Revised Section 2

2 The HIBEVA method

2.1 Theoretical model

The model chosen to represent and extrapolate extreme values of WL is the Generalised Pareto Distribution (GPD), applied to a Peaks-Over-Threshold (POT) sample. This extreme value model has been widely used and is most commonly recommended as it makes use of all the high values for the period under study to adjust the parametric distribution (Coles, 2001; Hawkes et al., 2008). Bernardara et al. (2014) recommend a double-threshold (u_p,u_s) approach to deal with auto-correlated environmental variables in a POT framework. First, physical declustering is performed by selecting a proper physical threshold u_p above which only the maximum WL value is selected for each event that exceeds this threshold. The independence of the maximum WL selected is ensured by setting a minimum interval between peak water levels. This interval is typically chosen to be representative of storm duration on the site under study. The value for u_p is set so that a sample of several hundred peak values can be selected to include both moderate and strong storm events. In practice, this corresponds to a number n of events per year between 5 and 10 in average. The second step of the double-threshold approach is a statistical optimization consisting in selecting a relevant value of the statistical threshold u_s ($u_s > u_p$), which is used in the formulation of the GPD, limiting both bias and variance (Bernardara et al., 2014). The choice of u_s is driven by classical visual tools such as mean residual life and parameters stability plots (see Coles (2001)).

The GPD is a distribution with two parameters (σ - scale parameter, and ξ - shape parameter). For a given threshold u_s, the cumulative distribution function (CDF) of the GPD is equal to the probability $P(X \leq x | X > u_s)$, where the random variable X describes observed peak water levels, and it can be written as follows:

$$G_{(\xi,\sigma)}(x) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u_s)}{\sigma}\right)_+^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - exp\left(-\frac{(x-u_s)}{\sigma}\right) & \text{if } \xi = 0 \end{cases} \qquad \text{for } x > u_s \qquad (1)$$

Where $\sigma > 0$ and the notation y_+ for $y \in \mathbb{R}$ is defined as $y_+ = max(y, 0)$. The support of the distribution is $u_s < x \leq u_s - (\sigma/\xi)$ if $\xi < 0$ and $x > u_s$ if $\xi \geq 0$. Whereas σ represents

the scale of the distribution (in units of x), ξ controls the behaviour of the distribution's tail. If $\xi < 0$, the distribution is bounded, we are in the Weibull domain. If $\xi > 0$ (resp. $= 0$), the distribution is unbounded, we are in the Fréchet (resp. Gumbel) domain. Contrary to the Weibull domain, a small change of ξ in the Fréchet domain involves significant changes of the distribution.

2.2 Bayesian framework

In contrast with classical statistical methods used to compute the parameters of the distribution and to derive extreme values (e.g., maximum likelihood, method of moments, probability weighted moments…), Bayesian techniques provide a natural framework to deal with uncertainties. They are designed to obtain the full posterior distribution of variables of interest and not only point estimates (Coles and Tawn, 2005).

Let's denote by θ the vector of parameters (ξ, σ). Its posterior distribution is related to the likelihood of data through Bayes' theorem:

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{f(D)} \quad (2)$$

Where $f(D|\theta)$ is the likelihood function of a set of observations D given the parameters vector, $f(\theta)$ is the prior distribution of the parameters and $f(D)$ is a normalising constant depending only on the observations. $f(\theta)$ translates the prior knowledge one may have about the parameters. In our study, we have no prior information about GPD parameters for our dataset. Consequently, we use a non-informative flat prior ($f(\theta) \propto 1$) (Payrastre et al., 2011). In that case, $f(\theta|D)$ is proportional to the likelihood function.

To sample effectively the posterior distribution of interest, we use a Markov Chain Monte Carlo (MCMC) algorithm. MCMC algorithms allow sampling values of the parameters from the posterior distribution, without computing the normalising constant. In this study, the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) is used to generate a set of 40,000 vectors θ with density $f(\theta|D)$. The convergence of the chain is checked numerically with the Geweke test (Geweke, 1992) and visually with trace plots. We can then compute the corresponding quantiles of WL according to the GPD. In particular, the mode of the set of vectors θ can be retrieved as the vector maximizing the likelihood function (because of the proportionality between $f(\theta|D)$ and $f(D|\theta)$). The associated quantiles x_T correspond therefore to the maximum likelihood estimates for WL. Credibility intervals on WL can also be estimated based on the large set of quantile values. Results can be displayed

on a return level plot once the correspondence between quantiles x_T $(x_T > u_s)$ and return periods T has been set up:

$$P(X > x_T) = \frac{1}{nT} \qquad (3)$$

where n is the mean number of exceedances of threshold u_p per year. The quantile x_T is said to be the standard estimative T-year return level and it is exceeded once on average by a peak event in T years. Conversely, T is said to be the standard estimative return period of level x_T. Since $P(X > x_T) = P(X > u_s)P(X > x_T|X > u_s)$, Eq. (3) can be rewritten in a more suitable form to construct a return level plot:

$$1 - G_\theta(x_T) = \frac{1}{\lambda T} \qquad (4)$$

where $\lambda = nP(X > u_s)$ is the mean number of exceedances of threshold u_s per year.

One main advantage of the Bayesian analysis is the possibility to integrate all the available information in a unique predictive distribution for extreme WL values (Coles and Tawn, 2005), which is defined as follows:

$$P(X \leq x|X > u_s, D) = \int_\theta G_\theta(x) f(\theta|D)\, d\theta \qquad (5)$$

Thus, the predictive distribution of a new observation x (given it is greater than u_s) can be easily estimated as the mean of GPD values calculated at x for the entire set of sampled parameters and can be represented on a return level plot after solving the equation $\tilde{T} = 1/\left(\lambda P(X > \tilde{x}_T|X > u_s, D)\right)$, where $\tilde{x}_T$ is the predictive return level associated with the predictive return period $\tilde{T}$. Although the terminology of predictive return period is loose, it is useful in order to maintain comparison with the standard analogue T (Coles and Tawn, 2005). Since all the uncertainty information has been integrated in the final result, credibility intervals are no longer defined. Instead, the value $p = 1/\left(n\tilde{T}\right)$ can be interpreted as the probability that, given all the available information, a future peak WL will exceed $\tilde{x}_T$.

Within the Bayesian framework, it is therefore possible to calculate and compare both standard estimative return levels x_T (equal to what would have been obtained using a classical maximum likelihood estimator) and predictive return levels $\tilde{x}_T$. While the predictive return levels incorporate all the uncertainty information, standard estimative return levels can be associated with credibility intervals which provide an overview of the uncertainty related to the quantiles x_T when visualised on a return level plot.

Finally, it is worth noting that for large return periods, the annual exceedance probability of a given level is directly available reading a return level plot constructed with peak event return periods contrary to the peak event exceedance probability of that level. Indeed, the former is equal to 1: T (or 1: $\tilde{T}$) whereas the latter is given by Eq. (3) 1: nT (or 1: $n\tilde{T}$) (cf Appendix A).

2.3 Likelihood formulation

The formulation of the likelihood function in Equation (2) depends on the characteristics of observations D (Payrastre et al., 2011). We can split the likelihood function into two parts, thus separating the systematic period (with systematic tide gauge records) and the historical period:

$$f(D|\theta) = \underbrace{f(D_{sys}|\theta)}_{\text{systematic likelihood}} \underbrace{f(D_{his}|\theta)}_{\text{historical likelihood}} \quad (6)$$

Let's assume we have a number s of systematic tide gauge observations above u_s $(x_1, \dots, x_s)$ and a historical period of n_y years with $H = h$ events above a perception threshold X_0 $(X_0 > u_s)$. The h events above X_0 during the historical period are supposed to be exhaustive. This is a necessary condition. Historical information can be of different types. The number h can thus be broken down into h_1 historical events whose water levels are known $(y_1, \dots, y_{h_1})$, a number h_2 of historical events that exceeded the perception threshold X_0 but whose exact water levels are not known and h_3 historical events whose water levels are known to be within a given range of values (lower bounds $y_1^{lb}, \dots, y_{h_3}^{lb}$ larger than X_0; upper bounds $y_1^{ub}, \dots, y_{h_3}^{ub}$). The general expression of the likelihood of systematic data is:

$$f\left(D_{sys}|\theta\right) = \prod_{i=1}^{s} g_\theta(x_i) \quad (7)$$

Where g_θ is the probability density function of the GPD for parameters θ.

The likelihood of historical data is:

$$f(D_{his}|\theta) = P(H|\theta) \prod_{j=1}^{h_1} g_{\theta,X|X>X_0}\left(y_j\right) \prod_{l=1}^{h_3} \left[G_{\theta,X|X>X_0}\left(y_l^{ub}\right) - G_{\theta,X|X>X_0}\left(y_l^{lb}\right)\right] \quad (8)$$

The first term of the right hand side is the probability of observing $h = h_1 + h_2 + h_3$ events above X_0 during n_y years whereas the two product terms specify the historical information for h_1 and h_3 historical events. Considering that the peaks exceeding u_s occur as a Poisson process (Coles, 2001), then the number of exceedances of u_s in n_y years follows a Poisson distribution of parameter λn_y. Consequently, the number of exceedances of X_0 in n_y years

follows a Poisson distribution of parameter $\lambda n_y P(X > X_0 | X > u_s) = \lambda n_y [1 - G_\theta(X_0)]$. Thus:

$$P(H|\theta) = \frac{\left(\lambda n_y [1 - G_\theta(X_0)]\right)^h}{h!} exp\left(-\lambda n_y [1 - G_\theta(X_0)]\right) \qquad (9)$$

Replacing Eq. (9) into Eq. (8) and since $g_{\theta,X|X>X_0}(x) = g_\theta(x)/\left(1 - G_\theta(X_0)\right)$ and $G_{\theta,X|X>X_0}(x) = G_\theta(x)/\left(1 - G_\theta(X_0)\right)$, Eq. (8) becomes:

$$f(D_{his}|\theta) = \frac{\left(\lambda n_y\right)^h}{h!} exp\left(-\lambda n_y [1 - G_\theta(X_0)]\right)\left(1 - G_\theta(X_0)\right)^{h_2} \prod_{j=1}^{h_1} g_\theta(y_j) \prod_{l=1}^{h_3} \left[G_\theta\left(y_l^{ub}\right) - G_\theta\left(y_l^{lb}\right)\right] \qquad (10)$$

So far, we have implicitly considered that the POT sample represents a stationary process. This assumption is systematically made in the hydrology field (Gaume et al., 2010). However, extreme WL exhibit long-term trends that cannot be ignored. Over the 20th century, these trends have been shown to be similar to those of mean sea-level (MSL) at most locations worldwide (Woodworth et al., 2011). To account for this behaviour in the systematic dataset, the linear trend is calculated for the entire tide gauge record and removed from the data. Then data are adjusted to have a mean sea-level equal to that of the reference year of interest. The historical perception threshold must also be corrected for the MSL rise (and called hereafter the adjusted perception threshold). Once this is done, Equation (10) becomes:

$$f(D_{his}|\theta) = \prod_{m=1}^{n_y} \left[\frac{\lambda^{h_m}}{h_m!} exp\left(-\lambda\left[1 - G_\theta\left(X_{0,m}\right)\right]\right)\left(1 - G_\theta\left(X_{0,m}\right)\right)^{h_{2,m}} \prod_{j=1}^{h_{1,m}} g_\theta(y_j) \prod_{l=1}^{h_{3,m}} \left[G_\theta\left(y_l^{ub}\right) - G_\theta\left(y_l^{lb}\right)\right]\right] \qquad (11)$$

Where $X_{0,m}$ is the adjusted perception threshold for historical year m and $h_{1,m}$, $h_{2,m}$, $h_{3,m}$ are respectively the numbers of historical events with known WL, with unknown WL and with WL within a range of values, that exceeded $X_{0,m}$ during year m. h_m is the total number of historical events that exceeded $X_{0,m}$ during year m ($h_m = \sum_{i=1}^{3} h_{i,m}$).

Appendix A

Appendix A: Relation between annual exceedance probability and peak event return period

Let's denote by maxy, the annual maximum. Using Eq. (3), the probability that the annual maximum of WL is greater than x_T is:

$$P(\text{maxy}(WL) > x_T) = 1 - P(X \leq x_T)^n = 1 - \left(1 - \frac{1}{nT}\right)^n \qquad \text{(A1)}$$

For large return periods T, more precisely when $nT \gg 1$, Eq. (A1) becomes:

$$P(\text{maxy}(WL) > x_T) \simeq \frac{1}{T} \qquad \text{(A2)}$$

which can be interpreted as follows: the standard estimation of the annual exceedance probability of level x_T is 1: T years. Thus, in that case, the annual exceedance probability is directly available reading a return level plot constructed with peak event return periods contrary to the peak event exceedance probability (cf Eq. (3)).

Similarly, in the case of the predictive distribution, we obtain:

$$P(\text{maxy}(WL) > \tilde{x}_T) \simeq \frac{1}{\tilde{T}} \qquad \text{(A3)}$$

which can be interpreted as follows: the probability that, given all the available information, next year's maximum WL will exceed $\tilde{x}_T$, is 1: $\tilde{T}$ years.