Technical Note: Comparison of methods for threshold selection for extreme sea levels

Technical Note: Comparison of methods for threshold selection for extreme sea levels

C. Caballero-Megido[1|2], J. Hillier[2], D. Wyncoll[3], L. Bosher[1] and B. Gouldby[3]

[1] School of Civil and Building Engineering, Loughborough University, LE11 3TU, UK

[2] Department of Geography, Loughborough University, LE11 3TU, UK

[3] Flood Management Group, HR Wallingford, Wallingford, OX10 8BA, UK

Abstract

Extreme value analysis is an important tool for studying coastal flood risk, but requires the estimation of a threshold to define an 'extreme', which is traditionally undertaken visually. Such subjective judgement is not accurately reproducible, so recently a number of quantitative approaches have been proposed. This paper therefore reviews existing methods, illustrated with coastal tide-gauge data and the Generalized Pareto Distribution, and proposes a new automated method that mimics the enduringly popular visual inspection method. In total five different types of statistical threshold selection and their variants are evaluated by comparison to manually derived thresholds, demonstrating that the new method is a useful, complementary tool.

1 Introduction

Extreme sea levels result when large storm surges coincide with high tides (Haigh et al., 2010). In coastal flood defence, it is important to have a good estimate of future extreme sea levels (McMillan et al., 2011) and this typically involves using extreme value theory (e.g. Li et al., 2014). Extreme value theory is used to develop techniques and models for describing the unusual rather than the usual (Coles, 2001).

A well-known problem in extreme value analysis (EVA) is threshold selection. Estimation of high water level events is commonly undertaken by fitting the annual maxima series (AMS), the r-largest values per year (Dixon et al., 1998; Haigh et al., 2010; McMillan et al., 2011) or

peaks over threshold (POT) (Bogner et al., 2012; Arns et al., 2013). AMS and r-largest value methods have their own complications; they are not recommended for analysing extremes storm surges as they are highly inefficient in their use of data (Bogner et al., 2012; Arns et al., 2013; Haigh et al., 2010), but these are not the subject of this paper. Skew surges (see Dataset) are usually assumed to be independent and identically distributed (i.i.d.) events (e.g. Bardet et al., 2011; Mazas et al., 2014) which will contain surges that are relatively common as well as those are truly exceptional and statistically extreme.

Once a sample of the i.i.d. events has been defined, the extreme subset of these events can be described and analysed. The POT approach consists of modelling exceedances above a pre-chosen 'statistical' threshold to distinguish the 'tail' of extreme values, which it is hoped are distinct and well characterized by a single statistical distribution such as the Generalized Pareto Distribution (GPD) (Coles, 2001; Ghil et al., 2011). In such an approach, a 'high' threshold (see Coles, 2001) is most likely to only contain data that are truly extreme. A lower threshold, however, will include more data. An optimal 'statistical' threshold is therefore a balance between ensuring that the chosen distribution well-fits all data defined as comprising the tail and retaining sufficient data for a meaningful analysis (Davison and Smith, 1990; Coles, 2001; MacDonald et al., 2011; Papalexiou et al., 2013; Wyncoll and Gouldby, 2013).

The literature suggests that most 'statistical' threshold selection methods are based on diagnostic plots (see in particular Coles (2001)). The first plot is the *Mean Residual Life* (MRL) plot. The second type of plots are the *Parameter Stability Plots*, created from fitting the GPD parameters (modified scale and shape) using a range of thresholds (Scarrott and MacDonald, 2012). Linearity above a threshold level in MRL plots indicates a consistency with a single GPD. Similar is indicated by a plateau above a threshold on the modified scale and shape stability plots, with the interpreter synthesising all the plots to select the lowest appropriate threshold such that information in the tail is not lost (see Scarrott and MacDonald, 2012). However, this subjectivity limits exact reproducibility, and interpreting the plots can be challenging (Davison and Smith, 1990; Coles, 2001; Solari and Losada, 2012a).

Consequently, a number of mathematically based threshold selection methods have been proposed, either based on parametric procedures (e.g. Rosbjerg et al., 1992), selecting a fixed percentile of data (e.g. Grabemann and Weisse, 2008; McMillan et al., 2011; Arns et al., 2013), or using 'mixture models' (e.g. Frigessi et al., 2002; Behrens et al., 2004; Mendes and Lopes, 2004; Tancredi et al., 2006; Carreau and Bengio, 2009). The first two of these methods

have the advantage of simplicity, whilst mixture models approximate the entirety of the dataset and estimate uncertainty in the threshold (Ghil et al., 2011; Scarrott and MacDonald, 2012; Solari and Losada, 2012b). However, despite the extent and rigor of this work, percentile selection (e.g. 95^{th} or 99^{th}) by parametric means or otherwise is ultimately a subjective procedure (Arns et al., 2013), and even mixture models do not produce a decisive verdict as evidenced by the variety of mixtures proposed (Scarrott and MacDonald, 2012; Solari and Losada, 2012a; Mazas et al., 2014). In parallel, Li et al. (2014) recently proposed a fourth way to select thresholds based on the Root Mean Square Error (RMSE) between empirical and analytical cumulative distribution functions (cdf). However, they only tested for thresholds over an arbitrary one year return period (RP) and their RMSE approach only optimises fit to the tail's shape. It is not intuitively clear how this acts to include maximum data from the tail; for example, why would more data be selected if the 5 most extreme data of 5,764 plausibly in the tail conformed perfectly to a GPD? The main pragmatic limitation to widespread uptake of the mathematical methods appears to be that they are less accessible to practitioners whilst producing a variety of results.

The aim of the paper is to determine how well existing threshold selection methods reproduce the visual based approach. It also proposes an automated procedure, which mimics the traditional graphical method, intending to benefit from its advantages and continued popularity with practitioners. The purpose of the Automated Graphic Threshold Selection (AGTS) method, in absence of *a priori* threshold value, is to guide in the choice of the threshold which requires judgment and expertise, making the process simple and approachable, whilst being reproducible and less subjective.

Firstly, the dataset is illustrated, and then the five threshold selection methods are briefly described: graphical, parametric, mixture, RMSE and the proposed AGTS method. After which they are evaluated by comparison to manually selected thresholds.

2 Dataset

Data from fourteen tide gauges (Fig. 1a) on the East Coast of the UK are used spanning the period 1920–2013, with an average of 35 years of coverage per gauge. Wick (1965-2013, 62% of complete data) and Dover (1924-2013, 80% of complete data) are used for illustration as examples of longer records in this dataset, and for their locations at the extremities of mainland UK (Fig. 1b, 1c).

The measure most useful for coastal flooding is arguably the skew surge (Lowe et al., 2009; Howard et al., 2010; McMillan et al., 2011). Skew surge is defined as the difference between the predicted astronomical high tide and nearest experienced high water (e.g. Bardet, 2011; McMillan et al., 2011; Mazas et al., 2014). There is one skew surge value per tidal cycle and there are two tides per day. This measure has two advantages. Firstly, of importance in this paper, a single, arguably independent (i.i.d.), value is calculated for each tide. Secondly, it removes all phase differences (timing differences) between astronomical and observed data, allowing for subsequent simpler probabilistic flood prediction. The skew surge data are from the UK National Tide Gauge Network, which are quality controlled and archived by the British Oceanographic Data Centre (NTSLF, 2014).

Additionally, four datasets provided within the 'ismev' (Stephenson, 2014) R package are used to demonstrate the general utility of the new threshold selection method. First dataset is 'rain' which is the daily rainfall accumulations at a location in south west England recorded over the period 1914-1962. Second, 'dowjones' data is the daily closing prices of the Dow Jones Index over the period 1996 to 2000. Third, 'euroex' is the daily exchange rates between the Euro and UK sterling. Finally, forth dataset is 'wavesurge' which is the surge heights (in metres) at a single location off south-west England.

3 Methods: Threshold selection techniques

If $X_1, \dots, X_n$ is an i.i.d. sequence of random variables then the GPD, above a threshold u (Eq. 1), defines a probability model for large values of the variable X (Coles, 2001). Its two parameters are a reparametrized scale parameter σ_u ($\sigma_u = \sigma - \xi u$) and a shape parameter ξ.

$$\mathrm{P}\{X > x | X > u\} = \left[1 + \xi \frac{(x-u)}{\sigma_u}\right]_+^{-1/\xi} \text{ for } x > u; \sigma_u > 0; \xi \in \Re \quad (1)$$

The selection of u is the subject of this note, and the methods below are used as described in order to compare them.

3.1 Graphical methods

Graphical methods are based on the MRL plot and the *Parameter Stability Plots*. MRL plots (e.g. Fig. 2a, 2b) were introduced by Davidson and Smith (1990). MRL plots involve plotting

u (x-axis) against the mean excess $\bar{v}$ (y-axis), which is the mean of the exceedances of the n i.i.d. observations (Eq. 2) for a sequence of thresholds u:

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n}(X_i - u) \quad (2)$$

If data are $X \sim \mathrm{GPD}(\sigma_u; \xi)$, then the empirical mean excess in Eq. 2 is $\mathrm{E}(X - u|X > u) = \sigma_u/(1-\xi)$, defined for $\xi < 1$ to ensure the mean exists. For any higher $\bar{v} > u$, the empirical mean excess becomes $\mathrm{E}(X - v|X > v) = (\sigma_u + \xi_v)/(1-\xi)$. Namely, the mean excess for data consistent with a GPD will be linear when seen on an MRL plot for $u > u_0$ where u_0 is an appropriate best threshold. Furthermore, if a distribution $\mathrm{GPD}(\sigma_u; \xi)$ is a valid model for $(X|X > u_0)$, estimates of ξ and $\sigma_u - \xi_{u_0}$ ought to be constant with respect to $u > u_0$. Thus, *Parameter Stability Plots* are simply estimates of σ_u and ξ (e.g. see Scarrott and MacDonald, 2012), calculated using all data above u as u varies, and a constant value is expected for $u > u_0$. So, in summary, it is expected of a tail well-fitted by a GPD that the MRL plot will be linear above a 'best' u (i.e. u_0), and that simultaneously the two GPD parameters will be constant above this best u_0. Practically, this involves attempting to place u_0 at the lower end of a reasonably well-defined plateau in the *Parameter Stability Plots* (Scarrott and MacDonald, 2012). These rules (as illustrated in Fig. 2) were applied independently by three practitioners to select the thresholds, and dubbed 'manual thresholds'.

Thompson et al. (2009) developed a less subjective, semi-automatic threshold selection procedure that uses elements of the manual selection approach, but does not replicate it. Thompson et al. (2009) difference the GPD parameter estimates (i.e., $\tau = \sigma - \xi$), effectively taking the derivative of this plotted against u, which they argue should be normally distributed. They then take the highest u that fulfils this criterion when tested using a χ^2 test for normality. Various parameters need to be set (e.g. test significance level), and a method that directly mimics the manual approach may prove more intuitive. Thompson's method is not implemented here.

3.2 Parametric methods

Rosbjerg et al. (1992) introduced a parametric procedure based on calculating the threshold as the mean value of the original dataset plus three standard deviations. This method assumes that the data are normally distributed. Other parametric methods (e.g. Grabemann and Weisse, 2008) are based on a fixed percentile of data, with the range of percentiles varying between

the 97.5th (McMillan, 2011) and the 99.7th (Arns et al., 2013). Pre-defined thresholds can be considered as an initial procedure to make an *a priori* threshold choice when dealing with multiple datasets is time-consuming. Another benefit of this approach is that it is easily automated. Each of the parametric methods is reproduced exactly here.

3.3 Mixture models

Mixture models estimate the domain in which a distribution fits the tail well via inference methods (e.g. Frigessi et al., 2002; Behrens et al., 2004; Mendes and Lopes, 2004; Tancredi et al., 2006; Carreau and Bengio, 2009). In this way, appropriate tail fits can be achieved using automated estimation of the 'statistical' threshold and, provided the 'bulk' (see Fig. 3) distribution model is sufficiently flexible and, by this or other means, the bulk and tail fit do not strongly influence each other (Scarrott and MacDonald, 2012). The threshold estimations are obtained as a by-product of the model fitting procedure. Overviews of possible approaches are given in Ghil et al. (2011) with a useful illustration in Fig. 4 of Scarrott and MacDonald (2012). The 'R' package 'evmix' (Scarrot and Hu, 2014) implements the main extreme value mixture models to make these tools accessible for researchers and practitioners (Hu and Scarrott, 2013); it also produces diagnostic plots of model fit and quantifies the uncertainty in threshold estimation. Following Hu and Scarrott (2013), five mixture models are therefore tested here using 'evmix' (Scarrot and Hu, 2014).

In the first method, Frigessi et al. (2002) designed a mixture model for datasets containing positive values only, and the full dataset is used for inference for GPD component. It is a dynamically weighted mixture model, where one part is the GPD and the other is a light-tailed density distribution such as the Weibull. There is no explicit threshold in this approach, the transition between GPD and Weibull is gradual following a Cauchy cdf weighting (see Scarrott and MacDonald, 2012). However, a threshold is assigned to be the point over which the weighted contribution of the GPD term is higher. This approach has two limitations (Scarrott and MacDonald, 2012); a lack of robustness in the inversion and a tendency for the bulk to at least partially influence the estimated character of the tail.

The method of Behrens et al. (2004) is arguably the simplest of the extreme value mixture models with which to fit the entirety of the dataset (Fig. 3), and the use of a threshold aims to decouple the bulk and the tail. This method combined a parametric form for the bulk distribution (e.g. normal or gamma) up to some threshold with a GPD for the tail above this

threshold, evaluated using Bayesian inference. The threshold is explicitly estimated as a parameter with its own probability density function (pdf), which is what allows uncertainty in its value to be assessed. In a related approach, Mendes et al. (2004) fitted models to both of a distribution's tails, even though this makes little difference for skew surge data where only the 'upper' tail (i.e., right-hand tail on Fig. 3) is considered. They used a normal distribution for the bulk and GPD models to fit the two tails. Thresholds for both tails are based on estimating the best proportion of observations for each tail by maximising the log-likelihood over all possible pairs of proportions. Carreau and Bengio (2009) further developed this approach by extending the GPD to a 'hybrid Pareto' and by placing a continuity constraint on the probability density function of the threshold's location and on its first derivative at the threshold. However, as they note, the continuity requirements effectively create some linkage between the bulk and the tail.

Alternatively, to avoid the influence of assuming a form for the bulk distribution, Tancredi et al. (2006) proposed a mixture model that combines non-parametric density estimation using an unknown number k of uniform distributions for the bulk (Scarrott and MacDonald, 2012). The bulk initial extends from a lower threshold u_{low}, which is known to be well below any reasonable estimate of a best threshold u_0, up to u above which a GPD applies. Variants of u are tested to determine u_0. This method, however, is computationally complex with difficulties of ensuring convergence (Thompson et al., 2009; Scarrott and MacDonald, 2012), and some subjectivity exists in the choice of Bayesian prior parameters (Tancredi et al., 2006).

3.4 RMSE methods

Li et al. (2014) proposed an RMSE measure (Eq. 3) of the difference between analytical and observed cdfs of *X* to select suitable thresholds.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{observed value}_i - \text{fitted value}_i)^2} \quad (3)$$

The RMSE observed values are the fitted cdfs assigned i/n at each of the observed values from the observed data, where i is the number of tied observation at that value. For these same data a GPD was fitted above u, whose parameters $(\sigma_u; \xi)$ were obtained by maximum likelihood. From this GPD, an analytical or fitted cdfs was constructed. Then, these analytical

cdfs were compared to the ones of the observed cdfs for each and every observation above u. Each observation is treated as a potential u, each u has a RMSE calculated for it, and the u with the lowest RMSE considered best. Arbitrarily, Li et al. (2014) also chose only to consider events with an RP greater than approximately 1 year.

3.5 Automated Graphic Threshold Selection

The proposed Automated Graphic Threshold Selection (AGTS) method is based on calculations and parameters that are as described in Section 3.1. AGTS has the following components (Fig. 4a, 4b), described initially for a single test threshold u: the RMSE between an estimated fitted line above u and mean excess $\bar{v}$ (E_{MRL}) plotted in the MRL plot (Fig. 5), the RMSE of deviations from a constant value of the two GPD parameters (E_{STAB}) above u, and the Exceedances Rate (E_{ER}) above u. These calculations are repeated for a range of potential u, equivalent to contemplating different values for u in a visual assessment. Thus, this method mimics the graphical method providing a computational support in the thresholds selection process (see section 3.1 and Fig. 2). Specifically, as Li et al. (2014), each observation in X is treated as a potential threshold u, and all components are calculated and summed to create the AGTS metric (Eq. 4) for each u.

Possible values of u are considered between the 0% and the 99.99% quantile of the dataset; this is arbitrary, but in line with published practice (Grabemann et al., 2008), and there are always > 2 values above 99.99% in the data used here. The metrics are combined to make the AGTS measure as in Eq. 4, with each component normalised to amplitude of 1.0 within the assessed range of u (Fig. 4c, 4d).

$$\mathrm{AGTS} = \frac{1}{3}(E_{\mathrm{MRL}} + E_{\mathrm{STAB}} + E_{\mathrm{ER}}) \quad (4)$$

In detail, the calculation of E_{MRL} is based on the MRL plot and 'observed' data points that comprise it (i.e., each point is a mean excess of observed skew surges over a threshold) and proceeds as follows for each potential optimal threshold u. Firstly, a line is set above u, using the parameters of the GPD ($\sigma_u; \xi$), namely the intercept is $\sigma_u/(1-\xi)$ and the gradient is $\xi/(1-\xi)$ as in Section 3.1 (e.g. Coles, 2001; Scarrot and MacDonald, 2012). Next, a RMSE as in Eq. 3 is calculated for the vertical difference (as plotted) between the observations (mean excesses) and the line for the region above or equal to u (see Fig. 5). The set of E_{MRL} values are then normalised such that the lowest one is set to 0 and the highest to 1.0. The E_{STAB} is

calculated in a similar way to the E_{MRL} but using both GPD parameters, the modified scale and shape; a line that is flat (see Section 3.1 above) with a constant value equal to that of the parameter at u is used to calculate an RMSE. These two RMSE values are then normalised (the highest values is set to one and the lowest to zero) before being added together to give E_{STAB}. E_{STAB} is itself then normalised in the same way to a values between 0 and 1. E_{ER} is simply the number of X_i above u divided by the total number of data considered (i.e., $X_i > 0$). Note that this implicitly normalises E_{ER} and that E_{ER} decreases as the threshold considered increases (see Fig. 4).

Finally, to stabilise the raw AGTS results, a polynomial is fitted. Then the final estimate of an optimal 'statistical' threshold is selected as the highest inflection point of the polynomial with a total AGTS error (Eq. 4) of less than 0.5. Polynomials of orders 2 to 10 were assessed for their suitability (see Table A1). The selected 5$^{\mathrm{th}}$ order polynomial is used as an illustration in Fig. 4c and 4d (solid black line). AGTS is implemented in 'R' (R Core Team, 2014) using several packages, notably 'evmix' (Hu and Scarrott, 2013). AGTS took < 1 minute to assess a threshold for these gauges, which each have ~50,000 data points.

3.6 Return values

A way to compare the threshold selection method is to calculate the return level H_m for the m-year return period such as 50, 100 and 250 years. Having identified the threshold and estimating the scale and shape parameters ($u,\ \sigma,\ \xi$) for each threshold selection method, the return value for the GPD can be estimated by Eq. 5 (Agarwal et al., 2013):

$$H_m = u + \frac{\sigma}{\xi}\left[(m\lambda)^{\xi} - 1\right] \quad (5)$$

where λ is the average number of exceedances per annum (Eq. 6), k is the size of the set of excesses above a high threshold and n is the number of years for which data is available.

$$\lambda = \frac{k}{n} \quad (6)$$

4 Results

This section presents comparisons of manually selected thresholds with those estimated from other methods. To simplify the comparison, the manual thresholds are averaged across

interpreters, percentile-based techniques are also averaged, and the proposed AGTS method is set to using a 5^{th} order polynomial. In addition, the thresholds estimated with the Carreau and Bengio (2009) method are not considered further as they are below the mean skew surge and so not appropriate for these data. Thus, the threshold selection methods considered are: manual (average), Rosbjerg et al. (1992), percentiles (average), both the Normal and Gamma of Behrens et al. (2004), Mendes and Lopes (2004), 1-year RP and AGTS (5^{th} order). For more detail see Table A2, which tabulates the thresholds obtained from all the threshold selection methods.

Fig. 6 shows thresholds selected by the different selection methods in Wick and Dover. Only the AGTS and 1-year RP thresholds are the inside the range of those estimated manually for both sites. This is a first indication of which methods have most capability to give estimates most similar to manual interpretation, which are taken here as 'correct' since the purpose of this study is to assess (semi-)automated methods' ability to reproduce manual ones. In addition, boxplots (Fig. 7) may also be used to gain an initial impression of the results taken collectively. The boxplots show that the mixture models estimate lower thresholds than manually determined. The range of thresholds (i.e. 2^{nd} to 3^{rd} quartiles) is most similar to the manual ones for the arbitrary 1-year RP (Li et al., 2014) and the AGTS thresholds. No clear indicators one way or the other are evident for the parametric methods.

Table 1 is the basis for a more detailed comparison of the visually selected manual thresholds (Section 3.1) and the (semi-)automated thresholds (Sections 3.2–3.5). Three metrics are used to evaluate the fit: coefficient of determination r^2 and its p-value (i.e., that $r^2 \neq 0$), the gradient of line fitted by an Ordinary Least Squares (OLS) regression, and an RMSE value of the differences between the manually estimated thresholds and the (semi-)automated thresholds. Three metrics are used as they reflect slightly different aspects of agreement or otherwise, and boxplots are used to gain an initial impression. Values of r^2 indicate how correlated the (semi-)automated estimates are with the manual ones. That is, when all the gauges are taken together, how strongly are the (semi-)automated estimates predictive of the 'correct' manual ones. Thus, values of r^2 close to 1 are desirable, with associated p-values < 0.05 indicating a relationship that is statistically significant at 95% certainty. In addition the OLS gradient measures the role of systematic biases, or how close to the true values the (semi-)automated estimates are. For instance, a perfect relationship with $r^2 = 1$, could still underestimate values by a half, with an OLS gradient of 0.5. Thus, gradients close to +1 are

desirable, with negative values indicating poor prediction by the (semi-)automated methods. The RMSE value is a measure of the absolute differences in the size of each pair of manual-automated estimates at the same gauge, and as such measures a combination of both these effects and whilst also explicitly requiring that the methods produce the same values at the same gauge. So, RMSE values close to zero are perhaps the most powerful indicator of a successful method.

Fig. 8 shows the scatterplots of the manual thresholds (average) and the main threshold selection methods: Rosbjerg et al. (1992), percentiles (average), Behrens et al. (2004) Normal and Gamma, Mendes and Lopes (2004), 1-year RP and AGTS. Positive relationships (i.e., OLS gradients > 0) exist between the methods tested and the manual control data for all approaches except Mendes and Lopes (2004) (Fig. 8). The slopes of the OLS regression lines range between 0.42 and 1.67, with the percentiles and AGTS methods giving regressions between the thresholds close to the 'correct' value of 1, but the 1-year RP is closest. r^2 values range between 0.014 and 0.778, with the lower values produced by the mixture model methods because their threshold are 'clustered' (i.e., have a very small range on the boxplots) and so cannot well match the range of the manual estimates. The percentiles, 1-year RP and AGTS methods produce thresholds whose relationships with the manually estimates ones are all significant ($p < 0.05$) and have $r^2 > 0.7$, but the AGTS method's r^2 is highest. RMSE errors range between 0.080 and 0.560, with only 1-year RP and AGTS thresholds presenting values < 0.1, closest to a perfect fit of 0.

In summary, the r^2 measure indicates that AGTS is the best method to replicate manual threshold selection, it comes second to the 1-year RP for OLS estimated slope, and both of these methods have the best (and very similar) RMSE values. So, the AGTS appears joint best at reproducing manual estimates with the arbitrary and ultimately subjective 1-year RP method (see Fig. 6).

Additionally, the sea level associated with a selection of return periods are used to illustrate the impact of choosing different threshold selection methods (see Fig. 9). The return levels for 50, 100 and 250 years in Wick (Fig. 9a) and Dover (Fig. 9b) increase with return period as expected for each method, but there are systematic differences across return levels that vary between sites. In Wick, for instance, manual estimates are near the mean, whilst in Dover they are near to the first quantile. In Wick, percentiles, 1-year RP and AGTS return levels are the

closest to the manual return levels. In Dover, the return values closest to the manual ones are Mendes and Lopes (2004), 1-year RP and AGTS.

Bootstrap testing, with replacement, of the data shows (Table A1) that the AGTS thresholds have relatively small uncertainty, and testing various polynomials orders for use in the method (i.e., 2^{nd} to 10^{th}) show that it is not overly sensitive to this choice; namely 4^{th} and 6^{th} order still perform well; see Table A1 for more detail on this comparison between the manual thresholds (average) and ATGS using polynomials of order 2 to 10.

AGTS also performs well on illustrative datasets provided within the 'ismev' (Stephenson, 2014) R package (see Fig. 10). Firstly, AGTS better replicates manual selection than the method of Thompson et al. (2009) for the 'rain' data (Fig. 10a). Both methods make estimates close to the value of $u \approx 30$ set by Coles (2001) that represents the manual, subjective procedure. However, by estimating 27 mm the ATGS is closer to 30 mm than the 20 mm of Thompson et al. (2009). Second, AGTS estimates a threshold of 1.84, close to the value of 2 recommended by Coles (2001) for the 'dowjones' data (Fig. 10b). A value of 0.9 is recommended by Coles (2001) and 0.993 is estimated by AGTS. Third, AGTS estimates a threshold of 0.392 mm for the 'wavesurge' data (Fig. 10d), which is marginally closer to that estimated by Mendes and Lopes (2004) using the visual procedure (0.50 mm) than their mixture model (0.368 mm).

5 Discussion

This research assesses various methods of estimating thresholds above which events are classed as extreme with the so-called POT approach used in EVA. The basis of the work is a comparison with manual thresholds selected by three practitioners who had no idea of the thresholds produced by any of the methods before they made their manual estimates. The intention is therefore not to determine 'objectively' correct thresholds, such as might be done with well-designed synthetic datasets (e.g. Coles, 2001; Thompson et al., 2009). The objective is to automatically create thresholds that reliably reflect those widely produced by the many practitioners who use the graphical approach (Solari and Losada, 2012b; Agarwal et al., 2013; Bernardara et al. 2014; Mazas et al., 2014).

With the various measures assessed in an overall sense, they show that the AGTS best replicates manual interpretations. Like the method of Thompson et al. (2009) it is

computationally efficient (see Fig. 4c and 4d), bootstrap estimates of uncertainty are possible and subjectivity is avoided. However, AGTS is also perhaps intuitively simpler as it replicates the manual process, and is relatively easy to implement as the R code is supplied as supplementary material. This is not to say that AGTS will be the most appropriate for practitioners in all circumstances. Percentile-based methods, and the very similar process of RP selection, perform really quite well, and perhaps because they are somewhat arbitrary and subjective are quick and easy to implement and easily comparable between studies. Possibly the approach used here (e.g. Fig. 7) could be a practical means of selecting an appropriate RP or percentile for wider use from manual interpretation of a few sites.

From the results of the return levels obtained by the different threshold selection methods (Fig. 9), the differences in estimated extreme sea levels is notable (i.e. 5–10%) and the influence of threshold selection method equates to a non-trivial difference in the RP selected. The influence of the spectrum of methods with respect to manual estimates varies by geographic location and is not necessarily predictable. So, since return levels estimates are incorporated into sea defence design such as along the East coast of the UK, the choice should be carefully considered.

The utility of AGTS as a complementary method is illustrated with a UK skew surge dataset, which is perhaps closer to normal that some data constraining extremes (Fig. 3). However, it also performs well on the ‘rain’, ‘dowjones’, ‘euroex’ and ‘wavesurge’ dataset from ‘ismev’ R package (Coles, 2001), and the estimates thresholds are within ±3% of the manually estimated values for all the datasets in Coles (2001). Note, only these four datasets provide thresholds to compare across (see Fig. 10). Note also that sea levels can act as a boundary condition for 1D and 2D flood routing models, and may have a significance comparable to that of the fluvial morphology (Bhuyian et al., 2014). Thus, the new proposed threshold selection method also has implications for hydraulic modelling, especially in areas such as coastal basins (Nardi et al., 2009).

Conclusions

By comparing a range of threshold selection methodologies to the visual selections of three independent interpreters, it is possible to conclude that AGTS is a useful complementary technique to estimate threshold for Extreme Value Analysis of sea levels, and is shown to be

applicable to a diverse range of other fields such as finance (Figs. 10b and 10c). AGTS generated favourable results in the majority of cases, both in the skew surge dataset and 'ismev' datasets. Including well-tested datasets from ismev validates the general utility of the new threshold selection method; no re-coding was necessary for this comparison as thresholds for the ismev datasets already existed, and so no doubt can arise about our application of the published threshold selection methods.

Regarding its application in practice to flood risk management, AGTS can enhance the coastal design process through the reproducible and rapid calculation of return levels consistent with traditional practice. Thresholds selected by AGTS might be useful indicators to be used in determining, for example, initial thresholds to define flood risk areas. AGTS proposes consistent thresholds that can be regularly updated in light of new data and do not vary from person to person as staff change roles, and so will help to improve coastal flood guidance for future coastal defence models. This increase in efficiency may also allow more regular updates, or more locally-applicable thresholds to be produced.

This paper provides an automated threshold selection method that effectively reproduces thresholds that are consistent with those obtained using the manual method. Correspondingly, the effect of uncertainty associated with threshold selection estimation using the bootstrap testing with replacement is measured, which reduces the subjectivity of the well-established manual approach. However, we note that other numerical methods will be equally valid, depending upon the circumstances. Indeed, by replicating the manual method it is possible that we have encoded some implicit failings that we are unaware of, although by making the manual method reproducible the work in this paper has at least opened the realm of manual threshold detection up to greater scrutiny.

Benefits of AGTS are that it is computationally efficient, reproducible, and of the methods tested most closely replicates manual thresholds estimated by practitioners. It may also be a useful tool for non-experts, or speed up work-flows that include manual threshold estimation. The amount of prior expertise required in the threshold selection is reduced considerably, bringing closer EVA to a broader range of users. Our new automated method is significantly easy and quick to implement (freely available in the appendix) from a practical point of view, for long datasets. The novelty of AGTS lies in it being the first procedure to replicate the traditional visual threshold selection method.

Further research might involve developing graphical user interface that can provide results in a more commercial way; a user friendly interface requires minimum technical skills for practitioners without knowledge of R software. Currently, AGTS has not been tested against synthetic time series of data, which would provide a further insight into its efficacy if the synthetic were appropriately designed, but we expect that threshold estimate superior to other methods could be commonly obtained.

Appendix A

	RMSE	Regression line	r^2	p
AGTS (2nd polynomial order)	0.181	y = 0.820x + 0.273	0.666	0.00037
AGTS (3rd polynomial order)	0.185	y = 0.964x + 0.203	0.695	0.00021
AGTS (4th polynomial order)	0.142	y = 0.756x + 0.276	0.664	0.00039
AGTS (5th polynomial order)	0.086	y = 0.668x + 0.286	0.779	0.00003
AGTS (6th polynomial order)	0.060	y = 0.745x + 0.221	0.741	0.00008
AGTS (7th polynomial order)	0.006	y = 0.605x + 0.269	0.737	0.00009
AGTS (8th polynomial order)	0.040	y = 0.551x + 0.288	0.737	0.00009
AGTS (9th polynomial order)	0.037	y = 0.530x + 0.305	0.735	0.00007
AGTS (10th polynomial order)	0.056	y = 0.560x + 0.272	0.747	0.00006

Table A1. Regression summary of the manual thresholds (average) and the polynomial of orders 2 to 10 assessed to calculated the automated threshold (AGTS).

	Manual				Parametric				
	Manual 1	Manual 2	Manual 3	Manual (average)	Rosbjerg et al. (1992)	97.5th	99.5th	99.7th	Percentiles (average)
LERWICK		0.480	0.510	0.495	0.432	0.307	0.429	0.459	0.398
WICK	0.460	0.660	0.630	0.583	0.490	0.357	0.507	0.559	0.474
MORAY FIRTH		0.490	0.480	0.485	0.490	0.345	0.490	0.527	0.454
ABERDEEN		0.680	0.660	0.670	0.456	0.334	0.488	0.540	0.454
LEITH		0.500	0.680	0.590	0.478	0.351	0.508	0.558	0.472
NORTH SHIELDS	0.650	0.540	0.660	0.617	0.479	0.337	0.516	0.579	0.477
WHITBY		0.500	0.840	0.670	0.563	0.400	0.597	0.666	0.554
IMMINGHAM	0.750	0.790	0.990	0.843	0.556	0.390	0.630	0.701	0.574
CROMER		0.660	0.950	0.805	0.607	0.438	0.697	0.806	0.647
LOWESTOFT		0.980	1.100	1.040	0.634	0.484	0.812	0.925	0.740
FELIXSTOWE	0.88	0.880	0.680	0.812	0.625	0.442	0.698	0.781	0.640
HARWICH		0.570	0.560	0.565	0.579	0.407	0.640	0.753	0.600
SHEERNESS	0.750	0.750	0.620	0.707	0.612	0.410	0.643	0.717	0.590
DOVER	0.700	0.740	0.900	0.780	0.565	0.400	0.652	0.725	0.592

	Mixture				RMSE	AGTS (Automated)			
	Behrens et al. (2004) Normal	Behrens et al. (2004) Gamma	Mendes and Lopes (2004)	Carreau and Bengio (2009)	1 year RP	AGTS (5th order)	AGTS (4th order)	AGTS (6th order)	AGTS (no smoothed)
LERWICK	0.092	0.251	0.079	-0.059	0.497	0.332±0.030	0.369	0.479	0.386
WICK	0.202	0.291	0.202	-0.074	0.632	0.514±0.062	0.487	0.530	0.641
MORAY FIRTH	0.172	0.281	0.207	-0.053	0.573	0.353±0.032	0.343	0.396	0.456
ABERDEEN	0.080	0.301	0.148	-0.080	0.624	0.427±0.038	0.415	0.484	0.573
LEITH	0.168	0.298	0.103	-0.067	0.644	0.451±0.068	0.437	0.516	0.845
NORTH SHIELDS	0.133	0.283	0.151	-0.066	0.671	0.530±0.071	0.478	0.529	0.493
WHITBY	0.169	0.331	0.202	-0.068	0.765	0.655±0.073	0.606	0.631	0.962
IMMINGHAM	0.210	0.333	0.145	-0.063	0.821	0.677±0.059	0.578	0.684	1.158
CROMER	0.180	0.360	0.070	-0.075	0.955	0.857±0.103	0.790	0.726	1.082
LOWESTOFT	0.136	0.740	0.092	-0.076	1.069	1.064±0.188	0.944	1.096	1.432
FELIXSTOWE	0.187	0.365	0.100	-0.027	0.919	0.644±0.061	0.506	0.659	0.784
HARWICH	0.213	0.388	0.188	-0.061	0.870	0.631±0.029	0.602	0.696	0.753
SHEERNESS	0.221	0.329	0.118	-0.057	0.892	0.533±0.084	0.463	0.572	1.137
DOVER	0.184	0.265	0.021	-0.080	0.856	0.798±0.095	0.655	0.825	0.972

Table A2. Overview of the thresholds selected (in meters) by the different methods (manual, parametric, mixture, RMSE and automated). Errors on AGTS 5th order are 1 standard deviation, derived by bootstrapping with replacement 30 times.

Appendix B

AGTS.txt

Acknowledgements

The dataset used in this publication was provided by HR Wallingford. The main author would like to thanks the co-authors whom have provided valuable feedback. Authors are particularly grateful to the anonymous reviewers for their useful comments.

References

Agarwal, A., Venugopal, V. and Harrison, G. P.: The assessment of extreme wave analysis methods applied to potential marine energy sites using numerical model data. Renew. Sust. Energ. Rev., 27, 244-257, 2013.

Arns, A., Wahl, T., Haigh, I. D., Jensen, J. and Pattiaratchi, C.: Estimating extreme water level probabilities: A comparison of the direct methods and recommendations for best practise, Coast. Eng., 81, 51-66, 2013.

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 Sys., 11, 1627–1639, 2011.

Behrens, C. N., Lopes, H. F. and Gamerman, D.: Bayesian analysis of extreme events with threshold estimation, Stat. Model., 4(3), 227–244, 2004.

Bernardara, P., Mazas, F., Kergadallan, X., and Hamm, L.: A two-step framework for over-threshold modelling of environmental extremes, Nat. Hazards Earth Sys., 14(3), 635-647, 2014.

Bogner, K., Pappenberger, F., and Cloke, H. L.: Technical Note: The normal quantile transformation and its application in a flood forecasting system, Hydrol. Earth Syst. Sc., 16(4), 1085-1094, 2012.

Bhuyian, Md. N. M., Kalyanapu, A. J., and Nardi F.: An Approach for Digital Elevation Models (DEM) Correction by Improving Channel Conveyance, J. of Hydrol. Eng., doi: 10.1061/(ASCE)HE.1943-5584.0001020, 2014.

Carreau, J. and Bengio, Y.: A hybrid Pareto model for asymmetric fat-tailed data: the univariate case, Extremes 12(1), 53–76, 2009.

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

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

Dixon, M. J., Tawn, J. A. and Vassie, J. M.: Spatial modelling of extreme sea-levels, Environmetrics, 9(3), 283-301, 1998.

Frigessi, A., Haug, O. and Rue, H.: A dynamic mixture model for unsupervised tail estimation without threshold selection, Extremes 5(3), 219–235, 2002.

Ghil, M., Yiou, P., Hallegatte, S., Malamud, B.D., Naveau, P., Soloviev, A., Friederichs P., Keilis-Borok, V., Kondrashov, D., Kossobokov, V., Mestre, O., Nicolis, C., Rust, H.W., Shebalin, P., Vrac, M., Witt, A. and Zaliapin, I.: Extreme events: dynamics, statistics and prediction, Nonlinear Proc. in Geoph., 18(3), 295-350, 2011.

Grabemann, I. and Weisse, R.: Climate change impact on extreme wave conditions in the North Sea: an ensemble study, Ocean Dynam., 58, 199–212, 2008.

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(9), 838-849, 2010.

Howard, T., Lowe, J. and Horsburgh, K.: Interpreting Century-Scales Changes in Southern North Sea Storm Surge Climate Derived from Coupled Model Simulation, J. Climate Vol.25, 2010.

Hu, Y. and Scarrott, C.J.: evmix: Extreme Value Mixture Modelling, Threshold Estimation and Boundary Corrected Kernel Density Estimation. Available on CRAN. http://www.math.canterbury.ac.nz/~c.scarrott/evmix, 2013.

Li, F., van Gelder, P. H. A. J. M., Ranasinghe, R., Callaghan, D. P. and Jongejan, R. B.: Probabilistic modelling of extreme storms along the Dutch coast. Coast. Eng., 86, 1-13, 2014.

Lowe, J. A., Howard, T. P., Pardaens, A., Tinker, J., Holt, J., Wakelin, S., Milne, G., Leake, J., Wolf, J., Horsburgh, K., Reeder, T., Jenkins, G., Ridley, J., Dye, S. and Bradley, S.: UK Climate Projections science report: Marine and coastal projections. Met Office Hadley Centre, Exeter, UK, 2009.

MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M. and Russell, G.: A flexible extreme value mixture model, Comput. Stat. Data An., 55(6), 2137-2157, 2011.

Mazas, F. and Hamm, L.: A multi-distribution approach to POT methods for determining extreme wave heights, Coast. Eng., 58, 385–394, 2011.

McMillan, A., Batstone, C., Worth, D., Tawn, J., Horsburgh, K. and Lawless, M.: Coastal flood boundary conditions for UK mainland and islands, Project SC060064/TR2: Design sea levels, Environment Agency, 2011.

Mendes, B. V. M. and Lopes, H. F.: Data driven estimates for mixtures, Comput. Stat. Data An., 47(3), 583–598, 2004.

Nardi, F., O'Brien, J.S., Cuomo, G. Garcia R., and Grimaldi S.: Updating flood maps using 2D models in Italy: A case study, Flood Risk Management: Research and Practice, Samuels et al. (Eds.), CRC Press - Taylor & Francis group, London, ISBN-13: 978-0-415-48507-4, 2009.

NTSLF (National Tidal and Sea Level Facility): http://www.ntslf.org/data, last access: 23 October 2014.

Papalexiou, S. M., Koutsoyiannis, D. and Makropoulos, C.: How extreme is extreme? An assessment of daily rainfall distribution tails, Hydrol. Earth Syst. Sc., 17(2), 851-862, 2013.

R Core Team: R, A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2014.

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.

Scarrott, C. J. and Hu, Y. (2014). evmix 0.2.6: Extreme Value Mixture Modelling, Threshold Estimation and Boundary Corrected Kernel Density Estimation. Available on CRAN. http://www.math.canterbury.ac.nz/~c.scarrott/evmix, 2014.

Scarrott, C. J. and MacDonald, A. E.: A review of extreme value threshold estimation and uncertainty quantification, REVSTAT Statistical Journal 10 (1), 33-60, 2012.

Solari, S. and Losada, M. A.: A unified statistical model for hydrological variables including the selection of threshold for the peak over threshold method. Water Resour. Res., 48(10), 2012a.

Solari, S. and Losada, M. A.: Unified distribution models for met-ocean variables: Application to series of significant wave height. Coast. Eng., 68, 67-77, 2012b.

Stephenson, A. G.: ismev 1.40: An Introduction to Statistical Modeling of Extreme Values. R package. http://CRAN.R-project.org/package=ismev, 2014.

Tancredi, A., Anderson, C. and O'Hagan, A.: Accounting for threshold uncertainty in extreme value estimation, Extremes 9(2), 87–106, 2006.

Thompson, P., Cai, Y., Reeve, D. and Stander, J.: Automated threshold selection methods for extreme wave analysis. Coast. Eng., 56(10), 1013-1021, 2009.

Wyncoll, D. and Gouldby, B. Integrating a multivariate extreme value method within a system flood risk analysis model, Journal of Flood Risk Management, 2013.

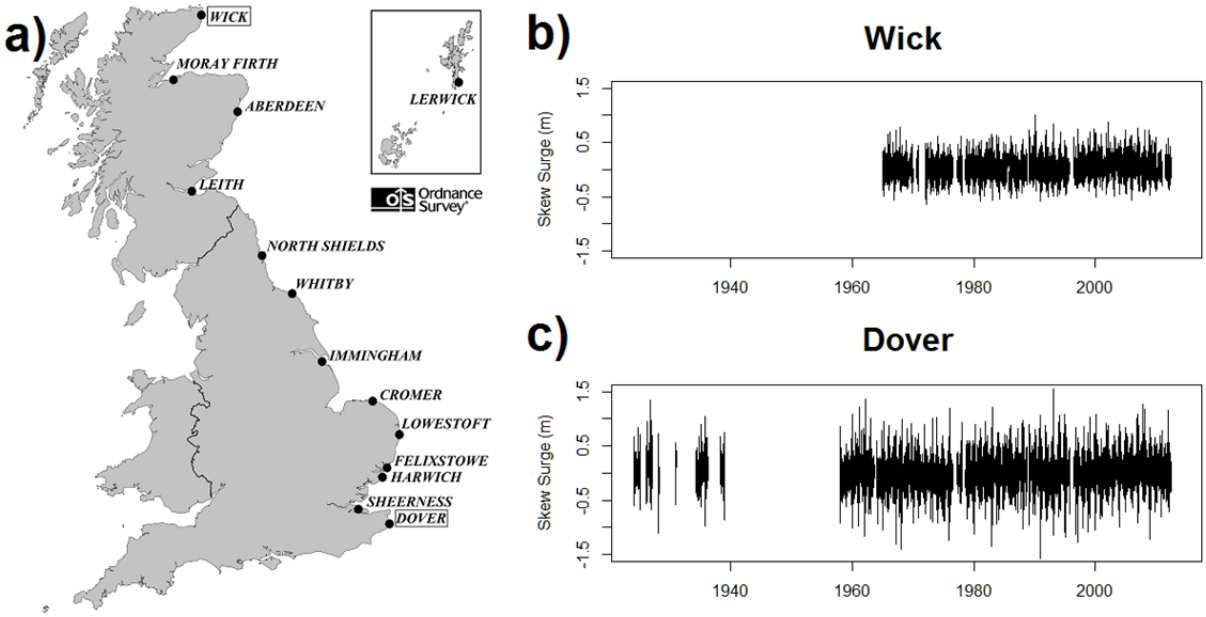

Figure 1. Study sites and examples of data. a) is a map of the study area, b) and c) are time series plots of skew surge (m) in Wick and Dover.

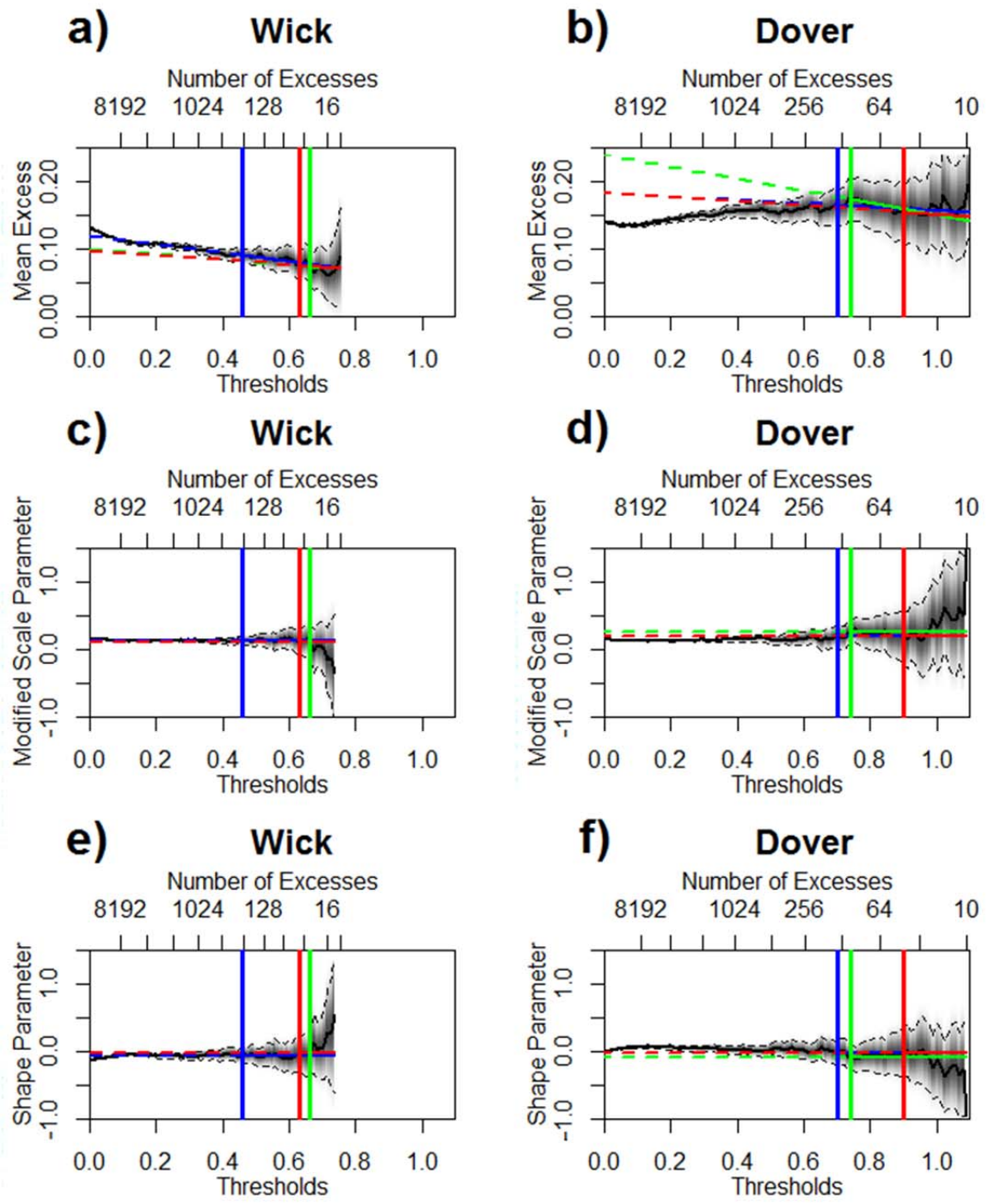

Figure 2. a) b) Mean residual life, c) d) modified scale threshold stability and e) f) shape threshold stability plots for Wick and Dover. Thresholds manually selected by three interpreters are shown (in blue, green and red lines): in Wick (0.46, 0.66 and 0.63) and Dover (0.70, 0.74 and 0.90). Symmetric confidence intervals are provided at the 95% level. The sampling density is shown by a greyscale, where lighter greys indicate low density. The plots were generated using the 'evmix' R package (Hu and Scarrott, 2013).

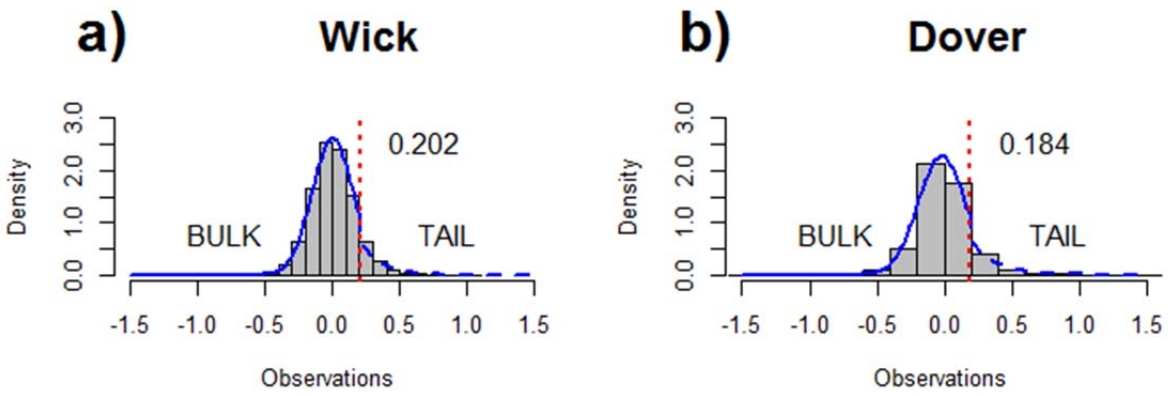

Figure 3. Histogram of skew surges of Wick (a) and Dover (b). The blue line shows the probability density function of the fitted mixture method (Behrens et al., 2004) with normal bulk (blue solid line) and parameterised tail (blue dashed line). The vertical red dotted line indicates the estimated threshold selected with Behrens et al. (2004) method.

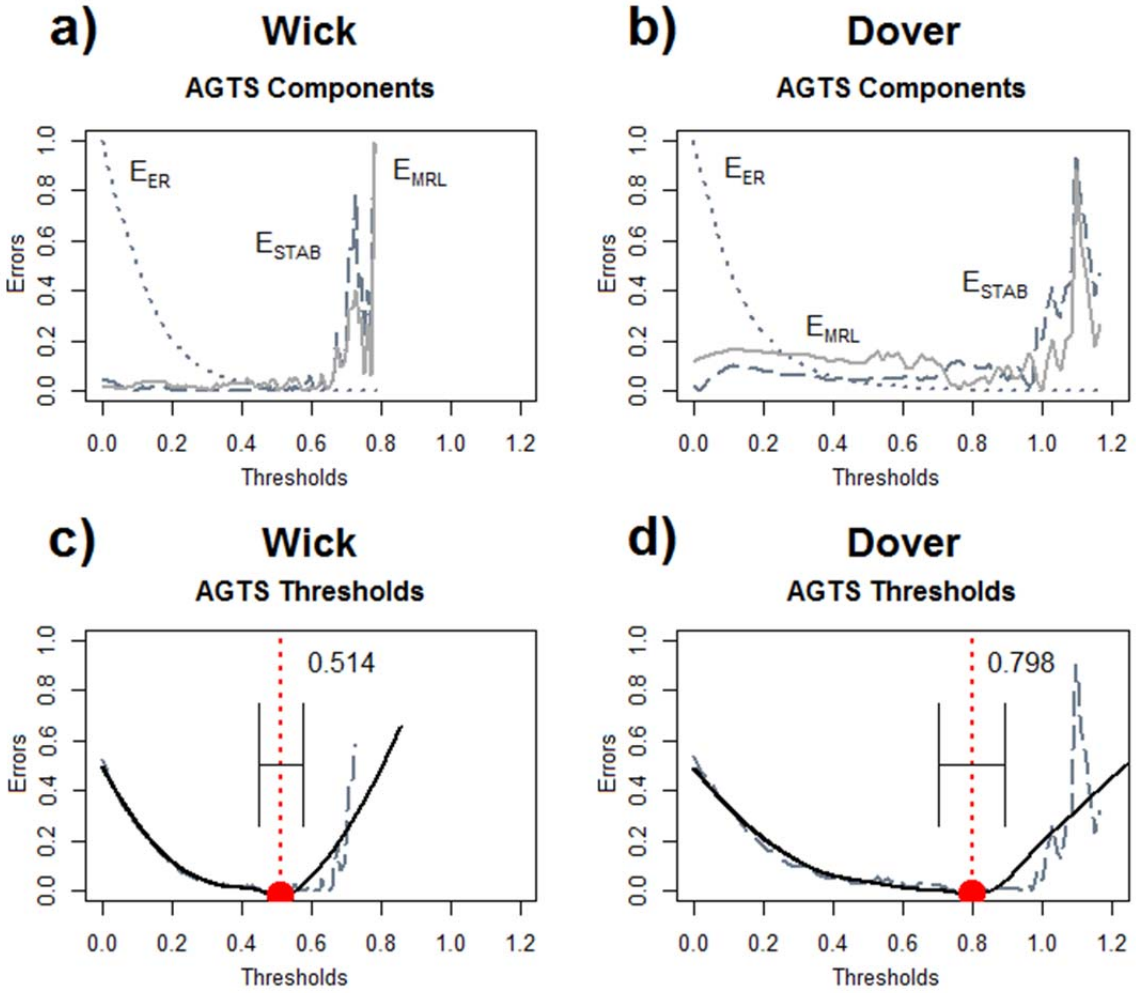

Figure 4. The AGTS components of Wick (a) and Dover (b). AGTS method in Wick (c) and Dover (d) where a smoothed solid line (black) is used to stabilise the output in the grey dashed line. The vertical dotted line indicates the estimated threshold at the red dot which is the highest inflection point of the AGTS. Error bars are shown from the bootstrap process.

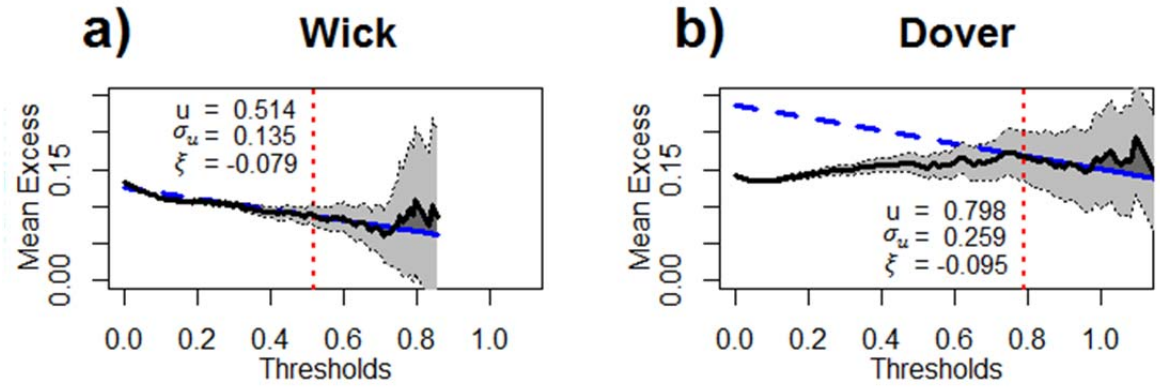

Figure 5. *Mean Residual Life* (MRL) plot for Wick (a) and Dover (b). The vertical red dotted line indicates the estimated threshold with the AGTS method. Symmetric confidence intervals are provided at the 95% level (in grey). The value of E_{MRL} for a specific threshold is the RMSE between the blue line (linear estimate constructed from the GPD parameters) and the mean excess (black line) for the region above or equal to u (areas in dark grey).

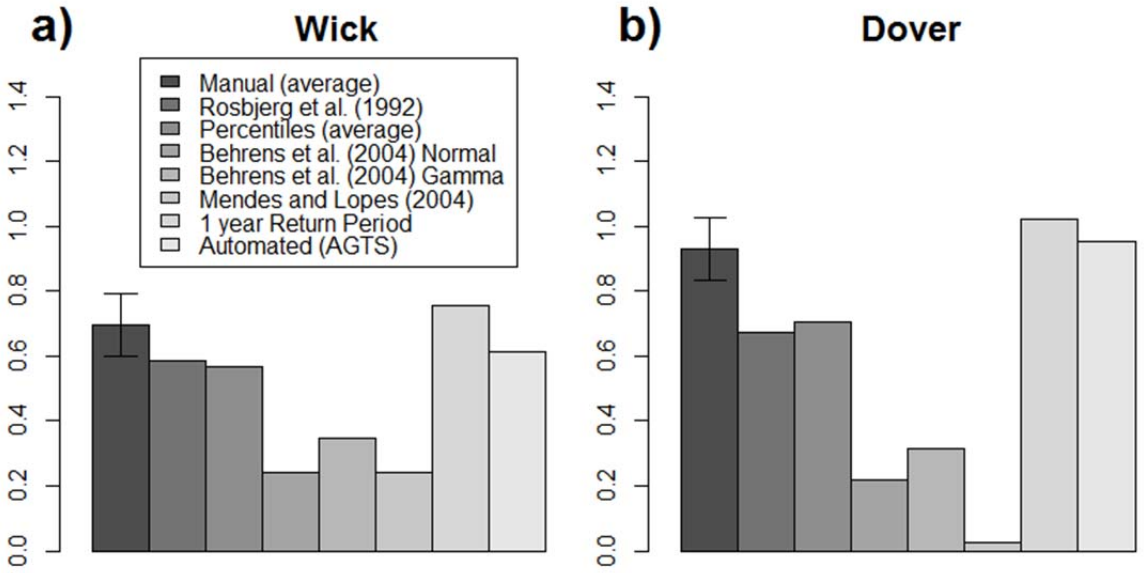

Figure 6. Thresholds selected by the different methods in Wick (a) and Dover (b). Error bars (standard error at 95% confidence interval) were added to demonstrate the variability in the manual thresholds.

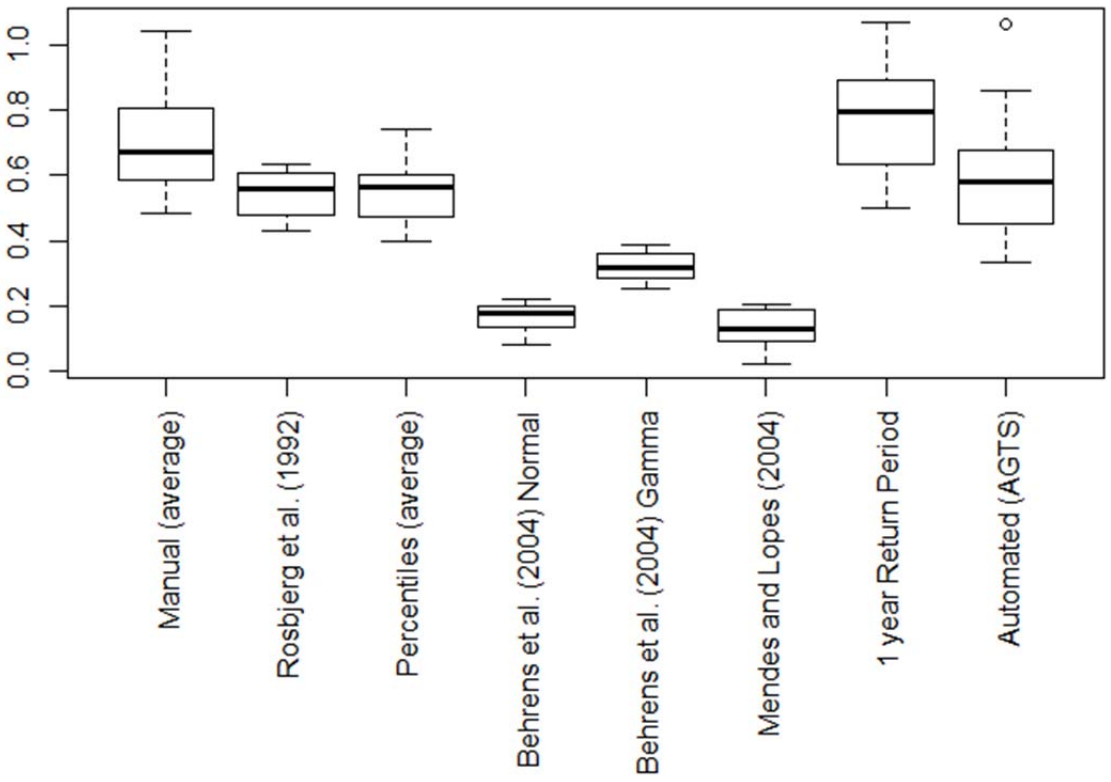

Figure 7. Boxplots of the thresholds selected in the 14 gauges. The boxplots describe the distribution of the thresholds, showing minimum, first quartile, mean, third quartile and maximum.

	RMSE	Regression line	r^2	p
Rosbjerg et al. (1992)	0.150	y = 1.677x - 0.216	0.552	0.00234
Percentiles (average)	0.143	y = 1.373x - 0.061	0.732	0.00010
Behrens et al. (2004) Normal	0.522	y = 0.424x + 0.619	0.014	0.68411
Behrens et al. (2004) Gamma	0.372	y = 1.985x + 0.059	0.307	0.03974
Mendes and Lopes (2004)	0.560	y =-1.264x + 0.855	0.219	0.09138
1 year Return Period	0.080	y = 0.779x + 0.090	0.715	0.00014
Automated (AGTS)	0.085	y = 0.668x + 0.286	0.778	0.00003

Table 1. Regression summary of the thresholds selected in the 14 gauges, a comparison between the manual thresholds (average) and the main threshold selection methods.

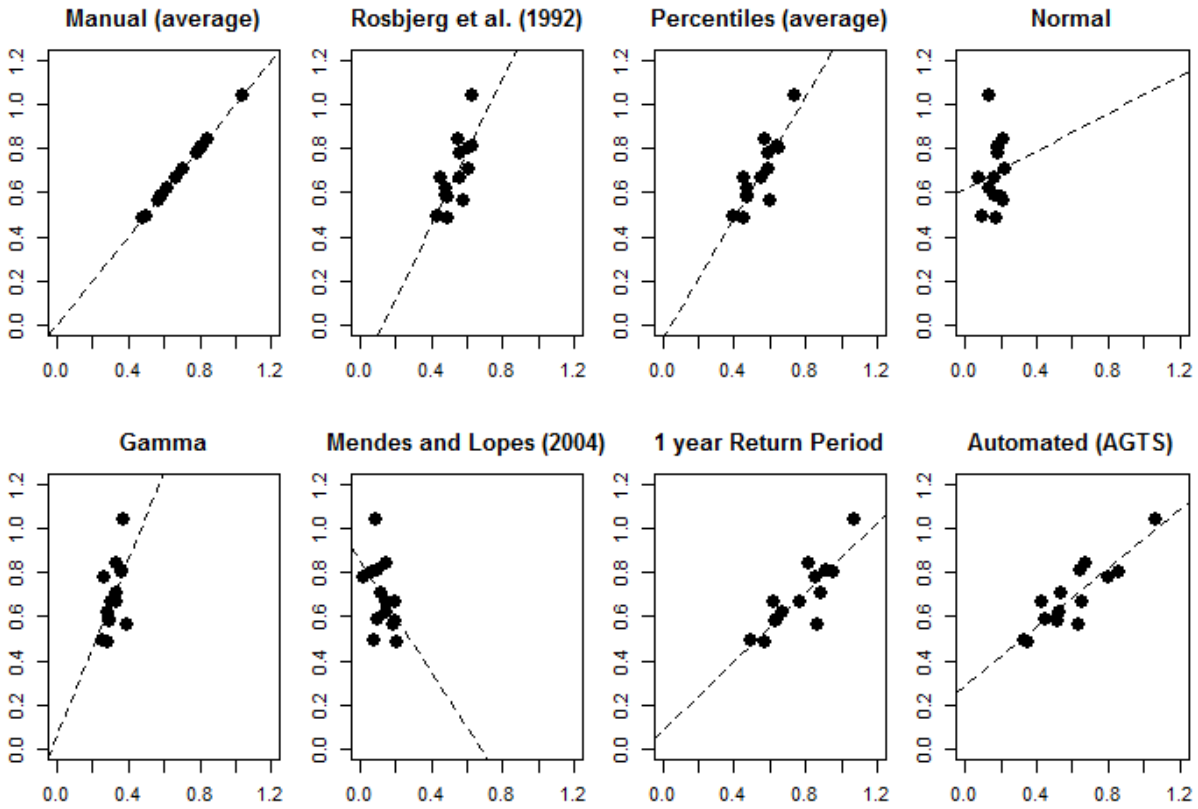

Figure 8. Scatterplots of the thresholds selected in the 14 gauges, a comparison between the manual thresholds (average) and the main threshold selection methods.

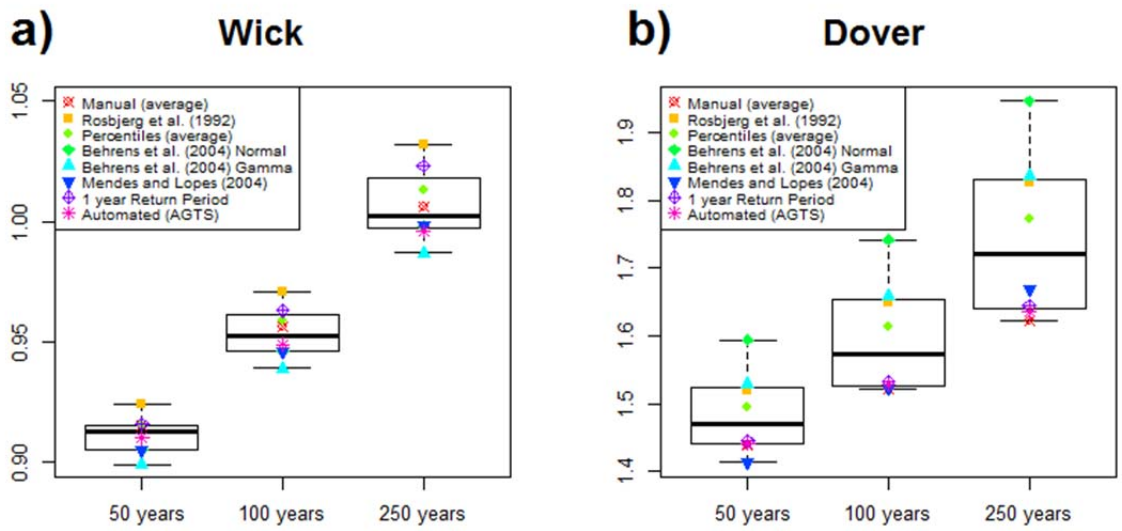

Figure 9. Return levels for 50, 100 and 250 years in Wick (a) and Dover (b) for the threshold selection methods analysed. Note that the y-axes are in different scales.

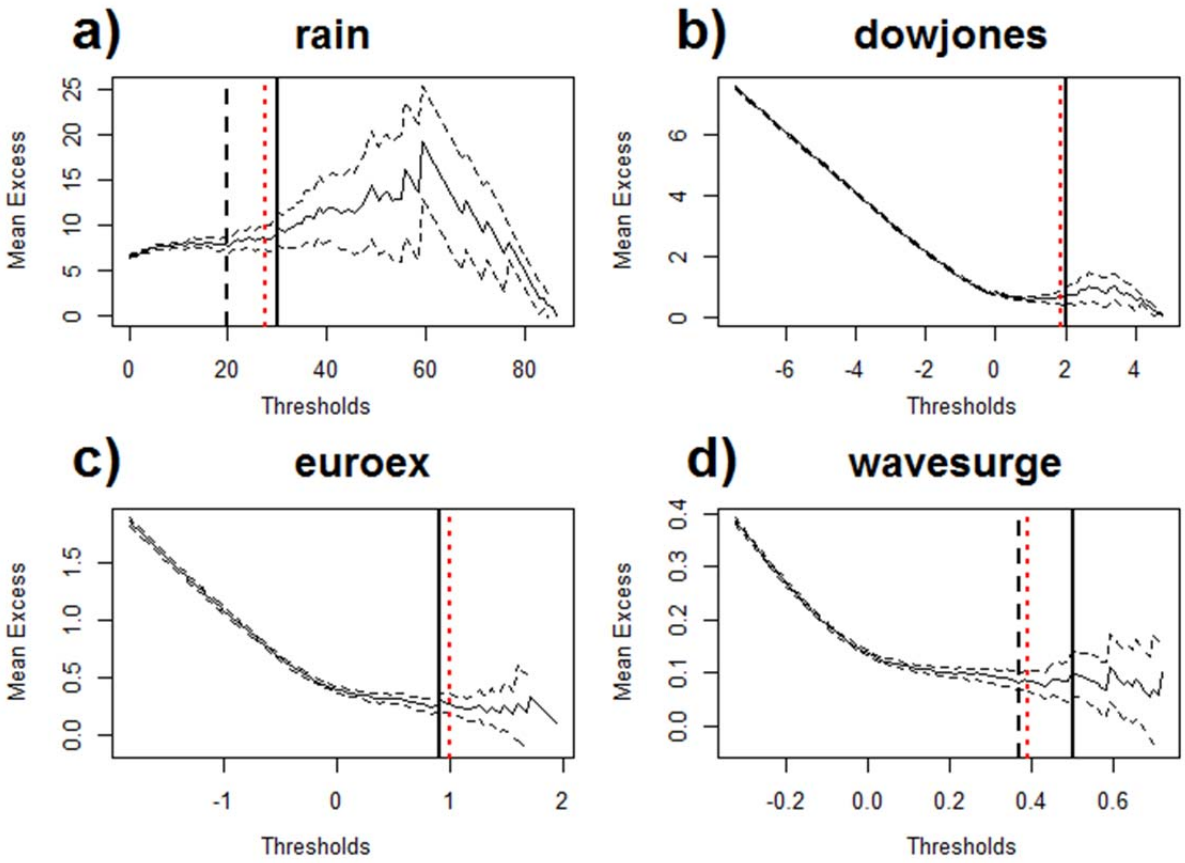

Figure 10. MRL plots from the 'ismev' R package (Stephenson, 2014) applied to the 'rain' (a), 'dowjones' (b), 'euroex' (c) and 'wavesurge' (d) datasets. The dashed black lines are the thresholds produced by Thompson et al. (2009) procedure, thee solid black lines are the threshold recommended by Coles (2001), and the dotted red lines are the AGTS threshold.