ENVIRONMETRICS
Environmetrics (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.810

Improved estimation for temporally clustered extremes

Lee Fawcett and David Walshaw*,†

University of Newcastle upon Tyne, Newcastle upon Tyne, U.K.

SUMMARY

In this paper, we investigate the effects of *declustering* applied to sequences of extreme observations. Through a simulation study, we demonstrate that the common practice of analysing *peaks over thresholds* (POT) is liable to incur serious bias in the estimation of parameters, as well as the *return levels* used as design specifications when building to withstand extremes of wind or rain, or river or sea level. We demonstrate that a much simpler approach, the direct analysis of *all* exceedances of a high threshold, can reduce this bias to negligible levels. This approach has, until now, been unpopular, because the data being analysed are not independent. The effect of this is to cause the standard errors associated with parameter estimates to underestimate the uncertainty attached to these estimates. We employ existing but little-used methodology to inflate these standard errors, and we demonstrate that the adjusted values are very good representations of the true uncertainty associated with maximum likelihood estimates. The overall approach has thus achieved the effect of eliminating the bias in estimation, while accounting for any undesirable effects caused by dependent data.

We apply our approach to a sequence of sea-surge data from southwest England, and illustrate the discrepancies between this and a POT approach, which are consistent with the POT approach underestimating long-period return levels. We also pay considerable attention to checking the robustness of our results, demonstrating that the problems of bias caused by the POT approach apply systematically over all of the declustering schemes we consider, as well as over the entire range of tail behaviours. When the primary interest is in return-level estimation, we recommend that our procedure will generally prove to be much more effective and reliable than the POT approach. Should there be a deeper interest in the serial dependence itself, then we recommend that this dependence is explicitly *modelled*, and we refer the reader to an earlier paper by the authors, published in this journal.

KEY WORDS: extreme value theory; temporal dependence; clusters; return levels; sea surges; peaks over thresholds

1. INTRODUCTION

In this paper, we study one of the most commonly employed procedures for the analysis of serially correlated environmental extremes, the peaks over threshold (POT) approach (see for example, Davison and Smith, 1990). Here, an appropriate limiting distribution is fitted to the largest exceedance selected

*Correspondence to: D. Walshaw, School of Mathematics and Statistics, Merz Court, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, U.K.
†E-mail: david.walshaw@ncl.ac.uk

Contract/grant sponsor: EPSRC Doctoral Training Award.

Received 24 March 2006
Accepted 8 July 2006

from each 'cluster' of values above some predetermined high threshold. We demonstrate, via a simulation study, that this approach leads to systematic bias in estimates for parameters and long-period return levels. We recommend a simpler analysis based on using *all* exceedances of the threshold. This appears to virtually eliminate bias, at the expense of causing standard errors obtained from a maximum likelihood estimation to be too small, as a result of the inherent, now incorrect, assumption that the exceedances are independent. We demonstrate the effectiveness of a procedure designed to adjust these standard errors to more realistic values, and suggest that this, used in combination with the analysis of all exceedances, leads to a far more effective approach than POT.

2. BACKGROUND

Threshold models for exceedances have been widely adopted in recent years in the study of extremes of environmental processes. The main advantage of such models over the so-called 'classical' extreme value models (in which a limiting distribution is fitted to the largest order statistics selected from fixed time intervals—the so-called 'block–maxima' approach) is their greater flexibility in the manner in which events are classified as 'extreme'. This generally leads to a larger number of extreme events being available for analysis, and this in turn to more precise estimates for particular parameters of interest.

Threshold models undoubtedly provide greater efficiency for data exploitation over the standard block–maxima approach, but at a cost: short-range serial correlation, almost always present in environmental time series, can no longer be ignored in the manner of a traditional block maxima analysis. The most popular approach to circumvent the problems caused by such temporal dependence is to employ some filtering scheme which identifies threshold exceedances which are far enough apart to be deemed independent; see, for example, Davison and Smith (1990). This pragmatic approach allows us to apply standard results with respect to, say, maximum likelihood estimation, so that standard errors and confidence intervals for parameter estimates are justified by the usual asymptotic arguments. However, 'declustering' the series in this way introduces some fairly arbitrary complexity to the analysis, the effects of which are difficult to predict, and is once again wasteful of data, since all but the filtered set of extremes are discarded.

Since information about extremes is almost always scarce, it would be highly desirable to press more than just a filtered set of independent extremes into use. In a recent paper, Fawcett and Walshaw (2006) demonstrated how all threshold exceedances can be used by explicitly modelling temporal dependence, and properly accounting for this through the likelihood function. By doing so, not only were they able to increase the precision of their analysis, but they were also able to investigate certain characteristics of extreme clusters which would have been impossible under the standard declustering approach. The increased complexity of their approach, however, which includes model selection issues for the temporal evolution of the process being studied, can only be justified if we are specifically interested in the clustering behaviour of the extremes.

With this in mind, we consider a much simpler analysis, which uses the entire set of threshold exceedances, and initially ignores dependence. The maximum likelihood approach, based on the assumption that these exceedances form an independent set, is liable to underestimate standard errors for model parameters. However, appropriate adjustments to these can be achieved by employing methods due to Smith (1991). Using simulated data, we compare the maximum likelihood estimators for parameters (and in particular *return levels*) under this approach, with the behaviour observed when the analysis is based on the usual filtered set of declustered exceedances. We find that by using *all* exceedances, and making appropriate adjustments to the standard errors, we achieve precision levels comparable to

DOI: 10.1002/env

those obtained when we model temporal dependence (see Fawcett and Walshaw, 2006). For the method which uses a filtered set of extremes, we also investigate the sensitivity of inference to the choice of filtering scheme used.

In Section 3, we analyse some real data on sea surges, and observe the differences in parameter estimation, and in particular *return level estimation*, when using (i) all threshold excesses and (ii) a filtered set of independent threshold excesses. In Section 4, we use simulated data to study the properties of estimators for model parameters and return levels when using all exceedances, in comparison with using a filtered set of declustered values. Data are simulated with a variety of attributes to cover the many situations which we might face with real life data. In particular, the strength of serial correlation between successive observations is controlled to vary from very strong dependence, right through to complete independence.

We find that when declustered extremes are used, the analysis performs poorly, and there is a substantial bias in the estimators for model parameters, which translates into systematic bias in return level estimation. The nature of this bias is such that, for strong serial correlation in extremes, return levels are systematically underestimated, whereas for weak serial correlation and complete independence, there is a small but systematic overestimation of return levels. By comparison, when using all extremes, the performance is vastly improved, and the bias in the estimators for model parameters and return levels is negligible for most levels of temporal dependence.

In conclusion, we suggest that by basing our analysis on all threshold exceedances, and making appropriate adjustments to allow for the dependence, we simultaneously achieve the aims of simplifying the analysis, and improving the performance of our estimators.

3. EXAMPLE: OCEANOGRAPHIC DATA

The aim of this section is to demonstrate differences in parameter estimation, and return level estimation, when an extreme value model is fitted to: (i) a temporally dependent series of oceanographic measurements and (ii) a filtered (and hence approximately independent) subset of these measurements.

3.1. The data

Figure 1 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. The practical motivation for the study of such data is that structural failure—probably a sea-wall in this case—is likely under the condition of extreme surges. Also shown in Figure 1 is a plot of the time series against the lag 1 time series. Table 1 gives summaries of this dataset; for a more detailed description of the oceanographic climate at this site and the UK in general see, for example, Coles (1991).

3.2. The model for threshold exceedances

A natural way of modelling extremes of time series such as the hourly maximum wind speeds is to use the Generalised Pareto Distribution (GPD) as a model for excesses over a high threshold. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables with common distribution function F. Then for a high value of the threshold, u, the conditional distribution $(X - u|X > u)$ will

DOI: 10.1002/env

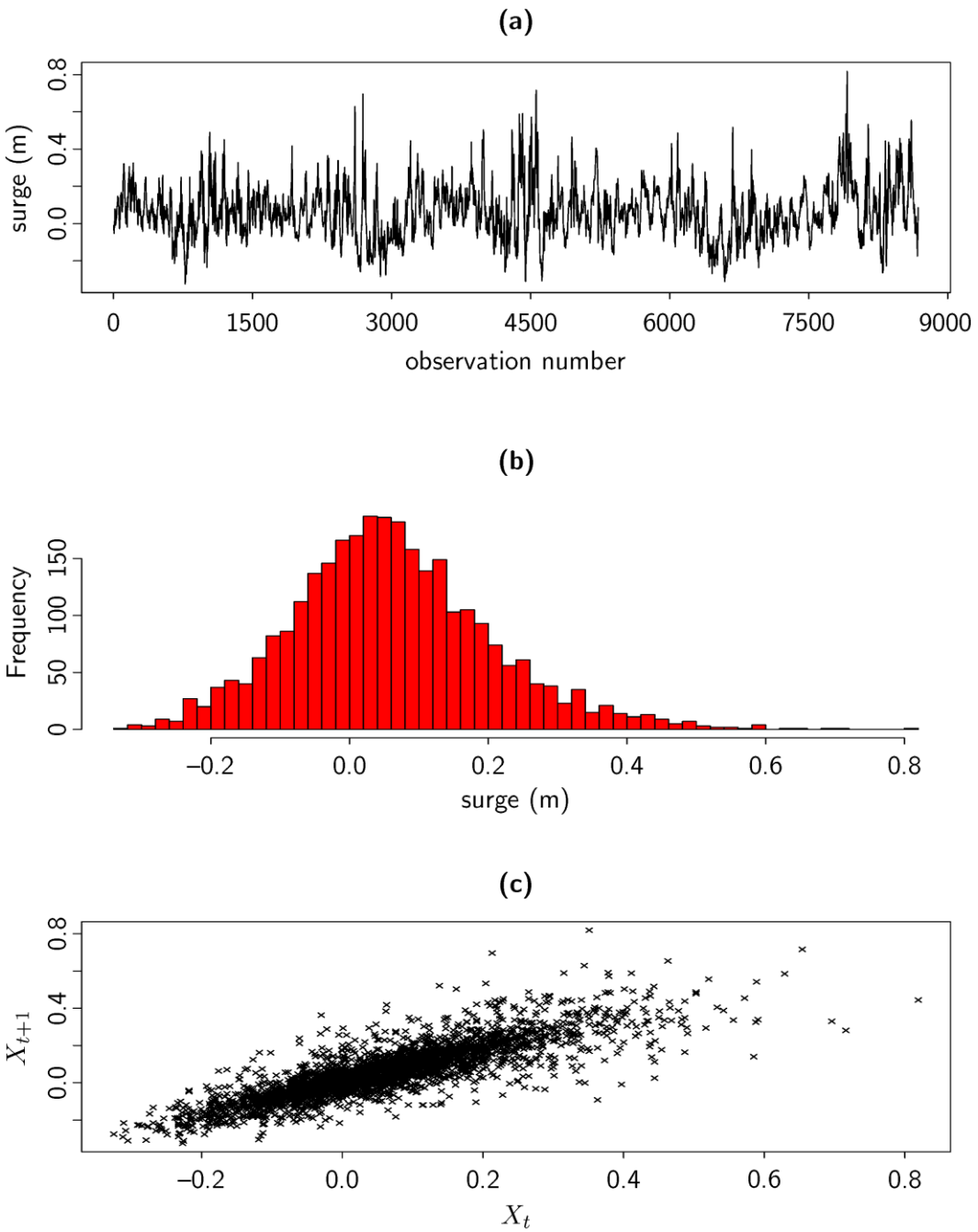

Figure 1. Newlyn sea-surge data: (a) time series plot; (b) histogram; (c) plot of the time series against the series at lag 1

DOI: 10.1002/env

Table 1. Summary statistics of Newlyn sea-surge data (in metres)

Mean	St. dev.	Median	LQ	UQ	Min	Max
0.062	0.144	0.052	−0.032	0.144	−0.325	0.819

be approximately of GPD form, with cumulative distribution function (c.d.f.):

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

where $a_+ = \max(0, a)$ and σ and ξ are scale and shape parameters, respectively. This arises because, if a limiting distribution exists for any normalisation of $(X - u|X > u)$, then this limit will be exactly of GPD form (Davison and Smith, 1990). The limit is taken as $u \nearrow u_{\mathrm{F}}$, where u_{F} is the finite upper endpoint of F if it exists, or ∞ otherwise.

3.3. Temporal dependence

Figure 1(c) shows the presence of substantial serial correlation in the sequence of 3-hourly sea surges. Indeed, the partial autocorrelation function for the data (not shown) shows a very large value at lag 1, 0.836, which indicates the significance of this (short-term) temporal dependence.

As already discussed, the most commonly adopted approach to circumvent the problems caused by such temporal dependence is to employ a declustering scheme to filter out a set of approximately independent threshold excesses. One method, which is often considered to be the most 'natural' way of identifying clusters of extremes, is 'runs-declustering'. Here, an auxiliary 'declustering parameter', which we denote by κ, is chosen prior to declustering, and a cluster of threshold excesses is deemed to have terminated as soon as at least κ consecutive observations fall below the threshold. The maximum (or 'peak') observation from each cluster is then extracted, and the GPD in Equation (1) fitted to the set of cluster peak excesses. This approach is often referred to as the 'peaks over threshold' approach (POT, Davison and Smith, 1990) and is widely accepted as the main pragmatic approach for dealing with clustered extremes. In this section, we compare this approach to that which fits the GPD to *all* threshold excesses.

3.4. Implementation

3.4.1. Selecting a threshold and declustering parameter. Here we choose our threshold u so that 5 per cent of the observations exceed it, giving a value $u = 0.3$ m. The rationale underlying this choice is that we need to have a value which is large enough so that the limiting GPD (1) is a good approximation for the exceedance distribution, whilst not so large as to reduce unnecessarily the number of exceedances available for the analysis. Here we use a standard exploratory technique, the *mean residual life plot* (see Coles, 2001, for example), to aid the choice. Here the plot (not shown) indicates a range of acceptable values for u, which includes 0.3 m. The precise value is chosen so as to be consistent with the simulation study of Section 4, where a threshold with a 5 per cent exceedance rate is used throughout. For the identification of clusters of extreme sea surges, we use a declustering parameter of $\kappa = 60$ h (i.e. 20 observations) to allow for wave propagation time, thereby following the example of Coles and Tawn (1991).

DOI: 10.1002/env

3.4.2. Parameter estimation and return levels. Suppose that a GPD with parameters (σ, ξ) is a suitable model for threshold excesses $(X - u)$ of a threshold u by a variable X. Then, from Equation (1), and for $x > u$,

$$\Pr(X < x) = 1 - \lambda_u \left[1 + \xi \left(\frac{x - u}{\sigma} \right) \right]_+^{-1/\xi}, \tag{2}$$

where $\lambda_u = \Pr(X > u)$; for a more detailed explanation, see Fawcett and Walshaw (2006). A typical application of the GPD would be to fit the distribution (1) to a series of threshold excesses using maximum likelihood estimation. The parameter λ_u would then be estimated empirically as the proportion of the data exceeding the threshold u. Estimates of an extreme quantile z_s would then be obtained by defining z_s by $\Pr(X < z_s) = 1 - s^{-1}$, and then solving (2) to obtain

$$z_s = u + \frac{\sigma}{\xi} \left[(\lambda_u s)^{\xi} - 1 \right], \tag{3}$$

In extreme value terminology, z_s is the *return level* associated with the *return period* s, and can be thought of as the level which we can expect to be exceeded once every s observations. In practice, it is common to extrapolate the relationship in Equation (3) to obtain estimates of return levels considerably beyond the range of the data to which the model has been fitted. These return level estimates are then often used as specifications for the design of sea-walls, or to estimate the level of protection offered by an existing wall. It is usually more practically convenient to give return levels on an annual scale, so that the r-year return level is the level expected to be exceeded once every r years. Here r is related to s by $r = Ns$, where N is the fixed number of time points occurring in a single year.

While point estimates for z_r can be obtained directly from Equation (3), standard errors can be estimated via techniques such as the delta method (Rao, 1973), but the construction of symmetrical confidence intervals, specified as a fixed number of standard errors either side of this point estimate ('Wald' intervals), is not recommended for return levels. Instead, we use the profile likelihood; rather than use the limiting quadratic form of the likelihood surface, profile likelihood confidence intervals make use of its actual shape for the model and the data in question (Venzon and Moolgavkar, 1988). The severe asymmetry of the surface often encountered when it is calculated for return levels suggests that conventional symmetrical confidence intervals are highly misleading (Walshaw, 1994).

3.4.3. Adjusting to account for dependence. The GPD in Equation (1) assumes our excesses are i.i.d., and so fitting to all exceedances of u when there is clearly evidence of short-term temporal dependence will result in underestimated standard errors. Smith (1991) suggests a procedure in which the usual asymptotic likelihood calculations are supplemented by empirical information on dependence, in order to produce a modified covariance matrix for the parameters, which is approximately correct after the dependence has been taken into account.

Under the model fitting procedure which assumes independence, denote the observed information matrix by $\mathbf{H}$. If independence were a valid assumption, then the covariance matrix of the maximum likelihood estimates (m.l.e.s) would be approximately $\mathbf{H}^{-1}$. Smith (1991) shows that to account for dependence, this approximation should be replaced by $\mathbf{H}^{-1}\mathbf{V}\mathbf{H}^{-1}$, where $\mathbf{V}$ is the covariance matrix of the likelihood gradient vector. Furthermore, $\mathbf{V}$ can be estimated by decomposing the log-likelihood sum into its contributions by year (which should be independent up to a good approximation) and obtaining the appropriate covariance matrix empirically.

DOI: 10.1002/env

Similar arguments can be applied to modify the procedure for testing hypotheses. Specifically, denoting model parameters by $\psi = (\rho, \zeta)$ where ρ and ζ are of dimensions p and q, respectively, suppose that a test of $H_0 : \rho = \rho_0$ against $H_1 : \rho \neq \rho_0$ is required, ζ being a nuisance parameter. Assuming independence, test procedures are usually based on the asymptotic distribution of

$$2\{\ell(\hat{\psi}_1) - \ell(\hat{\psi}_0)\}, \tag{4}$$

which is χ^2_p. Here, $\ell(\hat{\psi}_0)$ and $\ell(\hat{\psi}_1)$ denote the log–likelihood evaluated at the maximum likelihood estimate under H_0 and H_1 (respectively). Now suppose we wish to account for dependence. Partitioning

$$\mathbf{H} = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}, \tag{5}$$

where H_{11}, H_{12}, H_{21} and H_{22} are the appropriate sub-matrices of dimensions $p \times p$, $p \times q$, $q \times p$ and $q \times q$, respectively, then we partition the inverse of $\mathbf{H}$ as

$$\mathbf{H}^{-1} = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{pmatrix}, \tag{6}$$

where each sub-matrix $\mathbf{H}^{\cdot\cdot}$ has the same dimensions as $\mathbf{H}_{\cdot\cdot}$. Now let

$$\mathbf{C} = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} - H_{22}^{-1} \end{pmatrix}, \tag{7}$$

Then Smith (1991) shows that the approximate distribution of expression (4) is given by

$$\sum_{i=1}^{p} \lambda_i z_i^2 \tag{8}$$

where the z_i, $i = 1, \ldots, p$, are standard normal variates and the λ_i are the non-zero eigenvalues of $\mathbf{V}^{1/2}\mathbf{C}\mathbf{V}^{1/2}$. This replaces the usual χ^2_p-distribution, which is valid in the case of independence, and which would be recovered if all the λ_i were set equal to 1. It is then easy to simulate from the modified distribution (8) to estimate any required quantile of the test statistic. Profile likelihood confidence intervals then arise as the set of values of $\hat{\psi}_1$ such that the test statistic (4) is smaller than the quantile which represents the desired level of significance.

3.4.4. Results for the oceanographic data. Table 2 reports maximum likelihood estimates for the GPD scale and shape parameters, along with their Wald 95 per cent confidence intervals, for analyses using *all excesses* and just *cluster peak excesses*; in the analysis using information on all extremes, standard errors were inflated to account for temporal dependence via Smith's method (1991). Also shown is the threshold exceedance rate $\hat{\lambda}_u$ for each analysis.

Table 3 shows maximum likelihood estimates for return levels for four return periods—$s = 10$, 50, 200 and 1000 years—obtained by substituting the estimates shown in Table 2 into Equation (3). The corresponding 95 per cent confidence intervals have been obtained using the method of profile

DOI: 10.1002/env

Table 2. Maximum likelihood estimates, and associated Wald 95 per cent confidence intervals, for the GPD scale and shape parameters and the threshold exceedance rate when using all excesses and just cluster peak excesses

	$\hat{\sigma}$	$\hat{\xi}$	$\hat{\lambda}_u$
All excesses	0.104	−0.090	0.059
95% confidence interval	(0.082, 0.126)	(−0.217, 0.037)	(0.058, 0.060)
Cluster peaks	0.187	−0.259	0.013
95% confidence interval	(0.109, 0.265)	(−0.545, 0.027)	(0.012, 0.014)

Table 3. Maximum likelihood estimates, and associated 95 per cent profile likelihood confidence intervals, for four return levels (units are in metres)

	$\hat{z}_{10}$	$\hat{z}_{50}$	$\hat{z}_{200}$	$\hat{z}_{1000}$
All excesses	0.867	0.947	1.007	1.068
95% confidence interval	(0.736, 1.067)	(0.790, 1.193)	(0.844, 1.257)	(0.891, 1.335)
Cluster peaks	0.868	0.920	0.951	0.975
95% confidence interval	(0.770, 1.031)	(0.813, 1.099)	(0.838, 1.008)	(0.858, 1.063)

likelihood, where the appropriate cut–off for the test statistic (4) has been obtained using the modified distribution (8). In this way, the profile likelihood confidence intervals have been inflated to account for the dependence in a way which is consistent with the modifications proposed by Smith (1991). Figure 2 shows a plot of the profile likelihood for one of these return levels—z_{50}—illustrating the severe asymmetry which is commonly observed for return levels. This plot is for the analysis using all threshold exceedances. The 95 per cent profile likelihood confidence interval for z_{50}, after adjusting for

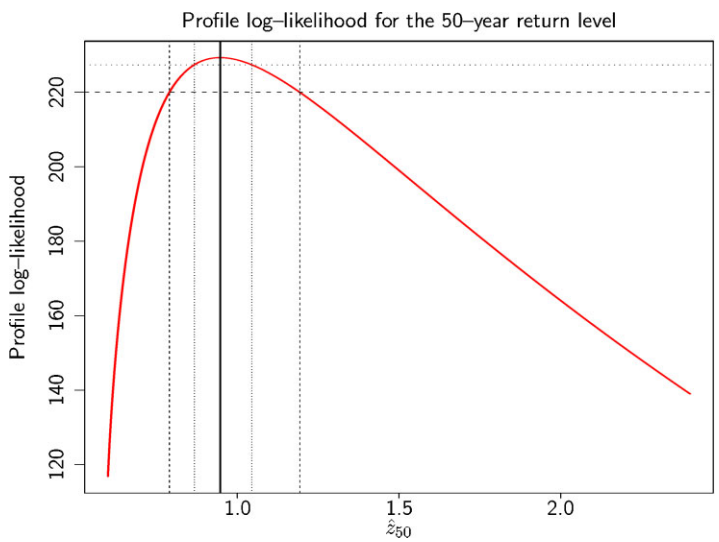

Figure 2. Profile log–likelihood surface, with corresponding 95 per cent confidence intervals, for the 50 year return level $\hat{z}_{50}$. The dashed lines show the construction of the interval which has been inflated to account for temporal dependence in the sea–surge data (since in this example all threshold excesses were used). The dotted lines show how the interval would be constructed if dependence had been ignored

DOI: 10.1002/env

dependence, is identified on the plot. Also shown is the much narrower interval which would have been obtained if dependence had been ignored.

Table 2 shows that, when the analysis is restricted to a set of cluster peak exceedances, the GPD scale parameter σ is overestimated, and the shape parameter ξ underestimated, relative to the approach which uses all exceedances. However, when we account for sampling variability, we see that these differences are not significant. The lower exceedance rate $\hat{\lambda}_u$ in the cluster peaks analysis is due to fewer extremes being used in the analysis.

Of greater practical interest are the estimated return levels. Table 3 shows that estimates barely differ for the 10-year return period, but are consistently smaller in the cluster peaks analysis for the other three periods studied—in fact, quite substantially so for the 200- and 1000-year return periods. Since estimates of such long–range return levels are often used as a design requirement in oceanographic situations (e.g. for the height of sea walls), designing to a level specified by an analysis based on cluster peak excesses could result in substantial under-protection. Of course, since the true return level is unknown, we do not know which of our analyses is providing the better estimates. We investigate the performance of the two approaches in Section 4.

4. SIMULATION STUDY

The analysis of the sea-surge data in Section 3 showed clear discrepancies in both parameter estimation and return level estimation between the method which used information on all extremes and that which used cluster peak extremes only. In this section, our aim is to use simulated data to compare the performance of estimators obtained under the two approaches. The bias and efficiency will be studied, and different methods of declustering will be used to identify the sensitivity of the cluster peaks analysis to any particular scheme.

This section is organised as follows. In Section 4.1, we discuss how the temporal evolution of an environmental process such as the sea-surge data can be captured by a parametric model, and we outline how to simulate a series from such a model. Section 4.2 gives some design details for our simulation study, and the results of the study are presented in Sections 4.3 and 4.4. Finally, Section 4.5 investigates the sensitivity of parameter estimation and return level estimation to different choices of declustering scheme, in order to show that our findings are a general feature of declustering, and not a function of some particular method of cluster identification.

4.1. Simulating temporally dependent extremes

We assume that our series of observations $X_1, X_2, \ldots, X_n$ forms a stationary first-order Markov chain. This assumption is true to a good approximation for many environmental variables, and provides a simple model for the serial correlation. The stochastic properties of such a chain are completely determined by the joint distribution of successive pairs of observations. Given a model $f(x_i, x_{i+1}; \boldsymbol{\psi})$, $i = 1, \ldots n-1$, specified by parameter vector $\boldsymbol{\psi}$, it follows that the joint density function for $x_1, \ldots, x_n$ is given by

$$f(x_1, \ldots, x_n; \boldsymbol{\psi}) = \prod_{i=1}^{n-1} f(x_i, x_{i+1}; \boldsymbol{\psi}) \Big/ \prod_{i=2}^{n-1} f(x_i; \boldsymbol{\psi}) \,. \tag{9}$$

DOI: 10.1002/env

A limiting model for $f(x_i; \boldsymbol{\psi})$, for the region (u, ∞), is the GPD given in Equation (1). For contributions to the numerator in Equation (9), we invoke bivariate extreme value theory considerations to obtain a corresponding model for $f(x_i, x_{i+1}; \boldsymbol{\psi})$, on the region $(u, \infty) \times (u, \infty)$.

4.1.1. Bivariate threshold excess model. Recall that under broad conditions, for an arbitrary distribution function F, then for a high enough threshold u, the GPD is justified as a model for the exceedances $(x - u | x > u)$, on the grounds that it is the only possible non-degenerate limiting form for such exceedances. To model the joint behaviour of (x_i, x_{i+1}) above high thresholds, we need a bivariate joint limiting distribution $F(x, y)$ on regions of the form $x > u_x$, $y > u_y$, for large enough u_x and u_y.

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 (1), with respective parameter sets $(\lambda_{u_x}, \sigma_x, \xi_x)$ and $(\lambda_{u_y}, \sigma_y, \xi_y)$ representing the exceedance rates and GPD parameters. After transforming the margins for (X, Y) to the standard Fréchet distribution (see Fawcett and Walshaw, 2006), it can be shown (Pickands, 1981) that the joint distribution function $G(x, y)$ for a bivariate extreme value distribution has the representation

$$G(x, y) = \exp\{-V(x, y)\}, \tag{10}$$

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

$$V(x, y) = 2 \int_0^1 \max(q/x, (1 - q)/y) \, \mathrm{d}H(q), \tag{11}$$

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

$$\int_0^1 q \, \mathrm{d}H(q) = \frac{1}{2}. \tag{12}$$

There is no characterisation of the complete family of distributions specified by Equation (10), and so model choice involves specifying an appropriate sub-family through the choice of H in Equation (11).

4.1.2. Model choice. For our simulated data, we consider three models for the *dependence function* H: the popular (symmetric) logistic model, the negative logistic model (also symmetric) and the bilogistic model, which is a generalisation of the logistic model that allows for asymmetry in the dependence structure (details of all these models can be found in Coles (2001)).

The logistic and negative logistic models are defined by Equation (10) with:

$$V(x, y) = \left(x^{-1/\alpha} + y^{-1/\alpha}\right)^{\alpha}, \qquad 0 < \alpha \leq 1, \qquad \text{and} \tag{13}$$

$$V(x, y) = -x - y + \left(x^{-r} + y^{-r}\right)^{-1/r}, \qquad r > 0, \tag{14}$$

respectively. Independence is obtained when $\alpha = 1$ and $r \searrow 0$ for the logistic and negative logistic models, respectively, while complete dependence is obtained when $\alpha \searrow 0$ and $r \to \infty$. The (asymmetric) bilogistic model has

$$V(x, y) = -x\gamma^{1-\alpha} - y(1 - \gamma)^{1-\beta}, \qquad 0 < \alpha, \beta < 1 \tag{15}$$

DOI: 10.1002/env

where $\gamma = \gamma(x, y; \alpha, \beta)$ is the solution of $(1-\alpha)x(1-\gamma)^{\beta} = (1-\beta)y\gamma^{\alpha}$. When $\alpha = \beta$, this model reduces to the (symmetric) logistic model. 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, β is fixed and the other approaches 1. Different limits occur when one of α or β is fixed and the other approaches zero.

4.1.3. Simulation. Replacing x and y in Equations (13–15) with x_i and x_{i+1}, $i = 1, \ldots, n-1$, respectively, gives the joint distribution function for successive pairs in a time series context. To simulate from one of the bivariate extreme value distributions, we use the following procedure:

1. Simulate the first observation, x_1, from the standard Fréchet distribution.
2. Use the simulated observation, x_1, to form the conditional distribution of $x_2 \mid x_1$ using the chosen model (13), (14) or (15). Simulate x_2 from this distribution.
3. Use the simulated observation, x_2, to form the conditional distribution of $x_3 \mid x_2$, again using the chosen model (13), (14) or (15). Simulate x_3 from this distribution.
4. Continue forming and sampling from new conditionals until the required number of observations (n) has been generated.

The marginals of a series simulated using the scheme above will be of standard Fréchet type. To assess the effects of using all threshold excesses and using just cluster peak excesses on GPD parameter estimation (and subsequent return level estimation), we transform the entire simulated series so that the margins are GPD. Thus all observations, not only those above a chosen threshold, will follow the model in Equation (1). For more information on the precise simulation details, see Fawcett (2005).

4.2. Simulation study details

The aim of the simulation study is to investigate differences in GPD parameter estimation and return level estimation between the methods which use all extremes, and a filtered set of extremes only. Thus, we simulate first-order Markov chains of extreme value type according to the three models outlined in Section 4.1.2. Within each dependence model, we simulate chains of varying degrees of temporal dependence according to the dependence parameter(s) for that model. We are able to carry out separate simulations for values of α (and β for the asymmetric model) over the whole range of values, on a fairly fine grid. We also choose five different values of the GPD shape parameter ξ: -0.4, -0.1, 0, 0.3 and 0.8, to reflect the various tails which might be observed in a real-life data set; the GPD scale parameter σ is held unit constant.

Although the simulated values will already follow a GPD (due to our transforming the series from standard Fréchet to GPD), to assess the effects of declustering on inference we need to specify a high threshold u, in order to observe realistic clustering behaviour. We set u so that $G(u) = 0.95$ in Equation (1), so that we would expect 5 per cent of simulated values to exceed u. This level is chosen to be representative of a typical high threshold used in real life situations. For our inferences, we replace the scale parameter σ by $\sigma^* = \sigma + \xi u$. The GPD parameters (σ^*, ξ) are then *threshold invariant*, in the sense that if exceedances over a threshold u^* are GPD distributed, then exceedances over all thresholds $u > u^*$ are GPD distributed with the same parameters (Coles, 2001). This arises as a result of the threshold stability property of the GPD (see for example Davison and Smith, 1990), and is very useful here in that it allows us directly to compare our inferences across the different thresholds u that arise when we set $G(u) = 0.95$ for different values of ξ, while keeping $\sigma = 1$.

DOI: 10.1002/env

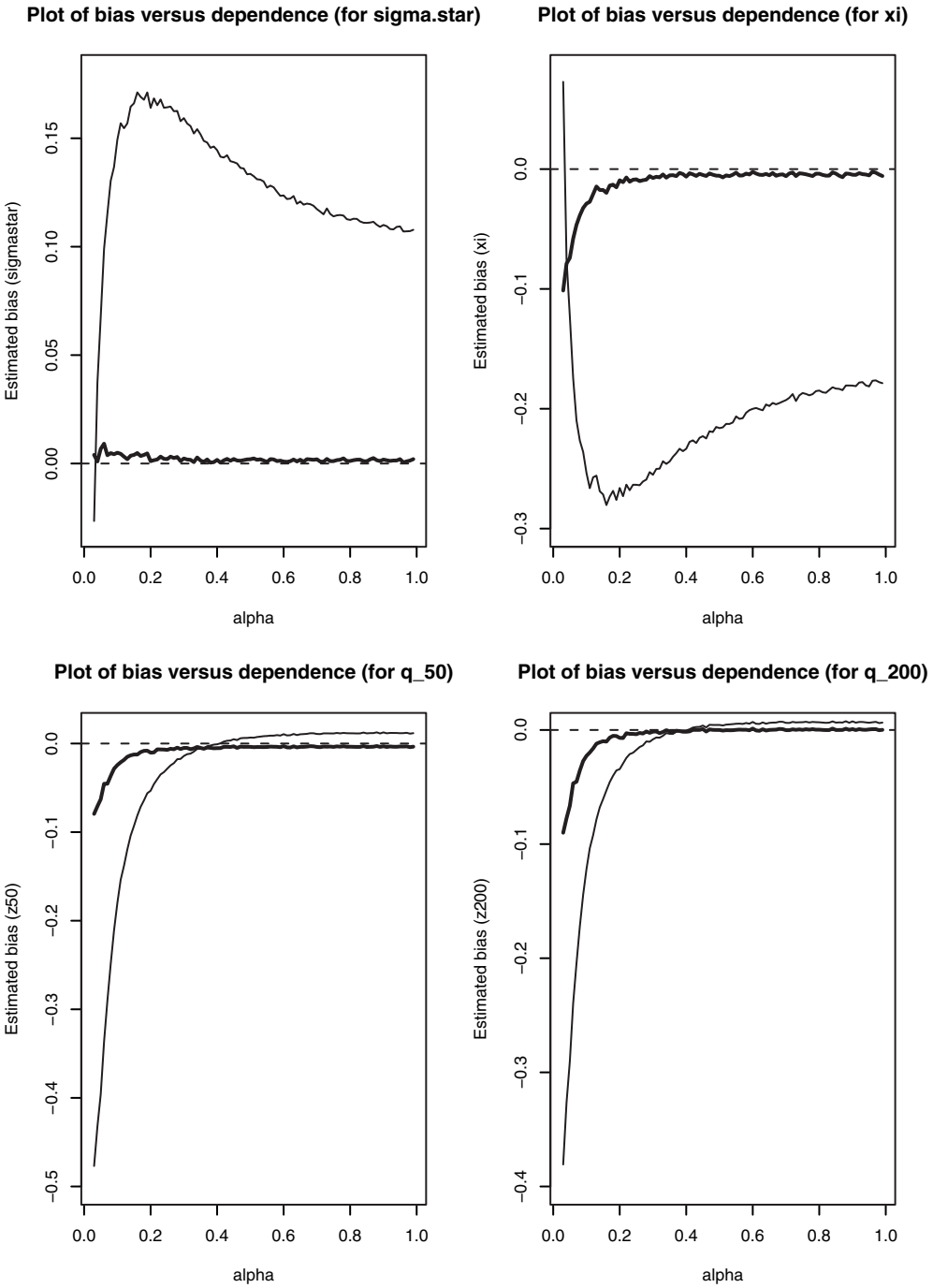

Figure 3. Plots of bias versus level of temporal dependence for the GPD parameters σ^* and ξ and the 50- and 200-year return levels z_{50} and z_{200}—the heavy line is for the analysis using all threshold excesses, the thin line for that using cluster peak excesses

DOI: 10.1002/env

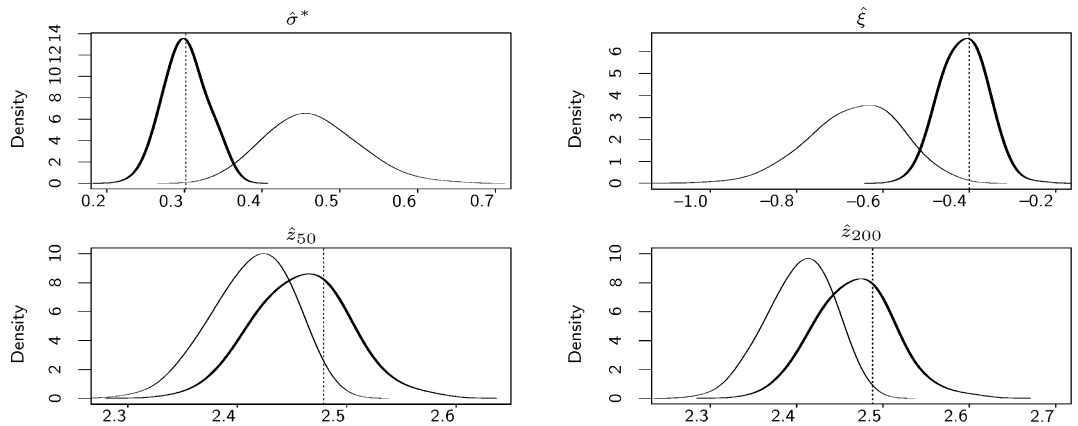

Figure 4. Sampling distributions of $\hat{\sigma}^*$, $\hat{\xi}$, $\hat{z}_{50}$ and $\hat{z}_{200}$ when $\alpha = 0.2$, and using all threshold excesses (heavy line) and cluster peak excesses (thin line). The *true* values for each parameter are shown by the vertical lines

For the simulated chain (of length n), we obtain estimates of the GPD parameters and return levels through maximum likelihood estimation using (i) all threshold excesses and (ii) cluster peak excesses. Cluster peak excesses are identified using runs-declustering. This simulation–estimation procedure is repeated N times for each strength of temporal dependence within each of the three models used (and for each of the pre-determined marginal GPD parameters) to create sampling distributions for the GPD scale and shape parameters, and also for some return levels. These sampling distributions can then be compared to the true parameter values used to simulate the chains in the first place. In this study, we simulate chains of length $n = 10\,000$ and repeat the simulation–estimation procedure $N = 1000$ times.

5. RESULTS

A full set of results for this simulation study is available in Fawcett (2005); in this paper, we highlight the most important outcomes of the study. We discuss the effects of using all excesses and just cluster peak excesses on the estimation of the GPD parameters σ^* and ξ, as well as the consequences of both approaches for return level inference.

Figure 3 shows a typical plot of the estimated bias for the GPD parameters σ^* and ξ, and the 50- and 200-year[‡] return levels, z_{50} and z_{200}. This plot shows the results from the simulation where $\xi = -0.4$, $\sigma^* = 0.302$, $z_{50} = 2.479$ and $z_{200} = 2.488$, using the logistic model for temporal dependence with a range of values for α—similar results were obtained for all other values of ξ in the study. Similar plots were also obtained when the other two models for temporal dependence were used (for similar strengths of serial correlation, and over a range of values of $\alpha - \beta$ in the asymmetric bilogistic model). Figure 3 shows that for all but the strongest levels of serial correlation, using cluster peak excesses in the analysis consistently overestimates the GPD scale parameter σ^* and underestimates the shape parameter ξ, with this bias decreasing as the temporal dependence weakens. We also see that for the cluster peaks approach, when there is strong temporal dependence in the series, both return levels are substantially underestimated, while when there is weak dependence, i.e. for values of α bigger than about 0.5, they are actually slightly *overestimated*.

[‡]For the simulated data, we fix the length of a 'year' to be the same as that for the (3-hourly) Newlyn sea-surge data, i.e. $365.25 \times 24/3 = 2922$ observations, in order to match the time scale with that of the real data.

DOI: 10.1002/env

Table 4. Sampling distribution means, and 95 per cent confidence intervals (parenthesised), for GPD parameters and two return levels for simulated data using (i) all threshold excesses and (ii) cluster peak excesses. In the last two rows of the Table, the values in parentheses (and italics) correspond to the estimated bias and MSE for estimates using cluster peak excesses

True value	$\sigma^* = 0.302$	$\xi = -0.4$	$z_{50} = 2.479$	$z_{200} = 2.488$
Estimate using all excesses	0.301 (0.251, 0.351)	−0.413 (−0.507, −0.323)	2.459 (2.378, 2.544)	2.568 (2.382, 2.561)
Estimate using cluster peaks	0.464 (0.358, 0.583)	−0.665 (−0.868, −0.480)	2.404 (2.331, 2.466)	2.413 (2.344, 2.472)
Estimated bias	0.000 (*0.162*)	−0.013 (*−0.265*)	−0.020 (*−0.066*)	−0.020 (*−0.084*)
MSE	0.001 (*0.030*)	0.003 (*0.080*)	0.002 (*0.006*)	0.002 (*0.008*)

Table 5. Comparison of estimated standard errors with standard deviations of estimators

		$\hat{\sigma}^*$	$\hat{\xi}$	$\hat{z}_{50}$	$\hat{z}_{200}$
$\alpha = 0.2$	St. dev.	0.039	0.075	0.041	0.047
	mean(e.s.e.)	0.024	0.053	0.019	0.026
	mean(*adjusted* e.s.e.)	0.038	0.076	0.040	0.045
	St. dev.	0.026	0.056	0.025	0.030
$\alpha = 0.5$	mean(e.s.e.)	0.023	0.051	0.019	0.026
	mean(*adjusted* e.s.e.)	0.026	0.056	0.024	0.030
	St. dev.	0.024	0.051	0.021	0.025
$\alpha = 0.8$	mean(e.s.e.)	0.023	0.050	0.019	0.024
	mean(*adjusted* e.s.e.)	0.024	0.051	0.021	0.025

By contrast, using all threshold excesses gives parameter estimates with virtually zero bias for all but the very strongest levels of dependence in the study, and even here the all-important return level estimates show much less bias than estimates obtained from cluster peak excesses.

Figure 4 shows the actual sampling distributions for $\hat{\sigma}^*$, $\hat{\xi}$, $\hat{z}_{50}$ and $\hat{z}_{200}$ for the component of the study shown in Figure 3 when $\alpha = 0.2$. The bias for the cluster peaks analysis indicated in Figure 3 at this level of dependence (which is similar to the temporal dependence observed for the Newlyn sea–surge data) is clearly seen—the sampling distributions are located well away from the true values, in contrast to the analysis using all exceedances, where the distributions locate their target values much more accurately. Table 4 reports the sampling distribution means and 95 per cent confidence intervals corresponding to the plots shown in Figure 4. The inadequacy of the cluster peaks approach relative to the approach which uses all excesses is obvious; in the former approach, the 95 per cent confidence intervals for some estimators even fail to include the true parameter values. From Table 4, we also see that as well as estimating with greater bias, estimates based on an analysis which uses cluster peak excesses have a larger mean squared error (MSE) than their counterparts from the analysis using all threshold excesses.

The inadequacy of the cluster peaks approach, particularly for return level estimation, could have major practical implications. In an oceanographic setting (as with the Newlyn sea-surge data analysed in Section 3), designing to a level specified by a cluster peaks analysis could result in considerable under–protection.

DOI: 10.1002/env

Table 6. Sampling distribution means, and 95 per cent confidence intervals, for GPD parameters and two return levels for declustered simulated data. Results are shown for two different values of cluster identification parameter ($\kappa = 30$ h and $\kappa = 90$ h, *c.f.* $\kappa = 60$ h as originally used). Also shown are results which decluster using randomly chosen cluster excesses. Results using cluster peaks identified using blocks-declustering, and 'automatic' declustering, are also shown

Declustering scheme	$\sigma^* = 0.302$	$\xi = -0.4$	$z_{50} = 2.479$	$z_{200} = 2.488$
Runs, $\kappa = 30$ h	0.476 (0.423, 0.511)	−0.668 (−0.766, −0.593)	2.389 (2.353, 2.438)	2.397 (2.342, 2.428)
Runs, $\kappa = 90$ h	0.478 (0.344, 0.608)	−0.680 (−0.981, −0.353)	2.394 (2.314, 2.482)	2.456 (2.397, 2.516)
Runs, $\kappa = 60$ h (random excesses)	0.466 (0.361, 0.587)	−0.621 (−0.863, −0.476)	2.408 (2.336, 2.471)	2.414 (2.345, 2.475)
Blocks (block length = 60 h)	0.409 (0.312, 0.506)	−0.610 (−0.833, −0.387)	2.365 (2.328, 2.402)	2.391 (2.354, 2.438)
Automatic	0.406 (0.308, 0.500)	−0.602 (−0.824, −0.378)	2.364 (2.326, 2.404)	2.389 (2.350, 2.440)

5.1. Assessing the adjustments for dependence

The simulation study provides us with a sample from the distribution of the m.l.e.s of the parameters and return levels. This allows us to assess the performance of our estimates for the *standard errors* associated with these m.l.e.s: a good estimate of the standard error associated with a particular maximum likelihood estimator should match well with the sample standard deviation of the m.l.e. in question, under repeated sampling.

Table 5 clearly shows the pitfalls of simply fitting to all threshold excesses and ignoring temporal dependence; the estimated standard errors for each parameter are considerably smaller than the standard deviation of the corresponding sampling distribution. However, Smith's adjustment is seen to perform well, inflating the estimated standard errors to a level which is close to that of the standard deviation of the sampling distributions.

5.2. Robustness of results

To demonstrate our results are not a function of any particular declustering scheme, we have also investigated the sensitivity of estimation to the choice of declustering scheme used to identify independent threshold excesses. Initially, we used runs-declustering with a cluster termination interval of $\kappa = 60$ h (equivalent to 20 3-hourly observations), in line with the choice of κ for the Newlyn data. However, estimated GPD parameters and return levels could be sensitive to the choice of this 'auxiliary parameter', and so the first extension to the study was to consider a range of different values for κ. Results for $\kappa = 30$ and $\kappa = 90$ are shown in Table 6.

Using runs–declustering is commonplace, and is possibly the most 'natural' way to identify clusters of extremes. However, not only is the choice of auxiliary parameter an arbitrary one, the declustering technique itself is only one of a range of filtering schemes available. The second extension to the study was to consider other declustering schemes, such as 'block–declustering' (see, for example, Smith and Weissman, 1994). Another scheme, proposed by Ferro and Segers (2003), is 'automatic declustering',

DOI: 10.1002/env

which does not require the specification of an auxiliary parameter but uses the inter–arrival times of threshold exceedances to identify clusters. Results for blocks–declustering with block length 60 h are shown in Table 6, along with results from the automatic method.

A third extension to the study was to investigate filtering schemes which choose a cluster inhabitant other than the maximum threshold excess; to this end, declustering schemes using the first observed threshold exceedance, a randomly chosen cluster exceedance and the mean cluster excess, were also used. Results for the randomly chosen exceedance approach are shown here; the results based on the other choices are very similar.

None of the extensions outlined above resulted in any change to the substance of our findings. All of the analyses based on cluster peaks gave rise to estimators with the characteristic bias we described earlier (as well as the increased MSE) relative to the estimators based on all exceedances. Return levels are consistently underestimated when using a filtered set of independent exceedances for analysing data which display strong serial correlation.

All of these findings support our main conclusion: the process of declustering applied to exceedances of a high threshold introduces unnecessary complexity which systematically incurs estimation bias. We recommend that *all* exceedances are modelled, and that the problems caused by the consequent dependence in the data are relatively easy to overcome, using the method we describe. This approach appears to work extremely well up to the requirements of estimating long-period return levels. For more complex requirements, particularly features of the serial correlation itself such as storm duration, we would recommend explicitly *modelling* the dependence. This is considered in Fawcett and Walshaw (2006).

ACKNOWLEDGEMENTS

Lee Fawcett's work was funded by an EPSRC Doctoral Training Award. We thank Stuart Coles for providing the sea–surge data used in Section 3.

REFERENCES

Coles SG. 1991. Statistical methodology for the multivariate analysis of environmental extremes. *PhD thesis*, University of Sheffield, Sheffield.

Coles SG. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer: London.

Coles SG, Tawn JA. 1991. Modelling extreme multivariate events. *Journal of Royal Statistical Society, Series B* **53**: 377–392.

Davison AC, Smith RL. 1990. Models for exceedances over high thresholds (with discussion). *Journal of Royal Statistical Society, Series B* **52**: 393–442.

Fawcett L. 2005. Statistical methodology for the estimation of environmental extremes. *PhD Thesis*. University of Newcastle, Newcastle.

Fawcett L, Walshaw D. 2006. Markov chain models for extreme wind speeds. *Environmetrics*, in press.

Ferro CAT, Segers J. 2003. Inference for clusters of extreme values. *Journal of Royal Statistical Society, Series B* **65**: 545–556.

Pickands J. 1981. Multivariate extreme value distributions. *Bullettin International Statistical Institute* **XLIX**(Book 2): 859–878.

Rao CR. 1973. *Linear Statistical Inference and its Applications*, 2nd edn. Wiley: New York.

Smith RL. 1991. Regional estimation from spatially dependent data. Preprint. (http://www.stat.unc.edu/postscript/rs/regest.pdf).

Smith RL, Weissman I. 1994. Estimating the extremal index. *Journal of Royal Statistical Society, Series B* **56**: 515–528.

Venzon DJ, Moolgavkar SH. 1988. Profile-likelihood–based confidence intervals. *Applications of Statistics* **37**: 87–94.

Walshaw D. 1994. Getting the most from your extreme wind data: a step by step guide. *Journal of Research in National Institute of Standards and Technology* **99**: 399–411.

DOI: 10.1002/env