output of a numerical model. Therefore it may be safe to select the maximum H_s value within a time window centred on the time of high water, +/− 3 h for instance. This sampling is akin both to POT and block maxima approaches: on the one hand, it takes the maximum in a constant time period (12 h 25), on the other hand, high water level can be seen as the peak level over a threshold applied to this variable (domains 2+3 on Fig. 1).

Another approach, more directly linked to the POT declustering framework, consists of using a multivariate threshold. Li et al. [48] applied a bivariate threshold for selecting 4-uplets of H_s, T_p, sea level Z and storm duration. An event (i.e. a storm) is defined when both H_s and the tidal anomaly TA (the sea level minus the astronomical tide level, i.e. the meteorological surge) exceed a threshold value. To guarantee

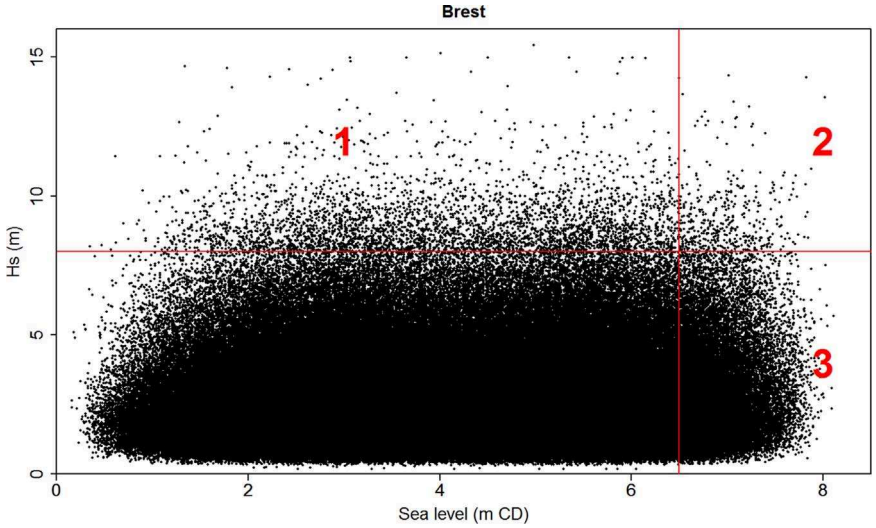

Fig. 1. scatterplot H_s/Z offshore Brest, France and associated possible thresholds.

the independency of selected storms, a minimum time interval between two events is set. The following parameters are selected for setting up the i.i.d. sample: the maximum H_s value over the storm, the associated T_p value, the maximum sea level over the storm and the event duration (including possible fluctuations). This method is illustrated in Fig. 2 that comes from Li et al. [48]. Of course, the difficulty of the method is choosing the values for the thresholds. Following the recommendations of Bernardara et al. [2] for the univariate case and transposing them to the multivariate case, we can consider these thresholds as multivariate physical thresholds (e.g. $\{u_p^{H_s}; u_p^{TA}\}$) and set their values based upon physical considerations. Li et al. [48] do not provide specific information about the size of the selected sample, but the number of storms per month seems to be around 3 in average. This sampling has the advantage of being event-based, a framework that allows for instance to work on storm duration and to set the parameters granting independence the most adapted to the site and to the physical phenomenon: see also Sanchez-Arcilla et al. [58] or Lin-Ye et al. [49].

We propose here to test a new sampling method. In contrast to the bivariate threshold method presented above that focuses on the events of the input variables, we consider the effect of the combination of variables while reducing them into a univariate response variable. The notion of event is shifted from the input to the output. This function can be overtopping over a breakwater, efforts on a seawall, beach erosion, etc.... A classical POT declustering can then be applied and the peak of the univariate response function gives the date and time of the event: the i.i.d. sample is made of the n-uplets of the variables corresponding to these timestamps. The main advantage of this sampling is that covariates can be accounted for. Let us consider the following case: we want to estimate the joint probabilities of wave height H_s and sea level Z for assessing coastal flooding hazard. Sea states data are available offshore and wave direction and periods are very variable. If we only consider offshore wave height, the analysis will be influenced by waves that propagate far from the location of the study. Wave refraction (that depends on wave direction θ and period T_p) can be accounted for by using the following univariate response function (among others): $f(Z,H_s,\theta,T_p)=Z+H_s'=Z+K_r(T_p,\theta,\ldots)K_s(T_p,\theta,\ldots)H_s$, where H_s' is the nearshore (refracted) wave height, K_r the refraction coefficient and K_s the shoaling coefficient. Gōda [28] provides analytical expressions of K_r and K_s (wave breaking is also accounted for). Fig. 3 illustrates the time series of this univariate response variable with a threshold set at 12 m and the selected peaks. The number of peaks is much lower than the number of high waters (in this example, around 18 per year vs 706 high waters per year). Thus the analysis will only consider independent conditions that have a significant effect at the coastline because they combine large nearshore wave heights with high sea levels. This "total water level" $Z+H_s$ was used by Hawkes et al. [35] but as a structure function (output of the analysis), and thus without including the effect of wave covariates.

Generally speaking, the sampling method is to be adapted to the aim of the study and to the available data. The simple univariate POT sampling can be applied for multivariate analyses of Type A, e.g. aiming

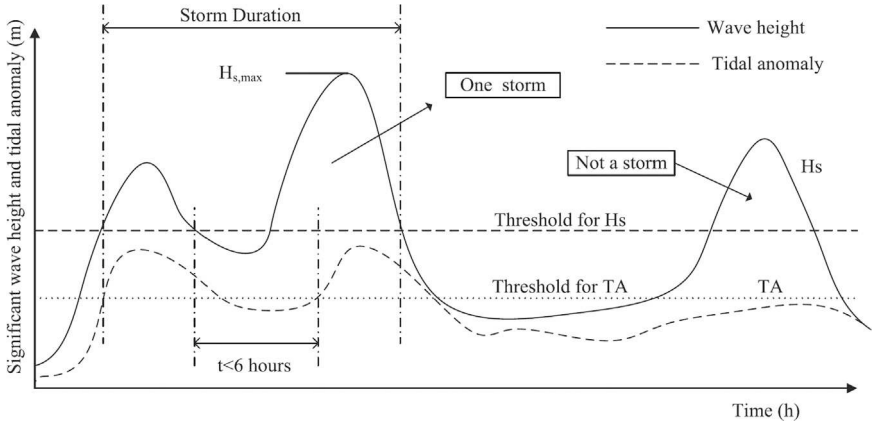

Fig. 2. Definition of independent storm events by a bivariate threshold, after Li et al. [48].

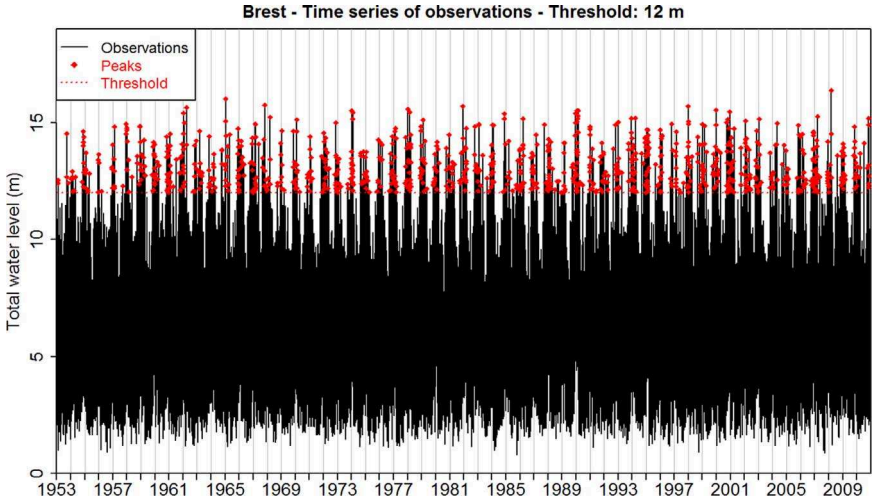

Fig. 3. Time series of a univariate response function (sum of sea level and nearshore wave height) with threshold and peak of the selected events.

at determining joint probabilities H_s/T_p, $H_s/Duration$, etc. For Type C analyses, the bivariate/multivariate threshold method is based upon physical considerations and greatly reduces the i.i.d sample compared with the high water sampling. The univariate response function method may be more appropriate when covariates play an important role and should be accounted for. Therefore, we will not recommend any of these methods; nevertheless the case study will consider the univariate response function as it is a newly proposed method.

The sampling step results in an i.i.d. multivariate sample of size N_p. $\lambda_p = N_p/K$ is the mean number of selected events per year, where K is the duration of the time series, in years. The size of this sample is matter to discussion. In the univariate case, Bernardara et al. [2] suggested that a value of λ_p between 5 and 10, in order to have a sample size N_p above 100, depending on the duration of time series (typically 20–25 years in engineering applications). In the multivariate case, it can be argued that more information is necessary in order to capture the dependence between variables. Therefore it is suggested to increase the sample size, with a value of λ_p in the range 15–25, to be adapted to the duration of the available time series, to the site and to the physical processes under study.

2.2. Marginal distributions

In contrast with univariate extreme analyses, multivariate studies must consider frequent values of the marginal variables. Thus the marginal distributions must cover both the frequent domain (bulk of the distribution) and the extreme domain (tail of the distribution). A logical solution consists of using a mixture distribution: an extreme value analysis is carried out for modelling the tail of the distribution while the bulk is modelled by the empirical (sample) distribution. This mixture model has been used, among others, by Hawkes et al. [34], Li et al. [48], Solari and Losada [64,65], Mazas et al. [51] and Kergadallan [44].

More specifically, for each variable X, a statistical threshold u_s is set above which the N exceedances are modelled by a Generalized Pareto Distribution (GPD). The value of this threshold can be determined using one or several of the methods reviewed by Bernardara et al. [2] in the second step of their framework. A multi-distribution approach may be adopted [52] in order to check whether other distributions (e.g. Weibull distribution) provide a better fit than the GPD.

The conditional distribution of the tail (distribution of X given $X>u_s$) is then weighed by $\zeta_{u_s}=\mathbb{P}[X\leq u_s]$, estimated by the empirical ratio of the number of observations below u_s over the sample size, and connected to the sample distribution used for $X<u_s$.

2.3. Dependence structure

2.3.1. Dependence coefficients

Several methods are available for analysing the dependence structure between two random variables X and Y. The most simple and obvious one is the scatterplot of the bivariate sample. A shapeless cloud will suggest independence, while an ovoid shape oriented towards the upper-right corner is typical of positive dependence.

This visual interpretation of the scatterplot may be complemented by the usual correlation coefficients. Pearson's r correlation coefficient describes the relationship between linear related variables; Spearman's ρ coefficient provides information about the possibility to model the dependence by a monotone function; Kendall's τ coefficient describes the correlation between the samples of the ranks of the variables. These coefficients tend to zero in the independence case (see Genest and Favre [26] for examples and interpretation of these coefficients).

For an extreme study, it is desirable to focus the analysis of the dependence on the extreme values of the sample. The usual correlation coefficients are less adapted to capture the asymptotic dependency and other measures should be used. The Chi-Plot $(\Lambda_i, \mathrm{X}_i)$ introduced by Fisher and Switzer [17] is a graphical tool for detecting dependence. X and Λ are data transforms defined as a combination of the sample bivariate and marginal distribution functions. Λ_i measures the distance of the pair (X_i, Y_i) from the pair of the medians $(\overline{X}, \overline{Y})$: a positive (resp. negative) value on the x-axis of the chi-plot means that both X_i and Y_i are on the same (resp. opposite) side of their respective medians, and a value close to 1 (resp. 0) means that are large or small relative to (resp. close to) their respective medians. X_i measures the dependence: a positive (resp. negative) value on the y-axis describes a positive (resp. negative) dependence, while a value close to zero suggests independence. Thus, it can be seen whether the dependence varies from the mean values to the tail regions. Illustrative examples of the chi-plot for samples with various types of independence are presented in Fisher and Switzer [17] and Fisher and Switzer [18]. However the plot does not make the distinction between pairs with both large values and those with both small values: this is discussed in the case study presented in Section 4.3.

The upper tail dependence coefficient λ_U was introduced by Sibuya [61]. It is defined by Joe [42] as follows:

$$\lambda_U = \lim_{t\to 1^-} \mathbb{P}[F_X(x) > t \mid F_Y(y) > t] \tag{1}$$

Eq. (1) describes the limit probability that a variable exceeds a threshold when the other does the same. λ_U approaches 0 for independence and 1 for complete dependence. It provides the analyst

with a coefficient dedicated to the area of interest, which will be linked to the copulas used for modelling the dependence structure (see below). Different methods are available for estimating λ_U from the sample. Poulin et al. [56] recommend using the non-parametric CFG estimator (especially in the case of extreme-value copulas). This approach, based on the empirical copula, has the benefit of avoiding any assumption on the copula and the marginal distribution and as such is much more general. The CFG (Capéraà-Fougères-Genest) estimator has been proposed by Frahm et al. [19], based on the work of Capéraà et al. [6]. It is defined by:

$$\hat{\lambda}_U^{CFG} = 2 - 2\exp\left\{\frac{1}{N}\sum_{i=1}^{N}\log\left[\sqrt{\log\frac{1}{U_i}\log\frac{1}{V_i}}\Big/\log\frac{1}{\max(U_i, V_i)^2}\right]\right\} \quad (2)$$

where (U_i, V_i) is the sample of the normalized ranks.

Another popular notation of λ_U is χ, following Coles et al. [11]. The authors also proposed a second dependence measure $\overline{\chi}$, sometimes called weak tail dependence coefficient, in order to provide information on the relative strength of dependence in the tails for asymptotically independent models, for which λ_U=0:

$$\overline{\chi} = \lim_{t\to 1^-}\frac{2\ln\mathbb{P}[U > t]}{\ln\mathbb{P}[U > t, V > t]} - 1 \quad (3)$$

$\overline{\chi}$ takes values in the range $[-1,1]$. Ledford and Tawn [46] also proposed the index of tail dependence κ.

2.3.2. Copulas

According to Sklar's theorem [62], the joint cumulative distribution function $H_{X,Y}(x, y)$=$\mathbb{P}[X \le x, Y \le y]$ can be linked to the marginal distributions of X and Y, F_X and F_Y, via a copula C:

$$H_{X,Y}(x, y) = C(F_X(x), F_Y(y)) \quad (4)$$

This result can be extended to the multivariate case.

Gudendorf and Segers [29] provide a sound theoretical justification for using extreme-value copulas in the domain of extreme value multivariate analysis. In particular, a key advantage of the extreme value copulas with respect to the popular class of the Archimedean copulas is that they are not symmetric (note that the class of the Archimax copulas contains both the Archimedean copulas and the EV copulas as a special case [7]. Parametric extreme-value copulas include the following distributions (among others): the Gumbel-Hougaard or logistic copula, the Galambos or negative logistic copula and the Hüsler-Reiss copula.

The distribution function of the Gumbel-Hougaard or logistic copula is given by (in the bivariate case):

$$C_\theta(u, v) = \exp\left\{-((-\log u)^\theta + (-\log v)^\theta)^{\frac{1}{\theta}}\right\} \quad (5)$$

where $(u, v) \in [0,1]^2$ and $\theta \ge 1$ is a parameter measuring the degree of dependence, ranging from independence ($\theta = 1$) to complete dependence ($\theta = \infty$). It allows positive dependence only (strong dependence for the right tail and weak dependence for the left tail). It is the only copula that is both Archimedean and extreme-value. Tawn [68] introduced a bivariate asymmetric logistic model, adding further flexibility to the basic logistic model.

The distribution function of the negative logistic or Galambos copula is given by (in the bivariate case):

$$C_\theta(u, v) = u \times v \times \exp\left\{((-\log u)^{-\theta} + (-\log v)^{-\theta})^{-\frac{1}{\theta}}\right\} \quad (6)$$

$\theta > 0$ approaches zero in the independent case and tends to infinity for complete dependence. It has been introduced by Galambos [20].

The distribution function of the Hüsler-Reiss copula is given by (in the bivariate case):

$$C_\theta(u, v) = \exp\left\{-\tilde{u}\Phi\left(\frac{1}{\theta} + \frac{\theta}{2}\log\left(\frac{\tilde{u}}{\tilde{v}}\right)\right) - \tilde{v}\Phi\left(\frac{1}{\theta} + \frac{\theta}{2}\log\left(\frac{\tilde{v}}{\tilde{u}}\right)\right)\right\} \quad (7)$$

where $\tilde{u}$=$-\log u$, $\tilde{v}$=$-\log v$, $\Phi()$ is the standard normal distribution function and the dependence parameter θ ranges from 0 (independence) to infinity (complete dependence).

These three copulas are fully described by a single parameter: the dependence coefficient θ (or association parameter). It can be analytically linked to the upper tail dependence coefficient as follows:

$$\lambda_U = \begin{cases} 2 - 2^{\frac{1}{\theta}} & \text{(Gumbel-Hougaard copula)} \\ 2^{-\frac{1}{\theta}} & \text{(Galambos copula)} \\ 2 - 2\Phi(\frac{1}{\theta}) & \text{(Hüsler-Reiss copula)} \end{cases} \quad (8)$$

Other extreme value copulas are available, such as the Tawn copula, the t-EV copula (**[29] The present study is limited to the three copulas detailed above but any extreme value copula could be used in the proposed methodology.

Asymptotically independent models are sometimes advocated (e.g. Coles et al. [11], Hawkes et al. [32], among others). The bivariate normal copula, or Gauss copula, is one of the most widely used models for this class. It is in particular used in the JOIN-SEA software [34]. However the multivariate methodology presented in Section 3 below aims at modelling the dependence structure between wave height and surges, two variables for which the underlying physics provide argument for asymptotic dependence in the case of mid-latitude storms. Indeed, although large wave heights can be observed without significant surges, a very extreme storm generating extreme wave heights will tend to induce large storm surges as well. Therefore the methodology will be applied to extreme value copulas only.

2.3.3. Estimation

Similarly to the univariate case, many estimators are available for copula estimation, i.e. for estimating the dependence parameter θ. These methods can be parametric, non-parametric or semi-parametric. The empirical copula, the method of moments and the Maximum Likelihood Estimator (provided the density of the copula can be differentiated) are available, among others.

In particular, maximum likelihood methods can be carried out in different ways, depending on the model used for the marginal distributions and on the confidence in the estimation of the parameters of these distributions. The classical maximum likelihood estimator groups the parameter θ and the parameters of the marginal distributions in a single vector to be estimated by maximisation of the complete likelihood function. The IFM (Inference From Margins) method [60] first estimates separately the marginal parameters by MLE, whose estimates are inserted into the copula likelihood: the dependence parameter can then be easily determined by simple maximisation. In order to avoid possible misspecification of marginal distributions, the Canonical Maximum Likelihood (CML) method, also known as Maximum Pseudo-Likelihood, estimates these by their (rescaled) empirical distribution functions then the dependence parameter is estimated by maximising the (pseudo-) likelihood [24].

A non-parametric method is also provided by the CFG estimator [6]. It is also possible to estimate the dependence parameter from Spearman's ρ and Kendall's τ coefficients, for which an analytical relationship with θ can be written [26].

Estimating θ from Kendall and Spearman dependence coefficients makes naturally arise the idea of using the upper tail dependence coefficient λ_U for the estimation of extreme value copulas. Such a fit may be less performing on the bulk of the data but may be expected to catch better the dependence in the extreme region. It is also very easy to compute when there is an analytical expression between λ_U and the copula parameters, as is the case for the extreme-value copulas:

$$\theta=\begin{cases}\frac{\log 2}{\log(2-\lambda_U)} & \text{(Gumbel-Hougaard copula)}\\ -\frac{\log 2}{\log \lambda_U} & \text{(Galambos copula)}\\ 1/\Phi^{-1}(1-\frac{\lambda_U}{2}) & \text{(Hüsler-Reiss copula)}\end{cases} \tag{9}$$

However, the literature is very poor about this possibility. While this paper does not intend to make a qualitative comparison of estimators (which would depend on sample size, copulas, marginal, type of dependence...), it is proposed to compare this estimator with another that uses a different approach, well documented in the literature, such as the CML estimator. Because the former focuses on the upper limit of the distribution in contrast with the latter that optimizes the fit over the entire distribution, such a comparison is valuable (see also the discussion section).

2.3.4. Goodness-of-fit and choice of the copula

When several copulas are tested against a sample, here the Gumbel-Hougaard, Galambos and Hüsler-Reiss copulas, it is necessary to assess the best-fitting one, thanks to goodness-of-fit tools. Genest and Favre [26] provide a wide review of available methods.

First, graphical diagnostics include the comparison of the copula fits with the empirical copula [66]; or generating large samples for comparing them with the base data.

Second, formal tests of goodness-of-fit may be used. Inspired by Genest and Rivest [23], Wang and Wells [75] proposed to compute a Cramér-von Mises statistic of the form:

$$S_n = \sum_{i=1}^{N} [C_n(U_i, V_i) - C_{\hat{\theta}}(U_i, V_i)]^2 \tag{10}$$

where N is the sample size, (U_i, V_i) is the sample of the normalized ranks, C_n is the empirical copula, $C_{\hat{\theta}}$ is the parametric copula. However, the p-values for the statistic cannot be computed and the authors recommended using the model yielding the lowest statistic. This difficulty was overcome by Genest et al. [25] who proposed a variant of the statistic with a bootstrap procedure allowing the computation of p-values. For the sake of simplicity, this procedure is not implemented in this study and only the statistic is examined.

When the copula parameter is not estimated by the upper tail dependence coefficient, it is also possible to compare the values of λ_U yielded by the fits (based on $\hat{\theta}$) with the sample λ_U: the closer the values, the better the fit.

In the following, the choice of the copula will be based on the Cramér-Von Mises (CVM) statistic.

2.4. Joint return period

The joint return period T of X and Y associated to the event ($X > x$ and $Y > y$) is given in the context of annual maxima widely used in hydrology by Yue and Rasmussen [77], and Poulin et al. [56]:

$$T(x, y) = \frac{1}{\mathbb{P}[X > x, Y > y]} \tag{11}$$

In a POT context, this formula has to be corrected by the mean number of events per year λ_p. The formula now reads:

$$T(x, y) = \frac{1}{\lambda_p \mathbb{P}[X > x, Y > y]} \tag{12}$$

Remembering that $\mathbb{P}[X \le x, Y \le y] = C_\theta(F_X(x), F_Y(y))$, $T(x, y)$ can be expressed as follows:

$$T(x, y) = \frac{1}{\lambda_p[1+C_{\hat{\theta}}(F_X(x), F_Y(y)) - F_X(x) - F_Y(y)]} \tag{13}$$

This expression links the return period to the copula and the marginal distributions and the pairs (x_T, y_T) can be determined for any value of return period.

Another methodology is used, in particular in the JOIN-SEA software: large samples of X and Y are simulated from the estimated copula and marginal distributions (equivalent to thousands of years of virtual observations) and the extreme values can be found by simple interpolation.

It may be noted that the very concept of return period may be confusing and even misleading, in particular in a multivariate framework. Serinaldi [59] presents a thorough analysis of the misconceptions associated to T. However, despite these true remarks, return period is still a basic tool of design in coastal engineering [22]. This is why we use it, while agreeing with Serinaldi. The bivariate return period in this paper is to be understood as the return period associated to the joint exceedances of both variables, as defined above.

Furthermore, other variables may be output from such a study rather than isolines of joint return period. Winterstein et al. [76] and Galiatsatou and Prinos [21] provide environmental contours, i.e. lines of equal density probability. It is also possible to draw these contours for density values corresponding to the return periods associated to one of the two variables.

3. Multivariate methodology for waves and sea level

The bivariate methodology presented above can be directly adapted to wave height H_s and sea level Z. In this case, the copula will model the dependence structure between these two variables. However, the physical processes that cause this dependence are not straightforward at all.

A first cause of dependence is the “storminess”: in mid-latitudes areas, a storm system will generate both large waves and a surge S, increasing the total sea level. In contrast, the astronomic tide level T is deterministic and does not depend on the weather conditions. Thus physical considerations advocate for focusing on the dependence between H_s and S only.

It should be noted that the remarks above hold in the offshore zone. Nearshore, propagation effects induce a new dependence. First, wave height will depend on the water depth, thus on the total sea level Z. Second, in shallow areas, there is an interaction between surge and astronomic tide because their celerity varies with the water depth: this is a dependence that is internal to sea level.

Hence, when the primary cause of dependence between waves and sea level is storminess, it is useful to refine the modelling of the dependence structure by analysing surge and wave height only. In this section we propose a combination of the bivariate methodology presented above with the POT-JPM approach for determining extreme sea levels presented by Mazas et al. [53] that allows accounting for tide-surge interaction.

Furthermore, this approach allows for a better estimation of the marginal distribution of the total sea level Z, using an indirect approach dealing separately with tide and surge: Haigh et al. [31] have shown the interest of this approach for tide-dominant areas.

Fig. 4 illustrates the methodology. First, the sampling step does not change and the discussion of Section 2.1 is still valid. Second, the steps 2 (marginal distributions F_X and F_S) and 3 (modelling of the dependence structure using a copula C) of the bivariate methodology are carried out for the wave height X and the meteorological surge S. The result is the joint distribution of wave height and surge $H_{X,S}$. In parallel, the POT-JPM approach is applied for determining the marginal distribution F_Z of sea level Z by convolution of the surge S and astronomical tide T. If needed, tide-surge interaction can be incorporated into the analysis (full details in Mazas et al. [53]).

The joint distribution of X and Z, $H_{X,Z}$ is still needed. It is determined by a bidimensional convolution of the joint distribution of X and S on the one hand, and of the astronomical tide T on the other hand. More precisely, this is a convolution of surge and tide conditional to the wave height. In practice, these distributions are discretised over their support. For each discrete bin of wave height δX_j, a classical

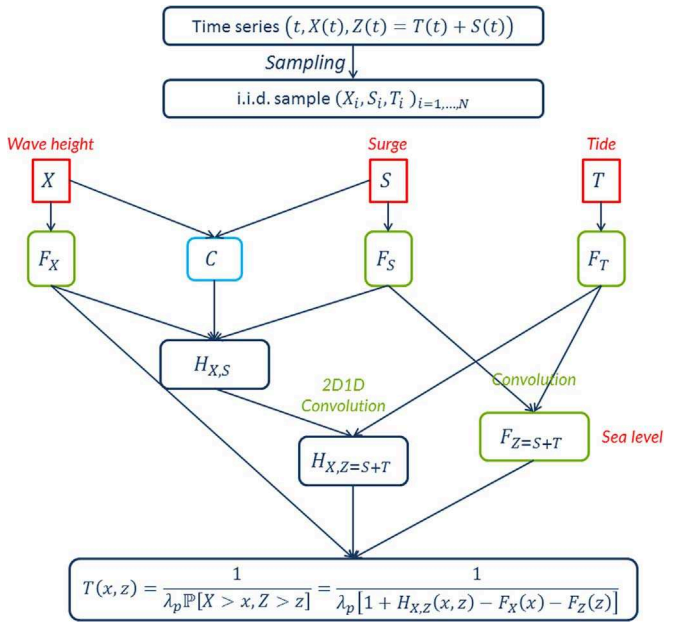

Fig. 4. Sketch of the multivariate methodology for determining extreme joint probabilities of wave height and sea level.

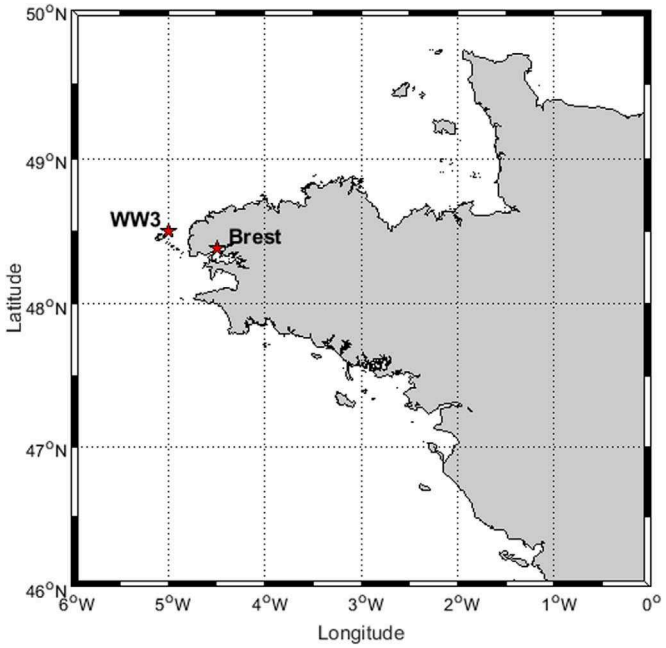

Fig. 5. Location of the case study.

convolution is performed between the (univariate) distribution of S given $X \in \delta X_j$ and the (univariate) distribution of T. Nowadays, a standard laptop is able to perform the computations for these convolutions within a few minutes with a satisfactorily fine bin width of 0.01 m for all variables, which corresponds more or less to the accuracy of available data.

At the end of the process, we have the marginal distribution of wave height F_X (determined as explained in Section 2.2), the marginal distribution of sea level F_Z (determined by the indirect approach described in Mazas et al. [53]) and the joint distribution of wave height and sea level $H_{X,Z}$, that includes both the dependence between X and S modelled by the copula C and the convolution between surge and tide.

It is then straightforward to calculate the joint return period of wave height and sea level, as presented in Section 2.4.

4. Case study

4.1. Presentation of the dataset

A dataset of sea states and sea level has been prepared for the case of Brest, France (Fig. 5).

The time series of hourly sea levels Z measured at the tide gauge of Brest from 1953/01/01 to today is available from REFMAR (*Réseaux de référence des observations marégraphiques*, http://refmar.shom.fr/). The exact coordinates of the station are 4.4950 °W; 48.3829 °N. The local datum is the *Zéro Hydrographique* (ZH), or Chart Datum (CD), defined in 1996 (0.5 m higher than the previous ZH). In the present study, the time series lasts until 2010/12/31; i.e. 58 years. The accuracy of the data is 0.01 m. The eustatic trend is analysed following the methodology presented by Bernardara et al. [1]. A positive trend of +1.48 mm/y is identified and removed from the analysis so as to get a stationary time series with a mean sea level (MSL) of +4.14 m ZH. The hourly tidal levels T have been computed with the SHOMAR software developed by the Service Hydrographique de la Marine (SHOM) and the residual $(Z - T)$ is considered as the meteorological surge S.

Xavier Bertin (University of La Rochelle) has kindly provided the authors with a 6-hourly database of sea states offshore Brest (output point at 48.5 °N, 5 °W, see Fig. 5) over the period 1948–2012 from a numerical model built with the WaveWatch III code [71], forced over the Atlantic ocean by NCEP wind fields [43] and run with the European "Cycle 4" parameterization (see Bertin et al. [4]). The following wave parameters are available: spectral significant wave height H_{m0} (or simply H_s), peak direction θ_p, peak period/frequency T_p / f_p. A linear interpolation of the wave parameters has been performed in order to get a time series of hourly sea states matching the sea level measurements. Because of this simple interpolation, the accuracy of H_s is poor; however this case study is for illustrative purpose only, not for design purpose.

As a result, a joint time series of hourly sea levels Z, tide level T, meteorological surge S, offshore wave height H_s, peak direction θ_p and peak period T_p from 1953/01/01 to 2010/12/31 (58 years, 501,363 data) is used as the dataset for the case study.

4.2. Bivariate methodology: comparison with JOIN-SEA software

4.2.1. JOIN-SEA analysis

First, a direct comparison with the JOIN-SEA software is carried out. It is thus decided to use the bivariate methodology (joint analysis of Z and H_s, without separate analysis of tide and surge) and to choose the high tide sampling used by JOIN-SEA.

Dr Peter Hawkes (HR Wallingford) kindly performed a JOIN-SEA analysis of the dataset presented above. The high tide sampling resulted in the selection of 40,694 pairs (H_s^{HT}, Z^{HT}). However, this sample, though i.i.d. and significantly smaller than the time series, was still too large for running JOIN-SEA with satisfactory results. In particular, the available joint distributions (a Bivariate Normal Distribution (BVN) and a mixture of two BVNs) were unable to capture the significantly higher dependence amongst the highest quartile of records. Consequently, Dr Hawkes decided to select a sub-set of the sample, conditional upon H_s being above a threshold value corresponding to an average of 100 records per year, that is 5753 records with H_s>4.12 m. The result of this hybrid sampling (high tide sampling associated with a threshold on one of the variables) is presented in Fig. 6 that is the scatterplot of both the original time series of hourly observations and the i.i.d. sample.

Based on Dr Hawkes's expertise, thresholds of 5% (top 288 records) were chosen for the marginal extremes (fitted by a GPD), and 1% (top 58 wave heights and top 58 sea levels) for the level above which the dependence is assumed to remain constant. The dependence was

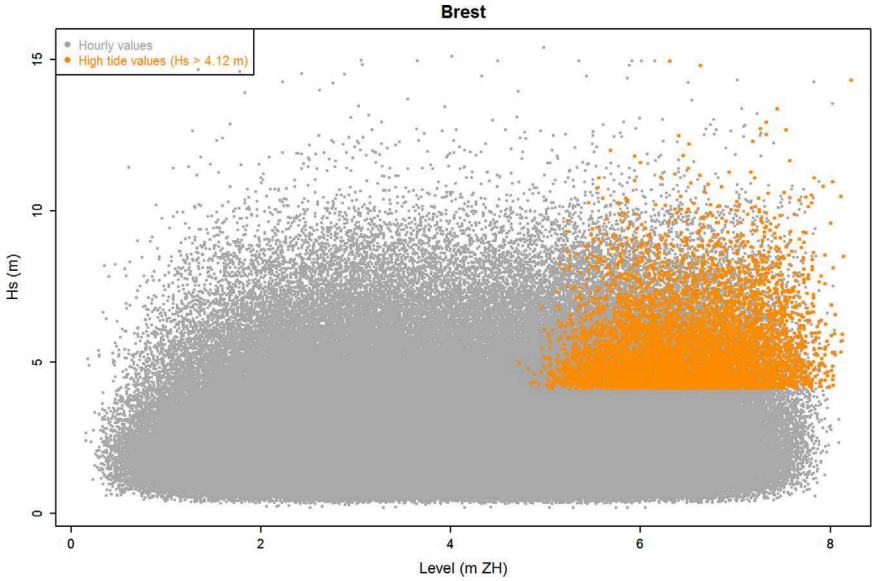

Fig. 6. scatterplot of the time series of hourly observations and of the iid sample (bivariate methodology).

modelled by a threshold Bi-Variate Normal distribution, with a user-chosen correlation coefficient. A 10,000 year sample (1,000,000 records of wave height, sea level, wave period) was simulated for the upper range of wave height (H_s>4.12 m).

4.2.2. Analysis using the bivariate methodology

For comparison purpose, the same sub-sample was used. Following the multi-distribution approach advocated in Mazas and Hamm [52], the extreme sea levels were fitted by a GPD, while the extreme wave heights were fitted by an Exponential distribution. In both cases, the choice of the statistical thresholds u_s was based upon statistical criteria such as those described by Bernardara et al. [2] and were close to the 95% percentile of the data. The extreme H_s and Z extrapolated by the marginal distributions are provided in Table 1.

The dependence between H_s and Z is illustrated by the chi-plot presented in Fig. 7. A positive association (positive values on the y-axis) seems to appear but to be more present for median values (values close to 0 on the x-axis) than for pairs with large or small values for both H_s and Z (values close to 1 on the x-axis). The upper tail dependence coefficient λ_U of the sample, estimated by the CFG estimator, is weak: $\hat{\lambda}_U$=0.069.

The copulas of Gumbel-Hougaard, Galambos and Hüsler-Reiss are estimated by the CML estimator and using λ_U. Table 2 below provides the results of the fits, along with the statistic of Cramér-von Mises and the upper tail dependence coefficient deduced from the estimation of the dependence parameter $\hat{\theta}$. When θ is estimated by λ_U, the result is the sample coefficient.

Both estimators provide very similar results for all the copulas. Regarding the Galambos and Hüsler-reiss copulas, the estimation by λ_U is slightly better than the CML estimator, according to the CVM statistic. The Galambos copula estimated by λ_U is chosen.

Because JOIN-SEA uses a Bivariate Normal distribution (with asymptotic independence), a Gaussian copula has also been fitted for comparative purposes. The statistic of Cramér-von Mises is larger (0.047), indicating poorer fitting. This result may be linked to the difficulties made by JOIN-SEA to capture the dependence amongst the highest quartile of records (see above).

Fig. 8 illustrates the fit of the three extreme value copulas estimated by CML and UTDC, along with the empirical copula and the Gaussian copula estimated by CML. It is shown that the extreme-value copulas provide a very similar fit whatever the estimator, all better than the Gaussian copula which is further from the empirical copula. This justifies using extreme value copulas for modelling the dependence in the tail region.

Table 1
Extreme wave height and sea level from the marginal distributions (bivariate methodology).

Return period (yr)	H_s (m)	Z (m CD)
1	10.14	+7.85
5	11.92	+8.04
10	12.68	+8.10
50	14.46	+8.20
100	15.23	+8.23

4.3. Comparison of joint return periods

Fig. 9 illustrates the difference between the results of the two methods. The lines (dashed lines for JOIN-SEA, plain lines for the bivariate methodology) represent the contours of equal joint exceedance probability corresponding to joint return periods of 1, 5, 10, 50, 100 and 500 years.

There is fair agreement for the marginal distributions (especially for H_s), but less good agreement for the dependence in the upper tail of the distribution. While JOIN-SEA is constrained to assume a constant level of dependence above a chosen threshold, the use of an extreme-value copula such as Gumbel-Hougaard copula makes the dependence between large wave heights and high sea levels continue to increase in the upper tail. It is to be noted that the correlation coefficient calculated by JOIN-SEA for the source data shows the coefficient to be still increasing into the upper tail: the use of extreme value copulas thus seems to be an appropriate approach for this dataset.

4.4. Multivariate methodology

A sampling based on the univariate response function $f(Z,H_s,T_p,\theta_p)=Z+H'_s=Z+K_rK_sH_s$ is used. K_r and K_s are calculated at a seabed level (−5 m CD) such that the nearshore wave height H_s' is of the same order of magnitude than Z, so that both variables have a similar weight in the sampling process. A physical threshold u_p=12 m is chosen (see Fig. 3), yielding an i.i.d. sample of the tuples (Z, S, T,H_s) of size N_p=1082, i.e. λ_p=18.7 events per year in average (see Fig. 10).

A GPD is chosen for modelling the upper tail of the marginal distributions of S and an exponential distribution for H_s. The tail distributions are connected to the bulk of the distributions modelled by

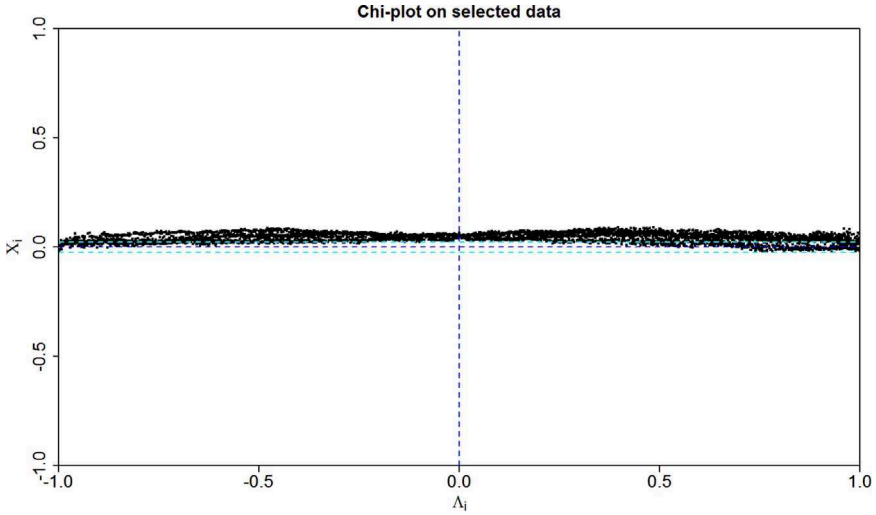

Fig. 7. Chi-plot Hs/sea level of the iid sample (bivariate methodology).

Table 2
Results of the fits (bivariate methodology).

	Estimation by Canonical Maximum Likelihood			Estimation by λ_U		
Copula	$\hat{\theta}$	λ_U	**CVM statistic**	$\hat{\theta}$	λ_U	**CVM statistic**
Gumbel-Hougaard	1.053	0.069	0.036	1.053	0.069	0.036
Galambos	0.254	0.065	0.038	0.259	0.069	0.034
Hüsler-Reiss	0.539	0.064	0.040	0.549	0.069	0.034

the empirical distribution of the observations. The astronomical tide T is modelled by the empirical distribution of the 1082 observations from the i.i.d. sample. A convolution yields the distribution of the sea level $Z = T + S$. Table 3 provides the return periods computed for H_s, S and Z.

The dependence is analysed between H_s and the surge component S. It is illustrated by the chi-plot presented in Fig. 11. The positive association is much more apparent than between H_s and Z. The upper tail dependence coefficient λ_U of the sample, estimated by the CFG estimator, is logically much larger: $\hat{\lambda}_U$=0.231. However two different populations seem to appear when it comes to the tail dependence: one which tends to independence, the other that exhibits positive dependence.

An in-depth analysis shows that the upper "lobe" (as named by Fisher and Switzer [18]) corresponds to pairs where both H_s and S are large while the lower lobe corresponds to pair where both H_s and S are small (relative to their medians). This is a further indication that positive dependence is stronger in the upper tail region, and a further justification for using extreme value copulas. In order to ease this graphical interpretation, we propose a new presentation of the chi-plot that clearly distinguishes the pairs depending on their position from the medians, as illustrated in Fig. 12.

Table 4 below provides the results of the fits for the three copulas. Both estimators provide similar results for each copula, and a graphical comparison (Fig. 13) shows that the six estimated copulas (three families, two estimators) are very close. The Galambos copula estimated by Canonical Maximum Likelihood is chosen: its Cramér-von Mises statistic is the smallest and its λ_U value is closer to the sample estimate. It has also been checked that the Gaussian copula provides a much poorer fit according to the CVM statistic (0.070), which is also visible on the graphical comparison.

The contours of equal return period associated to the joint exceedance probabilities of H_s and Z are provided in Fig. 14.

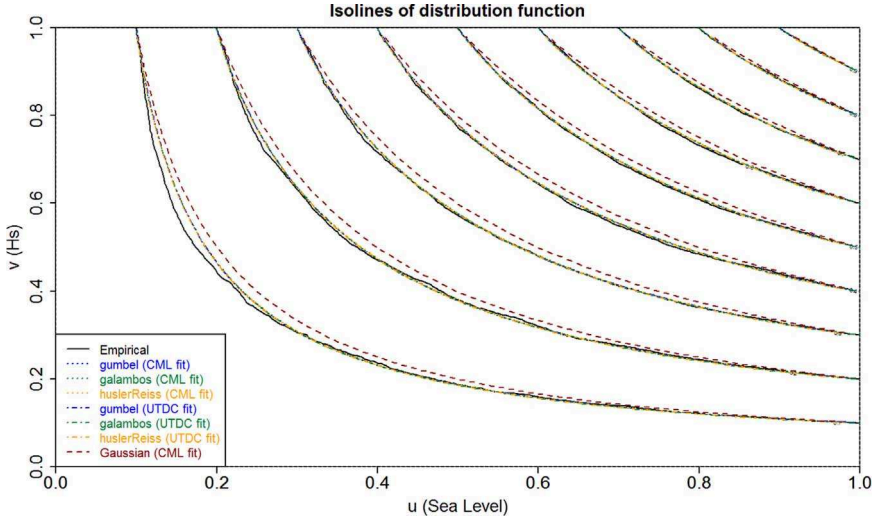

Fig. 8. Comparison of empirical copula with extreme value copulas estimated by CML and UTDC and Gaussian copula fitted by CML (bivariate methodology).

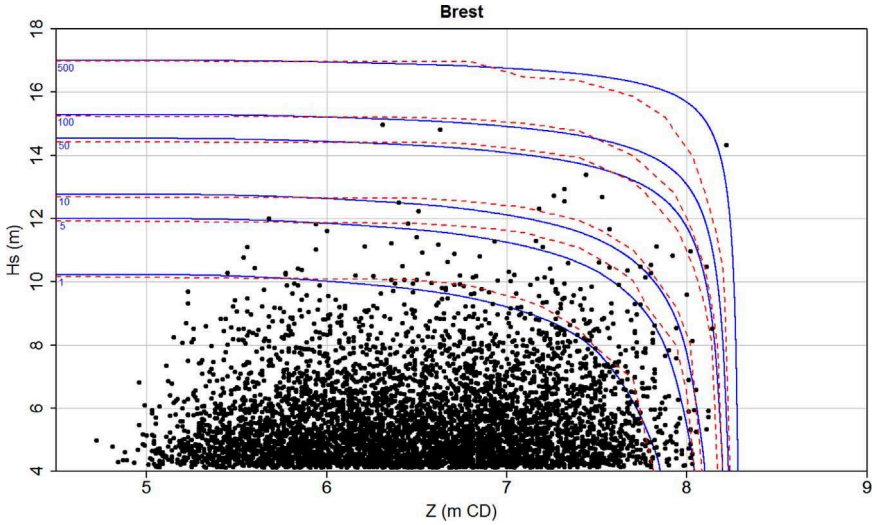

Fig. 9. Comparison of joint return periods between the JOIN-SEA simulations (dashed lines) and the bivariate methodology (plain lines).

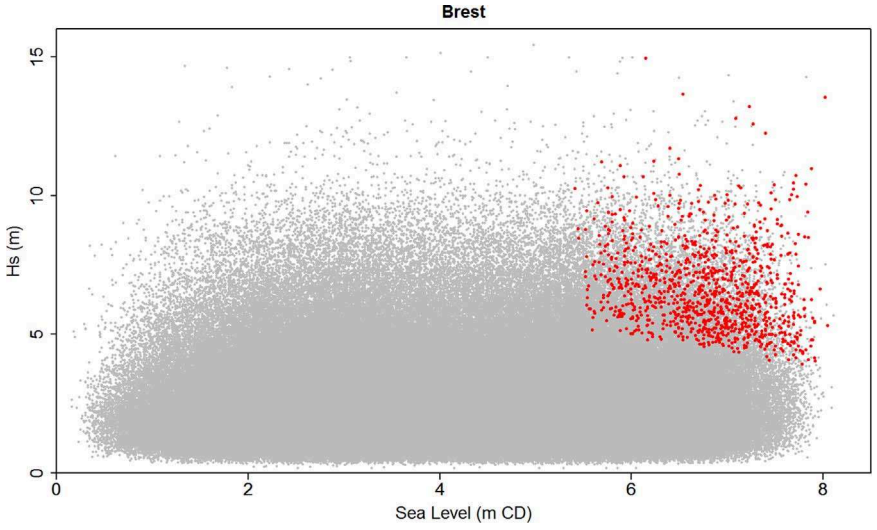

Fig. 10. scatterplot of the time series of hourly observations and of the iid sample (multivariate methodology).

5. Discussion

The use of extreme value copulas provide the analyst with mathematical functions that are adapted for modelling the structure of the dependence between two variables, particularly when the study aims at determining joint extremes. In particular, they allow for a varying level of dependence in the upper tail of the joint distribution. The modified version of the chi-plot proposed in this paper provides an easier visualisation of the variation of the dependence between the bulk and the tails. The upper tail dependence coefficient λ_U is also a useful tool for choosing the model.

The two estimators that were applied in this study, namely the Canonical Maximum Likelihood estimator and the estimation by λ_U, provided very similar results for the three families of extreme value copulas. This result was found for many samplings (high tide sampling, univariate response function such as "total water level" or wave overtopping, various sample size…). Thus this study suggests that λ_U allows for an easy analytic estimation of the copula parameter θ, getting rid of the sometimes laborious optimisation of a system of differential equations.

The multivariate methodology presented in this paper allows for a proper estimation of extreme sea levels Z in tide-dominant areas. Indeed, the extreme sea levels resulting from this methodology (marginal distribution for Z) are much closer to the results of Mazas et al. [53], who illustrated the POT-JPM methodology with the same case study: the difference is just 0.03 m for the 1-in-100-year sea level. In contrast, the bivariate methodology applies a simple direct extrapolation to the sea levels that yields an underestimation of extreme values [31]: the 1-in-100-year sea level is found to be 0.15 m lower than with the multivariate methodology.

A second advantage of this new multivariate methodology is that the dependence between surge and wave height is much stronger than between total sea level and wave height, at least for tide-dominant areas such as Brest. This can be seen on the chi-plots, the sample upper tail dependence coefficients and the values of the copula parameter θ that takes values close to the independence case when considering H_s and Z directly, whereas θ notably increases in the multivariate

Table 3
Extreme wave height, surge and sea level from the marginal distributions (multivariate methodology).

Return period (yr)	H_s (m)	S (m)	Z (m CD)
1	9.43	0.51	+7.75
5	11.24	0.67	+8.02
10	12.00	0.73	+8.12
50	13.75	0.84	+8.31
100	14.50	0.88	+8.38

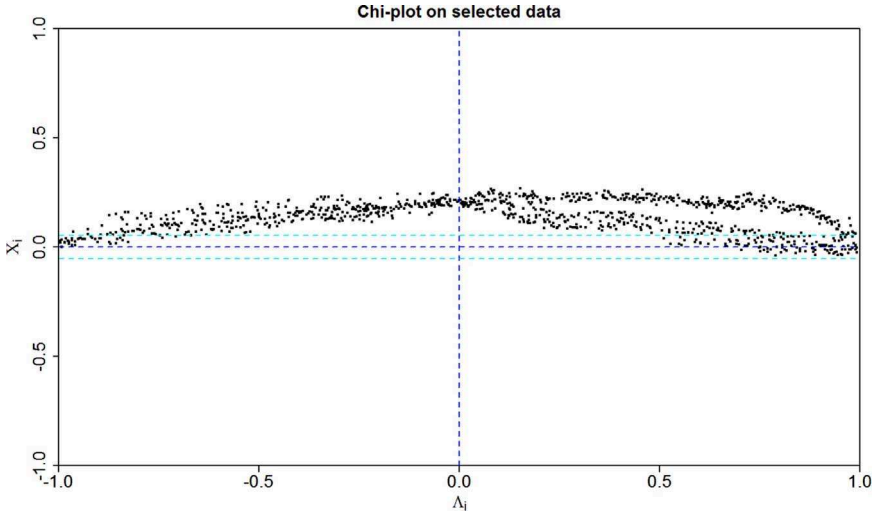

Fig. 11. Chi-plot Hs/surge of the iid sample (multivariate methodology).

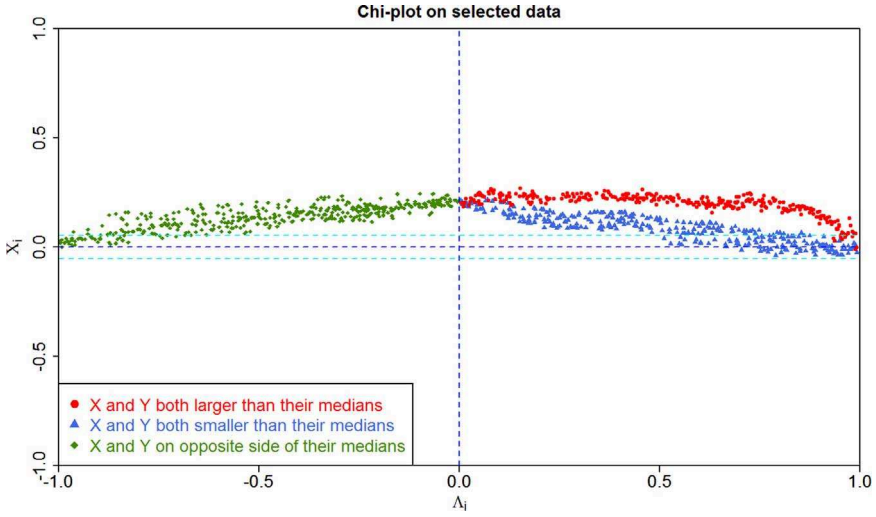

Fig. 12. Modified chi-plot Hs/surge of the iid sample (multivariate methodology).

approach. Modelling the dependence between surge and wave height has more sense and in more consistent with the physics.

The multivariate methodology has been built for a better modelling of the dependence and for a better estimation of the marginal distribution of the sea level. However, it can be seen that another major difference between the results of the bivariate and multivariate methodology lies in the marginal distribution of H_s. This is due to the uncertainties associated to the fit (choice of the distribution and of the statistical threshold), but also to the sampling. Indeed, the high tide sampling performed here does not make any distinction between the different wave populations and selects sea states coming from all directions: as a consequence, the extrapolation may be made on non-identically distributed observations [50]. In contrast, the sampling based on the "total water level" univariate response function selects only sea states coming from the directional sector [200 °N, 330 °N], corresponding to the Atlantic ocean. Sea states from the Channel, the Irish Sea and the Iroise Sea are excluded from the analysis. Thus it can be said that the modelling of H_s is limited to an homogeneous wave population, that is more consistent with the assumptions required by the Extreme Value Theory, and more generally by the fit of a distribution to a sample.

Sampling is most probably the hardest issue when determining joint probabilities. Sampling must be linked to the concept of event, which needs to be defined in a multivariate case. It was highlighted in Bernardara et al. [2] in the univariate case that the i.i.d. sample is made of a different random variable than the time series of sequential observations (e.g. H_s storm peak is a different random variable from hourly H_s). This is particularly visible for multivariate analysis of Type C: the statistical model is applied to pairs (or tuples) describing the events identified by the sampling. In our case study, we have thus determined the joint probabilities of sea levels and offshore waves which result in a noticeable "total water level" nearshore, as defined by the univariate response function used for the sampling. It is obvious that other response functions, high tide sampling or the bivariate threshold sampling would give different results. First, it must be made

Table 4
Results of the fits (multivariate methodology).

	Estimation by Canonical Maximum Likelihood			Estimation by λ_U		
Copula	$\hat{\theta}$	λ_U	**CVM statistic**	$\hat{\theta}$	λ_U	**CVM statistic**
Gumbel-Hougaard	1.221	0.235	0.036	1.215	0.231	0.041
Galambos	0.476	0.233	0.035	0.473	0.231	0.036
Hüsler-Reiss	0.829	0.227	0.038	0.835	0.231	0.035

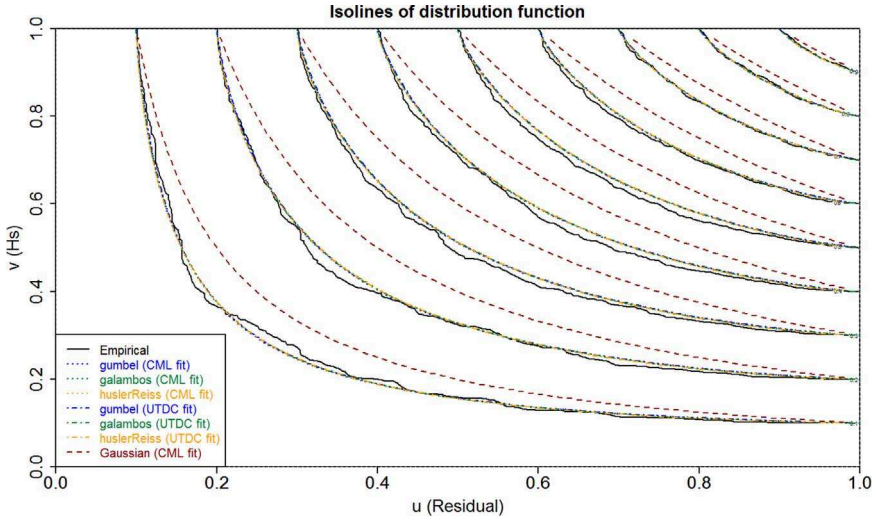

Fig. 13. Comparison of empirical copula with extreme value copulas estimated by CML and UTDC and Gaussian copula fitted by CML (multivariate methodology).

clear that sampling by a univariate response function has both its pros (it accounts for covariates; it reduces to the univariate case which makes it easy to select independent tuples; different functions can be chosen or even combined) and cons (it is quite difficult to find a function that represents moderate and high values for both variables; results may change significantly from one function to the other). The aim of the present paper is to introduce it as an additional tool available for the analyst. Indeed, it is the responsibility of the analyst to choose the sampling method that is the most adapted to his analysis and there is no universal solution.

Sample size is also to be considered. In Mazas et al. [52] the authors suggested to set the physical threshold u_p so as to get a mean number of event peaks per year λ_p between 5 and 10. In a multivariate analysis, more information is to be extracted from the sample in order to analyse and model the dependence structure. A sensitivity study was performed on sample size and a value of λ_p close to 18 was found to provide good results. Because the duration of this time series is quite long (58 years), it could be necessary to increase this value for shorter datasets and a range of 15 to 25 events per year is suggested.

Another question arises when considering the marginal distributions. Should the marginal distributions be modelled from the multivariate i.i.d. sample or should they come from a classical analysis of the time series? The question seems particularly relevant for the distribution of sea level when using the multivariate methodology. In the POT-JPM approach [53], the empirical distribution of tide T is derived from a time series over a saros period (223 synodic months, i.e. 6585.32 days or 18 years, 10–11 days, 8 hours: astronomical tide is assumed to be nearly identical after a saros), which should be more accurate than the distribution derived from the multivariate i.i.d. sample. However, as noted above, the distribution of sea level obtained from the i.i.d. sample is very close to that from the POT-JPM analysis and it does not seem necessary to further complicate the methodology. Regarding H_s, a classical univariate POT analysis from the time series would require working on the same wave population than the one selected by the multivariate sampling step. Without further investigation, it seems safer to work on the multivariate i.i.d. sample for modelling the marginal distribution of H_s.

Last, the interest of a bivariate or multivariate methodology when the design focuses on a single univariate response function such as overtopping or beach overwash may be questioned. Wouldn't it be more efficient to compute directly the time series of overtopping from the original time series of sea level and wave height, period and direction, before extrapolating the result? Different answers can be made to this question. First, the analytical formulas used for computing such complex and non-linear phenomena may be unable to cover complicated situations caused by complex bathymetry or coastal structures. In that case, the formulas provide a rough estimate that is sufficient for the sampling stage, then the multivariate methodology

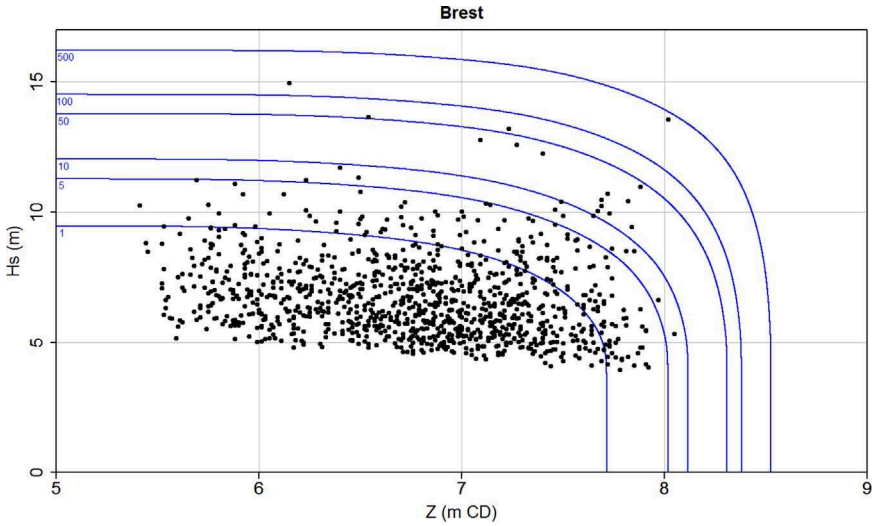

Fig. 14. joint return periods of wave height and sea level (multivariate methodology).

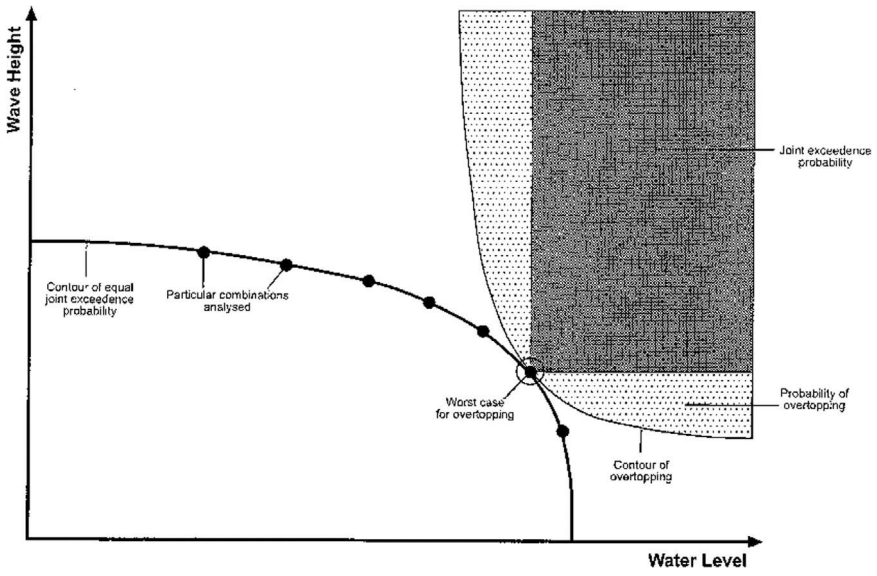

Fig. 15. joint exceedance probability and structure variable probabilities, after Hawkes et al. [34].

provides the contours of equal joint exceedance probability that can be used for defining a small set of input conditions for physical modelling in a wave basin or detailed numerical modelling. In that regard, the multivariate methodology may be useful for overcoming practical difficulties.

Still the respective pros and cons of the univariate vs. multivariate approaches cannot be properly comprehended without a full understanding of the concepts of return period and event. It has already been shown by others [34,44] that the return period associated to the response function is lower than the joint return period associated to the source variables. Let us consider the case of overtopping assessed from wave height/sea level joint probabilities, as illustrated in Fig. 15 from Hawkes et al. [34]. Let the design criteria be defined by the 10-year overtopping volume. Thanks to the methodology defined in this paper, we can precisely estimate the contour of the 10-year joint return period of H_s and Z. Along this contour, a single pair (H_s,Z) is associated to a worst case for overtopping (circled dot). We can then draw the contour of overtopping corresponding to this value and we see a discrepancy between the hatched area of H_s/Z joint exceedance probability and the dotted area of overtopping probability. The latter is larger: it means its probability of exceedance is higher and by consequence its return period lower.

What does this mean? The first lesson is that when defining design criteria, it must be decided whether the return period (and hence the encounter probability, which is the truly important indicator) is applied to the source variables or to the response function(s). An example for the former case may be the estimation of meteo-oceanic conditions (wind, waves, levels, currents) in an offshore wind farm for which multivariate analyses are useful for many purposes from choice of location to structural design. This leads us back to our definition of events: are they related to environmental conditions or to structural response, to what comes or to what results?

The following framework may be helpful to comprehend this: the source-pathway-receptor concept. It was forged originally to describe the flow of environmental pollutants from a source, through different pathways to potential receptors [38] and later applied to coastal flooding by the UK Environment Agency [40] to describe the propagation of a flood from a source through flood defences (pathways) to the floodplain beyond (receptors). We propose here to use this framework for visualizing the different possibilities to define the event. In our latest example, the sources may be the atmosphere (wind and pressure fields) and the astronomical forcing, the pathway would be the ocean (where the energy from the atmosphere and celestial bodies propagates through short and long waves) and the receptor would be the breakwater where overtopping occurs. Here it is more intuitive that the return period associated to a pathway event (joint probabilities of waves and sea level) will differ from the one associated to a receptor event (overtopping) or source event (storm or exceptional syzygy). Of course we may shift the terms of the SPR and consider the ocean as the source, the overtopping as the pathway and the buildings behind the levees as the receptors, but this does not change the idea.

Whatever the way to visualize this, it should always be made clear in engineering whether the design criteria are to be based upon the source variables or the response function. This will drive the choices of the event, sampling and possibly methodology.

6. Conclusions

The comparison between JOIN-SEA and the bivariate methodology presented in this paper has shown the interest in using extreme value copulas for modelling the dependence between two environmental variables. This methodology can be used applied for many cases such as the description of a storm event by two parameters (H_s/T_p, $H_s/Duration$, H_s/wind speed...), or the joint exceedances of wave height and sea level in surge dominant areas such as the Mediterranean Sea.

When considering the joint exceedances of wave height and sea levels in tide-dominant areas, the multivariate methodology introduced here provides two valuable advantages. First, it allows determining extreme sea levels using a specific indirect approach combining surge and astronomical tide. This approach avoids the underestimation of extreme sea levels induced by direct extrapolation, as highlighted by several authors such as Haigh et al. [31] (though it may overestimate them, but this yields conservative values). Second, the dependence is analysed between two fully stochastic values, which are usually generated by the same physical event, i.e. a storm. The comparison of both methods has shown the significant increase in the dependence levels. This methodology needs a bidimensional convolution operation, that is no more than a classical convolution conditional to the wave height.

A wide variety of sampling choices has shown that the three families of extreme value copulas considered in this paper, namely the Gumbel-Hougaard, Galambos and Hüsler-Reiss families, provide close results and are suitable for such analyses. Furthermore, it has shown that estimating the copula dependence parameter θ from the CFG estimate of the upper tail dependence coefficient λ_U of the sample yields values of $\hat{\theta}$ very close to the CML estimates. This estimator only needs the

solving of a simple analytic expression, without involving the implicit optimisation of a system of differential equations as requested by ML-based methods.

As often in such extreme studies of environmental variables, sampling is a key step of the study, and it may even be the most critical. While it is relatively easy to ensure the independence of the data, the assumption of identical distribution becomes quite difficult to ensure when studying the joint occurrence of two (or more) distinct metocean processes (Type C multivariate analysis). We have presented different possibilities in Section 2.1, though this list is far from being exhaustive: further work is necessary to better understand the advantages and drawbacks of each sampling method, depending on the type of analysis. But this is the responsibility of the analyst to choose the sampling that is the most adapted to the meteo-oceanic environment and to the final aim of his study. In particular, a clear understanding of the concept of event and return period is key to a proper definition of design criteria for an engineering application.

Acknowledgements

This work was first investigated by Vincent Auger, then student at the University of Lyon 1, under the supervision of the authors. Let him be thanked for his commitment and his results. He was greatly helped by Pr Anne-Catherine Favre (ENSE3/INPG, LTHE) and Pr Clémentine Prieur (University of Grenoble 1, LJK) who made them available to answer his questions and suggest him useful hints.

The authors also wish to thank Dr Peter Hawkes (HR Wallingford) for his support and his kind acceptance of running JOIN-SEA simulations and analysing them for our comparisons, and Dr Xavier Bertin (University of La Rochelle) for providing the sea state time series offshore Brest that was used for the case study.

Dr Xavier Kergadallan (CEREMA/DTecEMF) and Marc Andreewsky (EDF R & D/LNHE) provided helpful comments for improving the paper, and the reviewers provided multiple useful corrections and suggestions.

References

[1] P. Bernardara, M. Andreewsky, M. Benoit, Application of regional frequency analysis to the estimation of extreme storm surges, J. Geophys. Res. (2011) 116.
[2] P. Bernardara, F. Mazas, X. Kergadallan, L. Hamm, A two-step framework for over-threshold modelling of environmental extremes, Nat. Hazards Earth Syst. Sci. 14 (2014) 635–647.
[3] X. Bertin, N. Bruneau, J.F. Breilh, A.B. Fortunato, M. Karpytchev, Importance of wave age and resonance in storm surges: the case Xynthia, Bay of Biscay, Ocean Model. 42 (2012) 16–30.
[4] X. Bertin, E. Prouteau, C. Letetrel, A significant increase in wave height in the North Atlantic Ocean over the 20th century, Glob. Planet. Change 106 (2013) 77–83.
[5] D.P. Callaghan, P. Nielsen, A. Short, R. Ranasinghe, Statistical simulation of wave climate and extreme beach erosion, Coast. Eng. 55 (2008) 375–390.
[6] P. Capéraà, A.-L. Fougères, C. Genest, A nonparametric estimation procedure for bivariate extreme value copulas, Biometrika 843 (1997) 567–577.
[7] P. Capéraà, A.L. Fougères, C. Genest, Bivariate distributions with given extreme value attractor, J. Multivar. Anal. 72 (2000) 30–49.
[8] CETMEF, Analyse statistique des niveaux d'eau extrêmes – Environnements maritimes et Estuariens, Special Report, April 2013.
[9] S.G. Coles, An Introduction to Statistical Modeling of Extreme Values, Springer, London, 2001.
[10] S.G. Coles, J.A. Tawn, Statistical methods for multivariate extremes: an application to structural design, Appl. Stat. 43 (1994) 1–48.
[11] S. Coles, J.E. Heffernan, J.A. Tawn, Dependence measures for extreme value analyses, Extremes 2 (4) (1999) 339–365.
[12] S. Corbella, D.D. Stretch, Predicting coastal erosion trends using non-stationary statistics and process-based models, Coast. Eng. 70 (2012) 40–49.
[13] L. De Haan, J. De Ronde, Sea and wind: multivariate extremes at work, Extremes 1 (1998) 7–45.
[14] D. De Waal, P. Van Gelder, Modelling of extremewave heights and periods through copulas, Extremes 8 (4) (2005) 345–356.
[15] E. Di Bernardino, V. Maume-Deschamps, C. Prieur, Estimating a bivariate tail: a copula based approach, J. Multivar. Anal. 119 (2013) 81–100.
[16] M.J. Dixon, J.A. Tawn, The effect of non-stationarity on extreme sea-level estimation, Appl. Stat. 48 (2) (1999) 135–151.
[17] N.I. Fisher, P. Switzer, Chi-plots for assessing dependence, Biometrika 72 (1985) 253–265.
[18] N.I. Fisher, P. Switzer, Graphical assessment of dependence: is a picture worth 100 tests?, Am. Stat. 55 (3) (2001) 233–239.
[19] G. Frahm, M. Junker, R. Schmidt, Estimating the tail dependence coefficient: properties and pitfalls, Insur. Math. Econ. 371 (2005) 80–100.
[20] J. Galambos, Order statistics of samples from multivariate distributions, J. Am. Stat. Assoc. 70 (1975) 674–680 (351, Part 1).
[21] P. Galiatsatou, P. Prinos, Bivariate models for extreme of significant wave height and period. An application to the Dutch Coast, in: Proceedings of the 2nd IMA Conference on Flood Risk Assessment, 2007, Plymouth, Institute of Mathematics & its Applications, pp. 77–84, 2007.
[22] P. Galiatsatou, P. Prinos, Bivariate analysis of extreme wave and storm surge events. Determining the failure area of structures, Open Ocean Eng. J. 2011 (4) (2011) 3–14.
[23] C. Genest, L.P. Rivest, Statistical inference procedures for bivariate Archimedean copulas, J. Am. Stat. Assoc. 88 (3) (1993) 1034–1043.
[24] C. Genest, K. Ghoudi, L.P. Rivest, A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, Biometrika 82 (3) (1995) 543–552.
[25] C. Genest, J.F. Quessy, B. Rémillard, Goodness-of-fit procedures for copula models based on the probability integral transformation, Scand. J. Stat. 33 (2) (2006) 337–366.
[26] C. Genest, A.C. Favre, Everything you always wanted to know about copula modeling but were afraid to ask, J. Hydrol. Eng. 12 (4) (2007) 347–368.
[27] Y. Gōda, M. Kudaka, Bounded and unbounded distribution functions for extreme wave analysis, in: Scientific, W. (Ed.), Proceedings Asian and Pacific Coasts 2009. pp. 8–14, 2009.
[28] Y. Gōda, Random Seas and Design of Maritime Structures, World Scientific, 2010.
[29] G. Gudendorf, J. Segers, Extreme-value copulas, in: P. Jaworski, F. Durante, W.K. Härdle, T. Rychlik (Eds.), Copula Theory and Its Applications, Lecture Notes in Statistics, Springer, Berlin Heidelberg, 2010, pp. 127–145.
[30] E.J. Gumbel, Bivariate logistic distributions, J. Am. Stat. Assoc. 56 (1961) 335–349.
[31] I.D. Haigh, R. Nicholls, N. Wells, A comparison of the main methods for estimating probabilities of extreme still water levels, Coast. Eng. 57 (2010) 838–849.
[32] P.J. Hawkes, Use of Joint Probability Methods in Flood Management: A Guide to Best Practice, Defra/Environment Agency and Flood and Coastal Defence R & D Programme, 2005.
[33] P.J. Hawkes, Joint probability analysis for estimation of extremes, J. Hydraul. Res. 46 (2008) 246–256.
[34] P.J. Hawkes, B.P. Gouldby, J.A. Tawn, M.W. Owen, The joint probability of waves and water levels in coastal engineering design, J. Hydraul. Res. 40 (2002) 241–251.
[35] P.J. Hawkes, B.P. Gouldby, S. Weidong, J.A. Tawn, D. Hames, D. Reeve, D. Blackman, R. Sproson, K. Mavronasos, A comparison of marginal and joint extremes predicted from synthesised wave and water level data, in: Proceedings of the IMA Conference on Flood Risk Assessment. IMA, Southend, 2004, pp. 65–74.
[36] P. Hawkes, D. Gonzalez-Marco, A. Sánchez-Arcilla, P. Prinos, Best practice for the estimation of extremes: a review, J. Hydraul. Res. 46 (2) (2008) 324–332.
[37] J.E. Heffernan, J.A. Tawn, A conditional approach for multivariate extreme values (with discussion), J. R. Stat. Soc.: Ser. B Stat. Methodol. 66 (2004) 497–546.
[38] M.W. Holdgate, A Perspective of Environmental Pollution, Cambridge University Press, Cambridge, UK, 1979.
[39] J.R.M. Hosking, L-moments: analysis and estimation of distributions using linear combinations of order statistics, J. R. Stat. Soc. 52 (1990) 105–124.
[40] HR Wallingford, Risk, Performance and Uncertainty in Flood and Coastal Defence – A Review, 2nd Ed., H R Wallingford, Wallingford, UK, 2002 (SR 587).
[41] H. Joe, R.L. Smith, I. Weissman, Bivariate threshold methods for extremes, J. R. Stat. Soc. B 54 (1992) 171–183.
[42] H. Joe, Multivariate Models and Dependence Concepts, Chapman & Hall, London, 1997.
[43] E. Kalnay, M. Kanamitsu, R. Cistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woolen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. Mo, C. Ropelewski, J. Wang, A. Leetma, R. Reynolds, R. Jenne, D. Joseph, The NCEP/NCAR reanalysis project, Bull. Am. Meteorol. Soc. 77 (1996) 437–471.
[44] X. Kergadallan, Estimation Des Niveaux Marins Extrêmes Avec Et Sans L'action Des Vagues Le Long Du Littoral Métropolitain (Doctoral dissertation), Université Paris-Est, 2015.
[45] M.R. Leadbetter, Extremes and local dependence in stationary sequences, Probab. Theory Rel 65 (1983) 291–306.
[46] A.W. Ledford, J.A. Tawn, Statistics for near independence in multivariate extreme values, Biometrika 83 (1) (1996) 169–187.
[47] E.L. Lehmann, Theory of Point Estimation, Wiley & Sons, Australia, 1983.
[48] F. Li, P.H.A.J.M. van Gelder, R. Ranasinghe, D.P. Callaghan, R.B. Jongejan, Probabilistic modelling of extreme storms along the Dutch coast, Coast. Eng. 86 (2014) 1–13.
[49] J. Lin-Ye, M. Garcia-Leon, V. Gracia, A. Sanchez-Arcilla, A multivariate statistical model of extreme events: an application to the Catalan coast, Coast. Eng. 117 (2016) 138–156.
[50] M. Mathiesen, Y. Gōda, P.J. Hawkes, E. Mansard, M.J. Martin, E. Peltier, E.F. Thompson, G. Van Vledder, Recommended practice for extreme wave analysis, J. Hydraul. Res. 32 (1994) 803–814.
[51] F. Mazas, P. Garat, L. Hamm, Questioning MLE for the estimation of environmental extreme distributions, Ocean Eng. 92 (2014) 44–54.
[52] F. Mazas, L. Hamm, A multi-distribution approach to POT methods for determining extreme wave heights, Coast. Eng. 58 (2011) 385–394.
[53] F. Mazas, X. Kergadallan, P. Garat, L. Hamm, Applying POT methods to the

Revised Joint Probability Method for determining extreme sea levels, Coast. Eng. 91 (2014) 140–150.
[54] M.W. Owen, P.J. Hawkes, J.A. Tawn, P. Bortot, The Joint Probability of Waves and Water Levels: A Rigorous But Practical New Approach. MAFF (now DEFRA), in: Conference River and Coastal Engineers, University of Keele, B(4), 1997, pp. 1–10.
[55] J. Pickands, Statistical inference using extreme order statistics, Ann. Stat. 3 (1975) 119–131.
[56] A. Poulin, D. Huard, A. Favre, S. Pugin, Importance of tail dependence in bivariate frequency analysis, J. Hydrol. Eng. 12, SPECIAL ISSUE: Copulas Hydrol., 2007, pp. 394–403.
[57] D.T. Pugh, J.M. Vassie, Extreme sea levels from tide and surge probability, in: International Conference Coastal. Eng. 1, 1979.
[58] A. Sanchez-Arcilla, J. Gomez-Aguar, J.J. Egozcue, M.I. Ortego, P. Galiatsatou, P. Prinos, Extremes from scarce data. The role of Bayesian and scaling techniques in reducing uncertainty, J. Hydraul. Res. 46 (2) (2008) 224–234.
[59] F. Serinaldi, Dismissing return periods, Stoch. Environ. Res. Risk Assess. 29 (2015) 1179–1189.
[60] J.H. Shih, T.A. Louis, Inferences on the association parameter in copula models for bivariate survival data, Biometrics (1995) 1384–1399.
[61] M. Sibuya, Bivariate extreme statistics, Ann. Inst. Stat. Math. 11 (1960) 195–210.
[62] A. Sklar, Fonctions de répartition à n dimensions et leurs marges 8, Publications de l'Institut de Statistique de L'Université de Paris, 1959, pp. 229–231.
[63] O.P. Smith, Duration of extreme wave conditions, J. Waterw. Part. C – ASCE 114 (1988) 1–17.
[64] S. Solari, M. Losada, Unified distribution models for met-ocean variables: application to series of significant wave height, Coast. Eng. 68 (2012) 67–77.
[65] S. Solari, M. Losada, A unified statistical model for hydrological variables including the selection of threshold for the peak over threshold method, Water Resour. Res. 48 (2012) 10.
[66] M. Šraj, N. Bezak, M. Brilly, Bivariate flood frequency analysis using the copula function: a case study of the Litija station on the Sava River, Hydrol. Process. 29 (2) (2015) 225–238.
[67] J.A. Tawn, An extreme-value theory model for dependent observations, J. Hydrol. 101 (1988) 227–250.
[68] J.A. Tawn, Bivariate extreme value theory: models and estimation, Biometrika 75 (1988) 397–415.
[69] J.A. Tawn, Estimating probabilities of extreme sea-levels, Appl. Stat. 41 (1) (1992) 77–93.
[70] J.A. Tawn, J.M. Vassie, Extreme sea levels: the joint probabilities method revisited and revised, in: Proceedings – Institution of Civil Engineers, Part 2. 1989, pp. 429–442.
[71] H.L. Tolman, User manual and system documentation of WAVEWATCH III version 3.14 NOAA/NWS/NCEP/MMAB Tech. Note 276, 2009, pp. 194.
[72] G. Van Vledder, Y. Gōda, P. Hawkes, E. Mansard, M.J. Martin, M. Mathiesen, E. Peltier, E. Thompson, Case studies of extreme wave analysis: a comparative analysis, in: Proceedings of the Second Symposium on Ocean Wave Measurement and Analysis, New Orleans, Louisiana, USA, 1994, pp. 978-992.
[73] T. Wahl, C. Mudersbach, J. Jensen, Assessing the hydrodynamic boundary conditions for risk analyses in coastal areas: a multivariate statistical approach based on copula functions, Nat. Hazards Earth Syst. Sci. 12 (2) (2012) 495–510.
[74] T.L. Walton, Distribution for storm surge extremes, Ocean Eng. 27 (2000) 1279–1293.
[75] W. Wang, M.T. Wells, Model selection and semiparametric inference for bivariate failure-time data (with discussion), J. Am. Stat. Assoc. 95 (1) (2000) 62–76.
[76] S.R. Winterstein, T.C. Ude, C.A. Cornell, P. Bjerager S. Haver, Environmental Parameters for Extreme Response: Inverse Form with Omission Factors, in: Proceedings, ICOSSAR-93, Innsbruck, Austria, 1993.
[77] S. Yue, P. Rasmussen, Bivariate frequency analysis: discussion of some useful concepts in hydrological application, Hydrol. Process. 16 (2002) 2881–2898. http://dx.doi.org/10.1002/hyp.1185.
[78] S. Zachary, G. Feld, G. Ward, J. Wolfram, Multivariate extrapolation in the offshore environment, Appl. Ocean Res. 20 (5) (1998) 273–295.

2. R PACKAGE ARTEXTREME

2.1. R DOCUMENTATION

Package 'artextreme'

July 11, 2017

Type Package

Title EXTREMES FOR MARITIME ENGINEERING APPLICATIONS

Version 4.1.07

Date 2017-07-07

Author Franck Mazas

Maintainer Franck Mazas <franck.mazas@arteliagroup.com>

Depends
polynom, stats4, MASS, boot, copula, ggplot2, forcats, RColorBrewer, gridExtra, gtable, grid

Description Functions for univariate statistical extrapolation extremes using a POT approach
and a multi-distribution model (GPD, Weibull, Gamma);
for directional analysis with a Storms-Over-Threshold (SOT) sorting;
computation of parametric bootstrap confidence intervals;
convolution of tidal levels and surge values for determining extreme sea levels,
joint probability of waves and sea levels.

License GPL (>=3)

Repository
//arteliagroup.com/Missions_Global/FR/ECH/BRC/MAR/1715600_RD_Oceano_meteo/03_LOGICIELS_UTILITAIR
cran

LazyLoad yes

Encoding latin1

R topics documented:

`Bastia` *Wave Time Series off Bastia*

Description

Sea state 3-hourly time series (time, Hs, peak direction and peak period) of hindcast sea states off Bastia.

Usage

```
data(Bastia)
```

Format

A data frame with 29224 observations of the following 4 variables.

`Time` Julian time, Excel format.

`Hs` Spectral significant wave height H_m0, in m.

`Dir` Peak wave direction, in °N.

`Tp` Peak period, in s.

Details

Time series from GlobOcean for the Carbonite study (8713512) in 2015. Hindcast period reduced to 01/01/1992 to 31/12/2011.

Source

Hindcast by GlobOcean.

Examples

```
data(Bastia)
envir <- list(site = "Bastia", K = 23, origin = "1899-12-30", tz = "UTC1")
Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m", varcol = 2), Tr = c(1, 2, 5, 10,
20, 50, 100), conf.IC = 90, boot.iter = 1000)
Dir.envir <- list(var = list(id = "Dir", name = "Direction", unit = "°N", varcol = 3))
Tp.envir <- list(var = list(id = "Tp", name = "Tp", unit = "s", varcol = 4))
graphenvir <- list(lang = "EN", color = TRUE, saveplots = TRUE, fileformat = "png", cex = 1)
covar.plot(Bastia, xvar.envir = Dir.envir, yvar.envir = Hs.envir)
covar.plot(Bastia, xvar.envir = Tp.envir, yvar.envir = Hs.envir)
```

Brest *Sea Level and Wave Time Series off Brest*

Description

The dataset is a joint time series of sea level at Brest tide gauge and of sea states offshore Brest from 01/01/1953 to 31/12/2010 (58 years). The dataset provides a case study for extreme sea levels and joint occurrence of waves and sea levels.

Usage

```
data(Brest)
```

Format

A data frame with 505363 values of the following variables.

`Y` Year.

`M` Month.

`D` Day.

`h` Hour.

`Time` Julian time, Excel format.

`Z` Sea level, in m ZH.

`T` Level of astronomical tide, in m ZH.

`S` Non-tidal residual (meteorological surge), in m.

`Hs` Spectral significant wave height H_m0, in m.

`fp` Peak frequency, in Hz.

`Dir` Peak wave direction, in °N.

Details

Sea state data from X. Bertin (University of La Rochelle). 6-hourly time series of sea states offshore Brest (48.5°N, 5°W) over the period 1948-2012. Numerical model built on WaveWatch III forced by NCEP wind fields and run with the European 'Cycle 4' parameterization. Linear interpolation carried out to get hourly sea states from 01/01/1992 to 31/12/2011.

Sea level hourly time series from SHOM / REFMAR. Local datum: Zero Hydrographique (defined in 1996). Eustatic trend removed: +1.48 mm/yr. Mean sea level: +4.14 m ZH. Hourly tidal levels computed with SHOMAR software (SHOM). Hindcast period reduced to 01/01/1992 to 31/12/2011.

Source

See Details section.

Examples

```
data(Brest)
envir <- list(site = "Brest", K = 57, origin = "1899-12-30", tz = "UTC1")
Z.envir <- list(var = list(id = "Z", name = "Level", unit = "m ZH", varcol = 2))
T.envir <- list(var = list(id = "T", name = "Tide", unit = "m ZH", varcol = 3))
S.envir <- list(var = list(id = "S", name = "Surge", unit = "m", varcol = 4))
Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m", varcol = 5))
graphenvir <- list(lang = "EN", color = TRUE, saveplots = FALSE, fileformat = "png", cex = 1)
serie <- Brest[, -(1:4)]
covar.plot(serie, xvar.envir = Z.envir, yvar.envir = Hs.envir)
```

---

climat *Determination of the climate of meteo-oceanic phenomena*

Description

This function makes a discreta analysis of a phenomenon (sea states, wind, current...) described by two or three quantities: a magnitude (height, velocity...), a direction and possibly a third one (period...).

Usage

```
climat(dataset, mag.envir, dir.envir, zvar.envir = NULL, mag.bins = NULL,
dir.bins = NULL, z.bins = NULL, is.magcalm = TRUE, is.center = c(FALSE, TRUE, FALSE),
with_lowbound = c(FALSE, FALSE, FALSE), g.envir = envir, export.cli = FALSE)
```

Arguments

dataset	A table (numeric matrix or data frame) containing the data (by column). For tserie.plot, it must have a column with the time. The safest is to have the time in julian days in first column. The data frame may also have a column named "Horodate" in POSIX format. Other columns include the main variable and possibly covariates.
mag.envir, dir.envir, zvar.envir	A list regrouping the variable descriptions and properties for the magnitude variable, the direction and opssibly the third variable. The variable description is a sub-list var containing its symbol id, its name name, its unit unit, the column number in the dataset varcol, the number of sequential observations per year nu, the number of significant digits ndig...
mag.bins, dir.bins, z.bins	The discrete bins for the variable. It can be: - a sequency vector of length 3 with minimum, maximum, increment; - a numeric scalar: the increment (step), the minimum and maximum values being automatically determined; - NULL: the bins are fully automatically determined.
is.magcalm	Logical TRUE / FALSE. If TRUE, the magnitude variable is considered to be a calm when it is 0.
is.center	Logical vector TRUE / FALSE, for each variable. If TRUE, the bin limits given by mag.bins are considered as the center of the bins and not the bounds.
with_lowbound	Logical vector TRUE / FALSE, for each variable. Defines how the bounds of the bins are. If TRUE, the lower bound of the range is included and the upper bound is not.

`g.envir`	A list describing the global environment of the study: site name `$site`, duration of observations `K` (in years), origin `$origin` and time zone `$tz` of the time...
`export.cli`	Logical `TRUE` / `FALSE`. If `TRUE`, a `*.cli` file is output for Excel processing.

Details

-

Value

A list providing the arguments of the function, the discretization of the dataset (definition of bins) and the series of the two or three variables considered. If `export.cli=TRUE`, a .cli file is written in the working directory with the formatting allowing Excel to import it.

Author(s)

Franck Mazas

See Also

`dataplots`

Examples

```
data(Bastia)
envir <- list(site = "Bastia", K = 23, origin = "1899-12-30", tz = "UTC")

serie <- as.data.frame(Bastia)
serie$Horodate <- temps2horodate(Bastia[, 1], origin = envir$origin, tz = envir$tz)

Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m", varcol = 2))
Dir.envir <- list(var = list(id = "Dir", name = "Direction", unit = "°N"))
Tp.envir <- list(var = list(id = "Tp", name = "Tp", unit = "s"))
graphenvir <- list(lang = "EN", color = TRUE, saveplots = FALSE, fileformat = "png", cex = 1)

climat(dataset = serie, mag.envir = Hs.envir, dir.envir = Dir.envir, zvar.envir = Tp.envir,
mag.bins = c(0, 4, 0.5), dir.bins = 10, z.bins = NULL,
is.magcalm = TRUE, is.center = c(FALSE, TRUE, FALSE), with_lowbound = c(FALSE, FALSE, FALSE),
g.envir = envir, export.cli = FALSE)
```

`Data Plots` *Plots for Data Display and Characterisation*

Description

A set of plotting functions for displaying and characterizing the data, often from a time series: time series plot, scatterplot of two variables, logarithmic exceedance curve, histogram of occurrences, directional roses.

Usage

```
tserie.plot(dataset, tri = NULL, g.envir = envir, var.envir = varenvir,
graph.envir = graphenvir, saison = NA, with_points = FALSE, datesinflnm = FALSE, ...)
covar.plot(dataset, xvar.envir, yvar.envir, g.envir = envir, graph.envir = graphenvir,
dir.lim = c(0, 360), dir.tick = 22.5, dir.lab = 45, withdens = FALSE, ...)
logdep.plot(dataset, var.envir, res, coeff = 1, is.calm = TRUE, with_lowbound = FALSE,
g.envir = envir, graph.envir = graphenvir, ...)
histoapp.plot(dataset, var.envir, res, coeff = 1, is.calm = TRUE, with_lowbound = FALSE,
g.envir = envir, graph.envir = graphenvir, ...)
roseplot(dataset, var.envir, dir.envir, mag.bins, dir.res = 20, labels.dir = 0,
is.calm = TRUE, is.center = c(FALSE, TRUE), with_lowbound = c(FALSE, FALSE), pcmax = NULL,
delta.pc = NULL, g.envir = envir, graph.envir = graphenvir, palette = "RdYlGn")
roseplot.facet(rose, varname)
compar.plot(dataset, xvar.envir, yvar.envir, point1, point2, g.envir = envir,
graph.envir = graphenvir, scatterplot = TRUE, qqplot = FALSE, plotreg = FALSE, dp = 0.01, ...)
```

Arguments

`dataset`	A table (numeric matrix or data frame) containing the data (by column). For `tserie.plot`, it must have a column with the time. The safest is to have the time in julian days in first column. The data frame may also have a column named `"Horodate"` in POSIX format. Other columns include the main variable and possibly covariates.
`tri`	A list of POT or MAX data of class 'tri', as returned by the functions `triPOT`, `triMAX`or `dataset2tri`. If this argument is provided, the threshold and peaks (POT data) or maxima (MAX data) will be added to the plot of the time series. If `NULL`, just the time series is plotted.
`g.envir`	A list describing the global environment of the study: site name `$site`, duration of observations `K` (in years), origin `$origin` and time zone `$tz` of the time...
`var.envir, xvar.envir, yvar.envir, dir.envir`	A list regrouping the variable descriptions and properties. The variable description is a sub-list `var` containing its symbol `id`, its name `name`, its unit `unit`, the column number in the dataset `varcol`, the number of sequential observations per year `nu`, the number of significant digits `ndig`...
`graph.envir`	A list grouping different graphical options: which language (either `"FR"` or `"EN"`); should the plots be in colour (`color=TRUE`) or in black and white; should they be saved (`saveplots=TRUE`); in which format (to be chosen among `"wmf"`, `"emf"`, `"png"`, `"jpg"`, `"jpeg"`, `"bmp"`, `"tif"`, `"tiff"`, `"ps"`, `"eps"` or `"pdf"`); should they be magnified (use a numeric value for `cex`).
`saison`	If the time series has been declustered by block maximum within a season, a list with the name and time limits of the season.
`with_points`	Logical `TRUE` / `FALSE`. If `TRUE`, add points to the line at each data value.
`datesinflnm`	Logical `TRUE` / `FALSE`. If `TRUE`, add to the name of the exported plot file the initial and final dates pf the plot, in format YYYYMMDD.
`...`	Additional graphical parameters from function `par.graphe`.
`dir.lim`	The limits for the direction axis (whether it is x- or y-axis).
`dir.tick, dir.lab`	The step between two consecutive ticks / labels along the direction axis.
`withdens`	Logical `TRUE` / `FALSE`. If `TRUE`, the empirical bivariate density is estimated and visualized through a colour code for the points.

res	The resolution of the variable bins.
coeff	A coefficient applied to the percentage of the bins. Useful when studying a subsample, such as a directional sector that includes 30% of the total observations.
is.calm	Logical vector TRUE / FALSE, for each variable. If TRUE, the considered variable is considered to be a calm when it is 0.
with_lowbound	Logical vector TRUE / FALSE, for each variable. Defines how the bounds of the bins are. If TRUE, the lower bound of the range is included and the upper bound is not.
is.center	Logical vector TRUE / FALSE, for each variable. If TRUE, the bin limits given by mag.bins are considered as the center of the bins and not the bounds.
mag.bins	The sequency (vector of length 3 with minimum, maximum, increment) defining the bins of the variable magnitude (speed, height...). Note that the maximum of the sequency may be lower than the maximum of the variable.
dir.res	The resolution of the directional bins for the rose.
labels.dir	The direction (in °N) of the radial along which the magnitude labels should be drawn. Scalar between 0 and 360.
pcmax	The upper limit of the radial axis of the rose, i.e. the maximum percentage. Numeric between 0 and 1.
delta.pc	The step, in percentage, for the ticks of the radial axis of the rose. Numeric between 0 and 1.
palette	The palette to be used for the magnitude bins of the rose. See documentation of ggplot2 package.
rose	The rose plot object returned by function roseplot to be used for faceting.
point1	The name of the covariate used for rose faceting. It must be a column name of the input data frame.
point2	The name of the covariate used for rose faceting. It must be a column name of the input data frame.
scatterplot	Logical TRUE / FALSE. If TRUE, the scatterplot of both variables is provided.
qqplot	Logical TRUE / FALSE. If TRUE, the quantile-quantile plot of both variables is provided.
plotreg	Logical TRUE / FALSE. If TRUE, a linear regression is fitted and added to the plot.
dp	The step of quantile probabilities for the qqplot.

Details

tserie.plot may be used either for simple display of one variable with respect to time, or for showing the (POT or MAX) declustering with the i.i.d. sample of the peaks or maxima in addition to the sequential variable. logdep.plot and histoapp.plot characterize only one variable, plotting respectively the curve of exceedance frequency and the histogram of the occurrence frequency. covar.plot is a scatterplot useful for covariates. compar.plot proposes a scatterplot and a quantile-quantile plot for comparing the same kind of variable, from two different origins. roseplot is a directional rose for vectorial physical variables having a direction (waves, wind, current...). The rose can be faceted (display of multiple roses) against a covariate using roseplot.facet. For plots other than roses, the input data my be a numeric matrix. However good practice suggest using data frames. Data frames also allow for a POSIX-format column, that can be used for identifying months or seasons.

Value

These functions return a plot in a graphic device. Roses (functions `roseplot` and `roseplot.facet`) are displayed in a separate graphic device. These functions also return a list object, that contains the plot object characteristics (see package `ggplot2`).

Author(s)

Franck Mazas

See Also

`par.graphe`, `read.input`

Examples

```
data(Bastia)
envir <- list(site = "Bastia", K = 23, origin = "1899-12-30", tz = "UTC")

serie <- as.data.frame(Bastia)
serie$Horodate <- temps2horodate(Bastia[, 1], origin = envir$origin, tz = envir$tz)

Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m", varcol = 2))
Dir.envir <- list(var = list(id = "Dir", name = "Direction", unit = "°N"))
Tp.envir <- list(var = list(id = "Tp", name = "Tp", unit = "s"))
graphenvir <- list(lang = "EN", color = TRUE, saveplots = FALSE, fileformat = "png", cex = 1)

covar.plot(serie, xvar.envir = Dir.envir, yvar.envir = Hs.envir)
covar.plot(serie, xvar.envir = Dir.envir, yvar.envir = Tp.envir)
covar.plot(serie, xvar.envir = Tp.envir, yvar.envir = Hs.envir)

p.Hs <- roseplot(serie, var.envir = Hs.envir, dir.envir = Dir.envir, dir.res = 10,
mag.bins = seq(0, 4, 0.25), labels.dir = 0, g.envir = envir, graph.envir = graphenvir)
```

Dubai.SOT *SOT Data From A Location Off Dubai*

Description

Contains Time, Hs, Wave Direction and Storm Duration of hindcasted sea states above 1 meter for a loocation off Dubai.

Usage

```
data(Dubai.SOT)
```

Format

A data frame with 3557 observations on the following 4 variables.

`Time` A numeric vector.

`Hs` A numeric vector of the significant wave height in meters of the storms above 1 m.

`Dir` A numeric vector with the peak direction of the sea states.

`Dur` A numeric vector with the on-going storm duration.

Details

This data is to be sorted with SOrT function to identify storms and extract maxima for each considered directional sector. Location coordinates are (25°23'60.00"N, 54°17'60.00"E). The hindcast period is February 1958 - February 2008 (50 years).

Source

Hindcast by GlobOcean.

Examples

```
data(Dubai.SOT)
graphenvir <- list(lang = "EN", color = TRUE, saveplots = FALSE, fileformat = "png", cex = 1)
envir <- list(site = "Dubaï SOT", K = 50, origin = "1899-12-30", tz = "UTC")
Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m"))
Dir.envir <- list(var = list(id = "Dir", name = "Direction", unit = "°N"))
covar.plot(Dubai.SOT, xvar.envir = Dir.envir, yvar.envir = Hs.envir)
```

Exponential distribution

Functions for the Exponential distribution

Description

Density, cumulative distribution and quantile functions. Functions for random generation, quantile gradient, likelihood and derivatives.

Usage

```
dEXP(x, par = c(sigma, mu = 0), verif = TRUE)
pEXP(x, par = c(sigma, mu = 0), lower.tail = TRUE, verif = TRUE)
qEXP(p, par = c(sigma, mu = 0), lower.tail = TRUE, verif = TRUE)
rEXP(N, par = c(sigma, mu = 0), verif = TRUE)
EXP.grad.F(x, par, verif = FALSE)
EXP.hess.F(x, par, verif = FALSE)
EXP.grad.q(par, lambda, Tr = 100, verif = TRUE)
EXP.loglik(par, dat, verif = TRUE)
EXP.grad.loglik(par, dat, verif = TRUE)
EXP.hess.loglik(par, dat, verif = TRUE)
verif.EXP(par, dat = NULL, p = NULL)
```

Arguments

x	Vector of quantiles or support of the distribution.
p	Vector of probabilities.
par	Parameters of the distribution: scale parameter σ and location parameter μ.
N	The sample size or number of observations.
lambda	The Poisson rate parameter (mean number of events per year).
Tr	The return period, in years, associated to the quantile q.

dat	A sample to be fitted to the distribution.
lower.tail	Logical: if TRUE (default), probabilities are $\Pr[X \leq x]$, otherwise, $\Pr[X > x]$.
verif	Logical: if TRUE, arguments are checked by function verif.EXP.

Details

The cumulative distribution function of the Exponential distribution is given by:

$$F(x) = 1 - \exp\left(-\frac{x-\mu}{\sigma}\right)$$

with scale parameter σ and location parameter μ.

Value

dEXP returns the density of the quantiles x for Exponential distribution with parameters par. pEXP returns the non-exceedance probabilities (or exceedance probabilities if lower.tail = TRUE) of the quantiles x for Exponential distribution with parameters par. qEXP returns the quantiles associated to the probabilities p for Exponential distribution with parameters par. rEXP randomly generates a sample of size N for Exponential distribution with parameters par. EXP.grad.F returns the gradient of the cumulative distribution function, useful for the historical likelihood. EXP.hess.F returns the hessian of the cumulative distribution function, useful for the historical likelihood. EXP.grad.q returns the gradient of the quantile function, useful for computing the confidence intervals by the Delta method. EXP.loglik returns the likelihood of the sample dat following a Exponential distribution with parameters par. EXP.grad.loglik returns the gradient of the likelihood of the sample dat following a Exponential distribution with parameters par. EXP.hess.loglik returns the hessian of the likelihood of the sample dat following a Exponential distribution with parameters par. verif.EXP checks the arguments of the above functions and returns a logical value TRUE/FALSE. It checks whether the probabilities are in the range [0,1] and whether the parameters are within their interval of validity.

Author(s)

Franck Mazas

References

"Aide-mémoire statistique", ARTELIA

See Also

fit.MLE, fit.LMOM, fit.MOM, qIC.delta, qIC.boot, multiPOT

Examples

```
p <- seq(0, 1, 0.1)
x <- seq(0, 5, 0.2)
x2 <- seq(-5, 15, 1)
x3 <- seq(-5, 5, 0.25)
k1 <- -0.5
k2 <- 0.5
sigma1 <- 1.75
sigma2 <- 0.5
mu1 <- 0.1
mu2 <- 0.5
```

```
N <- 50

dEXP(x, par = sigma1)
dEXP(x, par = c(sigma = sigma1, mu = mu1))

pEXP(x, par = sigma1)
pEXP(x, par = c(sigma = sigma1, mu = mu1))

qEXP(p, par = sigma1)

Y2 <- rEXP(N, c(sigma = sigma1))

fit.MOM(Y2, "EXP", mu0 = TRUE)
fit.LMOM(Y2, "EXP", mu0 = TRUE)
fit.MLE(Y2, "EXP", optim.method = "Brent", lower = 0, upper = 10)

EXP.loglik(c(sigma = sigma1, mu = mu1), Y2)
EXP.grad.loglik(c(sigma = sigma1, mu = mu1), Y2)
EXP.hess.loglik(c(sigma = sigma1, mu = mu1), Y2)
```

`Fit Plots` *Plots for Fit Display and Characterization*

Description

A set of plotting functions for displaying and characterizing the fits to extreme peaks or maxima by a statistical distribution.

Usage

```
plots.fit(tri, var.envir = varenvir, tail = "upper", multiplots = TRUE, g.envir = envir,
Lois = NULL, saison = NA, graph.envir = graphenvir, coeff.p_empir = c(alpha = 0, beta = 1),
histo.arg = NA, Tr_Labels = FALSE, IC_Labels = FALSE, disable.saveplots = FALSE, ...)
dens.plot(fit.envir, loi, nbar = 20, var.envir = varenvir, graph.envir = graphenvir,
xlimit = NULL)
pp.plot(fit.envir, loi, graph.envir = graphenvir, colpoints = "darkblue")
qq.plot(fit.envir, loi, var.envir = varenvir, xylimit = NULL, graph.envir = graphenvir,
colpoints = "darkblue")
qT.plot(fit.envir, loi, coeff.p_empir = c(alpha = 0, beta = 1), histo.arg = NA,
var.envir = varenvir, xlimit = c(0.1, 100), ylimit = NULL, graph.envir = graphenvir,
colpoints = "darkblue", colpoints.histo = "darkred", pchpoints = 3, pchpoints.histo = 4,
print_titre = TRUE, Tr_Labels = FALSE, IC_Labels = FALSE)
```

Arguments

`tri`	A list of POT or MAX data of class 'tri', as returned by the functions `triPOT`, `triMAX`or `dataset2tri`.
`var.envir`	A list regrouping the variable descriptions and properties. It may contain a list of type `peaks.upper`, `peaks.lower` or `minima` with the declustered peaks / maxima and possibly fit details.
`tail`	The extrapolation tail (for POT approach only). If `"upper"`, selection of the maxima; if `"lower"`, selection of the minima.

multiplots	Logical TRUE/FALSE. If TRUE, four plots are displayed: the quantile vs. return period plot (qT.plot), the density plot (dens.plot), the probability-probability plot (pp.plot) and the quantile-quantile plot (qq.plot). If FALSE, only the first plot is displayed.
g.envir	A list describing the global environment of the study: site name $site, duration of observations K (in years), origin $origin and time zone $tz of the time...
Lois	A character vector with the statistical laws whose fit is to be plotted. There is one plotting window per law. Must match values of function liste.lois.
saison	If the time series has been declustered by block maximum within a season, a list with the name and time limits of the season.
graph.envir	A list grouping different graphical options: which language (either "FR" or "EN"); should the plots be in colour (color=TRUE) or in black and white; should they be saved (saveplots=TRUE); in which format (to be chosen among "wmf", "emf", "png", "jpg", "jpeg", "bmp", "tif", "tiff", "ps", "eps" or "pdf"); should they be magnified (use a numeric value for cex).
coeff.p_empir	The coefficients for the empirical plotting position formula. A numeric vector of length 2 c(alpha, beta. The epp formula is (i-alpha/(N+beta).
histo.arg	If historical data is used, a list of these data for plotting.
Tr_Labels	Logical TRUE/FALSE. If TRUE, a label is written with the quantile value for the set of return periods provided in var.envir.
IC_Labels	Logical TRUE/FALSE. If TRUE, labels are written with the bounds of the confidence intervals for the set of return periods provided in var.envir.
disable.saveplots	Logical. If TRUE, the plots are not saved, whatever the value of graph.envir$saveplots.
...	Other arguments to be passed for the functions qT.plot, dens.plot, qq.plot and pp.plot (see below).
fit.envir	A list of type peaks.upper, peaks.lower or minima with the declustered peaks / maxima and fit details. This list is a sub-list of var.envir.
loi	The statistical distribution whose fit is to be plotted. Must match a value of function liste.lois.
nbar	The number of bars for the histogram.
xlimit, xylimit, ylimit	The axis limits relative to the quantiles. For function qT.plot, xlimit is relative to the return period.
colpoints	The colour for the points of the plot.
colpoints.histo	The colour for plotting the historical data.
pchpoints	A pch parameter as defined by par(): the symbol to be used for plotting the peaks or maxima.
pchpoints.histo	A pch parameter as defined by par(): the symbol to be used for plotting the historical data.
print_titre	Logical: should the title of the qT-plot be printed?

Value

For the function plots.fit: one or several windows with either the qT-plot (if multiplots=FALSE or the four plots.

Author(s)

Franck Mazas

See Also

multiPOT

Examples

```
data(Bastia)
envir <- list(site = "Bastia", K = 23, origin = "1899-12-30", tz = "UTC1")
Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m", varcol = 2), Tr = c(1, 2, 5, 10, 20,
50, 100), conf.IC = 90, boot.iter = 1000)
graphenvir <- list(lang = "EN", color = TRUE, saveplots = TRUE, fileformat = "png", cex = 1)
Lois <- c("GPD", "WEI", "GAM", "EXP")

tri <- tri.POT(Bastia, up = 1.5, delta_t = 1, g.envir = envir, var.envir = Hs.envir, tail = "upper",
strict = FALSE, pbar = FALSE)
multiPOT(tri, us = 2.44, lois = Lois, est = "LMOM", IC_methode = "boot", envir, Hs.envir, graphenvir,
tail = "upper", smooth = FALSE)
plots.fit(tri, var.envir = Hs.envir, tail = "upper", g.envir = envir, Lois = Lois,
graph.envir = graphenvir)
graphenvir$xylimTplot <- list(NULL, c(0, 8))
plots.fit(tri, var.envir = Hs.envir, tail = "upper", g.envir = envir, Lois = Lois,
graph.envir = graphenvir, multiplots = FALSE, Tr_Labels = TRUE, IC_Labels = TRUE)
```

Gamma distribution *Functions for the Gamma distribution*

Description

Density, cumulative distribution and quantile functions. Functions for random generation, quantile gradient, likelihood and derivatives.

Usage

```
dGAM(x, par = c(k, sigma, mu = 0), verif = TRUE)
pGAM(x, par = c(k, sigma, mu = 0), lower.tail = TRUE, verif = TRUE)
qGAM(p, par = c(k, sigma, mu = 0), lower.tail = TRUE, verif = TRUE)
rGAM(N, par = c(k, sigma, mu = 0), verif = TRUE)
GAM.grad.F(x, par, verif = FALSE)
GAM.hess.F(x, par, verif = FALSE)
GAM.grad.q(par, lambda, Tr = 100, verif = TRUE)
GAM.loglik(par, dat, verif = TRUE)
GAM.grad.loglik(par, dat, verif = TRUE)
GAM.hess.loglik(par, dat, verif = TRUE)
verif.GAM(par, dat = NULL, p = NULL)
```

Arguments

x	Vector of quantiles or support of the distribution.
p	Vector of probabilities.
par	Parameters of the distribution: shape parameter k, scale parameter σ and location parameter μ.
N	The sample size or number of observations.
lambda	The Poisson rate parameter (mean number of events per year).
Tr	The return period, in years, associated to the quantile q.
dat	A sample to be fitted to the distribution.
lower.tail	Logical: if TRUE (default), probabilities are $\Pr[X \leq x]$, otherwise, $\Pr[X > x]$.
verif	Logical: if TRUE, arguments are checked by function verif.GAM.

Details

The cumulative distribution function of the Gamma distribution is given by:

$$F(x) = \frac{\gamma\left(\frac{x-\mu}{\sigma}\right)}{\Gamma(k)}$$

with shape parameter k, scale parameter σ and location parameter μ. $\Gamma()$ is the Gamma function and $\gamma()$ is the lower incomplete Gamma function.

Value

dGAM returns the density of the quantiles x for Gamma distribution with parameters par. pGAM returns the non-exceedance probabilities (or exceedance probabilities if lower.tail = TRUE) of the quantiles x for Gamma distribution with parameters par. qGAM returns the quantiles associated to the probabilities p for Gamma distribution with parameters par. rGAM randomly generates a sample of size N for Gamma distribution with parameters par. GAM.grad.F returns the gradient of the cumulative distribution function, useful for the historical likelihood. GAM.hess.F returns the hessian of the cumulative distribution function, useful for the historical likelihood. GAM.grad.q returns the gradient of the quantile function, useful for computing the confidence intervals by the Delta method. GAM.loglik returns the likelihood of the sample dat following a Gamma distribution with parameters par. GAM.grad.loglik returns the gradient of the likelihood of the sample dat following a Gamma distribution with parameters par. GAM.hess.loglik returns the hessian of the likelihood of the sample dat following a Gamma distribution with parameters par. verif.GAM checks the arguments of the above functions and returns a logical value TRUE/FALSE. It checks whether the probabilities are in the range [0,1] and whether the parameters are within their interval of validity.

Author(s)

Franck Mazas

References

"Aide-mémoire statistique", ARTELIA

See Also

fit.MLE, fit.LMOM, fit.MOM, qIC.delta, qIC.boot, multiPOT

Examples

```
parGAM <- c(k = 1.5, sigma = 1.5, mu = 2)
dGAM(seq(0, 8, 0.1), par = parGAM)       # Validé avec la fonction R dgamma
pGAM(seq(0, 8, 0.1), par = parGAM)       # Validé avec la fonction R pgamma
qGAM(seq(0, 1, 0.05), par = parGAM)      # Validé avec la fonction R qgamma
Y6 <- rGAM(500, parGAM)
fit.MOM(Y6, "GAM")
fit.MOM(Y6 - 2, "GAM", mu0 = TRUE)
fit.LMOM(Y6, "GAM")
fitGAM <- fit.MLE(Y6, "GAM", u = parGAM["mu"])#, optim.method = "Nelder-Mead")
GAM.loglik(Y6 - parGAM["mu"], par = fitGAM$par)
GAM.grad.loglik(Y6 - parGAM["mu"], par = fitGAM$par)
GAM.hess.loglik(Y6 - parGAM["mu"], par = fitGAM$par)
```

Generalized Extreme Value distribution

Functions for the Generalized Extreme Value distribution

Description

Density, cumulative distribution and quantile functions. Functions for random generation, quantile gradient, likelihood and derivatives.

Usage

```
dGEV(x, par = c(k = 0, sigma = 1, mu = 0), verif = TRUE)
pGEV(x, par = c(k = 0, sigma = 1, mu = 0), lower.tail = TRUE, verif = TRUE)
qGEV(p, par = c(k = 0, sigma = 1, mu = 0), lower.tail = TRUE, verif = TRUE)
rGEV(N, par = c(k = 0, sigma = 1, mu = 0), verif = TRUE)
GEV.grad.F(x, par, verif = FALSE)
GEV.hess.F(x, par, verif = FALSE)
GEV.grad.q(par, lambda, Tr = 100, verif = TRUE)
GEV.loglik(par, dat, verif = TRUE)
GEV.grad.loglik(par, dat, verif = TRUE)
GEV.hess.loglik(par, dat, verif = TRUE)
verif.GEV(par, dat = NULL, p = NULL)
```

Arguments

x	Vector of quantiles or support of the distribution.
p	Vector of probabilities.
par	Parameters of the distribution: shape parameter k, scale parameter σ and location parameter μ.
N	The sample size or number of observations.
lambda	The Poisson rate parameter (mean number of events per year).
Tr	The return period, in years, associated to the quantile q.
dat	A sample to be fitted to the distribution.
lower.tail	Logical: if TRUE (default), probabilities are $\Pr[X \leq x]$, otherwise, $\Pr[X > x]$.
verif	Logical: if TRUE, arguments are checked by function verif.GEV.

Details

The cumulative distribution function of the Generalized Extreme Value distribution is given by:

$$F(x) = \exp\left[-\left[1 + k\frac{x-\mu}{\sigma}\right]^{-1/k}\right]$$

with shape parameter k, scale parameter σ and location parameter μ.

Value

`dGEV` returns the density of the quantiles `x` for Generalized Extreme Value distribution with parameters `par`. `pGEV` returns the non-exceedance probabilities (or exceedance probabilities if `lower.tail = TRUE`) of the quantiles `x` for Generalized Extreme Value distribution with parameters `par`. `qGEV` returns the quantiles associated to the probabilities `p` for Generalized Extreme Value distribution with parameters `par`. `rGEV` randomly generates a sample of size `N` for Generalized Extreme Value distribution with parameters `par`. `GEV.grad.F` returns the gradient of the cumulative distribution function, useful for the historical likelihood. `GEV.hess.F` returns the hessian of the cumulative distribution function, useful for the historical likelihood. `GEV.grad.q` returns the gradient of the quantile function, useful for computing the confidence intervals by the Delta method. `GEV.loglik` returns the likelihood of the sample `dat` following a Generalized Extreme Value distribution with parameters `par`. `GEV.grad.loglik` returns the gradient of the likelihood of the sample `dat` following a Generalized Extreme Value distribution with parameters `par`. `GEV.hess.loglik` returns the hessian of the likelihood of the sample `dat` following a Generalized Extreme Value distribution with parameters `par`. `verif.GEV` checks the arguments of the above functions and returns a logical value `TRUE/FALSE`. It checks whether the probabilities are in the range [0,1] and whether the parameters are within their interval of validity.

Author(s)

Franck Mazas

References

"Aide-mémoire statistique", ARTELIA

See Also

fit.MLE, fit.LMOM, fit.MOM, qIC.delta, qIC.boot, multiPOT

Examples

```
p <- seq(0, 1, 0.1)
x <- seq(0, 5, 0.2)
x2 <- seq(-5, 15, 1)
x3 <- seq(-5, 5, 0.25)
k1 <- -0.5
k2 <- 0.5
sigma1 <- 1.75
sigma2 <- 0.5
mu1 <- 0.1
mu2 <- 0.5
N <- 50

dGEV(x3, c(k = k1, sigma = 1, mu = 0))
```

```
dGEV(x3, c(k = k2, sigma = 1, mu = 0))
dGEV(x3, c(k = 0, sigma = sigma1, mu = 0))
dGUM(x3, c(sigma = sigma1, mu = 0))

Y4 <- rGEV(500, c(k = -k1, sigma = sigma1, mu = 2))
fit.MOM(Y4, "GEV")
fit.LMOM(Y4, "GEV")
fit.MLE(Y4, "GEV")      # ATTENTION : ajustement MLE de la GEV non fiable
fit.MLE(Y4, "GEV", optim.method = "Nelder-Mead")
fit.MLE(Y4, "GEV", par.init = c(k = 0.1, sigma = sqrt(6 * var(Y4)) / pi,
mu = mean(Y4) - 0.57722 * sqrt(6 * var(Y4)) / pi), optim.method = "Nelder-Mead")
fit.MLE(Y4, "GEV", par.init = c(k = 0.1, sigma = sqrt(6 * var(Y4)) / pi,
mu = mean(Y4) - 0.57722 * sqrt(6 * var(Y4)) / pi), optim.method = "BFGS")
GEV.loglik(fit.MLE(Y4, "GEV", show = FALSE)$par, Y4)
GEV.grad.loglik(fit.MLE(Y4, "GEV", show = FALSE)$par, Y4)
GEV.hess.loglik(fit.MLE(Y4, "GEV", show = FALSE)$par, Y4)
```

`Generalized Pareto distribution`

Functions for the Generalized Pareto distribution

Description

Density, cumulative distribution and quantile functions. Functions for random generation, quantile gradient, likelihood and derivatives.

Usage

```
dGPD(x, par = c(k, sigma, mu = 0), verif = TRUE)
pGPD(x, par = c(k, sigma, mu = 0), lower.tail = TRUE, verif = TRUE)
qGPD(p, par = c(k, sigma, mu = 0), lower.tail = TRUE, verif = TRUE)
rGPD(N, par = c(k, sigma, mu = 0), verif = TRUE)
GPD.grad.F(x, par, verif = FALSE)
GPD.hess.F(x, par, verif = FALSE)
GPD.grad.q(par, lambda, Tr = 100, verif = TRUE)
GPD.loglik(par, dat, verif = TRUE)
GPD.grad.loglik(par, dat, verif = TRUE)
GPD.hess.loglik(par, dat, verif = TRUE)
verif.GPD(par, dat = NULL, p = NULL)
```

Arguments

`x`	Vector of quantiles or support of the distribution.
`p`	Vector of probabilities.
`par`	Parameters of the distribution: shape parameter k, scale parameter σ and location parameter μ.
`N`	The sample size or number of observations.
`lambda`	The Poisson rate parameter (mean number of events per year).
`Tr`	The return period, in years, associated to the quantile `q`.
`dat`	A sample to be fitted to the distribution.
`lower.tail`	Logical: if TRUE (default), probabilities are $\Pr[X \leq x]$, otherwise, $\Pr[X > x]$.
`verif`	Logical: if TRUE, arguments are checked by function `verif.GPD`.

Details

The cumulative distribution function of the Generalized Pareto distribution is given by:

$$F(x) = 1 - \left[1 + k\frac{x-\mu}{\sigma}\right]^{-1/k}$$

with shape parameter k, scale parameter σ and location parameter μ.

Value

`dGPD` returns the density of the quantiles `x` for Generalized Pareto distribution with parameters `par`. `pGPD` returns the non-exceedance probabilities (or exceedance probabilities if `lower.tail = TRUE`) of the quantiles `x` for Generalized Pareto distribution with parameters `par`. `qGPD` returns the quantiles associated to the probabilities `p` for Generalized Pareto distribution with parameters `par`. `rGPD` randomly generates a sample of size `N` for Generalized Pareto distribution with parameters `par`. `GPD.grad.F` returns the gradient of the cumulative distribution function, useful for the historical likelihood. `GPD.hess.F` returns the hessian of the cumulative distribution function, useful for the historical likelihood. `GPD.grad.q` returns the gradient of the quantile function, useful for computing the confidence intervals by the Delta method. `GPD.loglik` returns the likelihood of the sample `dat` following a Generalized Pareto distribution with parameters `par`. `GPD.grad.loglik` returns the gradient of the likelihood of the sample `dat` following a Generalized Pareto distribution with parameters `par`. `GPD.hess.loglik` returns the hessian of the likelihood of the sample `dat` following a Generalized Pareto distribution with parameters `par`. `verif.GPD` checks the arguments of the above functions and returns a logical value `TRUE / FALSE`. It checks whether the probabilities are in the range [0,1] and whether the parameters are within their interval of validity.

Author(s)

Franck Mazas

References

"Aide-mémoire statistique", ARTELIA

See Also

fit.MLE, fit.LMOM, fit.MOM, qIC.delta, qIC.boot, multiPOT

Examples

```
p <- seq(0, 1, 0.1)
x <- seq(0, 5, 0.2)
x2 <- seq(-5, 15, 1)
x3 <- seq(-5, 5, 0.25)
k1 <- -0.5
k2 <- 0.5
sigma1 <- 1.75
sigma2 <- 0.5
mu1 <- 0.1
mu2 <- 0.5
N <- 50

dGPD(x, c(k = k1, sigma = sigma1))
dGPD(x, c(k = k2, sigma = sigma1, mu = mu1))
```

```
pGPD(x, c(k = k1, sigma = sigma1))
pGPD(x, c(k = k2, sigma = sigma1))

qGPD(p, c(k = k1, sigma = sigma1))
qGPD(p, c(k = k2, sigma = sigma1))
qGPD(p, c(k = 0, sigma = sigma1))

Y <- rGPD(N, c(k = k1, sigma = sigma1))

GPD.loglik(c(k = k1, sigma = sigma1, mu = 0), Y)
GPD.grad.loglik(c(k = k1, sigma = sigma1, mu = 0), Y)
GPD.hess.loglik(c(k = k1, sigma = sigma1, mu = 0), Y)

fit.MOM(Y, "GPD")
fit.LMOM(Y, "GPD")
fit.MLE(Y, "GPD")
fit.MLE(Y, "GPD", optim.method = "Nelder-Mead")
test.chi2(fit.MLE(Y, "GPD", show = FALSE))
```

`Graphical Parameters` *Graphical Parameters for Plots Customization*

Description

A set of graphical parameters to adjust the settings of the plots.

Usage

```
par.graphe(marges = "blt", xlimit = NULL, ylimit = NULL, legend.loc = "topleft",
 cex = 1, cex.main = 1.75, cex.font = 1, cex.leg = 1.25,
 cex.lab = 1.75, cex.xlab = 1, cex.ylab = 1,
 cex.axis = 1.75, cex.xaxis = 1, cex.yaxis = 1, las = 1,
 col.line = "black", col.points = couleurs("ARTELIA")["Bleu"], col.box = "darkblue",
 lwd = 2, lty = 1,
 cex.points = 1, pch.points = 16, transp = 0.33,
 add = FALSE, disable.saveplots = FALSE)
couleurs(palette)
```

Arguments

`marges`	A combination of 'b' (bottom), 'l' (left), 't' (top) and 'r' right: for plot margins.
`xlimit`	The limits for the time axis, in POSIXct format. Vector of length 2.
`ylimit`	The limits for the y-axis. Vector of length 2.
`legend.loc`	The location of the legend from the list `"bottomright"`, `"bottom"`, `"bottomleft"`, `"left"`, `"topleft"`, `"top"`, `"topright"`, `"right"` and `"center"`.
`cex`	A numerical value giving the amount by which plotting text and symbols should be magnified relative to the default.
`cex.main`	Magnifying factor for main titles relative to the current setting of `cex`.
`cex.font`	Magnifying factor for text on the plot, such as `correlogram.plot` or `histoapp.plot`.
`cex.leg`	Magnifying factor for the legend.

`cex.lab, cex.xlab, cex.ylab`
: Magnifying factor for x and y labels, x label or y label relative to the current setting of `cex`.

`cex.axis, cex.xtick, cex.ytick`
: Magnifying factor for x and y axis, x axis and y axis annotation relative to the current setting of `cex`.

`las`
: The style of axis labels. 0: always parallel to the axis, 1: always horizontal, 2: always perpendicular to the axis, 4: always vertical.

`col.points, col.line, col.box`
: The color of the points, lines or boxes.

`lty`
: The line type. Line types can either be specified as an integer (0=blank, 1=solid (default), 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash) or as one of the character strings `"blank"`, `"solid"`, `"dashed"`, `"dotted"`, `"dotdash"`, `"longdash"`, or `"twodash"`, where `"blank"` uses 'invisible lines' (i.e., does not draw them).

`lwd`
: The line width, a positive number.

`transp`
: The transparency factor of the points of the scatter plot. Scalar between 0 and 1.

`add`
: Logical `TRUE/FALSE`. If `TRUE`, the line or points are added to an existing plot.

`disable.saveplots`
: Logical. If `TRUE`, the plots are not saved, whatever the value of `graph.envir$saveplots`.

`palette`
: The name of a list of predefined colours: - `"ARTELIA"`: the blue and green from ARTELIA graphic chart; - `"lois"`: default colours for statistical distributions; - `"mois"`: a set of 12 colours; - `"secteurs"`: a set of 8 colours.

Details

For each plot function detailed in `dataplots`, only some of the above arguments may be used. See also `help(par)` for further details.

Value

A list with list components returned invisibly for `par.graph`; a vector of colours for `couleurs`.

Author(s)

Franck Mazas

See Also

`dataplots`, `correlogram.plot`

Examples

```
data(Bastia)
envir <- list(site = "Bastia", K = 23, origin = "1899-12-30", tz = "UTC")

serie <- as.data.frame(Bastia)
serie$Horodate <- temps2horodate(Bastia[, 1], origin = envir$origin, tz = envir$tz)

Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m", varcol = 2))
Dir.envir <- list(var = list(id = "Dir", name = "Direction", unit = "°N"))
Tp.envir <- list(var = list(id = "Tp", name = "Tp", unit = "s"))
```

```
graphenvir <- list(lang = "EN", color = TRUE, saveplots = FALSE, fileformat = "png", cex = 1)

tserie.plot(serie, var.envir = Hs.envir, graph.envir = graphenvir)
tserie.plot(serie, var.envir = Hs.envir, graph.envir = graphenvir, col.line = "darkgreen",
cex.main = 1.25, cex.xtick = 0.8, cex.ytick = 0.8, cex.ylab = 0.75)

covar.plot(serie, xvar.envir = Dir.envir, yvar.envir = Hs.envir)
covar.plot(serie, xvar.envir = Dir.envir, yvar.envir = Hs.envir, transp = 1,
col.points = "royalblue", pch.points = 17)
```

`Gumbel distribution` *Functions for the Gumbel distribution*

Description

Density, cumulative distribution and quantile functions. Functions for random generation, quantile gradient, likelihood and derivatives.

Usage

```
dGUM(x, par = c(sigma = 1, mu = 0), verif = TRUE)
pGUM(x, par = c(sigma = 1, mu = 0), lower.tail = TRUE, verif = TRUE)
qGUM(p, par = c(sigma = 1, mu = 0), lower.tail = TRUE, verif = TRUE)
rGUM(N, par = c(sigma = 1, mu = 0), verif = TRUE)
GUM.grad.F(x, par, verif = FALSE)
GUM.hess.F(x, par, verif = FALSE)
GUM.grad.q(par, lambda, Tr = 100, verif = TRUE)
GUM.loglik(par, dat, verif = TRUE)
GUM.grad.loglik(par, dat, verif = TRUE)
GUM.hess.loglik(par, dat, verif = TRUE)
verif.GUM(par, p = NULL)
```

Arguments

`x`	Vector of quantiles or support of the distribution.
`p`	Vector of probabilities.
`par`	Parameters of the distribution: scale parameter σ and location parameter μ.
`N`	The sample size or number of observations.
`lambda`	The Poisson rate parameter (mean number of events per year).
`Tr`	The return period, in years, associated to the quantile `q`.
`dat`	A sample to be fitted to the distribution.
`lower.tail`	Logical: if TRUE (default), probabilities are $\Pr[X \leq x]$, otherwise, $\Pr[X > x]$.
`verif`	Logical: if TRUE, arguments are checked by function `verif.GUM`.

Details

The cumulative distribution function of the Gumbel distribution is given by:

$$F(x) = \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right]$$

with scale parameter σ and location parameter μ.

Value

dGUM returns the density of the quantiles x for Gumbel distribution with parameters par. pGUM returns the non-exceedance probabilities (or exceedance probabilities if lower.tail = TRUE) of the quantiles x for Gumbel distribution with parameters par. qGUM returns the quantiles associated to the probabilities p for Gumbel distribution with parameters par. rGUM randomly generates a sample of size N for Gumbel distribution with parameters par. GUM.grad.F returns the gradient of the cumulative distribution function, useful for the historical likelihood. GUM.hess.F returns the hessian of the cumulative distribution function, useful for the historical likelihood. GUM.grad.q returns the gradient of the quantile function, useful for computing the confidence intervals by the Delta method. GUM.loglik returns the likelihood of the sample dat following a Gumbel distribution with parameters par. GUM.grad.loglik returns the gradient of the likelihood of the sample dat following a Gumbel distribution with parameters par. GUM.hess.loglik returns the hessian of the likelihood of the sample dat following a Gumbel distribution with parameters par. verif.GUM checks the arguments of the above functions and returns a logical value TRUE/FALSE. It checks whether the probabilities are in the range [0,1] and whether the parameters are within their interval of validity.

Author(s)

Franck Mazas

References

"Aide-mémoire statistique", ARTELIA

See Also

fit.MLE, fit.LMOM, fit.MOM, qIC.delta, qIC.boot, multiPOT

Examples

```
p <- seq(0, 1, 0.1)
x <- seq(0, 5, 0.2)
x2 <- seq(-5, 15, 1)
x3 <- seq(-5, 5, 0.25)
k1 <- -0.5
k2 <- 0.5
sigma1 <- 1.75
sigma2 <- 0.5
mu1 <- 0.1
mu2 <- 0.5
N <- 50

dGUM(x2, c(sigma = sigma2, mu = mu2))
pGUM(x2, c(sigma = sigma2, mu = mu2))

Y3 <- rGUM(N, c(sigma = sigma2, mu = 2))
fit.MOM(Y3, "GUM")
fit.LMOM(Y3, "GUM")
fit.MLE(Y3, "GUM")
fit.MLE(Y3, "GUM", optim.method = "Nelder-Mead")
GUM.loglik(fit.MLE(Y3, "GUM", show = FALSE)$par, Y3)
GUM.grad.loglik(fit.MLE(Y3, "GUM", show = FALSE)$par, Y3)
GUM.hess.loglik(fit.MLE(Y3, "GUM", show = FALSE)$par, Y3)
```

multiPOT *Fitting of Extreme POT Values to Distributions*

Description

Fits POT data to distributions. Three estimators are available: Maximum Likelihood Estimator, L-moments estimator with Landwehr's unbiased formula (Landwehr *et al.*, 1979a) et the Method of Moments.

Usage

```
multiPOT(tri, us, var.envir, lois = c("GPD", "WEI", "GAM"), est = "LMOM", IC_methode = "boot",
g.envir = envir, graph.envir = graphenvir, tail = "upper", smooth = FALSE, round_quant = TRUE,
show = FALSE, ...)
```

Arguments

tri	A list of POT data of class 'tri', as returned by the functions triPOT or dataset2tri.
us	Numeric. The statistical threshold u_s above which the peak exceedances are fitted to the dsitributions.
lois	The distributions to fit to the sample of exceedances over threshold. To be chosen among "GPD", "WEI", "EXP", "GAM", "GEV" and "GUM".
est	The estimator: to be chosen among "MLE" (Maximum Likelihood Estimator), "LMOM" (L-moments) or "MOM" (Methods of Moments).
IC_methode	The method used for computing the confidence intervals. To be chosen among "boot" or "delta".
g.envir	A list describing the global environment of the study: site name, duration of observations K (in years), origin and time zone of the time...
var.envir	A list regrouping the variable descriptions and properties. The variable description is a sub-list var containing its symbol id, its name name, its unit unit, the column number in the dataset varcol, the number of sequential observations per year nu, the number of significant digits ndig... It must also include the numeric of return periods Tr, the level of confidence interval conf.IC, if needed the number of bootstrap iterations boot.iter...
graph.envir	A list grouping different graphical options: which language (either "FR" or "EN"); should the plots be in colour (color=TRUE) or in black and white; should they be saved (saveplots=TRUE); in which format (to be chosen among "wmf", "emf", "png", "jpg", "jpeg", "bmp", "tif", "tiff", "ps", "eps" or "pdf"); should they be magnified (use a numeric value for cex).
tail	For extrapolation of extreme maximal values, tail="upper". For extrapolation of extreme minimal values, tail="lower".
smooth	Logical TRUE / FALSE. If TRUE, the dataset is smoothed using the function dataset.smooth with default arguments.
round_quant	Logical TRUE / FALSE. Should the computed quantiles be rounded to the number of significant digits of the variable var$ndig?
show	Logical. If TRUE, print details of the fits on the R console.
...	Additional arguments for function fit.MLE

Details

If the MLE is chosen, two-parameter distributions are fitted, i.e. without location parameters. With L-moments estimator, this third parameter is also estimated.

Value

The function returns a list of the type `peak.envir`, either `peaks.upper` or peaks.lower, depending on the tail that is extrapolated. It is attached to the variable list `var.envir`. The list includes:

`tail`	Character. The extrapolated tail, either `"upper"` or `"lower"`.
`up`	Numeric. The physical threshold.
`Xp`	Numeric vector of the peaks above the physical threshold u_p.
`Np`	The number of peaks above the physical threshold u_p.
`lambda_p`	The mean number of peaks above the physical threshold u_p per year, with $\lambda_p = N_p/K$.
`us`	Numeric. The statistical threshold u_s above which the fits are carried out.
`X`	Numeric. The sample of the peaks above the statistical threshold u_s.
`Y`	Numeric. The sample of the exceedances above the statistical threshold u_s: $Y = X - u_s$.
`N`	Numeric. The number of exceedances above u_s.
`lambda`	Numeric. The mean number of peaks above the statistical threshold u_s per year, with $\lambda = N/K$.
`lois`	The distributions fitted to the peak exceedances.
`OKfit`	Logical `TRUE` / `FALSE`. Specifies whether the fit succeeded.
`est`	Character string. The estimator used for each distribution.
`crit`	Numeric. A 3 row matrix providing BIC and AIC criteria and the Kolmogorov-Smirnov p-value for each fit.
`quantiles`	Numeric matrix. For each fit, returns the quantiles corresponding to the return period `var.envir$Tr`.
`param`	Numeric 3-row matrix. A 3x3 matrix returning the shape, scale and location parameters for each fit.
`Lmoments`	Numeric 9x2 matrix. Returns the following L-moments of the sample `Y`: λ_1, λ_2, τ_3, τ_4, τ_5, LCV, λ_3, λ_4, λ_5. First column returns Landwehr's unbiased estimate (Landwehr *et al.*, 1979a); second column returns the estimate based on Landwehr's plotting position formula (Landwehr *et al.*, 1979b).
`smooth`	Logical. If `"TRUE"`, the dataset has been smoothed.

Author(s)

Franck Mazas

References

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

See Also

`us.range`, `dataset.smooth`, `fit.MLE`, `fit.LMOM`

Examples

```
data(Bastia)
envir <- list(site = "Bastia", K = 23, origin = "1899-12-30", tz = "UTC1")
Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m", varcol = 2), Tr = c(1, 2, 5, 10, 20,
50, 100), conf.IC = 90, boot.iter = 1000)
graphenvir <- list(lang = "EN", color = TRUE, saveplots = TRUE, fileformat = "png", cex = 1)
Lois <- c("GPD", "WEI", "GAM", "EXP")

tri <- tri.POT(Bastia, up = 1.5, delta_t = 1, g.envir = envir, var.envir = Hs.envir, tail = "upper",
strict = FALSE, pbar = FALSE)
multiPOT(tri, us = 2.44, lois = Lois, est = "LMOM", IC_methode = "boot", envir, Hs.envir, graphenvir,
tail = "upper", smooth = FALSE)
```

Negative Binomial distribution

Functions for the Negative Binomial distribution

Description

Probability mass, cumulative distribution and quantile functions. Functions for random generation, likelihood and derivatives.

Usage

```
dNBI(x, par = c(p, k), verif = TRUE)
pNBI(x, par = c(p, k), lower.tail = TRUE, verif = TRUE)
qNBI(p, par = c(p, k), lower.tail = TRUE, verif = TRUE)
rNBI(N, par = c(p, k), verif = TRUE)
NBI.loglik(par, dat, verif = TRUE)
NBI.grad.loglik(par, dat, verif = TRUE)
NBI.hess.loglik(par, dat, verif = TRUE)
verif.NBI(par, dat = NULL, p = NULL)
```

Arguments

x	Vector of quantiles or support of the distribution (integers).
p	Vector of probabilities.
par	Parameters of the distribution: p success probability in each experiment and $k > 0$ number of expected successes.
N	The sample size or number of observations.
dat	A sample to be fitted to the distribution.
lower.tail	Logical: if TRUE (default), probabilities are $\Pr[X \leq x]$, otherwise, $\Pr[X > x]$.
verif	Logical: if TRUE, arguments are checked by function verif.NBI.

Details

The probability mass function of the Negative Binomial distribution is given by:

Value

dNBI returns the probability of the quantiles x for Negative Binomial distribution with parameters par. pNBI returns the non-exceedance probabilities (or exceedance probabilities if lower.tail = TRUE) of the quantiles x for Negative Binomial distribution with parameters par. qNBI returns the quantiles associated to the probabilities p for Negative Binomial distribution with parameters par. rNBI randomly generates a sample of size N for Negative Binomial distribution with parameters par. NBI.loglik returns the likelihood of the sample dat following a Negative Binomial distribution with parameters par. NBI.grad.loglik returns the gradient of the likelihood of the sample dat following a Negative Binomial distribution with parameters par. NBI.hess.loglik returns the hessian of the likelihood of the sample dat following a Negative Binomial distribution with parameters par. verif.NBI checks the arguments of the above functions and returns a logical value TRUE/FALSE. It checks whether the probabilities are in the range [0,1] and whether the parameters are within their interval of validity.

Author(s)

Franck Mazas

References

"Aide-mémoire statistique", ARTELIA

See Also

fit.MLE, fit.LMOM, fit.MOM, qIC.delta, qIC.boot, multiPOT

Examples

```
parNBI <- c(p = 0.7, k = 5)
dNBI(0:10, parNBI)
pNBI(0:10, parNBI)
qNBI(seq(0, 1, 0.1), parNBI)
Y8 <- rNBI(500, parNBI)
fit.MOM(Y8, "NBI")
fit.MLE(Y8, "NBI")
fitNBI <- fit.MLE(Y8, "NBI", optim.method = "Nelder-Mead")
NBI.loglik(fitNBI$par, Y8)
NBI.grad.loglik(fitNBI$par, Y8)
NBI.hess.loglik(fitNBI$par, Y8)
```

Poisson distribution *Functions for the Poisson distribution*

Description

Probability mass, cumulative distribution and quantile functions. Functions for random generation, likelihood and derivatives.

Usage

```
dPOI(x, par = c(lambda = 1), verif = TRUE)
pPOI(x, par = c(lambda = 1), lower.tail = TRUE, verif = TRUE)
qPOI(p, par = c(lambda = 1), lower.tail = TRUE, verif = TRUE)
rPOI(N, par = c(lambda = 1), verif = TRUE)
POI.loglik(par, dat, verif = TRUE)
POI.grad.loglik(par, dat, verif = TRUE)
POI.hess.loglik(par, dat, verif = TRUE)
verif.POI(par, dat = NULL, p = NULL)
```

Arguments

`x`	Vector of quantiles or support of the distribution (integers).
`p`	Vector of probabilities.
`par`	Parameters of the distribution: λ rate parameter.
`N`	The sample size or number of observations.
`dat`	A sample to be fitted to the distribution.
`lower.tail`	Logical: if TRUE (default), probabilities are $\Pr[X \leq x]$, otherwise, $\Pr[X > x]$.
`verif`	Logical: if TRUE, arguments are checked by function `verif.POI`.

Details

The probability mass function of the Poisson distribution is given by:

$$f(n) = \exp(-\lambda)\frac{\lambda^n}{n!}$$

Value

`dPOI` returns the probability of the quantiles `x` for Poisson distribution with parameters `par`. `pPOI` returns the non-exceedance probabilities (or exceedance probabilities if `lower.tail = TRUE`) of the quantiles `x` for Poisson distribution with parameters `par`. `qPOI` returns the quantiles associated to the probabilities `p` for Poisson distribution with parameters `par`. `rPOI` randomly generates a sample of size `N` for Poisson distribution with parameters `par`. `POI.loglik` returns the likelihood of the sample `dat` following a Poisson distribution with parameters `par`. `POI.grad.loglik` returns the gradient of the likelihood of the sample `dat` following a Poisson distribution with parameters `par`. `POI.hess.loglik` returns the hessian of the likelihood of the sample `dat` following a Poisson distribution with parameters `par`. `verif.POI` checks the arguments of the above functions and returns a logical value `TRUE/FALSE`. It checks whether the probabilities are in the range [0,1] and whether the parameters are within their interval of validity.

Author(s)

Franck Mazas

References

"Aide-mémoire statistique", ARTELIA

See Also

fit.MLE, fit.LMOM, fit.MOM, qIC.delta, qIC.boot, multiPOT

Examples

```
parPOI <- 5
dPOI(0:10, parPOI)
pPOI(0:10, parPOI)
qPOI(seq(0, 1, 0.1), parPOI)
Y7 <- rPOI(50, parPOI)
fit.MOM(Y7, "POI")
fit.MLE(Y7, "POI")
fit.MLE(Y7, "POI", optim.method = "Brent", lower = 0, upper = 100)
(fit.MLE(Y7, "POI", show = FALSE)$par - mean(Y7)) == 0   # TRUE -> OK
POI.loglik(par = fit.MOM(Y7, "POI", show = FALSE)$par, Y7)
POI.grad.loglik(par = fit.MOM(Y7, "POI", show = FALSE)$par, Y7)
POI.hess.loglik(par = fit.MOM(Y7, "POI", show = FALSE)$par, Y7)
```

---

read.input *Reads Input Data Files*

Description

This function reads a file containing a table of input data (generally a time series).

Usage

```
read.input(file, g.envir = envir, header = NULL, col.nums = NULL, col.noms = NULL,
flag.NA = NULL, sep = "", col2horodate = 1, temps2julien = 1,
formathorodate = NULL, skip = 0, rounding = NULL, silent = FALSE, ...)
```

Arguments

file	Path of the input file.
g.envir	A list describing the global environment of the study: site name, duration of observations K (in years), origin and time zone of the time...
header	Either a logical value indicating whether the file has a header line, or the default NULL value for automatic detection.
col.nums	A numeric vector with the column numbers of the input table to import. If NULL, all the columns are imported and returned.
col.noms	A character vector with the names of the columns to set. If NULL, the headers of the files, if present, are used.
flag.NA	The flag used for the NA data in the input table. They are transformed into NA data in the output table.
sep	The field separator character. Values on each line of the file are separated by this character. If sep = "" (the default for read.table) the separator is "white space", that is one or more spaces, tabs, newlines or carriage returns.
col2horodate	The index(es) of the column(s) containing the time information and used for returning a timestamp ('Horodate') column in POSIX format added to the table. If NULL, nothing is done, or no such column. If formathorodate is specified, then the time information is a date or date-hour, either in character string or numeric (e.g. SWAN output). It is usually provided in a single column, but possibly in several. Then it is transformed into julian days (origin and time

	zone provided by g.envir) and a timestamp ('Horodate') column is binded to the table. If formathorodate is not specified, two possibilities. If the time information is provided in a single column, then it is a duration from an origin (julian days, hours, seconds...) and this column is made of numeric scalars. A transformation in julian days (origin and time zone provided by g.envir) is made, if necessary (see temps2julien) and a timestamp ('Horodate') column is binded to the table. If the input table has distinct columns for at least the year, month and day and possibly the hours, minutes and seconds, the argument is the column indexes of these, in this order, and its length is between 3 and 6.
temps2julien	To be used if the input file has a single time column that is not in julian days. This argument is the conversion factor between days and the unit of the time column. Use 86400 for seconds, 1440 for minutes, 24 for hours... col2horodate must be specified and of length 1.
formathorodate	The format of the date-hour (character string), if needed (see col2horodate). It must be a single character string, even when the horodate is provided in more than 1 column (use a blank " " to separate). See the documentation of function strptime for examples. Default value is "%Y-%m-%d %H:%M:%S".
skip	The number of lines to skip before beginning to read the file.
rounding	This argument is used for rounding the columns of the table. Nothing is done if NULL (by default). Otherwise, a numeric vector of length n with the number of digits four rounding the n-th first columns of the table. Use NA for not rounding a particular column.
silent	Logical TRUE / FALSE. If FALSE, information about the input series are printed in the console.
...	Additional arguments relative to the function read.table, such as dec, comment.char...

Details

The input file must be a table of data, possibly with a header line. If there are no headers, the column names can be set thanks to the col.noms argument. For *.wve, *.wco and *.ven files, if col.noms == NULL, column names are automatically set. If the durations of the time series is not set in the g.envir list, it is computed and the list is modified in the global environment. The statistics (number of data and quantiles) of the variables are printed in the R console.

Value

A data frame of the data. If col2horodate is not null, then a timestamp (Horodate) column is binded to the table.

Author(s)

Franck Mazas

Examples

```
# Cas simple : Tous arguments par defaut - Selection du fichier via le navigateur
#envir <- list(site = "", K = NA, origin = "1899-12-30", tz = "UTC")
#serie <- read.input(file.choose(), g.envir = envir)
#head(serie)
#envir

# Separateur point-virgule - Assignation des noms de colonnes
```

```
envir <- list(site = "", K = NA, origin = "1899-12-30", tz = "UTC")
serie <- read.input(file = system.file("extdata", "Bastia_Carbonite.csv",  package = "artextreme"),
g.envir = envir, sep = ";", col.noms = c("Temps", "Hs", "Dir", "Tp"))
head(serie)
envir

# Temps fourni sous 4 colonnes - Assignation des noms de colonnes (fichier GlobOcean)
envir <- list(site = "Hamrawein", point = "E3425N2625", K = NA, origin = "1899-12-30", tz = "UTC")
systems <- c("Global", "Ww", "Sw1", "Sw2")
params <- c("Hm0", "Tm02", "Tps", "Dirm", "Dirp", "Gamma", "s")
colonnes <- c("Year", "Month", "Day", "Hour", c(t(outer(systems, params, FUN = paste))), "Ws", "Wdir")
serie <- read.input(system.file("extdata", "E3425N2625.txt",  package = "artextreme"), g.envir = envir,
col.noms = colonnes, flag.NA = NULL, col2horodate = 1:4)
head(serie)
envir

# Importation de fichiers annuels de sortie SWAN (*.pnt)
envir <- list(site = "Tibar", K = 2, origin = "1899-12-30", tz = "UTC")
colonnes <- c("Temps", "Hm0", "Dirm", "Tp", "Dirp", "Tps", "s")
Annees <- 1992:1993
serie <- c()
for (annee in Annees) {
filename <- system.file("extdata", paste0("Tibar_", annee, "_EP.pnt"),  package = "artextreme")
serie0 <- read.input(file = filename, g.envir = envir, header = FALSE, col.nums = 3:9,
col.noms = colonnes, col2horodate = 1, formathorodate = "%Y%m%d.%H%M%S", skip = 7,
rounding = c(NA, 2, 1, 1, 0, 2, 0))
serie <- rbind(serie, serie0)
rm(serie0)
}
head(serie)

# Importation partielle des colonnes - Temps fourni en secondes
envir <- list(site = "", K = NA, origin = "2016-01-01", tz = "UTC")
serie <- read.input(file = system.file("extdata", "Martinique_courant.txt",  package = "artextreme"),
g.envir = envir, col.nums = c(1, 2, 6), col.noms = c("Temps", "U_surface", "V_surface"),
temps2julien = 86400)
head(serie)
envir

# Temps fourni sous forme d'une horodate
envir <- list(site = "Paris-Austerlitz", K = NA, origin = "1885-10-01", tz = "UTC")
serie <- read.input(system.file("extdata", "Paris_Austerlitz_QIX.csv",  package = "artextreme"),
g.envir = envir, col2horodate = 1, sep = ",", col.noms = c("Temps", "QIX"),
formathorodate = "%Y-%m-%d %H:%M:%S")
head(serie)
envir
```

`tri.POT` *Peaks-Over-Threshold Declustering of Time Series*

Description

This function performs POT declustering of a time series. An event is defined by the exceedance over a physical threshold u_p and it is described by its peak.

Usage

```
tri.POT(dataset, up, var.envir, g.envir = envir, delta_t = NA, delta_t_pics = NA,
alpha = NA, dur.min = NA, strict = TRUE, tail = "upper", pbar = TRUE)
```

Arguments

dataset	Numeric matrix or data frame. A time series of the sequential observations of the variable to decluster. It can include covariates in other columns. Julian time (in days) must be in first column (see Details section). The column of the variable must be given in var.envirvarvarcol.
up	The physical threshold u_p for event (storm) identification and declustering.
var.envir	A list regrouping the variable descriptions and properties. The variable description is a sub-list var containing its symbol id, its name name, its unit unit, the column number in the dataset varcol, the number of sequential observations per year nu, the number of significant digits ndig...
g.envir	A list describing the global environment of the study: site name $site, duration of observations K (in years), origin $origin and time zone $tz of the time...
delta_t	Criterion of temporal separation of clusters (events / storms). A fluctuation shorter than delta_t is allowed within a cluster. Also useful in case of irregular time sries: if two successive values are separated by more than delta_t, they cannot belong to the same event. Numeric value, in days. If the criterion is not used, delta_t = NA.
delta_t_pics	Criterion of temporal separation of cluster peaks. Two peaks must be separated by a time interval longer than delta_t_pics, otherwise the highest peak only is kept. If the criterion is not used, delta_t_pics = NA.
alpha	Criterion of down-crossing event separation. The end of a cluster is determined by the down-crossing of $\alpha * u_p$. alpha is thus a fraction of the physical threshold. When the sequential values of the variable stay between $\alpha * u_p$ and u_p, this is considered as a simple fluctuation. If the criterion is not used, alpha = NA.
dur.min	Criterion of minimal event duration. A cluster shorter than dur.min is not accepted. If the criterion is not used, dur.min = NA.
strict	Logical TRUE/FALSE. Should the peaks be strictly above up?
tail	The extrapolation tail (for POT approach only). If "upper", selection of the maxima; if "lower", selection of the minima.
pbar	Logical TRUE/FALSE. Should a progress bar be displayed during the declustering?

Details

The time is in column 1, in julian days. The origin and time zone must be specified in the list g.envir. It is also possible to add a first column named "Horodate", in POSIXlt format (e.g. from function temps2horodate). If so, julian time must be in column 2. The variable to decluster must be in the column var.envirvarvarcol. It is possible to add covariates.

The parameters granting independence: delta_t, delta_t_pics, alpha are not to be used together. If so, the algorithm works as follows: 1. Identification of a new cluster by 1st exeedance of the physical threshold u_p 2. For the following values: 2.1 Check of the criterion delta_t 2.2 Check of the criterion alpha 3. Extraction of cluster peaks 4. Check of the criterion delta_t_pics 5. Check of the criterion dur.min

Value

A list of class `triPOT` with the following elements:

`up`	The physical threshold.
`Np`	The number of peaks.
`lambda_p`	The mean number of peaks per year $\lambda_p = N_p/K$.
`pics`	A Np-row matrix listing the peaks, the associated timestamp and (numeric) covariates, the cluster size and duration, and the number of days since the beginning of the time series.
`occurrence`	A matrix returning the number of peaks for each year.

This list is attached to `var.envir` as a sub-list, called either `peaks.upper` or `peaks.lower`, depending on the tail.

Author(s)

Franck Mazas

References

Bernardara P., Mazas F., Kergadallan X., Hamm L., 2014. A two-step framework for over-threshold threshold modelling of environmental extremes. *Natural Hazards and Earth System Sciences* 14, 635-647.

See Also

`tri.MAX`

Examples

```
data(Bastia)
envir <- list(site = "Bastia", K = 23, origin = "1899-12-30", tz = "UTC1")
Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m", varcol = 2),
Tr = c(1, 2, 5, 10, 20, 50, 100), conf.IC = 90, boot.iter = 1000)

tri <- tri.POT(Bastia, up = 1.5, delta_t = 1, g.envir = envir, var.envir = Hs.envir,
tail = "upper", strict = FALSE, pbar = FALSE)
```

---

us.range *Determination of the statistical threshold*

Description

This function provides tools to help the user to choose the statistical threshold in a sample. The sample is sorted in the ascending order and the threshold u_s is set at each (unique) observation. Properties and tests are carried out for each value and transformed into plots.

Usage

```
us.range(tri, var.envir, lois = c("GPD", "WEI", "GAM"), est = "LMOM",
g.envir = envir, graph.envir = graphenvir, bornes = "auto", N.min = 10,
tail = "upper", Tr.plot = 100, smooth = FALSE, us = NULL, identify = FALSE,
remove.max = FALSE, disable.saveplots = FALSE, show = FALSE, qlim = NULL,
mle.method = "Newton-Raphson", ...)
```

Arguments

`tri`	A list of POT data of class 'tri', as returned by the functions `triPOT` or `dataset2tri`.
`lois`	The distributions to fit to the sample of exceedances over threshold. To be chosen among `"GPD"`, `"WEI"`, `"EXP"`, `"GAM"`, `"GEV"` and `"GUM"`.
`est`	The estimator: to be chosen among `"MLE"` (Maximum Likelihood Estimator), `"LMOM"` (L-moments) or `"MOM"` (Methods of Moments).
`g.envir`	A list describing the global environment of the study: site name `$site`, duration of observations `K` (in years), origin `$origin` and time zone `$tz` of the time...
`var.envir`	A list regrouping the variable descriptions and properties. The variable description is a sub-list `var` containing its symbol `id`, its name `name`, its unit `unit`, the column number in the dataset `varcol`, the number of sequential observations per year `nu`, the number of significant digits `ndig`...
`graph.envir`	A list grouping different graphical options: which language (either `"FR"` or `"EN"`); should the plots be in colour (`color=TRUE`) or in black and white; should they be saved (`saveplots=TRUE`); in which format (to be chosen among `"wmf"`, `"emf"`, `"png"`, `"jpg"`, `"jpeg"`, `"bmp"`, `"tif"`, `"tiff"`, `"ps"`, `"eps"` or `"pdf"`); should they be magnified (use a numeric value for `cex`).
`bornes`	The ranges of the interval within which the statistical threshold varies. If `"auto"`, they are automatically computed: the minimum tested threshold value is the physical threshold u_p and the maximum value is the K-th largest value of the dataset, where `envir$K` is the time series duration. Otherwhise, a numeric vector of size 2: `c(us_min, us_max)`.
`N.min`	The minimum sample size. If the number of exceedances over the tested threshold is lower, the sensitivity study stops.
`tail`	For extrapolation of extreme maximal values, `tail="upper"`. For extrapolation of extreme minimal values, `tail="lower"`.
`Tr.plot`	The return period (in years) to be used for the plot displaying the evolution of the quantiles with resoect to u_s.
`smooth`	Logical `TRUE` / `FALSE`. If `TRUE`, the dataset is smoothed using the function `dataset.smooth` with default arguments.
`us`	Null or numeric. If numeric, the value of the statistical threshold to be plotted on graphics.
`identify`	Logical. If `TRUE`, graphical function `identify` is enabled for the L-moments plot.
`remove.max`	Logical. If `TRUE`, the largest peak value of the dataset is not accounted for threshold selection. Useful in case of outlier.
`disable.saveplots`	Logical. If `TRUE`, the plots are not saved, whatever the value of `graph.envir$saveplots`.
`show`	Logical. If `TRUE` and if MLE has been chosen, print details of each fit.

qlim	Either NULL or numeric vector of length 2 providing the y-limits for the quantile / threshold plot.
mle.method	The numerical method for maximization of the model likelihood when the ML estimator is used. Default is "Newton-Raphson", otherwise the classical methods of function fit.MLE are used (see also function optim).
...	Additional arguments for function fit.MLE

Details

Use of MLE along with two-parameter distributions implies to be very careful about the thresholds used. Thus, only thresholds meeting data values (and excluding them) are studied, as they give better estimates. When the L-moments estimator is used, a three-parameter GPD is estimated, although the location parameter is not displayed here.

Value

A variety of plots are produced, helping the user to choose the optimal threshold u_s. First panel gives the evolution of the sample: sample size, χ_2 statistic for the fit of Poisson and Negative Binomial distributions to the occurrence process, mean excess life plot. Second panel provides the evolution of the fits of the distributions to the exceedances over the threshold: evolution of Tr-year quantile, χ_2 statistic of the fit, p-value of the Kolmogorov-Smirnov test. The evolution of the GPD parameters k (shape parameter) and σ^* (modified scale parameter) is provided, along with their confidence intervals. If the estimator is the L-moments, the L-moments plot is returned. If the estimator is MLE, a panel with the evolution of AIC, BIC and their normalized values is returned.

Author(s)

Franck Mazas

References

Bernardara P., Mazas F., Kergadallan X., Hamm L., 2014. A two-step framework for over-threshold threshold modelling of environmental extremes. *Natural Hazards and Earth System Sciences* 14, 635-647. Mazas F., Hamm L., 2011. A multi-distribution approach to POT methods for determining extreme wave heights. *Coastal Engineering* 58, 385-394.

See Also

multiPOT, fit.MLE, fit.LMOM

Examples

```
data(Bastia)
envir <- list(site = "Bastia", K = 23, origin = "1899-12-30", tz = "UTC1")
Hs.envir <- list(var = list(id = "Hs", name = "Hs", unit = "m", varcol = 2),
Tr = c(1, 2, 5, 10, 20, 50, 100), conf.IC = 90, boot.iter = 1000)
graphenvir <- list(lang = "EN", color = TRUE, saveplots = FALSE, fileformat = "png", cex = 1)
Lois <- c("GPD", "WEI", "GAM", "EXP")

tri <- tri.POT(Bastia, up = 1.5, delta_t = 1, g.envir = envir, var.envir = Hs.envir, tail = "upper",
strict = FALSE, pbar = FALSE)

us.range(tri = tri, lois = Lois, var.envir = Hs.envir)
```

`Weibull distribution` *Functions for the Weibull distribution*

Description

Density, cumulative distribution and quantile functions. Functions for random generation, quantile gradient, likelihood and derivatives.

Usage

```
dWEI(x, par = c(k, sigma, mu = 0), verif = TRUE)
pWEI(x, par = c(k, sigma, mu = 0), lower.tail = TRUE, verif = TRUE)
qWEI(p, par = c(k, sigma, mu = 0), lower.tail = TRUE, verif = TRUE)
rWEI(N, par = c(k, sigma, mu = 0), verif = TRUE)
WEI.grad.F(x, par, verif = FALSE)
WEI.hess.F(x, par, verif = FALSE)
WEI.grad.q(par, lambda, Tr = 100, verif = TRUE)
WEI.loglik(par, dat, verif = TRUE)
WEI.grad.loglik(par, dat, verif = TRUE)
WEI.hess.loglik(par, dat, verif = TRUE)
verif.WEI(par, dat = NULL, p = NULL)
```

Arguments

`x`	Vector of quantiles or support of the distribution.
`p`	Vector of probabilities.
`par`	Parameters of the distribution: shape parameter k, scale parameter σ and location parameter μ.
`N`	The sample size or number of observations.
`lambda`	The Poisson rate parameter (mean number of events per year).
`Tr`	The return period, in years, associated to the quantile `q`.
`dat`	A sample to be fitted to the distribution.
`lower.tail`	Logical: if TRUE (default), probabilities are $\Pr[X \leq x]$, otherwise, $\Pr[X > x]$.
`verif`	Logical: if TRUE, arguments are checked by function `verif.WEI`.

Details

The cumulative distribution function of the Weibull distribution is given by:

$$F(x) = 1 - \exp\left[-\left(\frac{x-\mu}{\sigma}\right)^k\right]$$

with shape parameter k, scale parameter σ and location parameter μ.

Value

dWEI returns the density of the quantiles x for Weibull distribution with parameters par. pWEI returns the non-exceedance probabilities (or exceedance probabilities if lower.tail = TRUE) of the quantiles x for Weibull distribution with parameters par. qWEI returns the quantiles associated to the probabilities p for Weibull distribution with parameters par. rWEI randomly generates a sample of size N for Weibull distribution with parameters par. WEI.grad.F returns the gradient of the cumulative distribution function, useful for the historical likelihood. WEI.hess.F returns the hessian of the cumulative distribution function, useful for the historical likelihood. WEI.grad.q returns the gradient of the quantile function, useful for computing the confidence intervals by the Delta method. WEI.loglik returns the likelihood of the sample dat following a Weibull distribution with parameters par. WEI.grad.loglik returns the gradient of the likelihood of the sample dat following a Weibull distribution with parameters par. WEI.hess.loglik returns the hessian of the likelihood of the sample dat following a Weibull distribution with parameters par. verif.WEI checks the arguments of the above functions and returns a logical value TRUE/FALSE. It checks whether the probabilities are in the range [0,1] and whether the parameters are within their interval of validity.

Author(s)

Franck Mazas

References

"Aide-mémoire statistique", ARTELIA

See Also

fit.MLE, fit.LMOM, fit.MOM, qIC.delta, qIC.boot, multiPOT

Examples

```
p <- seq(0, 1, 0.1)
x <- seq(0, 5, 0.2)
x2 <- seq(-5, 15, 1)
x3 <- seq(-5, 5, 0.25)
k1 <- -0.5
k2 <- 0.5
sigma1 <- 1.75
sigma2 <- 0.5
mu1 <- 0.1
mu2 <- 0.5
N <- 50

dWEI(x, c(k = k2, sigma = sigma1))
dWEI(x, c(k = k2, sigma = sigma1, mu = mu1))
pWEI(x, c(k = k2, sigma = sigma1))
qWEI(p, c(k = k2, sigma = sigma1))

Y5 <- rWEI(500, c(k = 1.5, sigma = sigma1, mu = 0))
fit.MOM(Y5, "WEI")
fit.MOM(Y5, "WEI", mu0 = TRUE)
fit.LMOM(Y5, "WEI")
WEI.loglik(fit.LMOM(Y5, "WEI")$par, Y5, verif = FALSE)
WEI.grad.loglik(fit.LMOM(Y5, "WEI")$par, Y5, verif = FALSE)
fit.MLE(Y5, "WEI", optim.method = "Newton-Raphson")
```

```
WEI.grad.loglik(fit.MLE(Y5, "WEI", optim.method = "Newton-Raphson")$par, Y5, verif = FALSE)
fit.MLE(Y5, "WEI", optim.method = "Nelder-Mead")
fit.MLE(Y5, "WEI", par.init = fit.LMOM(Y5, "WEI", mu0 = TRUE)$par, optim.method = "Nelder-Mead")
WEI.grad.loglik(fit.MLE(Y5, "WEI", par.init = c(k = 1.5, sigma = sigma1),
optim.method = "Nelder-Mead")$par, Y5, verif = FALSE)
```

Index

2.2. USER MANUAL (IN FRENCH)

L'ombre
Suit
Sombre
Nuit ;
Une
Lune
Brune
Luit.

Tranquille
L'air pur
Distille
L'azur ;
Le sage
Engage
Voyage
Bien sûr !

L'atmosphère
De la fleur
Régénère
La senteur,
S'incorpore,
Evapore
Pour l'aurore
Son odeur.

Parfois la brise
Des verts ormeaux
Passe et se brise
Aux doux rameaux ;
Au fond de l'âme
Qui le réclame
C'est un dictame
Pour tous les maux !

Un point se déclare
Loin de la maison,
Devient une barre ;
C'est une cloison ;
Longue, noire, prompte,
Plus rien ne la dompte,
Elle grandit, monte,
Couvre l'horizon.

L'obscurité s'avance
Et double sa noirceur ;
Sa funeste apparence
Prend et saisit le cœur !
Et tremblant il présage
Que ce sombre nuage
Renferme un gros orage
Dans son énorme horreur.

Au ciel, il n'est plus d'étoiles
Le nuage couvre tout
De ses glaciales voiles ;

Il est là, seul et debout.
Le vent le pousse, l'excite,
Son immensité s'irrite ;
A voir son flanc qui s'agite,
On comprend qu'il est à bout !

Il se replie et s'amoncelle,
Resserre ses vastes haillons ;
Contient à peine l'étincelle
Qui l'ouvre de ses aquilons ;
Le nuage enfin se dilate,
S'entrouvre, se déchire, éclate,
Comme d'une teinte écarlate
Les flots de ses noirs tourbillons.

L'éclair jaillit ; lumière éblouissante
Qui vous aveugle et vous brûle les yeux,
Ne s'éteint pas, la sifflante tourmente
Le fait briller, étinceler bien mieux ;
Il vole ; en sa course muette et vive
L'horrible vent le conduit et l'avive ;
L'éclair prompt, dans sa marche fugitive
Par ses zigzags unit la terre aux cieux.

La foudre part soudain ; elle tempête, tonne
Et l'air est tout rempli de ses longs roulements ;
Dans le fond des échos, l'immense bruit bourdonne,
Entoure, presse tout de ses cassants craquements.
Elle triple d'efforts ; l'éclair comme la bombe,
Se jette et rebondit sur le toit qui succombe,
Et lé tonnerre éclate, et se répète, et tombe,
Prolonge jusqu'aux cieux ses épouvantements.

Un peu plus loin, mais frémissant encore
Dans le ciel noir l'orage se poursuit,
Et de ses feux assombrit et colore
L'obscurité de la sifflante nuit.
Puis par instants des Aquilons la houle
S'apaise un peu, le tonnerre s'écoule,
Et puis se tait, et dans le lointain roule
Comme un écho son roulement qui fuit ;

L'éclair aussi devient plus rare
De loin en loin montre ses feux
Ce n'est plus l'affreuse bagarre
Où les vents combattaient entre eux ;
Portant ailleurs sa sombre tête,
L'horreur, l'éclat de la tempête
De plus en plus tarde, s'arrête,
Fuit enfin ses bruyants jeux.

Au ciel le dernier nuage
Est balayé par le vent ;
D'horizon ce grand orage
A changé bien promptement ;
On ne voit au loin dans l'ombre
Qu'une épaisseur large, sombre,

Qui s'enfuit, et noircit, ombre
Tout dans son déplacement.

La nature est tranquille,
A perdu sa frayeur ;
Elle est douce et docile
Et se refait le cœur ;
Si le tonnerre gronde
Et de sa voix profonde
Là-bas trouble le monde,
Ici l'on n'a plus peur.

Dans le ciel l'étoile
D'un éclat plus pur
Brille et se dévoile
Au sein de l'azur ;
La nuit dans la trêve,
Qui reprend et rêve,
Et qui se relève,
N'a plus rien d'obscur.

La fraîche haleine
Du doux zéphir
Qui se promène
Comme un soupir,
A la sourdine,
La feuille incline,
La pateline,
Et fait plaisir.

La nature
Est encor
Bien plus pure,
Et s'endort ;
Dans l'ivresse
La maîtresse,
Ainsi presse
Un lit d'or.

Toute aise,
La fleur
S'apaise ;
Son cœur
Tranquille
Distille
L'utile
Odeur.

Elle
Fuit,
Belle
Nuit ;
Une
Lune
Brune
Luit.

Jules Verne, Tempête et Calme

Résumé

Cette thèse sur travaux vise à rassembler et unifier les travaux réalisés sur le sujet des évènements météo-océaniques extrêmes depuis 2009, dans le cadre de mon travail à SOGREAH, devenu depuis ARTELIA. à mesure que progressaient ces travaux, un thème central a progressivement apparu : la notion d'évènement, tel qu'une tempête. Ce concept fournit un cadre robuste et pertinent, en particulier dans le cas des extrêmes multivariés (par exemple, la probabilité d'occurrence conjointe des vagues et des niveaux marins), ainsi qu'une meilleure compréhension de la notion de période de retour, très utilisée dans le domaine de l'ingénierie. Les principaux résultats des travaux réalisés au cours de la décennie écoulée sont les suivants :

- mise à jour de la méthodologie de détermination des houles ou vents extrêmes :
  - développement et justification d'un cadre en deux étapes pour la modélisation sup-seuil des extrêmes univariés (méthode du renouvellement), introduisant la notion d'évènement et la séparation des seuils physique et statistique,
  - proposition d'outils pratiques pour le choix du seuil statistique,
  - introduction de la méthode du bootstrap paramétrique pour le calcul des intervalles de confiance,
  - identification d'un comportement problématique de l'Estimateur du Maximum de Vraisemblance et proposition d'une solution : utilisation de distributions à trois paramètres avec l'estimateur des L-moments,
- application du cadre POT (Peaks-Over-Threshold) à la Méthode des Probabilités Jointes (JPM) pour la détermination des niveaux marins extrêmes :
  - distinction entre les valeurs séquentielles et les pics des évènements à l'aide d'indices extrémaux pour les surcotes et les niveaux marins,
  - construction d'un modèle mixte pour la distribution des surcotes,
  - raffinements pour le traitement de la dépendance marée-surcote,
- application du cadre POT-JPM pour l'analyse conjointe des hauteurs de vagues et des niveaux marins :
  - proposition d'une procédure alternative d'échantillonnage,
  - analyse séparée de la marée et de la surcote dans le but de modéliser la dépendance entre la hauteur de vagues et la surcote ; avec incorporation dans la distribution conjointe de la hauteur de vagues et du niveau marin à l'aide d'une opération de convolution 2D1D,
  - utilisation de copules des valeurs extrêmes,
  - présentation améliorée du chi-plot,
- introduction d'une nouvelle classification pour les analyses multivariées :
  - **Type A** : un phénomène unique décrit par différentes grandeurs physiques qui ne sont pas nécessairement du même type,
  - **Type B** : un phénomène fait de différentes composantes, décrits par des grandeurs physiques du même type d'un composant à l'autre,
  - **Type C** : plusieurs phénomènes décrits par des grandeurs physiques qui ne sont pas nécessairement du même type,
- interprétation de la signification des évènements multivariés :
  - lien avec l'échantillonnage,
  - lien avec les différentes définitions de la période de retour,
  - dans le cas bivarié : transformation d'une distribution conjointe de variables descriptives de l'évènement vers la distribution des couples de variables séquentielles,
  - génération de graphes de sortie alternatifs tels que les contours d'iso-densité pour les couples de variables séquentielles,
- un package R dédié, artextreme, pour l'implémentation des méthodes ci-dessus.

Mots clés : évènements extrêmes, météo-océanographie, probabilités jointes, submersion marine, ingénierie côtière

Abstract

This PhD on published works aims at unifying the works carried out on the topic of extreme metocean events since 2009, while working for SOGREAH then ARTELIA. As these works went along, a leading theme progressively appeared: the notion of event, such as a storm. This concept provides a sound and relevant framework in particular in the case of multivariate extremes (such as joint probabilities of waves and sea levels), as well as a better understanding of the notion of return period, much used for design in the field of engineering. The main results of the works carried out in the last decade are as follows:

- updating of the methodology for determining extreme wave heights or wind speeds:
  - development and justification of a two-step framework for extreme univariate over-threshold modelling introducing the concept of event and the separation of the physical and statistical thresholds,
  - proposal of practical tools for choosing the statistical threshold,
  - introduction of the parametric bootstrap approach for computing confidence intervals,
  - identification of a problematic issue in the behaviour of the Maximum Likelihood Estimator and proposal of a solution: use of 3-parameter distributions along with the L-moments estimator,
- application of the POT framework to the Joint Probability Method for determining extreme sea levels:
  - distinction between sequential values and event peaks through extremal indexes for surge and sea level,
  - construction of a mixture model for the surge distribution,
  - refinements for handling tide-surge dependence,
- application of the POT-JPM framework for the joint analysis of wave height and sea level:
  - proposal of an alternative sampling procedure,
  - separate analysis of tide and surge in order to model the dependence between wave height and surge to be incorporated in the joint distribution of wave height and sea level thanks to a 2D1D convolution operation,
  - use of extreme-value copulas,
  - improved presentation of the chi-plot,
- introduction of a new classification for multivariate analyses:
  - **Type A**: a single phenomenon described by different physical quantities that are possibly not of the same kind,
  - **Type B**: a phenomenon made of different components, described by physical quantities of the same kind between one component and another,
  - **Type C**: several phenomena described by physical quantities that are possibly not of the same kind,
- interpretation of the meaning of multivariate events:
  - link with the sampling procedure,
  - link with the different definitions of the return period,
  - in the bivariate case: transformation of the joint distribution of event-describing variables into the joint distribution of sequential pairs,
  - generation of alternative output plots such as contours of density for sequential pairs;
- a dedicated R package, artextreme, for implementing the methodologies presented above.

Keywords: extreme events, meteo-oceanography, joint probabilities, coastal flooding, coastal engineering