OECD/NEA/CSNI/WGEV Benchmark: External Events Hazard Frequency and Magnitude Statistical Modelling - IRSN approach

Y. HAMDI, V. REBOUR

Institute for Radiological Protection and Nuclear Safety, Fontenay aux Roses, France,
yasser.hamdi@irsn.fr

Abstract: A general feature of external hazards is that they produce off-normal conditions that can impact nuclear installations. Scenarios- and frequencies-based risk analysis allows a more reliable practice by allowing key stakeholders to make risk informed choices rather than simply relying on traditional deterministic estimates of risk, with a brief description of **uncertainty**. This benchmark is conducted to facilitate an exercise on the estimation of extreme external events. The aim of the benchmark is to better understand the technical aspects and processes used for the characterization of natural hazards. Data and overall objectives for the benchmarking exercise are presented for a hypothetical external event (e.g., precipitation, extreme temperatures, high winds). Our analysis steps, assumptions and modelling results were summarized. Uncertainties are generated by the fact that the provided synthetic data (SD) is relatively uninformative. The use of this SD-based approach allows to evaluate the proposed statistical model and add known uncertainty to the data. Two cases are provided: (i) with a given generating model, (ii) a "blind-test case" where only the data is provided.

To make the exercise as useful as possible and to present the work in the most comprehensible way, some theoretical and technical background for the selected approaches and the assumptions are provided. After recalling the data provided in the exercise, the underlying theory and assumptions are presented. The results of the proposed statistical models are then presented for each test case together with the associated uncertainties as well as the model adequacy assessment. This work is concluded with a summary of the most essential results.

1. INTRODUCTION

Risk associated to natural hazards is often invisible on a day-to-day basis and it's hard for example to prevent where a flood might happen or anticipate and identify what could get damaged during a flooding or an earthquake. Hazard is one of the components of risk, with exposure and vulnerability. In the nuclear safety field, an initiating event (IE) is an event that challenges the normal operation of a nuclear power plant (NPP) and could lead to fuel damage, or the failure of containment for a radiation source.

The identification of the most representative variable (or variables) of the hazard IEs constitutes the first step of the risk analysis modelling and simulation. Data associated to this variable of interest, or data from each other category of the risk, can be used to paint a picture of risk in a certain location and over time. Natural hazard events can be characterized by their magnitude, duration and frequency. They occur at different intensities over different time scales. Indeed, one can talk about the occurrence of hazards of different intensities in terms of probabilities or return periods, within the context of uncertainty. In general, the less frequent the hazard (the longer the return period) the greater its intensity. In addition, it is well established today that, due to the global warming, weather-related extreme events have become more frequent over many areas over the world. In this context, time- and climate indices- varying statistical models must be used. Hazards also occur over different spatial scales. For instance, the occurrence and impact of a river or pluvial flooding tend to be quite localized, whereas droughts can occur over several thousands of kms.

The statistical data-driven effective modelling of the magnitude, duration and frequency of a hazard IE is then a key issue since it both contributes to the risk evaluation and provides the boundary conditions for the hazard scenario-based probabilistic approaches. Relating hazard IEs to their intensity and frequency of occurrence using probability distributions has been a common issue since 1950s (e.g., [1,2]). This relationship corresponding to long return periods is based on the extreme value theory [3]. Indeed, a variety of technical approaches and models are presented in the literature. Some of them, which are the current practice in the field of the characterization of natural hazards, have their limitations especially when data sets are short and characterized by the presence of gaps and outliers. Three approaches to extreme value analysis including the Annual Maxima (AM), the Peaks-Over Threshold (POT) and the r-Largest Order Statistics (r-LOS) are used in The Institute for Radiation Protection and Nuclear Safety (IRSN).

The AM sampling method is a straightforward and simple and approach, used in a large number of design codes worldwide, in which a Generalized Extreme Value (GEV) distribution is applied to fit AM observations. However, if data are available for a relatively limited period, the statistical extrapolation to estimate events corresponding to high return periods, is seriously contaminated by this sampling method and model uncertainty. Another major limitation of the AM samples is the fact that all the other high observations occurring during the whole year are not considered. A way to enlarge the sample beyond the AM is to use a Point-Process Method (PPM) by extracting a number of high observations in each year (r-LOS approach) or by setting an exceedance high threshold above which observations are considered as extremes. This latter is the so-called Peaks-Over-Threshold (POT) method and it is the most commonly used in IRSN. The Generalized Pareto Distribution (GPD) is the most adapted theoretical distribution to fit POT series. However, the choice of a threshold value is subjective and there is no a theoretical based mathematical approach designated for this purpose. A bias in the estimation is automatically introduced when an unreasonably low threshold is used. Indeed, observations which may not be extreme can creep into the sample and this violates the principle of the extreme value theory. On another hand, the sample of extremes will be reduced if the selected threshold is too high. The reader is referred to Hamdi et al. [4] and [3] for more details about AM and the PPM methods presented above.

This paper provides details (overall objectives, assumption, and results) for a univariate frequency analysis using synthetic data (SD) for a hypothetical external event (e.g., extreme temperatures, high winds, precipitation). The particularity of SD is the fact that its nature, generating process and uncertainties are unknown. The use of a SD-based approach is motivated by the need to evaluate the predictive performance of a suggested frequency model and the possibility to control the random nature and known uncertainty of data.

It begins by outlining the rationale for this paper in further detail. Section 2 provides the synthetic models (SMs) used to generate data (three SD cases are presented). This section also provides a general introduction to key features and ideas related to data that can be used to characterize the external hazards. The underlying theory and crucial assumptions together with the framework for the model adequacy assessment are presented in Section 3. It also covers the sampling methods and statistical data-driven models used by IRSN to estimate hazards IE. Section 4 covers the results of the SD-driven models and the adequacy assessment followed by a brief discussion of their predictive performance, the estimation of uncertainties for each SD case. The discussion is also devoted to highlight the usefulness or not of SD sets and put in light the advantages and drawbacks of their use in the framework of the extreme value theory. This paper is concluded with a summary of the most essential results. The appendix presents all the scripts used in this work.

2. THE USE OF SYNTHETIC MODELS AND DATA

The information on past hazard IEs is important to characterize the hazard and evaluate the associated risk for future events. Historical information is important to generate local hazard curves, by combining scenario-based statistical estimations (several return hazard levels with associated probabilities) with numerical models for example. Historical information on past hazardous events is aleatory, always incomplete and subject of considerable uncertainty but, interestingly, a general and common feature of the data-driven hazard assessment models is that they would produce larger hazard return levels for longer return frequencies.

SD of a hypothetical variable representing a physical process (e.g., local precipitation, high winds) are used in this paper rather than gathering real datasets. Unlike the real cases found in the external hazard community, the underlying data-generating process can be known exactly and controlled. This will allow us to evaluate the predictive performance of data-driven models by simulating a SD set with the characteristics (time periods, sample size, etc.) we select. Uncertainty present in the SD can also be considered in the analysis. Indeed, the generating model output is function of both the inputs and associated uncertainties. Two SD cases are used: one with a given generating SM and a "blind-test case" where only the data is provided.

2.1. Case 1 – the model producing the synthetic data is known

The SM used for this first Case is:

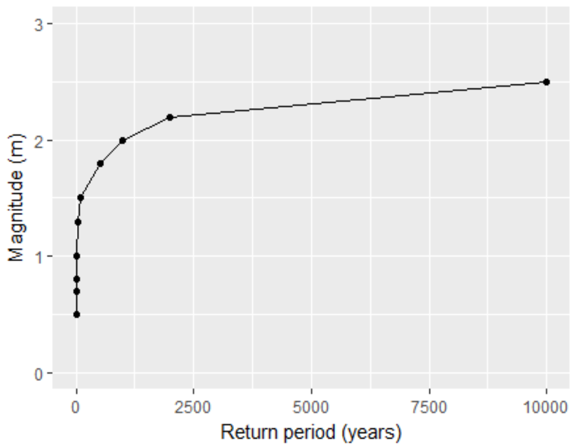
$$M = 0.5 + 0.5 \times \log_{10}(t) \quad (1)$$

This equation is of the form that different time values t produce hypothetical magnitudes M for an AM event for example. As used herein, this SM serves as surrogate for a complex phenomenological process producing SD used to predict different event outcomes as a function of time. As with all the statistical models used to characterize natural hazards, the proposed SM produces SD such that the event magnitude increases as the return periods becomes large (the larger the event, the less frequent it is). Table 1 shows the events magnitudes produced by the SM-case1 for selected return time intervals ranging from 1 to 10,000 years. Figure 1 shows the plot of these events.

Table 1. SD- Case1

t	M
1	0.50
2	0.65
5	0.85
10	1.00
50	1.40
100	1.50
500	1.90
1000	2.00
2000	2.20
10000	2.50

Figure 1. Hazard curve (SM-Case1).



2.2. Case 2 – the model producing the synthetic data is unknown

For the 2nd case, the SM is not provided. Two SD sets (10 and 26 data points) are used with no uncertainty provided. Further, three parts to this example are provided. The objective of this synthetic case study is to provide, using the two SD sets separately, a model that best describes the frequency/magnitude relationship and the associated analysis and insights. The SD-cas2a and SD-Cas2b output from the unknown model are shown in Tables 2 and 3, respectively. Figures 2 and 3 show the plots of the cas2a and cas2b events.

Table 2. SD-Case2a

t	M
1	0.53
2	0.53
5	0.54
10	0.55
50	0.59
100	0.62
500	0.79
1000	0.95
3000	1.60
10000	4.00

Figure 2. Hazard curve (SD-Case2a).

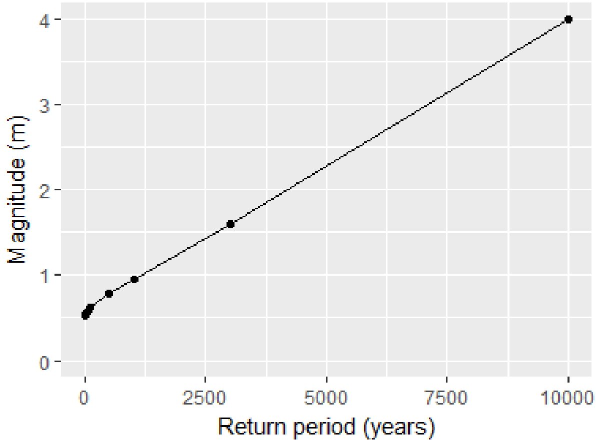
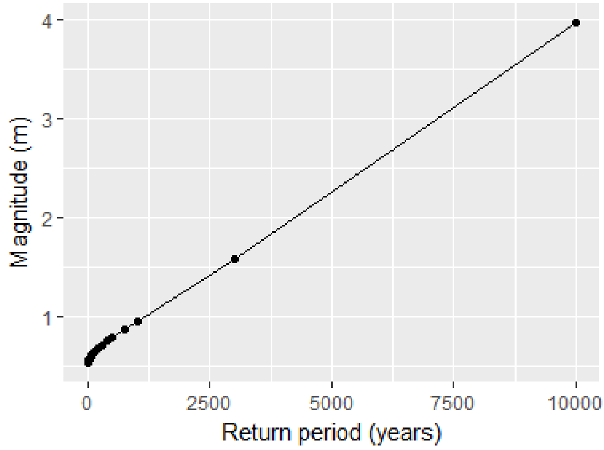


Table 3. SD-Case2b

t	M	t	M
1	0.53	90	0.62
2	0.53	100	0.62
5	0.54	125	0.63
10	0.55	150	0.65
15	0.56	175	0.66
20	0.56	200	0.67
25	0.57	300	0.71
30	0.57	400	0.75
40	0.58	500	0.79
50	0.59	750	0.87
60	0.60	1000	0.95
70	0.60	3000	1.57
80	0.61	10000	3.97

Figure 3. Hazard curve (SD-Case2b).



3. EXTREME VALUE FREQUENCY ESTIMATION

A standard frequency analysis includes the following steps: (i) verification of homogeneity, randomness and stationarity of data; (ii) empirical probabilities computation; (iii) Fitting to these data a theoretical curve (a distribution function). Adequacy tests can be used to select the more appropriate distribution function and (iv) extrapolating to a high return period of the extreme value of interest (say 10,000 years). Regardless the sampling method and frequency model used, randomness, homogeneity and stationarity of data are necessary conditions in a frequency analysis.

As mentioned in the introductory section, of the many statistical distributions commonly used for extremes, the Generalized Extreme Value GEV function was retained for the AM sampling method and the Generalized Pareto GP distribution function (for the intensity) and the a Poisson process (for the number of exceedances) were used to apply the POT approach. The GEV distribution introduced by Jenkinson (1955) [5] combines three asymptotic extreme value distributions, identified by Fisher and Tippet (1928) |6], into a single form with the following cumulative distribution function F:

$$F(x) = \begin{cases} e^{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}} & \xi \neq 0 \\ e^{-e^{-(x-\mu)/\sigma}} & \xi = 0 \end{cases} \quad (2)$$

Where μ, $\sigma > 0$ and ξ are the location, scale and shape parameters, respectively. Depending on the value of the shape parameter ξ, the GEV can take the form of the reverse Weibull function when $\xi < 0$, the Fréchet distribution if $\xi > 0$ and the Gumbel distribution which has an exponential tail when $\xi = 0$. The first one has a finite and short theoretical upper tail $\left(\infty < x < \mu - \sigma/\xi\right)$ that may be useful for estimates of specific cases of extreme values, which may have an upper bound. Let $\hat{z}_p$ be the $1/p$ return level and it's given by

$$\hat{z}_p = \begin{cases} \hat{\mu} - \hat{\sigma} log\{y_p\} & \xi = 0 \\ \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\{1 - y_p{}^{-\hat{\xi}}\} & \xi \neq 0 \end{cases} \quad (3)$$

Where $y_p = -log\{1-p\}$ and the GEV distribution parameters estimated with the maximum likelihood method.

On the other hand, the GP distribution has the following cumulative distribution function:

$$G(x) = \begin{cases} 1 - (1 + \xi\, x/\sigma)^{-1/\xi} & \xi \neq 0 \\ 1 - e^{-x/\sigma} & \xi = 0 \end{cases} \quad (4)$$

The GP distribution corresponds to a long tailed ordinary Pareto distribution for positive values of ξ. It takes the form of a Pareto Type II distribution with a short tail upper bounded by $\mu + \sigma/\xi$ when $\xi < 0$ and finally to an exponential distribution with a medium-size tail when $\xi = 0$.

4. RESULTS OF THE ASSESSMENT (SM-CAS1)

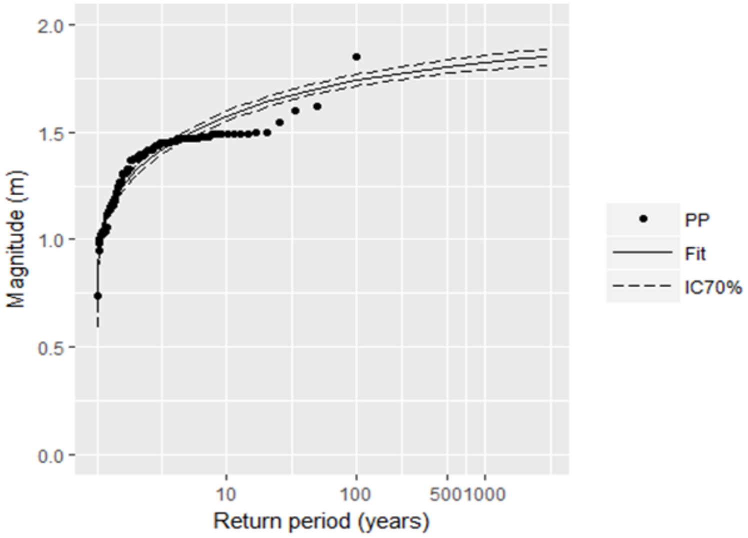
In the SM-Cas1, a set of 100 return periods is randomly sampled. Magnitudes are calculated with the SM and classic frequency estimation is then performed with a GEV and GPD distributions. To take this issue one step further, a simple analogy of the synthetic model with GEV and GP distribution functions is performed. In SM-cas1, a magnitude M is produced by the SM with a rate of events equal to one ($\lambda = 1$). Let x_T be the event with the return period T and F the non-exceedance probability function. In its general definition, x_T corresponds to the quantile of probability of exceedance equal to $1/\lambda T$.

4.1 AM - GEV frequency model:

A sample of 100 magnitudes (AM events with an effective duration equal to 100 years) is generated and analyzed. The hypothetical observational dataset is selected in such a way their empirical distribution looks as natural as possible (close to the natural behavior and variability of external hazards such as floods, extreme temperatures or high winds, etc.). More concretely, the sampling is done as follow: randomly sample 100 values of "return times" t from 1 to 100 years. 95% of the values have empirical return periods less than 100 years. It is worth noting that the case with aleatory sampling in a period of 10,000 years is not realistic and cannot represent natural hazards.

The fitting is bounded to a final value equal to $\mu - \sigma/\xi$ (equal to 1.91 m). The 500,000-year return level, for example, is equal to this end value (1.90 m) which is much lower than the 3.35 m calculated with the synthetic model for the same return period. Paradoxically, it can be seen from figure 4 that the adequacy of the GEV distribution is quite good and the uncertainty (in term of confidence intervals) is very low for high return levels.

Figure 4. SM-Case1: fitting with a GEV distribution
($\mu = 1.26; \sigma = 0.19; \xi = -0.29$)

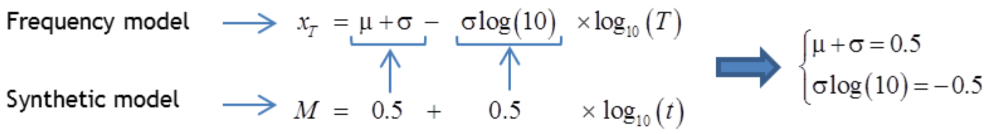


A linear relationship between magnitudes M and the logarithm of return periods t could be obtained if the x-axis was plotted in a logarithmic scale. Note that this linear relationship cannot be obtained with all the asymptotic extreme value distributions and only the exponential form (the Gumbel distribution) might be suitable (but with the limitation of bad fitting of small magnitudes associated to small return periods up to 50-100 years).

As can be seen from the probability equation, the GEV distribution can as well have the form of the synthetic model when the shape parameter is equal to -1. With such a value, the theoretical upper tail can only be finite and bounded (as may be useful for estimates of specific cases of extreme values which may have an upper bound, as is the case here). This hypothesis is in line with the bounded theoretical fitting presented in Figure 4.

$$x_T = \mu + \sigma\{1 + \log(1-F)\} = \mu + \sigma - \sigma\log(T) = \mu + \sigma - \sigma\log(10) \times \log_{10}(T) \quad (5)$$

This last equation is similar to the proposed SM-cas1 as follows:



It can be then concluded from these two conditions that, with a shape parameter equal to -1, the scale parameter can only be negative. But obviously, a GEV distribution cannot be used with a negative scale parameter. Consequently, it cannot describe the frequency/magnitude relationship for the SM-Cas1.

4.2 POT - GPD/Exponential frequency model

The GP distribution for the return levels (with an exponential distribution for the exceedances over the threshold u). The exceedances over the threshold u follow an exponential distribution of parameter ρ :

$$F_u(X) = 1 - \exp\left[-\rho(x-u)\right] \tag{6}$$

Therefore, the $T-year$ return level can be written as:

$$x_T - u = -\frac{1}{\rho}\log(1-F) = u - \frac{1}{\rho}\log\left(\frac{1}{\lambda T}\right) = u + \frac{1}{\rho}\ln(\lambda) + \frac{1}{\rho}\log(10)\log_{10}(T) \tag{7}$$

This equation is similar to the proposed SM-cas1 as follows:

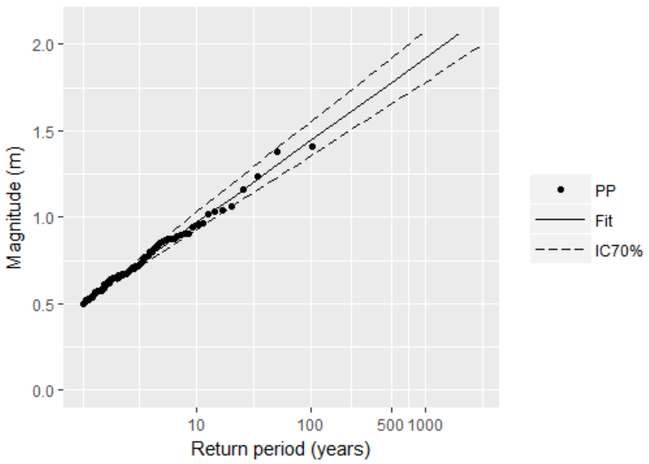
Frequency model $\longrightarrow$ $x_T = \underbrace{u + \frac{1}{\rho}\ln(\lambda)} + \underbrace{\frac{1}{\rho}\log(10)} \times \log_{10}(T)$

Synthetic model $\longrightarrow$ $M = 0.5 + 0.5 \times \log_{10}(t)$

$$\Rightarrow \begin{cases} u + \frac{1}{\rho}\ln(\lambda) = 0.5 \\ \frac{1}{\rho}\log(10) = 0.5 \end{cases}$$

Considering the case in which the rate of events is equal to one ($\lambda = 1$), the first condition gives a threshold $u = 0.5$. It can also be easily concluded from the second condition that $\rho = \log(10)/0.5 = 4.605$. In the first step, the frequency model is used to describe the frequency/magnitude relationship. Magnitudes are then sampled on a period of time w , say 100 years. In the next step, the fitting with the exponential distribution is performed (Figure 5). The results are also shown in Table 4. A good and visually adequate fitting is obtained. Indeed, all the observed probabilities are in in the 70% confidence interval. As shown in Table 4, the 500,000-year return level is equal to 3.19 m which is close to the 3.35 m calculated by the synthetic model for the same return period. On the other hand, the uncertainty (in term of confidence intervals) is reasonably low for high return levels.

Figure 5. Fitting of the SM-Case1 data sets (w=100 years) with GPD distribution ($u = 0.5$ & $\lambda = 1$) for the return levels (the exceedances over u are Exponential).



All the simulations were carried out within the R environment (open-source software for statistical computing: http://www.r-project.org/). The Renext library (IRSN and Alpstat, 2013) (developed by the French Institute for Radiological Protection and Nuclear Safety - IRSN) was used for the frequency estimations. The Renext package was specifically developed for flood frequency analyses using the POT method.

Table 4. Comparison of magnitudes (m) calculated with the synthetic data for Case 1. The values in brackets correspond to the 70% confidence intervals.

Return Period (years)	500	5,000	50,000	500,000
SD-Case1	**1.85**	**2.35**	**2.85**	**3.35**
POT model	1.78 (1.66-1.92)	2.25 (2.08-2.45)	2.72 (2.51-2.98)	3.19 (2.94-3.50)

5. RESULTS OF THE ASSESSMENT (SM-CAS2)

As it was mentioned in section 2, the synthetic model used was not provided to the participants for this second SM-case2. However, three parts to this case are proposed: (i) SD-cas2a that provides the ten SD; (ii) SD-cas2b that provides additional SD (26 data points) and (iii) and SD-cas2c that has uncertainty estimates on some of the data. SD-Cases 2a et 2b are provided with no uncertainty provided. Only the first two parts are evaluated in this work. The same assumptions are used for this second case: the magnitudes are assumed to be stationary, independent and homogeneous.

As in SM-case1, return periods can then be estimated with one of the AM/GEV and POT/GPD frequency models. For both cases 2a & 2b, non-linear least-squares estimates of the GEV and GPD parameters are performed. It is worth noting that the GPD must give almost the same parameters and fitting (results not presented hereafter). The fitting with confidence intervals is presented in Figures 6 and 7. The adequacy of the theoretical distribution in both cases 2a & 2b is visually quite good with heavy tails (very high shape parameter $\xi = 0.96$). A comparison of *T*-year return levels (corresponding to 500-, 5000-, 50,000- and 500,000-year return periods) for both Case 2a and Case 2b are presented in Table 5 (the values in brackets correspond to the absolute and relative widths of the 70% confidence intervals). These results indicate that the relative widths of confidence intervals in case 2b (with 26 SD points) are 1.5 times narrower than those obtained in case 2a (with 10 SD points) for the GEV distribution.

Table 5. The quantiles with absolute and relative widths of their 70% confidence intervals.

T (years)	500	5,000	50,000	500,000
Case 2a	0.75 (0.68-0.84) (21.3%)	2.32 (1.72-3.21) (64.2%)	16.79 (10.75-26.44) (93.4%)	150.34 (89.35-254.16) (109.6%)
Case 2b	0.76 (0.71-0.82) (14.5%)	2.31 (1.88-2.87) (42.8%)	16.50 (12.17-22.46) (62.4%)	146.53 (102.70-209.56) (73.0%)

Figure 6. Fitting SD-case2a
GEV($\mu = 0.5593; \sigma = 0.0005; \xi = 0.9650$)

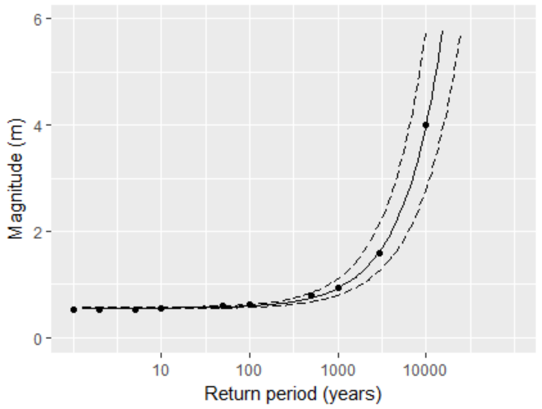
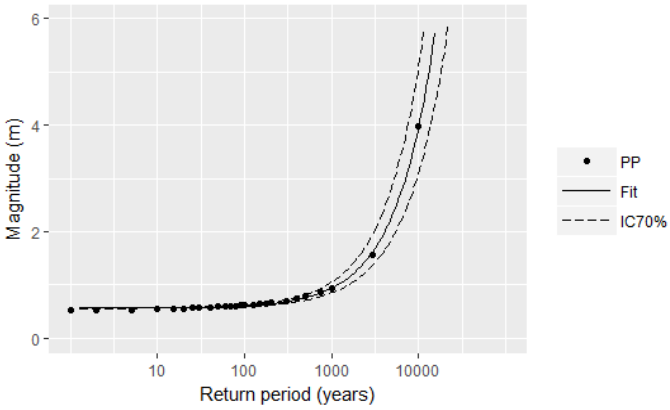


Figure 7. Fitting SD-case2b
GEV($\mu = 0.5709; \sigma = 0.0005; \xi = 0.9619$)



To recap, therefore, commendable practices in formulating and assessing the quantification of external IEs when using statistical models are needed. Nevertheless, the issue that received our attention in this paper is: how hazard risk assessment is performed in IRSN and what are the quantitative technical analysis steps and processes used in a probabilistic hazard modelling framework? and How associated uncertainties are estimated.

5. CONCLUDING REMARKS

The principal objective of this study was to perform a statistical modelling using synthetic models and data for assessing hazard frequency and magnitude for external events risk assessment. Synthetic models and data for a hypothetical external event (e.g., precipitation, extreme temperatures, high winds) are then used instead of real observations. Two synthetic data cases are used: In this first case, the hypothetical data and the generating synthetic model are known, while the second case is a blind-test one (two synthetic data of two different sizes are used but the synthetic model are unknown). The annual maxima (AM) and peaks-over-threshold (POT) sampling methods are used to extract extreme values. The analysis steps and modelling results are provided and discussed.

It was concluded that the frequency model using the GEV distribution (using the AM sampling method) is not very suitable to describe all the frequency/magnitude relationship for the case with known synthetic model. These first case synthetic data are rather best described when a POT sampling method (with an exponential behaviour of the exceedances over the threshold) is used. Indeed, the uncertainty is reasonably low and the normalized confidence interval did not exceed 18% for the 500,000-year return level. On the other hand, when the generating process of data is unknown (the second case), both frequency models using the GEV and the GPD distribution functions (using AM and POT sampling methods, respectively) best describe the frequency/magnitude relationship with heavy tails. Moreover, as less data are provided in this second case, the uncertainty is larger (the confidence intervals are narrower).

It is obvious that, with such heavy tails (high shape parameter), these data reflect what can be observed for natural hazards up to a certain return period. Beyond this, return magnitudes increase much more quickly to very high and unrealistic values. This is the main characteristic of models with very heavy tails (as is the case here).

Acknowledgements. The authors gratefully acknowledge Yves Deville (consulting engineer) and Dr. Yann Richet (IRSN) for their collaboration and for the scientific support they have given to this work.

References

[1] V.T. Chow. "*Frequency analysis of hydrologic data with special application to rainfall intensities*", Eng. Exp Sta. Bulletin, 414, Univ. Illinois (1953).
[2] T. Dalrymple. "*Flood Frequency Analyses. Manual of Hydrology*", Part 3, Flood Flow Techniques USGS Water Supply Paper, 1543-A, pp. 51-77, (1960).
[3] S. Coles. "*An Introduction to Statistical Modeling of Extreme Values*", Springer, 2001, Berlin.
[4] Y. Hamdi, L. Bardet, C.-M. Duluc, C.-M., and V. Rebour. "*Extreme storm surges: a comparative study of frequency analysis approaches*", Nat. Hazards Earth Syst. Sci., 14, pp. 2053–2067, (2014).
[5] A. F. Jenkinson. "*The frequency distribution of the annual maximum (or minimum) values of meteorological elements*". Q J Roy Meteor Soc, 81, pp. 158−171, (1955).
[6] R. A. Fisher, and L. H. C. Tippett. "*Limiting forms of the frequency distribution of the largest or smallest member of a sample*", in: Proceedings of the Cambridge Philosophical Society, 24, pp.180−190, (1928).