Research Article Environmetrics

Received: 30 November 2010, Revised: 30 November 2011, Accepted: 7 December 2011, Published online in Wiley Online Library: 27 March 2012

(wileyonlinelibrary.com) DOI: 10.1002/env.2133

Estimating return levels from serially dependent extremes

Lee Fawcett[a]* and David Walshaw[a]

In this paper, we investigate the relationship between *return levels* of a process and the strength of serial correlation present in the extremes of that process. Estimates of long period return levels are often used as design requirements, and *peaks over thresholds* analyses have, in the past, been used to obtain such estimates. However, analyses based on such declustering schemes are extremely wasteful of data, often resulting in great estimation uncertainty represented by very wide confidence intervals. Using simulated data, we show that—provided the *extremal index* is estimated appropriately—using *all* threshold excesses can give more accurate and precise estimates of return levels, allowing us to avoid altogether the sometimes arbitrary process of cluster identification. We then apply our method to two data examples concerning sea-surge and wind-speed extremes. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: bootstrap; clusters; extremal index; extreme value theory; peaks over thresholds; return levels; sea surges; temporal dependence; wind speeds

1. INTRODUCTION

Statistical modelling of environmental extremes has a very practical motivation—reliability: anything that is built needs to have a good chance of withstanding the weather/environment it is exposed to for the duration of its working life. This has implications for civil engineers. For example, they need to know how strong to make buildings, or how high to build sea walls, motivating the need to estimate the strongest wind or the highest tide. Thus, estimating such *return levels* is often the primary objective in an analysis of extreme values. A commonly employed estimation procedure is the peaks over threshold (POT) approach (e.g. Davison and Smith 1991). Here, an appropriate limiting distribution for independent excesses over a high threshold u is fitted to the largest exceedance selected from each 'cluster' of values above u—these *cluster peak* excesses are used instead of *all* excesses to avoid the issue of serial correlation, usually present in environmental extremes. Anderson (1990), in his discussion of Davison and Smith, gave a theoretical basis for the use of POT, explaining how, *in the limit*, the distribution of excess for a randomly chosen threshold exceedance has exactly the same distribution as that of a cluster peak—a consequence of length-biased sampling. Return levels are then estimated using high quantiles of the limiting distribution, fitted values of which are obtained by inversion of the corresponding distribution function.

Data on extremes are, by their very nature, scarce. Although methods such as POT *do* make use of more information on extremes than the classical annual maxima approach (e.g. Coles, 2001, Ch. 3), discarding all but the cluster maxima is still wasteful of precious data. This usually manifests itself when we come to quantify our uncertainty in parameter estimates. For example, standard errors attached to estimates of return levels as a result of a POT analysis are often so large that confidence intervals constructed in the usual way (estimate $\pm$ 1.96 $\times$ standard error) can be quite impractical (e.g. the analysis of river flow extremes in Davison and Smith, 1991). Using the profile log-likelihood (e.g. Coles, 2001, Ch. 2) improves return level inference (the log-likelihood surface for return levels is rarely symmetric), but resulting 95% confidence intervals, for example, can still be very wide.

Extensive simulations in Fawcett and Walshaw (2007) also show that using cluster peak excesses results in significantly biased estimates of model parameters and associated quantiles. Findings from their simulations are supported by an analysis of sea-surge extremes in the same paper, as well as analyses of wind speed extremes in Fawcett (2005), river flow extremes in Eastoe and Tawn (2012), and hurricane-induced wave heights in Northrop and Jonathan (2011). Fawcett and Walshaw (2007) showed that, by initially ignoring any serial correlation between threshold excesses, bias in parameter estimates is virtually eliminated. However, the strength of serial correlation present needs to be taken into account when estimating return levels—something that Fawcett and Walshaw overlooked.

In this paper, we investigate the relationship between return levels of extremes of a process and the strength of serial correlation present in that process. We argue that, provided this dependence can be quantified in an appropriate way, inference for return levels using all extremes can improve over a typical POT analysis, increasing estimation accuracy and precision.

* *Correspondence to: Dr Lee Fawcett, School of Mathematics and Statistics, Herschel Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.. E-mail: lee.fawcett@ncl.ac.uk*

a *School of Mathematics and Statistics, Herschel Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.*

2. BACKGROUND

2.1. Motivating examples

Figure 1 (top left) shows a series of 3-hourly measurements of sea-surge heights at Newlyn, a coastal town in the southwest of England, collected over a 3-year period. The sea surge is the meteorologically induced non-tidal component of the still-water level of the sea. Figure 1 (top right) shows the first 3 years of a series of hourly gust maximum wind speeds recorded at High Bradfield, a high altitude site in the Pennines. The practical motivation for the study of such data is that structural failure of, perhaps, a sea wall or a building, is possible if extreme surges or extreme wind speeds (respectively) are observed. The plots in the bottom row of Figure 1 show each series plotted against the version at lag 1 (used to highlight serial dependence—see Section 2.3.2); the green lines in these plots represent high thresholds above which events are classified as extreme (see Section 2.3.1).

2.2. Statistical modelling of extremes

Let $\{X_n\}$ denote a stationary sequence of random variables with common distribution function F, and let $M_n = \max\{X_1, \ldots, X_n\}$. It is typically the case that, as $n \to \infty$,

$$\Pr(M_n \leq x) \approx F^{n\theta}(x) \tag{1}$$

where $\theta \in (0, 1)$ is known as the *extremal index*; for example, Leadbetter and Rootzén (1988). As $\theta \to 0$, there is increasing dependence in the extremes of the process; for an independent process, $\theta = 1$. In practice, F is unknown, and very small discrepancies in estimates of F obtained from observed data can lead to rather substantial discrepancies for F^n (and thus $F^{n\theta}$). Initially concerned with the independent case (i.e. $\theta = 1$), classical extreme value theory sought families of limiting models for F^n for large n. This leads to the generalised extreme value (GEV) distribution (e.g. Jenkinson 1955), with distribution function

$$\mathcal{G}(y; \mu, \varsigma, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{y-\mu}{\varsigma}\right)\right]^{-1/\xi}\right\} \tag{2}$$

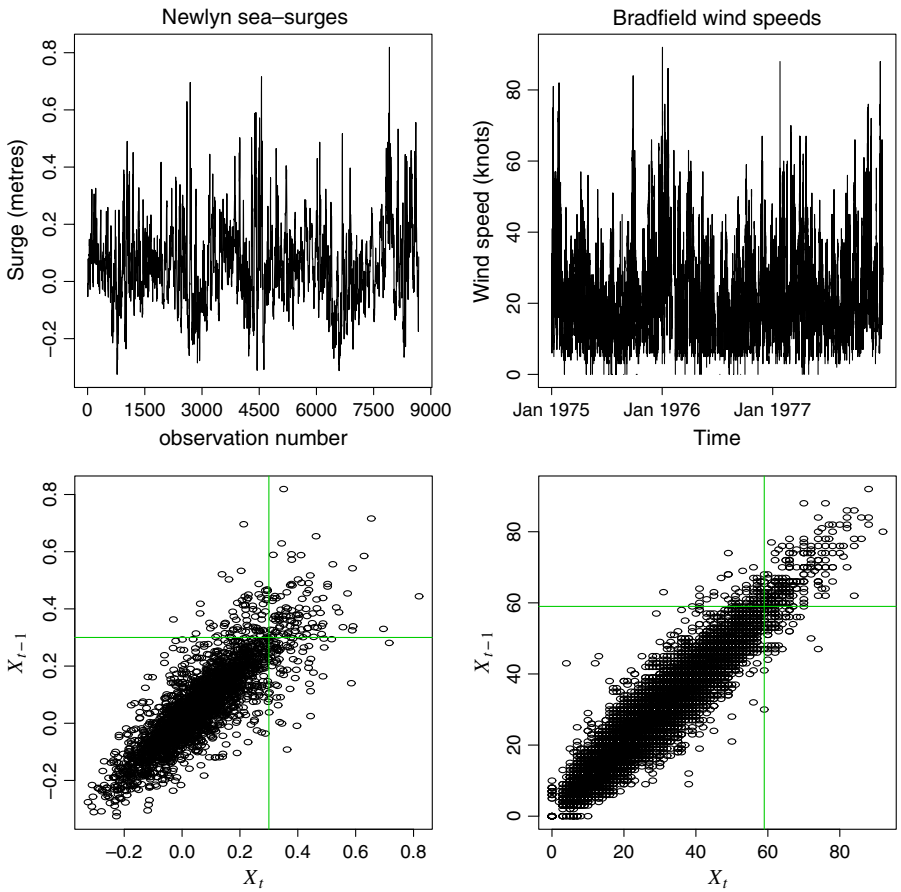

Figure 1. Left-hand-side: Newlyn sea-surge data. Right-hand side: Bradfield wind speed data. Top: time series plots. Bottom: plots of time series against series at lag 1, with thresholds

defined on $\{y : 1+\xi(y-\mu)/\varsigma > 0\}$, where $-\infty < \mu < \infty$, $\varsigma > 0$ and $-\infty < \xi < \infty$ are location, scale and shape parameters, respectively. The GEV can be used to model a set of block maxima $\{M_\tau\}$ with block length τ; the calendar year is often used for τ, giving the set of *annual maxima*.

Pickands (1975) showed that for large enough u, the distribution of $(X-u)$, conditional on $X > u$, is approximately generalised Pareto with distribution function

$$\mathcal{H}(y;\sigma,\xi) = 1-\left(1+\frac{\xi y}{\sigma}\right)^{-1/\xi} \tag{3}$$

defined on $\{y : y > 0 \text{ and } (1+\xi y/\sigma) > 0\}$, where $\sigma = \varsigma + \xi(u-\mu)$ and ξ are the generalised Pareto distribution (GPD) scale and shape parameters (respectively). The GPD, being the limiting distribution for excesses over a high threshold u, provides a natural way of modelling extremes of time series such as those shown in Figure 1. Modelling extremes in this way can be less wasteful than a block maxima approach using the GEV, as more extremes are usually pressed into use. In the rest of this paper, we shall focus on the use of the GPD in (3) as a model for extremes.

2.3. Modelling threshold exceedances in practice

2.3.1. Threshold selection

A *mean residual life plot* (e.g. Coles, 2001, Ch. 4) can be employed to select a suitably high threshold u (see the green lines in Figure 1). For independent threshold excesses (i.e. when $\theta = 1$), maximum likelihood (for example) can then be applied directly to (3) to estimate the GPD scale and shape parameters for excesses over u. The initial choice of threshold can then be assessed by fitting the GPD to other thresholds $u^* > u$ and checking for stability in estimates of $\sigma^{*\dagger}$ and ξ.

2.3.2. Temporal dependence and non-stationarity

Figure 1 reveals substantial short-term serial correlation in both the Newlyn sea surges and Bradfield wind speeds (the lag 1 autocorrelations are 0.856 and 0.957, respectively). Above u, serial correlation still seems to exist, indicating the presence of *extremal* dependence. The tendency for extremes of environmental time series to 'cluster' gives $\theta < 1$ in (1); this means that either (i) a suitable method for estimating θ needs to be considered or (ii) a scheme which filters out a set of approximately independent threshold excesses needs to be employed. The most commonly used approach is (ii), which ensures that $\theta \approx 1$. Specifically, 'runs declustering' is most often used: a cluster of extremes is deemed to have terminated as soon as κ consecutive observations fall below u. From each cluster, the maximum is then extracted, and the GPD is fitted to the set of cluster peak excesses (a POT analysis).

Various approaches have been developed to deal with issues of seasonal variability, such as that apparent in Figure 1 for the Bradfield wind speed data. Walshaw (1991) investigated the use of Fourier forms for the GPD parameters, allowing them to vary continuously; Fawcett and Walshaw (2006) used a piecewise seasonality approach fitting a separate GPD to each set of monthly excesses; often, attention is restricted to the season for which extremes are largest (and approximately stationary within; e.g. Smith *et al.*, 1997).

2.3.3. Return level estimation

Suppose the GPD(σ,ξ) is a suitable model for threshold exceedances $(X-u)$ of a threshold u by a variable X. Then, from Equation (3), it can easily be shown that

$$\Pr(X \le x) = 1-\lambda_u\left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi} \tag{4}$$

where $\lambda_u = \Pr(X > u)$ is the 'threshold exceedance rate' and can be estimated empirically as the proportion of observations above u. Estimates of an extreme quantile z_s can then be obtained by equating the right-hand side of (1) to $1-s^{-1}$ (where $F^n(x)$ is given by (4)) and then solving for $x = z_s$, where z_s is the *s-observation return level* with associated *return period* s; this can be thought of as the level that is exceeded once, on average, every s observations. It is usually more convenient to work with return levels on an annual scale; the *r-year* return level, z_r, is then given by

$$z_r = u+\frac{\sigma}{\xi}\left[\left(\lambda_u^{-1}\left\{1-\left[1-1/(rn_y)\right]^{\theta^{-1}}\right\}\right)^{-\xi}-1\right] \tag{5}$$

where n_y is the number of observations per year. In a POT analysis, it is assumed that $\theta = 1$; r-year return level estimates $\hat{z}_r$ are thus obtained by replacing (λ_u,σ,ξ) in Equation (5) with (for example) their maximum likelihood estimates $(\hat{\lambda}_u,\hat{\sigma},\hat{\xi})$. Confidence intervals for estimated return levels are usually constructed via profile likelihood, owing to the severe asymmetry of the likelihood surface often encountered for return levels; that is, we obtain the set of values z_0 for which $2\{\ell_p(\hat{z}_r)-\ell_p(z_0)\}$ is not significant when compared to a χ^2_1 distribution, where ℓ_p is the associated profile log-likelihood.

$^\dagger\sigma^* = \sigma+\xi u^*$; then (σ^*,ξ) are *threshold invariant*—if exceedances over a threshold u are generalised Pareto distributed, then exceedances over all $u^* > u$ are also GPD with the same parameters (e.g. Coles, 2001, Ch.4).

2.3.4. Oceanographic example: Newlyn sea surges

A mean residual life plot (not shown) suggests a threshold of $u = 0.3$ m to identify extreme sea surges at Newlyn. Implementing runs declustering, we use a cluster separation interval of 60 h—or $\kappa = 20$ observations—as suggested by Coles and Tawn (1991), to allow for wave propagation time. This gives maximum likelihood estimates of $\hat{\lambda}_u = 0.013$ (0.002), $\hat{\sigma} = 0.187$ (0.040) and $\hat{\xi} = -0.259$ (0.146), with estimated standard errors shown in parentheses. Substituting these estimates into Equation (5), and assuming $\theta = 1$, we can obtain estimates of z_r (see the last row of Table 1), along with associated 95% profile likelihood confidence intervals. For example, the 95% confidence interval for the 10-year return level z_{10} is (0.765, 1.569) m; z_{1000} has a range of (0.835, 6.452) m.

Although Coles and Tawn (1991) suggested $\kappa = 20$, κ is often chosen arbitrarily or simply by visual inspection of the data. It is worth noting at this point the sensitivity of return level estimation—in particular the upper bounds of 95% profile likelihood confidence intervals for z_r—to the choice of κ: for example, the upper bound for z_{1000} when using the suggested $\kappa = 20$ is 6.452 m, compared with 3.365 m when using $\kappa = 10$ and 33.143 m when using $\kappa = 30$. Other declustering schemes (e.g. 'blocks declustering' in Smith and Weissman, 1994) are also available, making return level estimation sensitive to both the choice of scheme and the choice of auxiliary parameter within that scheme. From a practitioner's point of view, this is worrying: designing a sea wall, for example, to a height specified by POT analysis could result in substantial under-protection or over-protection. Ideally, we would like to remove the need to decluster altogether by pressing into use *all* threshold excesses, which could also increase estimation precision. This will require careful consideration of the clustering behaviour of our process at extreme levels—in particular, estimation of the extremal index θ for use in Equation (5).

3. SIMULATION STUDY

In this section, we use simulated data to investigate the use of *all* excesses over a high threshold u instead of the standard POT approach. We will: briefly discuss the models from which we will simulate our data; investigate the relationship between return levels and the strength of serial correlation present in the extremes of our simulated chains; consider methods for estimating the extremal index θ from real data; give some design details for our simulation study; and present some results, which highlight the main conclusions of our investigation.

3.1. Simulated data

We simulate Markov chains with joint density given by

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2|x_1) \cdots f(x_n|x_{n-1}) = \prod_{i=1}^{n-1} f(x_i, x_{i+1}; \boldsymbol{\psi}) \Big/ \prod_{i=2}^{n-1} f(x_i; \boldsymbol{\phi}) \tag{6}$$

where $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ are specified parameter vectors for models for $f(x_i, x_{i+1})$ and $f(x_i)$, respectively. A limiting model for $f(x_i; \boldsymbol{\phi})$ for the region (u, ∞) is the density of the GPD function given in Equation (3). The first-order dependence structure given by (6) also requires a model for consecutive pairs in the process $f(x_i, x_{i+1}; \boldsymbol{\psi})$; we invoke bivariate extreme value theory (e.g. Coles, 2001, Ch. 8) to obtain contributions to the numerator in (6) on the region $(u, \infty) \times (u, \infty)$.

In particular, we consider the *logistic* and *negative logistic* models, with dependence parameters α and ρ (respectively), wherein the dependence structure between components x_i and x_{i+1} is symmetric; we also consider a model that allows for asymmetry in this dependence structure—the *bilogistic* model, with dependence parameters (α, β). For the logistic and negative logistic models, respectively, independence is obtained when $\alpha = 1$ and $\rho \searrow 0$, whereas complete dependence is obtained when $\alpha \searrow 0$ and $\rho \to \infty$. The bilogistic model reduces to the logistic model when $\alpha = \beta$; the value $\alpha - \beta$ determines the extent of asymmetry in the dependence structure. Independence is obtained when $\alpha = \beta \to 1$ or when one of α or β is fixed and the other approaches 1. Different limits occur when one of α or β is fixed and the other approaches 0. See Appendix A for more information on these dependence models.

3.2. Relationship between return levels and extremal dependence

When attempting to use all threshold excesses, it is clear from Equation (5) that return levels will depend on the degree of clustering present. Figure 2 shows how return levels can vary with θ when the marginal distribution is held fixed. Here, in line with one component of the simulation study (see Section 3.4), we use $(\lambda_u = 0.05, \sigma = 0.302, \xi = -0.4)$ in Equation (5). For the simulated data in the corresponding component, setting the threshold at the 95% marginal quantile gives $u = 1.746$; we also use $n_y = 2922$, the average number of yearly observations in the sea-surge data.

We can investigate the relationship between return levels and the parameters of the dependence models in (11)–(13) (see Appendix B) in the following way. Define (arbitrarily) x_n such that $F^n(x_n) = 1/2$ in Equation (1). Then, using Equation (1), we can define

$$\theta_n = -\frac{\log \Pr(\max\{X_1, \ldots, X_n\} \leq x_n)}{\log 2} \tag{7}$$

and so $\theta_n \to \theta$ as $n \to \infty$. We can use (7) to investigate the relationship between the extremal index θ and the dependence parameter(s) in the models we consider, via simulation. The extremal index is deterministically related to the dependence parameter(s) for any given model for dependence—the need for simulation arises because this relationship is analytically intractable. For example, if the logistic model (11) is assumed for the joint distribution of consecutive pairs, we can simulate N first-order Markov chains each of length n with dependence parameter $\alpha \in (0, 1)$ controlling the strength of serial correlation; then, the probability in the numerator of Equation (7) can be estimated as the proportion of simulated chains whose maximum value does not exceed x_n. The left-hand plot in Figure 3 shows the results of such

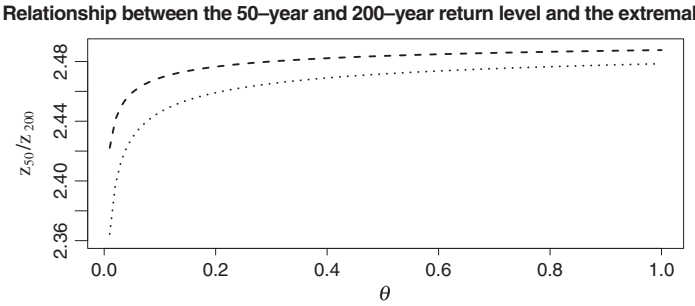

Figure 2. Plot of return level against extremal index θ: the dotted line corresponds to the 50-year return level and the dashed line corresponds to the 200-year return level

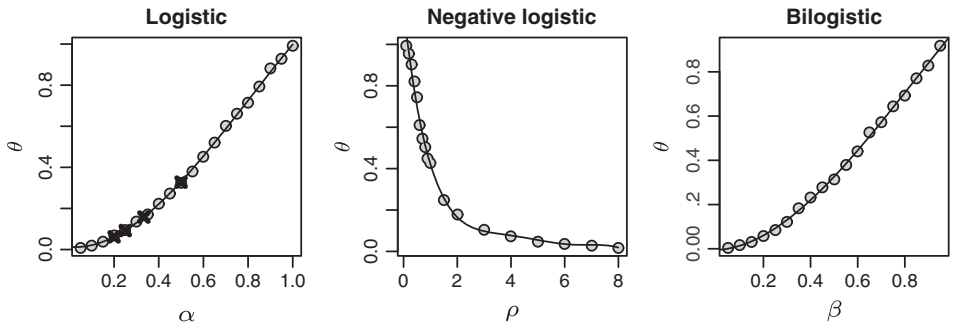

Figure 3. Simulated values of the extremal index θ for α (logistic), ρ (negative logistic) and β (bilogistic with $\alpha = 0.6$). The curved lines correspond to fitted polynomials: $\theta = 0.013 - 0.092\alpha + 1.833\alpha^2 - 0.756\alpha^3$; $\theta = 1.153 - 1.107\rho + 0.463\rho^2 - 0.096\rho^3 + 0.010\rho^4 - 0.0004\rho^5$; $\theta = -0.005 + 0.045\beta + 1.539\beta^2 - 0.607\beta^3$. The crosses in the first plot show limiting values of θ for some values of α in the logistic model, as derived in Smith (1992)

simulations for $\alpha = 0.05, 0.10, \ldots, 1$ in the logistic model. Here, we use $N = n = 10\,000$. The curved line shows the fit from a cubic regression of θ on α; substituting $\alpha = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}$ and $\frac{1}{5}$ into this cubic gives $\theta \approx 0.331, 0.158, 0.0925$ and 0.0616 (respectively), which match the limiting values of θ for these values of α, as reported in Smith (1992), almost perfectly. Smith uses a numerical procedure to find these limiting values; this procedure is computationally expensive, and the results differ negligibly to the fitted values obtained directly from our cubic. Similar polynomial relationships between the extremal index and the dependence parameters for the negative logistic and bilogistic models are obtained in the same way (see the second and third plots in Figure 3, respectively). We use polynomial relationships rather than smoothing splines, for example, because they are more than adequate and extremely simple.

3.3. Extremal index estimation

Section 3.2 illustrates the dependence of return levels on the strength of short-term serial correlation present in our series, and a link is made between the extremal index and the dependence parameters of the models considered. In real-life applications, we might not be able to rely on this link as the precise form of temporal structure might not be known. Indeed, basing our estimate of θ on a particular model would require us to first check the appropriateness of that model. Where models are nested (e.g. logistic within the bilogistic), likelihood ratio tests can be performed to choose between models; otherwise, more informal ad hoc procedures have to be used. Although this can be done (e.g. for the Bradfield wind speed data in Fawcett and Walshaw 2006), such checks can be rather subjective. We now turn our attention to different extremal index *estimators*, of which we consider five.

We consider a *polynomial estimator* of the extremal index, labelled $\hat{\theta}^{[1]}$, which assumes one of the extremal models given by Equations (11)–(13) for consecutive extremes in our series; the polynomial relationships shown in Figure 3 are then used to obtain $\hat{\theta}^{[1]}$ from $\hat{\alpha}$, $\hat{\rho}$ or $\hat{\beta}$. We consider two *cluster size methods* based on runs and blocks declustering (with cluster separation interval κ when using runs declustering and l blocks of length τ when using blocks declustering), and label these $\hat{\theta}^{[2]}$ and $\hat{\theta}^{[3]}$, respectively. We consider a *maxima method*, $\hat{\theta}^{[4]}$, which is based on estimating θ from separate fits of the GEV to sets of block maxima $\{M_\tau\}$ and $\{M_\tau'\}$ from our stationary series and an *independent* series with the same marginal distribution, respectively. We also use the *intervals estimator* of Ferro and Segers (2003), on the basis of a model for the inter-arrival times of threshold exceedances; we label this $\hat{\theta}^{[5]}$. Other methods have been proposed (e.g. Ancona-Navarrete and Tawn, 2000; Northrop, 2005; Süveges, 2007—the work in this paper simply serves to illustrate the use of all threshold excesses in return level estimation. For full details of the five estimators we use, see Appendix B.

3.4. Simulation study details

We simulate N stationary first-order Markov chains, each of length n, of extreme value type according to the three models given by Equations (11)–(13). Within each model, we simulate chains with varying degrees of serial correlation according to the parameter(s) for that model. Specifically, for the logistic model (11), we use $\alpha = 0.10, 0.11, \ldots, 1$ to cover almost the entire range of serial correlation, from very strong dependence to independence; for the negative logistic model (12), we use $\rho = 0.10, 0.15, \ldots, 1, 1.1, 1.2, \ldots, 6.9, 7$; for the bilogistic model (13), we fix α at 0.6 and use $\beta = 0.10, 0.11, \ldots, 0.99$. We use five different values of GPD shape parameter ξ to transform the marginals of the simulated chains to GPD (from standard Fréchet—see Appendix A): -0.4, -0.1, 0, 0.3 and 0.8, to reflect the various tails which might be observed in real life; the GPD scale parameter σ is held unit constant. We set u at the 95% marginal quantile (similar to thresholds suggested by mean residual life plots for the Newlyn and Bradfield data sets) giving $\lambda_u = 0.05$ and $\sigma^* = 1 + \xi u$. The true return levels for each arm of the simulation study are found via Equation (5), where the extremal index θ is obtained using one of the polynomial relationships shown in Figure 3.

For each simulated chain, the GPD is fitted, via maximum likelihood, to all threshold excesses, giving triples $\left(\hat{\lambda}_u^{(j)}, \hat{\sigma}^{*(j)}, \hat{\xi}^{(j)}\right)$ at each replication $j = 1, \ldots, N$. At each replication, we also estimate the extremal index, giving $\hat{\theta}^{[1]^{(j)}}, \ldots, \hat{\theta}^{[5]^{(j)}}$. Each set $\left(\hat{\lambda}_u^{(j)}, \hat{\sigma}^{*(j)}, \hat{\xi}^{(j)}, \hat{\theta}^{[1]^{(j)}}\right), \ldots, \left(\hat{\lambda}_u^{(j)}, \hat{\sigma}^{*(j)}, \hat{\xi}^{(j)}, \hat{\theta}^{[5]^{(j)}}\right)$ is then used to estimate the r-year return level via Equation (5); in line with the Newlyn sea-surge data, we use $n_y = 365.25 \times (24/3) = 2922$. For each arm of the study, we use $n = 10\,000$ and $N = 1000$. As an illustration of the sensitivity of return level estimation to the choice of declustering interval κ, at each replication, we also estimate return levels from the set of cluster peak excesses obtained using runs declustering with various values of κ.

3.5. Results

The plots shown in Figure 4 summarise results for when $\xi = -0.4$, although similar findings were observed for other values of ξ. Plots for the 50-year return level show dotted lines at $\alpha = 0.577$, $\rho = 1.022$ and $\beta = 0.544$, which correspond to the fitted values for these dependence parameters when the logistic, negative logistic and bilogistic[‡] models are applied to consecutive extremes in the Newlyn data set (we will return to the viability of these models for the sea-surge extremes in Section 4.1).

3.5.1. Polynomial estimators

As might be expected, the polynomial estimators of θ perform very well, giving (for all but the strongest levels of serial correlation) accurate estimates of return levels. The only source of error here is the estimation of the dependence parameters—once we have the estimate for α, ρ or β, we use the fitted polynomials to find the corresponding estimate of θ. A first-order Markov structure is assumed with extremal dependence given by (11), (12) or (13), which is exactly how the data have been constructed! Of course, we would not expect $\hat{\theta}^{[1]}$ to perform quite so well in practice: with real data, both the model and the model order might be inappropriate.

3.5.2. Cluster size methods

The performance of both cluster size methods $\hat{\theta}^{[2]}$ and $\hat{\theta}^{[3]}$ is variable, with substantial deviations from the fitted polynomials for some levels of dependence in all three models used. Here, the cluster separation interval was set at $\kappa = 20$ observations for $\hat{\theta}^{[2]}$; the block length for $\hat{\theta}^{[3]}$ was set at $\tau = 100$ so that $l = \tau$. In practice, $\hat{\theta}^{[2]}$ and $\hat{\theta}^{[3]}$ can be sensitive to the choice of κ and τ, respectively (see Fawcett and Walshaw, 2008); without good physical grounds leading to an 'optimal' choice for κ or τ, estimating θ in this way can be prone to severe bias. One of the advantages of using all threshold exceedances as opposed to working with cluster *peak* exceedances is that we avoid having to decluster— estimating θ using a cluster size method once again introduces the problem of declustering. Poor estimation of the extremal index here then leads to poor return level estimates.

3.5.3. Maxima method and intervals estimation

Both the maxima method ($\hat{\theta}^{[4]}$) and the intervals estimator ($\hat{\theta}^{[5]}$) perform very well, showing little deviation from the fitted polynomials in all three dependence models used; this, in turn, provides accurate return level estimates. This is promising. In particular, the intervals estimator does not rely on any choice of auxiliary parameter (as do the maxima and cluster size methods); neither does it rely on any assumptions regarding the form of dependence structure in the extremes (as do the polynomial estimators). This could make $\hat{\theta}^{[5]}$ the best candidate for use with real data.

3.5.4. POT estimation of return levels

Shown for comparison in the third column of plots are estimates of r-year return levels obtained from POT analyses with varying declustering intervals κ. This shows the sensitivity of estimates to the choice of κ when using cluster peaks, with $\kappa = 20$, 30, 50 and 60 often giving estimated return levels that fall substantially short of the true values for return periods greater than about $r = 10$ years.

[‡]Fitting the bilogistic model gives $(\hat{\alpha}, \hat{\beta}) = (0.608, 0.544)$; the fitted value for α is close to the fixed value used in the simulation study.

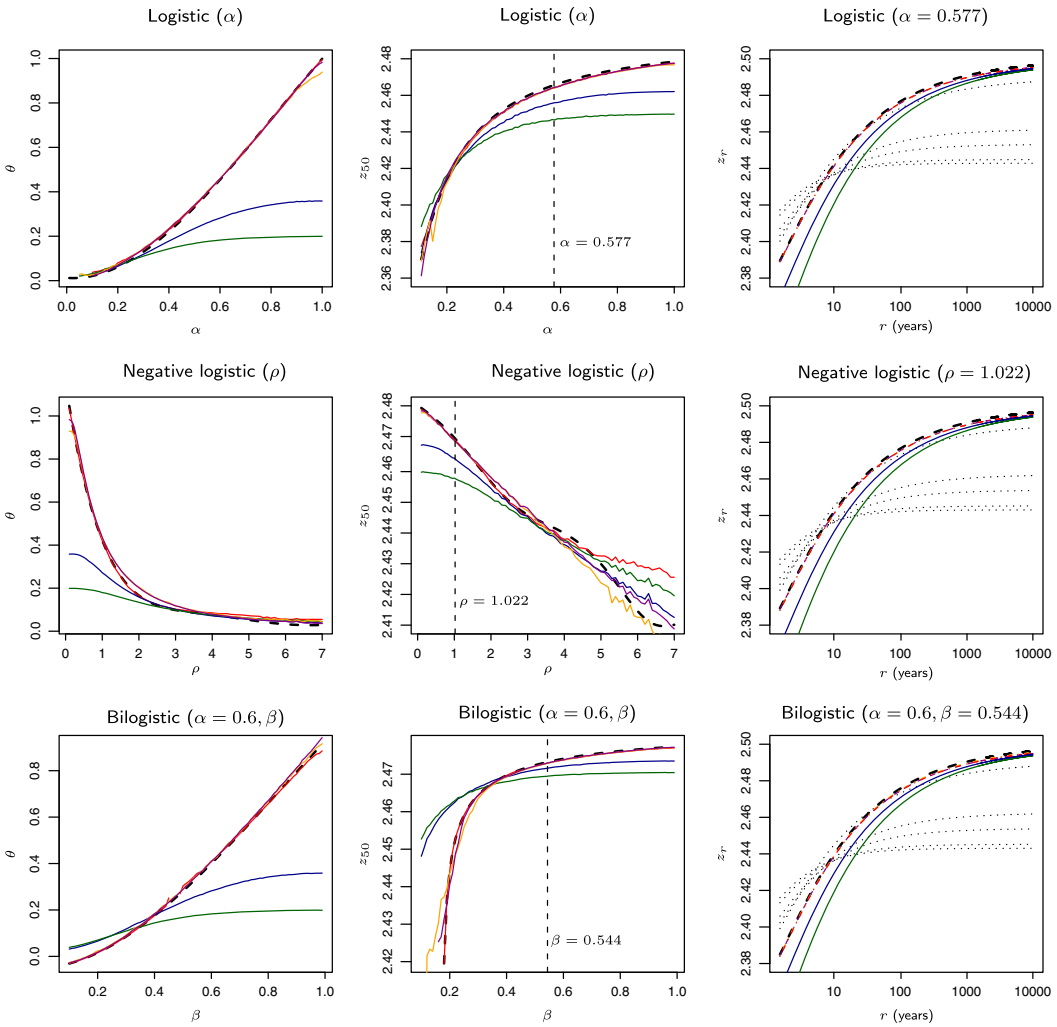

Figure 4. Plots of sampling distribution means for the extremal index θ (left) and the 50-year return level (middle), for different strengths of temporal dependence; r-year return levels (right) for fixed strengths of temporal dependence. Successive extremes have been simulated from the logistic (top), negative logistic (middle) and bilogistic (bottom) models. The curved dashed lines represent the true values. Red lines, polynomial estimator $\hat{\theta}^{[1]}$; green lines, cluster size estimator $\hat{\theta}^{[2]}$; blue lines, cluster size estimator $\hat{\theta}^{[3]}$; orange lines, maxima method $\hat{\theta}^{[4]}$; purple lines, intervals estimator $\hat{\theta}^{[5]}$. For comparison, plots down the right-hand side also show estimated return levels from POT analyses using various values of cluster separation interval κ (dotted lines; from top to bottom, $\kappa = 5$, 20, 30, 50 and 60)

3.5.5. Estimation precision

Although we can obtain standard deviations and 95% confidence intervals for extremal index estimators and their corresponding return levels empirically, from their sampling distributions, including these in Figure 4 would make the plots very difficult to read. Instead, we refer the reader to Fawcett (2005, Ch. 3). Here, full tables of results are given for components of this simulation study that use the logistic model for the generation of successive extremes, giving sampling distribution standard deviations, mean squared errors (MSEs) and relative efficiencies of our return level estimators—using both cluster peaks and all threshold excesses. Typically, for levels of serial correlation often observed in real-life environmental series (e.g. the sea-surge and wind speed extremes shown in Figure 1), the MSE for the 50-year return level from POT analyses is more than twice that from analyses using all threshold excesses; this does, however, vary depending on the value of κ used to identify clusters.

3.5.6. Asymptotic independence

To imply that the results discussed so far are relevant to real data would require that the models used are approximately correct at the chosen thresholds. In fact, with real data, it is often the case that we observe dependence above levels of practical interest, but *asymptotic independence*. To this end, we also simulate extremes from Gaussian AR(1) processes in place of the extreme value Markov chains for which we see results in Figure 4 (results not shown). Again, POT estimation of return levels appears sensitive to the choice of κ. In contrast, using all threshold exceedances gives more accurate estimates of return levels, and with increased precision.

4. DATA APPLICATIONS: SEA SURGES AND WIND SPEEDS

4.1. Newlyn sea surges

We now estimate return levels for the Newlyn sea-surge data using *all* excesses above a threshold of $u = 0.3$ m, accounting for the dependence on serial correlation by estimating the extremal index. Note that preliminary investigations using plots of the χ and $\bar{\chi}$ dependence measures (not shown here; see Coles *et al.* (1999) for more details) support the assumption of asymptotic dependence in the sea-surge data set and hence the use of models such as those outlined in Appendix A.

Table 1 shows estimates of the extremal index for the Newlyn sea-surge data using the five estimators $\hat{\theta}^{[1]}, \ldots, \hat{\theta}^{[5]}$. Both $\hat{\theta}^{[1]}$ and $\hat{\theta}^{[4]}$ are functions of other parameters, and so we have used the delta method (e.g. Coles, 2001, Ch. 2) to obtain estimated standard errors here. The standard error for $\hat{\theta}^{[5]}$ has been estimated using the bootstrap procedure outlined in Ferro and Segers (2003). Here, sets of *intercluster* and *intracluster* times $\{T_i\}$ (see Appendix B) are resampled, with replacement, B times to form a bootstrap replication of the process, within which the clustering behaviour has been preserved. This yields a collection of estimates $\{\hat{\theta}^{[5]}_{(1)}, \ldots, \hat{\theta}^{[5]}_{(B)}\}$ for which the sample standard deviation can be calculated. This procedure was repeated for increasing B, which showed convergence to about 0.05. A block bootstrap procedure (e.g. Efron and Tibshirani, 1993, Ch. 8) was used to estimate the standard errors for $\hat{\theta}^{[2]}$ and $\hat{\theta}^{[3]}$. The sensitivity of extremal index estimation to the choice of estimation procedure is clear, with estimates ranging from 0.106 to 0.425.

The simulation study revealed that both cluster size estimators ($\hat{\theta}^{[2]}$ and $\hat{\theta}^{[3]}$) perform poorly at some strengths of temporal dependence, which casts doubt on their reliability here. Fawcett and Walshaw (2008) also showed that both these methods are sensitive to the choice of κ or τ (for $\hat{\theta}^{[2]}$ and $\hat{\theta}^{[3]}$, respectively). Using a polynomial estimator ($\hat{\theta}^{[1]}$) relies on us being able to confirm a first-order Markov structure for the sea-surge extremes, as well as the suitability of the chosen model. Using a diagnostic plot to assess the validity of a first-order temporal dependence, relative to higher-order dependencies (see the 'simplex plots' in Coles and Tawn, 1991), reveals that a second-order dependence structure might be more plausible for the Newlyn extremes. For the logistic and bilogistic models, referring $2\{\ell(\hat{\alpha}, \hat{\beta}; \text{bilogistic}) - \ell(\hat{\alpha}; \text{logistic})\}$ to χ^2_1 tables (where ℓ is the log-likelihood) shows no significant model improvement by including the asymmetry parameter β. However, comparing other symmetric models with the logistic model (e.g. the negative logistic) is not so straightforward and would rely on various ad hoc checks such as those used in Fawcett and Walshaw (2006). Although the logistic model might be appropriate here, the assumption of a first-order temporal structure is questionable, possibly casting doubt on the reliability of any of our polynomial estimators.

The simulation study showed that, regardless of the structure of temporal dependence, both the maxima method ($\hat{\theta}^{[4]}$) and the intervals estimator ($\hat{\theta}^{[5]}$) were robust across all strengths of serial correlation. The similarity of both $\hat{\theta}^{[4]}$ and $\hat{\theta}^{[5]}$ (notwithstanding the large standard error for the former) might lend support to either of these being used to obtain estimates of return levels. Of the two, the intervals estimator is probably preferable—unlike the maxima method, which requires a block length τ to be chosen, $\hat{\theta}^{[5]}$ is completely automatic.

The five extremal index estimators $\hat{\theta}^{[1]}, \ldots, \hat{\theta}^{[5]}$, along with the GPD parameter estimates for all threshold excesses, were then used to estimate return levels via Equation (5). Table 1 shows some numerical results for $r = 10$, 50 and 1000, along with estimates from a POT analysis using $\kappa = 20$ (see Section 2.3.4). For each analysis using all threshold excesses, standard errors for return levels have been estimated using the delta method, incorporating uncertainty in the value of θ; we assume that the estimate of θ is uncorrelated with the estimated GPD marginal parameters.

Although we can see that, when attempting to make use of all extremes, return level estimates are sensitive to the choice of extremal index estimator, we have already discussed that we might prefer to use $\hat{\theta}^{[5]}$. The simulations show that using all excesses, having appropriately accounted for the dependence on serial correlation, can give more accurate return level estimates than those from a standard POT analysis; we also see an increase in precision here, with smaller standard errors attached to estimates for the periods considered owing to the inclusion of more data (e.g. 0.062 for all excesses using $\hat{\theta}^{[5]}$ cf. 0.106 for the POT analysis, for the 10-year return level).

Table 1. Maximum likelihood estimates for the extremal index and three return levels for the Newlyn sea surges (units for return levels are in metres)

	$\hat{\theta}$	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
All excesses				
Using $\hat{\theta}^{[1]}_{\text{log}}$	0.425 (0.045)	0.817 (0.073)	0.903 (0.107)	1.034 (0.179)
Using $\hat{\theta}^{[1]}_{\text{neglog}}$	0.413 (0.037)	0.816 (0.073)	0.902 (0.107)	1.033 (0.178)
Using $\hat{\theta}^{[1]}_{\text{bilog}}$	0.377 (0.020)	0.810 (0.071)	0.897 (0.105)	1.029 (0.176)
Using $\hat{\theta}^{[2]}$	0.182 (0.047)	0.767 (0.059)	0.860 (0.090)	1.000 (0.159)
Using $\hat{\theta}^{[3]}$	0.106 (0.032)	0.732 (0.052)	0.830 (0.079)	0.978 (0.146)
Using $\hat{\theta}^{[4]}$	0.282 (0.206)	0.793 (0.078)	0.883 (0.105)	1.024 (0.171)
Using $\hat{\theta}^{[5]}$	0.223 (0.050)	0.779 (0.062)	0.870 (0.094)	1.018 (0.163)
Cluster peak excesses	—	0.868 (0.106)	0.920 (0.144)	0.975 (0.202)

Estimated standard errors are given in parentheses.

Table 2. Maximum likelihood estimates for three return levels for the Bradfield wind speeds (units are in knots)

	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
All excesses			
Using $\hat{\theta}^{[1]}_{\log,m}$	88.463 (5.520)	96.071 (9.967)	107.644 (22.435)
Using $\hat{\theta}^{[5]}_{m}$	84.885 (6.151)	92.882 (8.873)	105.003 (19.745)
Cluster peak excesses	96.556 (13.527)	102.537 (22.776)	107.143 (43.052)

Estimated standard errors are given in parentheses.

4.2. Bradfield wind speeds

An extensive study of the Bradfield wind speeds in Fawcett and Walshaw (2006) suggests that a first-order Markov structure, with logistic dependence, is a suitable assumption for the temporal evolution of the process (again, plots of the χ and $\bar{\chi}$ dependence measures support the assumption of asymptotic dependence). Model-based estimates of various quantities, including the *mean storm length* and the *mean inter-storm duration*, compared quite well with their empirical counterparts. Thus, we might trust our polynomial estimator of the extremal index here ($\hat{\theta}^{[1]}_{\log}$) to obtain estimates of return levels. As discussed in Section 2.3.2, a piecewise seasonality approach can be adopted to circumvent the problem of seasonal variability. The monthly varying GPD parameters can then be recombined to estimate overall return levels by solving

$$\prod_{m=1}^{12} \mathcal{H}_m(x)^{n_m \theta_m} = 1 - r^{-1}, \qquad m = 1, \ldots, 12, \tag{8}$$

for $x = z_r$, where $\mathcal{H}_m$, n_m and θ_m are the GPD distribution function, number of observations and extremal index in month m (respectively).

Table 2 shows return level estimates obtained using all threshold exceedances—this time using only $\hat{\theta}^{[1]}_{\log,m}$ and $\hat{\theta}^{[5]}_{m}$. Shown for comparison are results from a standard POT analysis, where *reclustered excess plots* (Walshaw, 1994) have been used to simultaneously choose monthly varying thresholds and a cluster separation interval, u_m and κ, respectively. Although results are not shown here, return level estimates from a POT analysis are sensitive to the choice of cluster separation interval κ, with estimates of the 50-year return level varying between 95.303 and 106.100 when κ varies between 5 and 40. As with the sea-surge example, we see that pressing all extremes into use increases the precision of our estimates considerably.

4.3. Confidence intervals for return levels

As discussed in Section 2.3.3, it is preferable to provide confidence intervals for return levels that are not constrained to be symmetric. Thus, although the standard errors for return levels shown in Tables 1 and 2 are useful for highlighting the gain in precision when using all threshold excesses, we would probably rather *not* use these standard errors to construct confidence intervals in the usual way. However, as we are now acknowledging the presence of serial correlation by using all threshold excesses, we cannot use the standard profile likelihood approach to obtain confidence intervals for z_r.

Instead, we extend the bootstrap procedure described by Ferro and Segers (2003) and implemented for the extremal index estimator $\hat{\theta}^{[5]}$ in Sections 4.1 and 4.2. We used this procedure to form B bootstrap replications of the sea-surge and wind-speed extremes, yielding $\{\hat{\theta}^{[5]}_{(1)}, \ldots, \hat{\theta}^{[5]}_{(B)}\}$. For each of these replications, we now also estimate the marginal GPD parameters; substitution of $(\lambda_u, \sigma, \xi, \theta)$ in Equation (5) with $(\hat{\lambda}_{u_{(b)}}, \hat{\sigma}_{(b)}, \hat{\xi}_{(b)}, \hat{\theta}^{[5]}_{(b)})$, $b = 1, \ldots, B$, then gives a collection of estimates $\{\hat{z}_{r_{(1)}}, \ldots, \hat{z}_{r_{(B)}}\}$ from which we can construct confidence intervals.

We *could* record the 2.5 and 97.5 percentiles from the bootstrap sample for $\hat{z}_r$; however, such *bootstrap percentile intervals* are known to perform poorly in cases where the bootstrap distribution is asymmetric. Rather, we use the *bias-corrected, accelerated* ($\mathrm{BC_a}$) *intervals*, as proposed in Efron (1987). This method corrects for bias owing to non-normality; it also accelerates convergence to a solution by correcting for the rate of change of the normalised standard error of $\hat{z}_r$ relative to the true value of z_r in constructing the confidence bounds of the percentile method. In practice, we estimate the bias-correction $\mathcal{Z}$ as

$$\hat{\mathcal{Z}} = \Phi^{-1}\left[\frac{1}{B}\sum_{b=1}^{B} I_{(b)}\right]$$

where $I_{(b)}$ takes the value 1 if $\hat{z}_{r_{(b)}} < \hat{z}_r$ and 0 otherwise, and $\Phi(\cdot)$ is the standard Normal distribution function. We then estimate the 'acceleration constant' a using

$$\hat{a} = \frac{\sum_{\forall d}(\tilde{z}_{r_{(\cdot)}} - \tilde{z}_{r_{(d)}})^3}{6\{\sum_{\forall d}(\tilde{z}_{r_{(\cdot)}} - \tilde{z}_{r_{(d)}})^2\}^{3/2}}$$

Table 3. Bootstrapped 95% (BC_a) confidence intervals for three return levels for the Newlyn sea surges (units are in metres) and the Bradfield wind speeds (units are in knots)

	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{1000}$
Newlyn sea surges	(0.657, 0.872)	(0.708, 1.019)	(0.772, 1.306)
Bradfield wind speeds	(80.847, 87.749)	(86.088, 98.540)	(90.623, 116.103)

where $\tilde{z}_{r_{(d)}}$ is the jackknife value of $\hat{z}_r$ at deletion d (e.g. Efron and Tibshirani, 1993), and $\tilde{z}_{r_{(\cdot)}}$ is the sample mean of all the jackknife values. Specifically, sets of intercluster and intracluster times $\{T_i\}$ are deleted sequentially giving, at each deletion d, $(\tilde{\lambda}_{u_{(d)}}, \tilde{\sigma}_{(d)}, \tilde{\xi}_{(d)}, \tilde{\theta}^{[5]}_{(d)})$; we then apply Equation (5) to obtain $\tilde{z}_{r_{(d)}}$. The position of the lower and upper bounds of the BC_a intervals in the ordered set of bootstrap estimates of z_r are then $[[Ba_1]]$ and $[[Ba_2]]$, respectively (where [[ ]] indicates rounding to the nearest integer), where

$$a_1 = \Phi\left[\hat{\mathcal{Z}} + \frac{\hat{\mathcal{Z}} - \mathcal{Z}_{1-A/2}}{1 - \hat{a}(\hat{\mathcal{Z}} - \mathcal{Z}_{1-A/2})}\right] \quad \text{and}$$

$$a_2 = \Phi\left[\hat{\mathcal{Z}} + \frac{\hat{\mathcal{Z}} + \mathcal{Z}_{1-A/2}}{1 - \hat{a}(\hat{\mathcal{Z}} + \mathcal{Z}_{1-A/2})}\right]$$

and $\mathcal{Z}_{1-A/2}$ is the $1 - A/2$ quantile of the standard normal distribution (e.g. 1.96 for a 95% confidence interval, where $A = 0.05$). Using $B = 5000$, we obtain the 95% BC_a confidence intervals shown in Table 3 for both the Newlyn sea surges and Bradfield wind speeds. Notice that the resulting intervals for the sea-surge data are considerably narrower here than in the POT analysis of Section 2.3.4.

In a simulation study, we reproduce the BC_a confidence intervals for simulated data where the true values of z_r are known, in line with the simulation study detailed in Section 3.4. We repeat the bootstrapping procedure 10 000 times to estimate one-sided coverage probabilities—for example, $\Pr(z_r < \hat{z}_{r_{up}})$, where $\hat{z}_{r_{up}}$ is the upper bound of the bootstrapped 95% confidence interval for z_r; the *intended* coverage probability here is $1 - A/2 = 0.975$. For comparison, at each replication in the study, we also obtain the simpler percentile intervals. Typically, we find that the estimated coverage probabilities are closer to the intended values when using the BC_a intervals: for example, for one arm of the study, $\widehat{\Pr}(z_r < \hat{z}_{r_{up}}) = 0.976$, 0.977 and 0.979 for $r = 10$, 50 and 1000 (respectively) for the BC_a intervals, compared with 0.978, 0.988 and 1.000 (respectively) for the simpler percentile intervals.

5. CONCLUDING REMARKS

In this paper, we demonstrate how information on *all* extremes of a data set can be pressed into use, regardless of the strength of extremal serial dependence present, in order to estimate return levels.

As we discuss, return level inference via the standard POT approach circumvents the problem of dependence; however, using only a filtered set of threshold excesses reduces estimation precision, which can result in very wide, sometimes impractical, confidence intervals. As we demonstrate, the upper end-point of the 95% (profile log-likelihood) confidence interval for a return level can increase dramatically with relatively small changes in the POT declustering parameter (see the example in Section 2.3.4); unsuitable choices of this parameter can also lead to return level estimation bias as shown in the plots down the right-hand side of Figure 4.

Once a suitable threshold has been selected for identifying extremes, we recommend using Equation (5) to estimate return levels: all threshold excesses should be used to estimate the marginal GPD parameters (λ_u, σ, ξ), and a suitable estimator for the extremal index θ should be sought. But what constitutes a 'suitable estimator' for θ? As we demonstrate in our simulation study and in the sea-surge analysis in Section 4.1, return level estimation can be sensitive to the choice of estimator for θ. If an assumption of first-order extremal dependence can be verified, with a specific form of dependence structure as provided by a known bivariate extreme value model, then the relationship between the parameter(s) in this model and the extremal index can be exploited. Alternatively, a more flexible, non-parametric approach can be used in the form of Ferro and Segers' intervals estimator for θ (see Appendix B). Tables 1 and 2 indicate the gain in precision of such approaches over the standard POT approach for return level inference.

We report estimated standard errors for return levels in Tables 1 and 2 to demonstrate the gain in precision when using our approach, although we do *not* recommend using these standard errors to form confidence intervals in the usual way (e.g. estimator $\pm$ 1.96 $\times$ standard error). We advocate the use of Efron's bias-corrected, accelerated bootstrap confidence intervals (using a bootstrap procedure that preserves the clustering of extremes), with a practical calculation procedure being outlined in Section 4.3. We discuss that for return levels, these intervals give coverage that is appreciably closer to the intended value than the simpler percentile intervals.

Acknowledgements

We would like to thank Professors Anthony Davison and Rob Henderson for helpful discussions about this work. We are also indebted to two referees for making some excellent suggestions.

REFERENCES

Ancona-Navarrete MA, Tawn JA. 2000. A comparison of methods for estimating the extremal index. *Extremes* **3**: 5–38.
Anderson CW. 1990. Discussion of the paper by Davison and Smith. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **52**: 393–442.
Coles SG. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer: London.
Coles SG, Heffernan J, Tawn JA. 1999. Dependence measures for extreme value analyses. *Extremes* **2**: 339–365.
Coles SG, Tawn JA. 1991. Modelling extreme multivariate events. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **53**: 377–392.
Davison AC, Smith RL. 1991. Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **52**: 393–442.
Eastoe EF, Tawn JA. 2012. The distribution for the cluster maxima of exceedances of sub-asymptotic thresholds. *Biometrika* **99**(1): 43–55. submitted.
Efron B, Tibshirani RJ. 1993. *An Introduction to the Bootstrap*. Chapman & Hall: London.
Efron B. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**(397): 171–185.
Fawcett L. 2005. Statistical Methodology for the Estimation of Environmental Extremes. *PhD Thesis*, Newcastle Univ.
Fawcett L, Walshaw D. 2008. Bayesian inference for clustered extremes. *Extremes* **11**: 217–233.
Fawcett L, Walshaw D. 2007. Improved estimation for temporally clustered extremes. *Environmetrics* **18**(2): 173–188.
Fawcett L, Walshaw D. 2006. Markov chain models for extreme wind speeds. *Environmetrics* **17**(8): 795–809.
Ferro CAT, Segers J. 2003. Inference for clusters of extreme values. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **65**: 545–556.
Gomes MI. 1993. On the estimation of parameters of rare events in environmental time series. In *Statistics for the environment 2: Water Related Issues*. (Barnett and Turkman); 225–241.
Hsing T, Hüsler J, Leadbetter MR. 1988. On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields* **78**: 97–112.
Jenkinson AF. 1955. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society* **81**: 158–171.
Leadbetter MR, Rootzén H. 1988. Extremal theory for stochastic processes. *Ann. Probab.* **16**: 431–476.
Northrop PJ, Jonathan P. 2011. Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics* **22**(7): 799–809.
Northrop PJ. 2005. Semiparametric Estimation of the Extremal Index Using Block Maxima. *Preprint*, University College London.
Pickands J. 1981. Multivariate extreme value distributions. *Bulletin International Statistical Institute* **XLXI**(Book 2): 859–878.
Pickands J. 1975. Statistical inference using extreme order statistics. *Annals of Statistics* **3**: 119–131.
Smith RL. 1992. The extremal index for a Markov chain. *Journal of Applied Probability* **29**: 37–45.
Smith RL, Tawn JA, Coles SG. 1997. Markov chain models for threshold exceedances. *Biometrika* **84**: 249–268.
Smith RL, Weissman I. 1994. Estimating the extremal index. *Full journal title for Smith RL, Weissman I. 1994: Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **56**: 515–528.
Süveges M. 2007. Likelihood estimation of the extremal index. *Extremes* **10**: 41–55.
Walshaw D. 1994. Getting the most from your extreme wind data: a step by step guide. *Journal of Research of the National Institute of Standards and Technology* **99**: 399–411.
Walshaw D. 1991. Statistical Analysis of Extreme Wind Speeds. *PhD Thesis*, Univ. of Sheffield.

APPENDIX A. BIVARIATE THRESHOLD EXCESS MODELS

Suppose $(x_1, y_1), \ldots, (x_n, y_n)$ are independent realisations of a random variable (X, Y) with joint distribution function F. For suitably large thresholds u_x and u_y, the marginal distributions of F each have an approximation of the form given by Equation (4), with respective parameter sets $(\lambda_{u_x}, \sigma_x, \xi_x)$ and $(\lambda_{u_y}, \sigma_y, \xi_y)$. After transforming the margins for (X, Y) to standard Fréchet (see Coles, 2001), it can be shown (Pickands, 1981) that the joint distribution function $G(x, y)$ for a bivariate extreme value distribution has the representation

$$G(x, y) = \exp\{-V(x, y)\} \tag{9}$$

for $x > 0$, $y > 0$, where

$$V(x, y) = 2\int_0^1 \max(q/x, (1-q)/y)\mathrm{d}H(q) \tag{10}$$

and H is a distribution function on [0, 1], which satisfies the mean constraint

$$\int_0^1 q\mathrm{d}H(q) = \frac{1}{2}.$$

There is no characterisation of the complete family of distributions specified by (9), and so model choice involves specifying an appropriate sub-family through the choice of H in (10).

There are various possibilities for the dependence function H; two commonly used *symmetric* models are the logistic and negative logistic models, where

$$V(x, y) = \left(x^{-1/\alpha} + y^{-1/\alpha}\right)^{\alpha}, \qquad 0 < \alpha \leq 1, \qquad \text{and} \tag{11}$$

$$V(x, y) = -x - y + \left(x^{-\rho} + y^{-\rho}\right)^{-1/\rho}, \qquad \rho > 0, \tag{12}$$

respectively. The *asymmetric* bilogistic model has

$$V(x, y) = -x\gamma^{1-\alpha} - y(1-\gamma)^{1-\beta}, \qquad 0 < \alpha, \beta < 1, \tag{13}$$

where $\gamma = \gamma(x, y; \alpha, \beta)$ is the solution of $(1-\alpha)x(1-\gamma)^{\beta} = (1-\beta)y\gamma^{\alpha}$.

APPENDIX B. EXTREMAL INDEX ESTIMATORS

1. One method for estimating the extremal index is to fit an extreme value Markov chain model (e.g. the logistic, negative logistic or bilogistic) to successive pairs of extremes in our series; then, we can use the polynomial relationships shown in Figure 3 to obtain our estimate of θ. We call the extremal index estimator obtained in this way a *polynomial estimator*, and we label it $\hat{\theta}^{[1]}$ (with subscripts 'log', 'neglog' or 'bilog' for each of the models considered in this paper). As discussed, using this estimator with real data would require us to check the suitability of the model used.
2. We give a formal definition of the extremal index in Section 2.2; an alternative characterisation, provided by Hsing *et al.* (1988), is that θ^{-1} is the limiting mean cluster size in the point process of exceedance times over a high threshold. This suggests that a suitable way to estimate the extremal index can be found through methods which identify clusters of extremes, the estimate itself being found as the reciprocal of the mean cluster size. We call such an estimator a *cluster size method*; when runs declustering is used to identify clusters of extremes, we label this estimator $\hat{\theta}^{[2]}$.
3. Another cluster size method uses 'blocks declustering' to identify clusters of extremes and again estimates θ as the reciprocal of the mean cluster size. This method of cluster identification partitions the data into approximately l blocks of length τ, and the threshold exceedances within each block are treated as a single cluster of extremes. We label this estimator $\hat{\theta}^{[3]}$.
4. Gomes (1993) proposed a method for estimating θ on the basis of separate model fits to the two sets of block maxima $\{M_\tau\}$ and $\{M'_\tau\}$. The former are block maxima from a stationary series and the latter are block maxima from an *independent* series with the same marginal distribution as the stationary series, having been obtained after randomising the index of the original observations. The GEV (2) is used to model both sets of block maxima. If the extremal index for the block maxima from the stationary series $\{M_\tau\}$ is equal to θ, then, using Equations (1) and (2), the $\{M_\tau\}$ have distribution function $\mathcal{G}^\theta(y;\mu,\varsigma,\xi)$, which is easily seen to be a GEV distribution $\mathcal{G}(y;\mu_\theta,\varsigma_\theta,\xi_\theta)$, where

$$\begin{aligned}\mu_\theta &= \mu - \varsigma(1-\theta^\xi)/\xi,\\ \varsigma_\theta &= \varsigma\theta^\xi \qquad \text{and}\\ \xi_\theta &= \xi.\end{aligned}$$

   Gomes (1993) suggested estimating θ from estimates $(\hat{\mu},\hat{\varsigma},\hat{\xi})$ and $(\hat{\mu}_\theta,\hat{\varsigma}_\theta,\hat{\xi}_\theta)$ obtained from the separate fits to the two sets of block maxima. A pooled estimate of ξ is calculated as

$$\tilde{\xi} = \frac{\hat{\varsigma}-\hat{\varsigma}_\theta}{\hat{\mu}-\hat{\mu}_\theta};$$

   then an estimate of θ, which we label $\hat{\theta}^{[4]}$, is given by

$$\hat{\theta}^{[4]} = \left(\frac{\hat{\varsigma}}{\hat{\varsigma}_\theta}\right)^{-1/\tilde{\xi}}.$$

   We call this approach the *maxima method*.
5. We also consider an *intervals estimator* for θ on the basis of the inter-arrival times of threshold exceedances, proposed by Ferro and Segers (2003). Suppose K exceedance times are observed: $S_1 < S_2 < \cdots < S_K$, and $T_i = S_{i+1} - S_i$ for $i = 1,\ldots,K-1$ are the inter-arrival times. Ferro and Segers (2003) derived the distribution of the T_i and showed that a moments-based estimator of θ is

$$\hat{\theta}^{[5]} = \min\left(1, \frac{2\left\{\sum_{i=1}^{K-1}(T_i-a)\right\}^2}{(K-1)\sum_{i=1}^{K-1}(T_i-b)(T_i-c)}\right)$$

   where $a=b=c=0$ if the largest inter-arrival time is no greater than 2, and $a=b=1$ and $c=2$ if the largest inter-arrival time is greater than 2.

