Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights

Journal:	*Environmetrics*
Manuscript ID:	env-10-0019.R2
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Northrop, Paul; University College London, Department of Statistical Science Jonathan, Philip; Shell Research Limited
Keywords:	extreme value regression modelling, dependent data, quantile regression, threshold selection, wave heights

Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights

Paul Northrop and Philip Jonathan

January 12, 2011

Abstract

In environmental applications it is common for the extremes of a variable to be non-stationary, varying systematically in space, time or with the values of covariates. Multi-site datasets are common, and in such cases there is likely to be non-negligible inter-site dependence. We consider applications in which multi-site data are used to infer the marginal behaviour of the extremes at individual sites, while adjusting for inter-site dependence. For reasons of statistical efficiency, it is standard to model exceedances of a high threshold. Choosing an appropriate threshold can be problematic, particularly if the extremes are non-stationary. We propose a method for setting a covariate-dependent threshold using quantile regression. We consider how the quantile regression model and extreme value models fitted to threshold exceedances should be parameterized, in order that they are compatible. We adjust estimates of uncertainty for spatial dependence using methodology proposed recently. These methods are illustrated using time series of storm peak significant wave heights from 72 sites in the Gulf of Mexico. A simulation study illustrates the applicability of the proposed methodology more generally.

Keywords: Extreme value regression modelling; dependent data; quantile regression; threshold selection; wave heights.

1 Introduction

This article deals with regression modelling of extreme values in the presence of spatial dependence. In environmental applications it is common for the extremes of a variable to be non-stationary, varying systematically in space, time or with the values of covariates. Regression modelling is a natural way to account for such effects. For reasons of statistical efficiency, modern extreme value analyses model the number of magnitude of exceedances of a high threshold. We argue that if there is clear non-stationarity in extremes a non-constant threshold should be set that reflects the non-stationarity. We propose a method for setting covariate-dependent threshold using quantile regression. We also consider how the quantile regression model and extreme value models fitted to threshold exceedances should be parameterized, in order that they are compatible. Often response data are not independent, even once the effects of covariates have been taken into account. For example, in multi-site data-sets extremal behaviour, and systematic effects on extremal behaviour, will tend to be similar at neighbouring sites. It is advantageous to model simultaneously data from different sites and thus improve precision of estimation. However, inferences must take proper account of spatial dependence.

1.1 Threshold modelling of non-stationary extremes

Extreme value theory provides asymptotic justification for particular families of models for extreme data. A key result for stationary extremes (Leadbetter et al., 1983) suggests a GEV distribution as a working model for the maximum of a large number of identically distributed random variables. Thus we might model annual maxima as having a GEV(μ, σ, ξ) distribution. Further theory (Pickands, 1975) suggests that, when a large threshold u is exceeded, the amount by which it is exceeded be modelled by a Generalized

Pareto, GP(σ_u, ξ) distribution, where the scale parameter $\sigma_u = \sigma + \xi(u - \mu)$. An equivalent formulation is the non-homogeneous Poisson process characterization of threshold exceedances, developed by Pickands (1971) and first used for practical application by Smith (1989). The parameterization of this model, in terms of the GEV parameters μ, σ and ξ, is invariant to u, a property which is advantageous if a non-constant threshold is used. For threshold models local dependence produces threshold exceedances that occur in clusters in time. Standard methods for non-stationary extremes use regression modelling (Davison and Smith, 1990), in which the parameters of the extreme value model depend on the values of covariates, although typically ξ is taken to be constant. The main aim of an extreme value analysis is the estimation of extreme quantiles. Suppose that, conditional on the values of covariates, annual maxima Y_m have a GEV(μ, σ, ξ) distribution. The conditional N year return level q_N, exceeded approximately once on average every N years, satisfies $P(Y_m > q_N) = 1/N$ and is given by $q_N = \mu + \sigma \left[\{ -\log(1 - 1/N) \}^{-\xi} - 1 \right] / \xi$.

There are strong arguments against using a constant threshold for non-stationary extremes. For a given covariate value we need the threshold to be large enough that exceedance magnitudes can be modelled using a GP distribution. A threshold that is sufficiently large for this purpose at one covariate value may be rather low for another covariate value. Another argument is one of statistical design: to improve the precision of estimation of a covariate effect we should aim to have exceedances spread as far across the observed values of the covariate as possible. Setting a constant threshold will tend to narrow the range of covariates for which there are exceedances.

In this paper we use quantile regression (Koenker and Bassett, 1978; Koenker, 2005) to set a threshold for which the probability p of threshold exceedance is approximately constant across different values of the covariates, in this case across different spatial locations. We argue that it is more logical to model the threshold for constant p than to model how

the probability of exceedance of a constant threshold varies with covariates. Theoretical work will quantify the extent to which this is an optimal strategy. This is described in the discussion section of this paper.

Covariate-dependent thresholds for non-stationary extremes have previously been set in an ad hoc manner (Smith, 1989; Coles, 2001). Eastoe and Tawn (2009) set a threshold based on a model fitted to all the data, extreme and non-extreme, so that the threshold is affected by trends in typical values of the response. In contrast quantile regression models directly covariate effects at the desired level of extremity, does not require a distributional assumption for the response data and is robust to the presence of outliers. The parameters of quantile regression retain their statistical properties under any monotonic transformation of the response, which may be useful in applications where a non-linear transformation is applied to the response data (Wadsworth et al., 2010).

1.2 Spatial dependence

One way to account for spatial dependence is to model it explicitly. Recent work in this area includes Naveau et al. (2009), Casson and Coles (1999), Cooley et al. (2007). The multivariate extreme value methodology of Heffernan and Tawn (2004) is also applicable to spatial problems. If interest is in the marginal distributions at the sites, it may not be necessary or desirable to model explicitly the spatial dependence (Fawcett and Walshaw, 2007) and we may carry out a marginal analysis: ignoring the dependence initially, and then making adjustments to test statistics and estimates of parameter uncertainty to account for the dependence.

In section 2 we describe the GOMOS wave hindcast dataset and produce some preliminary plots to guide later analyses. In section 3 we consider how to use the point process model

to analyse the hindcast data. Section 3.1 gives a summary of the inferential adjustments made for spatial dependence in the data. Sections 3.2.1 , 3.2.2 and 3.2.3 consider how to use quantile regression to set an appropriate threshold, how to parameterize appropriately the quantile regression model and how to choose a suitable overall level for the threshold. The extreme value of analysis of the hindcast data is presented in section 3.3. Inferences about return levels are presented in section 4. All the calculations and plots are produced using R (R Development Core Team, 2009).

2 Wave hindcast data

When developing environmental design criteria for marine structures it is important to quantify the stochastic behaviour of extreme sea states. A standard measure of sea surface roughness is significant wave height (H_s), defined as the mean of the highest one third of waves. When modelling H_s it is important that we capture systematic spatial effects and that we account of the strong spatial dependence in the data (Jonathan and Ewans, 2007; Jonathan et al., 2008; Jonathan and Ewans, 2011).

The data (Oceanweather Inc., 2005) cover the period September 1900 to September 2005 inclusive, at 30 minute intervals. The hindcasts are produced by a physical model, calibrated to observed hurricane data. For a typical Gulf of Mexico location we selected 72 sites arranged on a 6 $\times$ 12 rectangular lattice with spacing 0.125 degrees (approximately 14km). For reasons of confidentiality we code longitude as 1,...,12 and latitude as 1,...,6, rather than giving their true values. Over this region there are 315 hurricane events during the period in question: an average of three events per year. For each event for each site, we isolate the storm peak significant wave height H_s^{sp}, that is, the largest H_s value observed during the event, and treat the data at each site as a discrete time series

indexed by event number.

Figure 1 shows how the site maximum varies with longitude and latitude. The plot suggests that the data may vary non-linearly over space. Similar effects are observed in plots of at-site estimates of high quantiles.

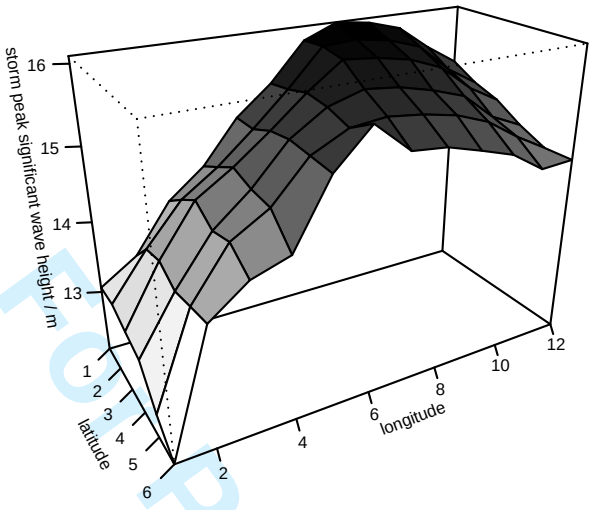

Figure 1: Site maximum of H_s^{sp} against longitude and latitude. The grey scale indicates the value: the larger the storm peak the darker the shading.

Figure 2 is a plot of the network maximum against event number. A robust LOESS

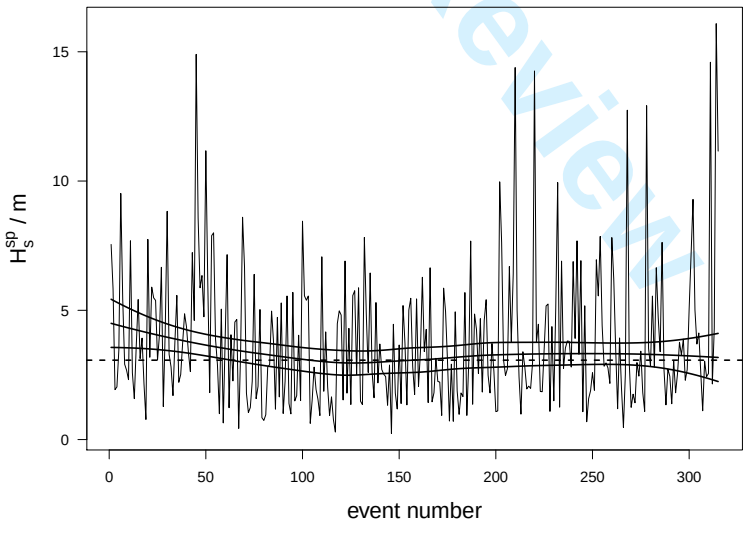

Figure 2: Network maximum of H_s^{sp} against hurricane event number. The horizontal dashed line indicates the median of the series. The solid lines are a robust LOESS smooth of the series and 95% confidence intervals.

smooth (using Tukey's biweight function) of the series suggests a slight decrease in average storminess over the earliest third of the data, but any trend in time appears less

pronounced than over space. We will assume in this paper that the data are temporally homogeneous. However, there is interest in examining temporal variability of wave hindcasts, in part because methods for measuring hurricane strength have changed over the 20th century (see, for example, Killick et al. (2010)).

Figure 3 shows that, even for pairs of sites situated at opposite points of the network, there is strong spatial association in the storm peak values.

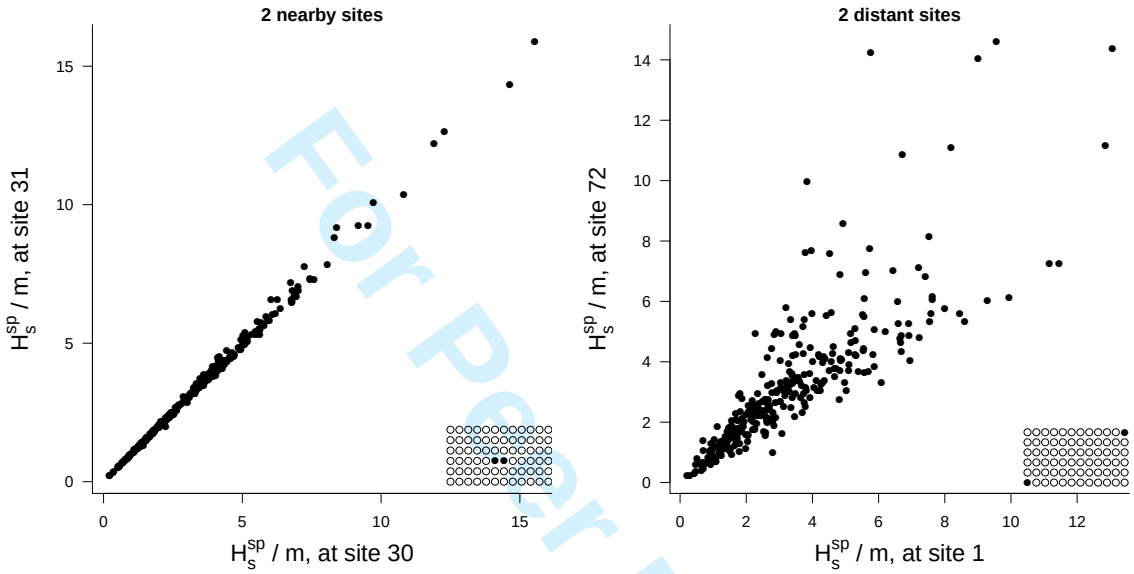

Figure 3: Scatter plots of contemporaneous H_s^{sp} values at pairs of sites. Left: contiguous sites (site 30: latitude=3, longitude=6 and site 31: latitude=3 and longitude=7). Right: distant sites (site 1: latitude=1, longitude=1 and site 72: latitude=6, longitude=12).

Storm peaks have been isolated with the aim of eliminating temporal dependence in the data. We check this informally at each site by plotting successive H_s^{sp} values against each other. Inspection of these plots supports the assumption that H_s values from different hurricane events are approximately independent. The Spearman rank correlation ranges from 0.028 to 0.070 across sites, with median value 0.047.

3 Extreme value modelling

We fit to the GOMOS wave data spatial regression models based on the point process characterization of exceedances of a high threshold developed by Pickands (1971). We consider models in which the parameters μ, σ and perhaps ξ are related to longitude and latitude. We use quantile regression to set a threshold for which the probability p of threshold exceedance is constant for all sites.

Let the response Y_{ij} be the storm peak significant wave height at site i in storm j, where $i = 1, \ldots, 72$ and $j = 1, \ldots, 315$, and let $\boldsymbol{x}_{ij}$ be a vector of covariates relevant to Y_{ij}. In particular, $\boldsymbol{x}_{ij}$ contains functions of the longitude and latitude of site i. Let $u(\boldsymbol{x}_{ij})$ be the threshold applied for a given $\boldsymbol{x}_{ij}$. Let θ be a vector containing the parameters that relate the GEV parameters $\mu(\boldsymbol{x}_{ij})$, $\sigma(\boldsymbol{x}_{ij})$ and $\xi(\boldsymbol{x}_{ij})$ to the covariates and let $f_{ij}(y_{ij} \mid \boldsymbol{x}_{ij}; \theta)$ be the conditional density of y_{ij} given $\boldsymbol{x}_{ij}$. If, conditional on the covariate values, the responses are independent, the likelihood function under the point process model is given by

$$\begin{aligned} L_I(\theta) &= \prod_{j=1}^{315} \prod_{i=1}^{72} f_{ij}(y_{ij} \mid \boldsymbol{x}_{ij}; \theta) \\ &= \prod_{j=1}^{315} \prod_{i=1}^{72} \exp\left\{ -\frac{1}{\lambda} \left[1 + \xi(\boldsymbol{x}_{ij}) \left(\frac{u(\boldsymbol{x}_{ij}) - \mu(\boldsymbol{x}_{ij})}{\sigma(\boldsymbol{x}_{ij})} \right) \right]^{-1/\xi(\boldsymbol{x}_{ij})} \right\} \\ &\quad \times \prod_{j=1}^{315} \prod_{i : y_{ij} > u(\boldsymbol{x}_{ij})} \frac{1}{\sigma(\boldsymbol{x}_{ij})} \left[1 + \xi(\boldsymbol{x}_{ij}) \left(\frac{y_{ij} - \mu(\boldsymbol{x}_{ij})}{\sigma(\boldsymbol{x}_{ij})} \right) \right]^{-1/\xi(\boldsymbol{x}_{ij})-1}, \end{aligned} \tag{1}$$

where λ is the mean number of observations per year, $1 + \xi(\boldsymbol{x}_{ij}) \left[y_{ij} - \mu(\boldsymbol{x}_{ij}) \right] / \sigma(\boldsymbol{x}_{ij}) > 0$ for all $y_{ij} > u(\boldsymbol{x}_{ij})$ and $1 + \xi(\boldsymbol{x}_{ij}) \left[u(\boldsymbol{x}_{ij}) - \mu(\boldsymbol{x}_{ij}) \right] / \sigma(\boldsymbol{x}_{ij}) > 0$ and $\sigma(\boldsymbol{x}_{ij}) > 0$ for all i, j.

We fit the point process model underlying (1) by maximizing $\log L_I(\theta)$ with respect to θ. In section 3.1 we consider how to adjust the 'independence' likelihood $L_I(\theta)$ to account for spatial dependence.

Following Chandler (2005) we represent systematic spatial variation using a basis of Legendre polynomials: $\phi_i(\cdot), i = 0, 1, \ldots$ (Abramowitz and Stegun, 1965). For example, if l_x represents longitude and l_y represents latitude, a quadratic representation of the effects of longitude and latitude on the GEV parameter μ is given by

$$\mu = \mu_0 + \mu_1\,\phi_1(l_x) + \mu_2\,\phi_1(l_y) + \mu_3\,\phi_2(l_x) + \mu_4\,\phi_1(l_x)\,\phi_1(l_y) + \mu_5\,\phi_2(l_y),$$

where $\phi_1(l_x) = (l_x - 6.5)/5.5$, $\phi_1(l_y) = (l_y - 3.5)/2.5$ and $\phi_2(\cdot) = 3(\phi_1^2(\cdot) - 1)/2$. A linear representation is obtained if μ_3, μ_4 and μ_5 are set to zero.

An alternative is to use non-parametric regression modelling, for example by extending the spline-based generalized additive modelling approach of Chavez-Demoulin and Davison (2005) to the spatial situation. However, in Chavez-Demoulin and Davison (2005) the increased complexity and computational demands mean that thresholds were chosen in a rather arbitrary way. The extension to the spatial context is non-trivial as smoothing over two dimensions is more demanding of data and an additive representation. Another possibility is the local-likelihood approach used by Butler et al. (2007) to smooth temporal trends over space. The simpler polynomial basis approach used in the current paper is designed to produce a reasonable and computationally tractable approximation to the spatial structure seen in figure 1. The model diagnostics described in section 3.4 suggest that this is adequate for the data considered in this paper.

3.1 Adjusting for spatial dependence

We estimate model parameters using independence estimating equations and a robust estimate of the variance matrix of the resulting estimators to adjust for spatial dependence

(Liang and Zeger, 1986). The 'independence' log-likelihood $l_I(\theta) = \log L_I(\theta)$ is given by

$$l_I(\theta) = \sum_{j=1}^{315} \sum_{i=1}^{72} \log f_{ij}(y_{ij} \mid \boldsymbol{x}_{ij}\ ; \theta) = \sum_{j=1}^{315} l_j(\theta; \boldsymbol{y}_j),$$

where $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{72\,j})'$. Each of the 315 hurricane events constitutes a cluster of 72 observations. We assume that, given $\boldsymbol{x}_{ij}$, data from different clusters (hurricane events) are independent.

The maximum likelihood estimator (MLE) $\widehat{\theta}$ maximizes $l_I(\theta)$, that is, it is the root of

$$U(\theta) = \frac{\partial l_I(\theta)}{\partial \theta} = \sum_{j=1}^{315} \frac{\partial}{\partial \theta} l_j(\theta; \boldsymbol{y}_j) = \sum_{j=1}^{315} U_j(\theta) = 0.$$

Suppose that θ_0 is the true value of θ. We define H_I to be the expected Hessian of $l_I(\theta)$ at θ_0 and V to be the covariance matrix of $U(\theta_0)$. H_I is estimated by the observed Hessian, $\widehat{H}_I$, at $\widehat{\theta}$. V is estimated by $\widehat{V} = \sum_j U_j(\widehat{\theta})$, justified by the independence of the score contributions from different clusters. In regular problems as the number of clusters tends to infinity, in distribution,

$$\widehat{\theta} \to N(\theta_0, H_I^{-1} V H_I^{-1}). \tag{2}$$

To account for dependence within clusters Chandler and Bate (2007) scale $l_I(\theta)$ so that it has Hessian $\widehat{H}_A = (\widehat{H}_I^{-1} \widehat{V} \widehat{H}_I^{-1})^{-1}$ at $\widehat{\theta}$. $l_I(\theta)$ has the appropriate curvature at $\widehat{\theta}$ while retaining the general shape of $l_I(\theta)$. For the extreme value models considered in this paper, (2) holds provided that $\xi(\boldsymbol{x}_{ij}) > -1/2$ for all $\boldsymbol{x}_{ij}$ (Smith, 1994).

In this article we use the vertically-adjusted log-likelihood

$$l_A(\theta) = l_I(\widehat{\theta}) + \frac{(\theta - \widehat{\theta})' \widehat{H}_A (\theta - \widehat{\theta})}{(\theta - \widehat{\theta})' \widehat{H}_I (\theta - \widehat{\theta})} \left(l_I(\theta) - l_I(\widehat{\theta}) \right),$$

proposed in the discussion section of Chandler and Bate (2007). This has the advantage, over a horizontal adjustment, that the adjusted log-likelihood is always defined, even though there are constraints on the parameter space of extreme value models.

The covariance matrix in (2) is estimated by $\widehat{H}_I^{-1}\widehat{V}\widehat{H}_I^{-1}$, from which adjusted standard errors for the parameter estimates can be obtained. Approximate confidence intervals for model parameters can be obtained by profiling $l_A(\theta)$, if a symmetric interval based on (2) is not appropriate. Nested models, with a difference in dimensionality q, can be compared using a χ^2_q distribution for the adjusted likelihood ratio statistic (ALRS) $\Lambda_A = 2\left\{l_A(\widehat{\theta}) - l_A(\tilde{\theta})\right\}$, where $\tilde{\theta}$ maximizes $l_A(\theta)$ subject to the constraint imposed on the unrestricted model to obtain the restricted model.

Previous approaches (Smith, 1990; Fawcett and Walshaw, 2007) to adjusting inferences from extreme value models for dependence modify the null distribution of $l_I(\theta)$. In contrast, $l_A(\theta)$ is defined in order to preserve the usual asymptotic distribution of the likelihood ratio statistic. An advantage of $l_A(\theta)$ is that it is multi-dimensional and, in vector-parameter situations, can be expected to perform better than adjusting the null distribution of $l_I(\theta)$. In general neither method will recover the full log-likelihood. However, Chandler and Bate (2007) and other authors have found that adjustments of this type work well in a wide range of problems unless the number of clusters is small. For the data in this paper we have a large number (315) of clusters.

3.2 Modelling of threshold exceedances

In the presence of covariates, standard methods to choose a threshold depend on which covariates are in the model for threshold exceedances. To inform this process we define, at each site, approximate annual maxima as the largest value of successive triplets of storm

peaks (there are on average 3 hurricane events per year) and fit GEV regression models with covariates of Legendre polynomials of longitude and latitude. We find that a model in which μ is quadratic in longitude and latitude, but σ and ξ are constant, is suggested.

3.2.1 Quantile regression

We wish to set a threshold for which the probability of exceedance is constant across different values of covariates. Quantile regression (Koenker and Bassett, 1978) is used to quantify how a given conditional quantile (or quantiles) of a response variable Y depends on the observed values of covariates. Let y^τ denote the conditional τ quantile of Y, satisfying $P(Y \leqslant y^\tau) = \tau$. In the current example, we assume that

$$y^\tau = \beta_0 + \beta_1\,\phi_1(l_x) + \beta_2\,\phi_1(l_y) + \beta_3\,\phi_2(l_x) + \beta_4\,\phi_1(l_x)\,\phi_1(l_y) + \beta_5\,\phi_2(l_y), \tag{3}$$

that is, we assume that y^τ is quadratic in longitude and latitude. Suppose that we have responses $y_1, \ldots, y_n$ with associated predictors $y_1^\tau, \ldots, y_n^\tau$. The regression parameters are estimated by minimizing

$$\min_{\boldsymbol{\beta}} \left\{ (1-\tau) \sum_{y_i < y_i^\tau} (y_i^\tau - y_i) + \tau \sum_{y_i > y_i^\tau} (y_i - y_i^\tau) \right\},$$

with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$. We use the R package quantreg (Koenker, 2009) to estimate $\boldsymbol{\beta}$. The presence of spatial dependence complicates model selection and estimation of parameter uncertainty. This is the subject of current research (see, for example Hallin et al. (2010)). As we will see in section 3.2.2, if the probability of threshold exceedance is constant, the form of the point process model fitted implies a particular form for the quantile regression model used to set the threshold. Therefore, for a given point process model, we know which quantile regression model to fit, but

subsequent inferences neglect the uncertainty due to the estimation of $\boldsymbol{\beta}$.

3.2.2 Parameterization of covariate effects

The assumption that the conditional probability of threshold exceedance is constant has consequences for the parameterization of the effects of the covariates on the extreme value parameters. Let $Y(\boldsymbol{x}_{ij})$ denote the response, and $u(\boldsymbol{x}_{ij})$ the threshold applied, for covariate $\boldsymbol{x}_{ij}$. The probability of exceedance $p(\boldsymbol{x}_{ij}) = P(Y(\boldsymbol{x}_{ij}) > u(\boldsymbol{x}_{ij}))$ for covariate $\boldsymbol{x}_{ij}$ is related approximately to the GEV parameters via

$$p(\boldsymbol{x}_{ij}) \approx \frac{1}{\lambda}\left[1 + \xi(\boldsymbol{x}_{ij})\left(\frac{u(\boldsymbol{x}_{ij}) - \mu(\boldsymbol{x}_{ij})}{\sigma(\boldsymbol{x}_{ij})}\right)\right]^{-1/\xi(\boldsymbol{x}_{ij})}. \tag{4}$$

Suppose that $\xi(\boldsymbol{x}_{ij}) = \xi$ is constant. If we use quantile regression to set $u(\boldsymbol{x}_{ij})$ so that $p(\boldsymbol{x}_{ij}) = p$ is constant then, inverting (4),

$$u(\boldsymbol{x}_{ij}) = \mu(\boldsymbol{x}_{ij}) + c\,\sigma(\boldsymbol{x}_{ij}), \tag{5}$$

where $c = [(\lambda\, p)^{-\xi} - 1]/\xi$.

Suppose that we set $u(\boldsymbol{x}_{ij})$ using a linear quantile regression on $\boldsymbol{x}_{ij}$. To have the correct functional form on the right hand side of (5), $\mu(\boldsymbol{x}_{ij})$ and $\sigma(\boldsymbol{x}_{ij})$ must be linear in $\boldsymbol{x}_{ij}$, e.g. if $\mu(\boldsymbol{x}_{ij})$ and/or $\sigma(\boldsymbol{x}_{ij})$ are quadratic in longitude and latitude then quantile regression model (3) is indicated. The constraint $\sigma(\boldsymbol{x}_{ij}) > 0$ is imposed at each covariate value. If we take $\log \sigma(\boldsymbol{x}_{ij})$ to be linear in $\boldsymbol{x}_{ij}$, to constrain functionally $\sigma(\boldsymbol{x}_{ij})$ to be positive, we should use non-linear quantile regression (Koenker and Park, 1994) to set $u(\boldsymbol{x}_{ij})$.

3.2.3 Threshold selection

Since the threshold $u(\boldsymbol{x}_{ij})$ is set to achieve constant exceedance probability p, choosing the level of the threshold amounts to choosing p. We choose the largest value of p above which, taking into account the uncertainty in the estimates summarized by 95% confidence intervals, the estimates appear approximately stable. We use the point process model suggested by the preliminary GEV analysis in section 3: $\mu(\boldsymbol{x}_{ij})$ is quadratic in longitude and latitude and σ and ξ are constant. Having identified a suitable threshold p, we then revisit this choice of exceedance model.

Figure 4, for different thresholds p, shows MLEs for this initial point process model. Inevitably there is some subjectivity in the choice of p. Here we choose $p = 0.4$. This is a relatively large value but we should bear in mind that the raw data contain only the largest values from each hurricane event. Indeed Jonathan and Ewans (2007) use a constant threshold of 2.5m for these data, which equates to an exceedance probability of 0.54.

If $\mu(\boldsymbol{x}_{ij})$ is quadratic in $\boldsymbol{x}_{ij}$ then we should set the threshold using quantile regression model (3), using $\tau = 0.6$. The resulting threshold

$$\widehat{u}_{QR} = 3.310 + 0.106\,\phi_1(l_x) - 0.153\,\phi_1(l_y) - 0.081\,\phi_2(l_x) + 0.098\,\phi_1(l_x)\,\phi_1(l_y) - 0.025\,\phi_2(l_y), \quad (6)$$

varies between 2.85m and 3.48m and produces 9075 threshold exceedances out of 22680 observations. The number of exceedances per site varies between 123 and 130, and shows no obvious systematic variation with longitude and latitude.

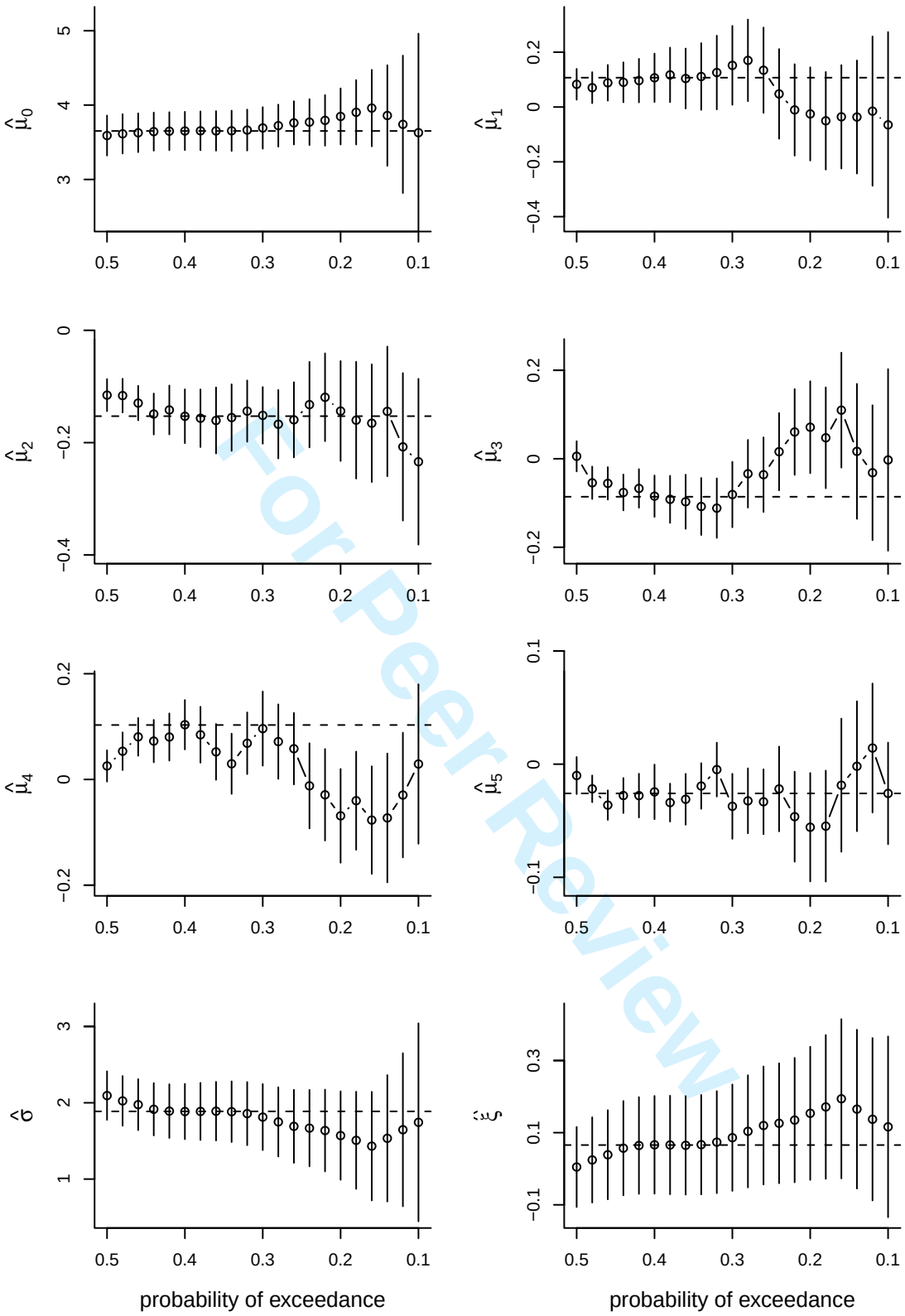

Figure 4: MLEs against probability of exceedance p for a point process model in which μ is quadratic in longitude and latitude. The vertical lines indicate approximate (symmetric) 95% confidence intervals, based on adjusted standard errors. The threshold, also quadratic in longitude and latitude, is estimated using quantile regression with exceedance probability p. The horizontal dashed lines indicate the MLEs for $p = 0.4$.

3.3 Modelling spatial variation

We consider exceedance models in which the location parameter $\mu(\boldsymbol{x}_{ij})$ is modelled as a Legendre polynomial function of longitude and latitude and the scale σ and shape parameters are constant. The results in table 1 suggest that we use a model in which $\mu(\boldsymbol{x}_{ij})$ is quadratic in longitude and latitude. The effect of adjusting the independence log-likelihood for spatial dependence is to make the regression effects in μ more statistically significant. For example, when comparing the quadratic model to the linear model, the unadjusted LR statistic is 11.00, compared to the ALRS of 20.50. Extension to include a quadratic form for $\sigma(\boldsymbol{x}_{ij})$ has some supported from an ALRS (p-value 0.10), but the resulting fits made little difference to the diagnostic plots presented in section 3.4. Therefore, we proceed with the model in which $\sigma(\boldsymbol{x}_{ij})$ is constant. The estimates and adjusted standard errors are given in table 2. The adjusted standard errors for $\widehat{\mu}_0, \widehat{\sigma}$ and $\widehat{\xi}$ are much larger than their unadjusted versions of 0.022, 0.031 and 0.011 respectively. We found that the additional complexity of allowing the shape parameter ξ to vary spatially seems not to be warranted (for example, extension to include a quadratic form for $\xi(\boldsymbol{x}_{ij})$ produced an ALRS with an associated p-value of 0.30). We have checked that these conclusions are not sensitive to the choice of exceedance probability. An approximate 95% confidence interval for ξ, calculated by profiling the adjusted log-likelihood, is $(-0.052, 0.223)$. Jonathan and Ewans (2007) obtained a point estimate of ξ of -0.098 (compared to 0.066 in the current analysis) and a bootstrap 95% confidence interval of $(-0.164, 0.015)$ using the same raw data. The latter confidence interval is narrower because the threshold is set lower than in the current analysis and because such intervals tend to be wider when $\widehat{\xi}$ is positive. In addition Jonathan and Ewans (2007) use a constant threshold and do not model spatial variation.

3.4 Model checking

We check three aspects of the model fit: the compatibility of the threshold with the fitted threshold exceedance model, where threshold exceedances occur and the values of threshold exceedances.

In section 3.2.2 we observed that if the threshold $u(\boldsymbol{x}_{ij})$ is set so that the probability p of exceedance is constant for all observations then $u(\boldsymbol{x}_{ij}) = \mu(\boldsymbol{x}_{ij}) + c\,\sigma$, where $c = [(\lambda\,p)^{-\xi} - 1]/\xi$. Comparing equation (6) and table 2, and noting that $\widehat{\mu}_0 + \widehat{c}\widehat{\sigma} = 3.310$, shows that the threshold $\widehat{u}_{PP}$ implied by the fitted point process model is approximately equal to the threshold $\widehat{u}_{QR}$ estimated using quantile regression.

We treat the threshold $\widehat{u}_{QR}$ as fixed and examine whether $\widehat{u}_{PP}$ is significantly different from $\widehat{u}_{QR}$. An adjusted likelihood ratio test of $\mu_0 + c\,\sigma = \widehat{\beta}_0, \mu_i = \widehat{\beta}_i, i = 1, \ldots, 5$ produces an ALRS of 0.07 on 6 d.f. and a p-value of 0.99999. This suggests that the threshold $\widehat{u}_{QR}$ is compatible with the inferences made about the parameters of the point process model, based on this threshold. This comparison of $\widehat{u}_{PP}$ to $\widehat{u}_{QR}$ could be used to select the level of the threshold.

At each site, the numbers of events between successive threshold exceedances should be geometrically distributed with mean 1/0.4. Performing Pearson chi-squared goodness-of-fit tests at each of the sites results in p-values that vary from 0.18 to 0.87. Due to the strong dependence between the time series at the sites, we expect these p-values to vary less than a random sample from a standard uniform distribution. The p-values show no obvious systematic variation with longitude and latitude.

Figure 5 shows how the number of exceedances increases with event number n_s for site 1 (longitude=1, latitude=1), with 95% envelopes based on a binomial$(n_s, 0.4)$ distribution. Due to a series of large events around event number 50, evident in figure 2, a few points

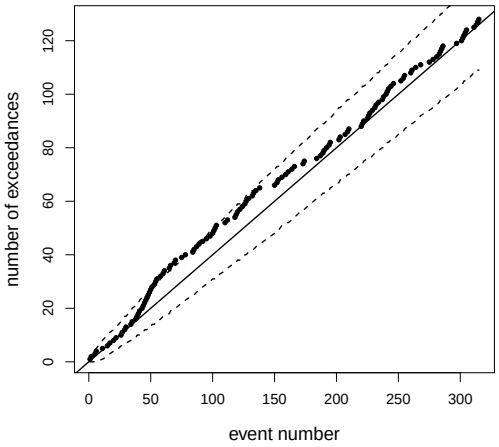

Figure 5: Number of exceedances against event number n_s for site 1. The solid line gives the expected number ($0.4n_s$) of exceedances under the fitted model. The dashed lines give the 97.5% and 2.5% quantiles of the binomial($n_s, 0.4$) distribution.

lie above the envelope. This may be suggestive a local periods of increased storminess (Killick et al., 2010) not captured by the model. Otherwise, this plot does not reveal any clear time trend. The plots for the other sites are very similar.

Figure 6 shows QQ-plots from six of the sites. The 95% envelopes have been adjusted for the uncertainty in estimating the GEV parameters, using simulation from the estimated normal distribution in (2). The strong spatial dependence means that neighbouring sites have very similar QQ plots. None of the QQ-plots reveal any clear lack-of-fit.

4 Return level estimation

The solid lines in figure 7 show how the estimates, and 95% confidence intervals, of the conditional 105 year return level q_{105} vary with site longitude and latitude under the fitted regression model. To provide an additional model diagnostic we have superimposed the largest of the raw data values. Since the data span 105 years we expect to see approximately 72 raw values exceed the estimates of q_{105}: we observe 113. Strong spatial dependence means that the distribution of this statistic is more widely dispersed than un-

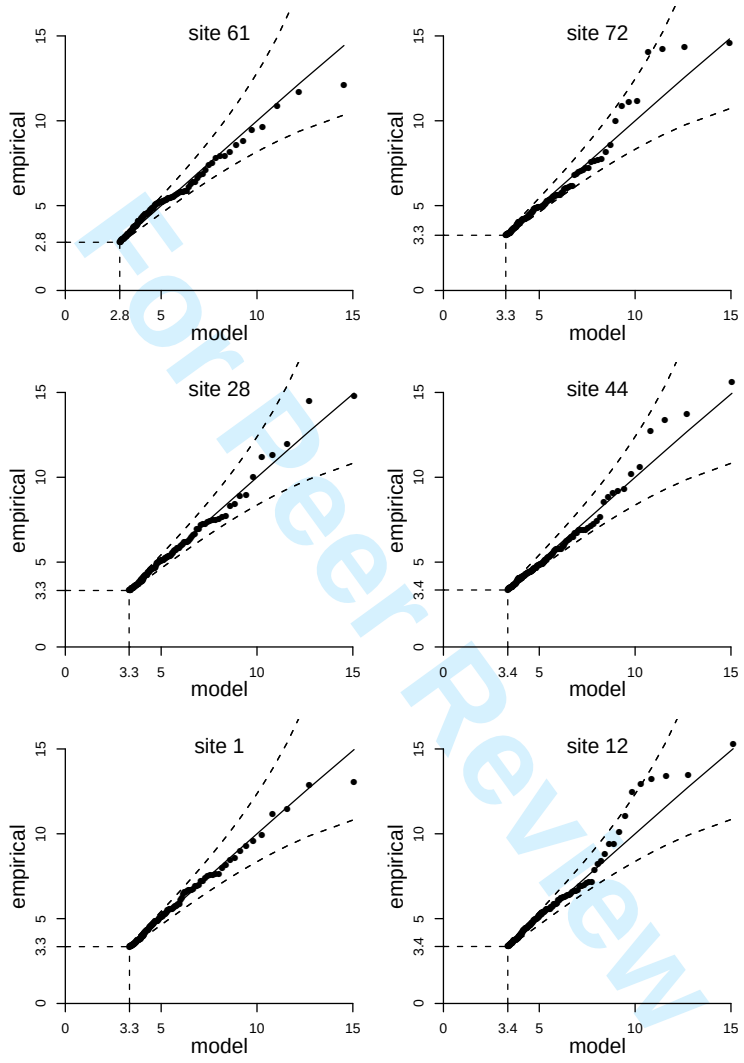

Figure 6: GP QQ-plots at 'corner' sites 1, 12, 61 and 72, and two sites, 28 and 44, nearer the centre of the region. The dashed curves give the 2.5% and 97.5% quantiles of the fitted distributions of the order statistics. The threshold is shown using dashed lines.

der independence, so 113 exceedances does not seem excessive. The dashed lines in figure 7 give the corresponding values obtained by fitted the point process model individually at each of the sites, using the same threshold values set for the regression model. The uncertainty in estimating the threshold, and setting its general level, is neglected. Figure 7 shows the main advantage of pooling information over sites: typically the confidence intervals are narrower under the regression model than the at-site intervals. Exceptions to this occur where the at-site estimate of ξ is negative, because smaller estimates of ξ tend to produce narrower confidence intervals for large return levels.

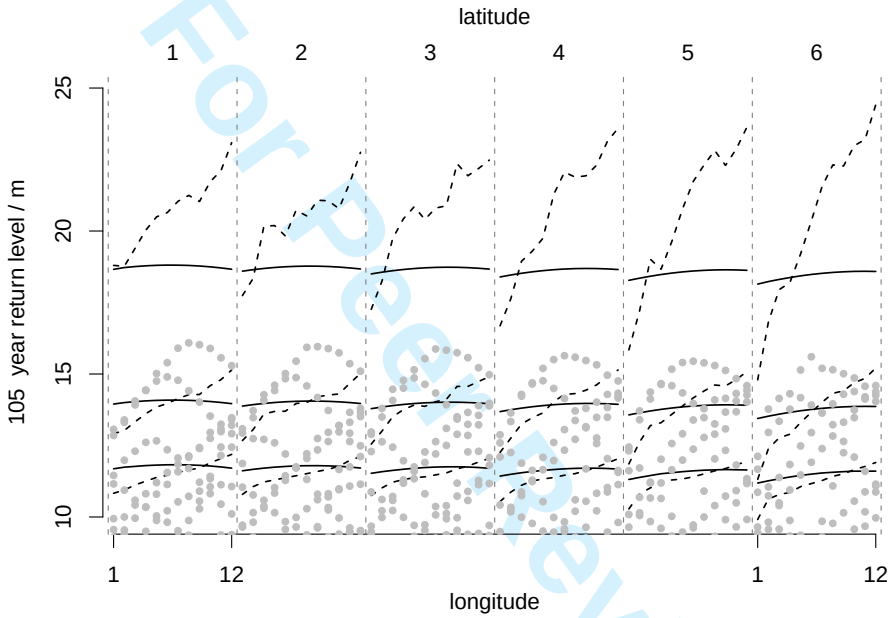

Figure 7: Point estimates and 95% confidence intervals for 105 year return levels, by longitude and latitude, based on the regression model (solid lines) and at-site estimates (dashed lines). The 95% confidence intervals are produced by profiling the (adjusted) log-likelihood. The grey dots are largest of the raw data values.

5 Simulation study

The wave height dataset considered in this paper is atypical because sequences of independent storm maxima have been formed by 'declustering' sequences of temporally dependent values. To illustrate the applicability of the methodology described in this paper we carry out a simulation study in which the response data are dependent in space *and* time.

The spatial dependence is based on a Schlather max-stable process (Schlather, 2002), with Cauchy correlation structure $\rho(h) = \left[1 + (h/24.90)^2\right]^{-0.67}$, estimated from the wave height data. We simulate spatially-dependent unit Fréchet variates from this fitted process. We simulate 30 years of daily values at each site on a rectangular 4 by 4 grid that spans the 12 by 6 grid of the wave height dataset. Model fitting and simulation are performed using the SpatialExtremes R package (Ribatet, 2010). At each site we induce temporal dependence using a moving maxima process $Y_t = \max\left(\frac{1}{2}U_t, \frac{1}{3}U_{t-1}, \frac{1}{6}U_{t-2}\right)$, where $U_1, U_2, \ldots$ are the unit Fréchet variates at a given site. The series $\{Y_t\}$ has extremal index 1/2. Then the $\{Y_t\}$ series are transformed to have GEV$(\mu(l_x, l_y), \sigma^d, \xi)$ margins, where, using the notation of section 3, $\mu(l_x, l_y)$ depends linearly on the location (l_x, l_y) of the site via $\mu(l_x, l_y) = \mu_0^d + \mu_1\,\phi_1(l_x) + \mu_2\,\phi_1(l_y)$.

The daily GEV parameters μ_0^d and σ^d are held constant (at 0 and 1 respectively) across all simulations as is the form of the temporal dependence at each site. Otherwise, we consider two levels of spatial dependence: strong (based on wave height data) and spatial independence; 2 types of covariate effects: $\mu_1 = 2$ and $\mu_2 = -3$; $\mu_1 = \mu_2 = 0$ (i.e. no covariate effects); 3 levels of threshold: exceedance probability $p = 0.1, 0.05$ and 0.01; and 4 values for the GEV shape parameter: $\xi = -0.2, 0.1, 0.4$ and 0.7. The 'no covariate effects' simulations are included to provide a comparison between using quantile regression to set a threshold for non-stationary data, and setting a constant threshold for stationary data. The 'spatial independence' simulations are included because the covariate effects in the quantile regression will be less precisely estimated than in the presence of (positive) spatial dependence.

To each set of simulated data we fit the point process model (1) with the same covariate effects (if any) for μ as the model from which the data were simulated. A threshold of the form implied by (5), i.e. linear in $\phi_1(l_x)$ and $\phi_1(l_y)$ if there are covariate effects in μ and

constant otherwise, is set using quantile regression with the probability of exceedance set at p. The standard errors of the GEV parameters are adjusted for spatial and temporal dependence by assuming that data from distinct years are independent, using the methods of section 3.1.

5.1 Results

Table 3 summarizes the results (based on 1000 simulated datasets) for $\xi = 0.1$ for datasets with spatial dependence and $\mu(l_x, l_y)$ linear in l_x and l_y. Similar results were obtained for the other values of ξ in the study. The parameters in table 3 relate to annual maxima, that is, adjustment of the location μ_0 and scale σ has been made for the extremal index of 1/2.

The regression parameters μ_1 and μ_2 are very precisely estimated, with very small bias. The large precision is the result of the strong spatial dependence in the simulated data. The negative bias observed for $\widehat{\xi}$ has been found in similar studies, see, for example Fawcett and Walshaw (2007). As we would hope the 95% confidence intervals contain comfortably the true values.

An important comparison is between the sample standard deviation of the estimates of a parameter (st.dev.) and the mean of the adjusted standard errors (adj.se.): if the standard errors are appropriately estimated st.dev. and adj.se should be similar. We find that, on average, the standard errors are slightly underestimated because the uncertainty in estimating the three parameters of the threshold has been ignored.

There is uncertainty associated with threshold selection even when setting a constant threshold. The results (not shown) obtained when no covariates are present and a constant threshold is set differ very little from those in table 3: the corresponding values of

st.dev. and adj.se differ by at most one digit in their second significant digit. This is also true when the comparison is made for the simulations in which there is no dependence in space. Thus the predominant source of ignored threshold uncertainty is the general level of the threshold rather than its regression parameters. For each set of simulated data, the quantile regression estimates $(\widehat{\beta}_1, \widehat{\beta}_2)$ are remarkably close to corresponding estimates $(\widehat{\mu}_1, \widehat{\mu}_2)$. This is the case for both the spatially dependent and independent simulations. This close agreement occurs because, using the same set of data, both the quantile regression used to set the threshold and maximization of the likelihood (1) estimate how extreme quantiles vary with covariates. Close agreement between $(\widehat{\beta}_1, \widehat{\beta}_2)$ and $(\widehat{\mu}_1, \widehat{\mu}_2)$ means that uncertainty in $(\widehat{\beta}_1, \widehat{\beta}_2)$ is, to a large extent, accounted for because uncertainty in $(\widehat{\mu}_1, \widehat{\mu}_2)$ is quantified.

6 Discussion

We have proposed the use of quantile regression to set a threshold for extreme value regression models. The quantile regression model is determined by the extreme value model to be fitted subsequently, using the arguments in section 3.2.2. In the current example extreme value parameters are modelled as a smooth function of space and therefore it makes sense to approach the setting of a threshold in the same way, rather than setting a threshold separately at each location. Recently, Kyselý et al. (2010) used quantile regression to set a time-dependent threshold for GP modelling of declustered daily temperatures.

On-going theoretical work will investigate the efficiency of quantile regression as a threshold selection strategy. For example, suppose that for given x responses are sampled independently from a GEV$(\mu_0 + \mu_1 x, \sigma\, \xi)$ distribution and we set a threshold of the form $u_0 + u_1 x$. Standard likelihood calculations show that, if the values of x are symmetri-

cally distributed, the asymptotic variance of the maximum likelihood estimator of μ_1 is minimized for $u_1 = \mu_1$. If the x data are positively/negatively skewed then a slightly smaller/larger value of u_1 is indicated. However, if $u_1 = \mu_1$ is used the loss in efficiency is small even for highly skewed x data.

For the hindcast data we have checked that there are no significant time trends in the location parameter μ of the model. We have also checked that for these data the Poisson approximation to the temporal arrival process of exceedances underlying the likelihood (1) is appropriate despite the large exceedance probability of 0.4.

Acknowledgements

The authors acknowledge discussions with Kevin Ewans (Shell) and Vince Cardone (Oceanweather) and thank Richard Chandler (UCL) for his extremely helpful comments on a first draft of this paper. We thank two anonymous referees whose comments and suggestions have resulted in improvements to the original paper.

References

Abramowitz, M. and I. Stegun (1965). *Handbook of Mathematical Functions: With Formulas, Graphs and Mathematical Tables*. New York: Dover.

Butler, A., J. E. Heffernan, J. A. Tawn, and R. A. Flather (2007). Trend estimation in extremes of synthetic North Sea surges. *Appl. Statist. 56*(4), 395–414.

Casson, E. and S. Coles (1999). Spatial regression models for extremes. *Extremes 1*, 449–468.

Chandler, R. E. (2005). On the use of generalized linear models for interpreting climate variability. *Environmetrics 16*, 699–715.

Chandler, R. E. and S. B. Bate (2007). Inference for clustered data using the independence loglikelihood. *Biometrika 94*(1), 167–183.

Chavez-Demoulin, V. and A. C. Davison (2005). Generalized additive modelling of sample extremes. *Appl. Statist. 54*(1), 207–222.

Coles, S. G. (2001). *An introduction to statistical modeling of extreme values*. London: Springer.

Cooley, D., D. Nychka, and P. Naveau (2007). Bayesian spatial modelling of extreme precipitation return levels. *J. Am. Stat. Assoc. 102*(479), 824–840.

Davison, A. C. and R. L. Smith (1990). Models for exceedances over high thresholds (with discussion). *J. R. Statist. Soc. B52*, 393–442.

Eastoe, E. F. and J. A. Tawn (2009). Modelling non-stationary extremes with application to surface level ozone. *Appl. Statist. 58*, 25–45.

Fawcett, L. and D. Walshaw (2007). Improved estimation for temporally clustered extremes. *Environmetrics 18*, 173–188.

Hallin, M., Z. Lu, and K. Yu (2010). Local linear spatial quantile regression. *Bernoulli*.

Heffernan, J. E. and J. A. Tawn (2004). A conditional approach for multivaiate extremes. *J. Roy. Statist. Soc. B, 66*, 497–546.

Jonathan, P. and K. C. Ewans (2007). Uncertainties in wave height estimates for hurricane-dominated regions. *J. Offshore Mechanics and Arctic Engineering 129*(4), 300–305.

Jonathan, P. and K. C. Ewans (2011). A spatiodirectional model for extreme waves in the Gulf of Mexico. *J. Offshore Mechanics and Arctic Engineering 133*(1).

Jonathan, P., K. C. Ewans, and G. Z. Forristall (2008). Statistical estimation of extreme ocean environments: the requirement for modelling directionality and other covariate effects. *Ocean Engineering 35*, 1211–1225.

Killick, R., I. A. Eckley, K. Ewans, and P. Jonathan (2010). Detection of changes in the characteristics of oceanographic time-series using changepoint analysis. *Ocean Engineering 37*(13), 1120–1126.

Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.

Koenker, R. (2009). *quantreg: Quantile Regression*. R package version 4.44.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Koenker, R. and B. Park (1994). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics 71*(1–2), 265–283.

Kyselý, J., J. Picek, and R. Beranová (2010). Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold. *Global and Planetary Change 72*, 55–68.

Leadbetter, M., G. Lindgren, and H. Rootzén (1983). *Extremes and related properties of random sequences and series*. New York: Springer Verlag.

Liang, K.-Y. and S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Naveau, P., A. Guillou, D. Cooley, and J. Diebolt (2009). Modelling pairwise dependence of maxima in space. *Biometrika 96*(1), 1–17.

Oceanweather Inc. (2005). GOMOS – USA Gulf of Mexico Oceanographic Study, Northern Gulf of Mexico Archive, Oct.

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *J. App. Prob. 8*, 745–756.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Stat. 3*, 119–131.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ribatet, M. (2010). *SpatialExtremes: Modelling Spatial Extremes*. R package version 1.5-1.

Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes 5*(1), 33–44.

Smith, R. L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science 4*, 367–377.

Smith, R. L. (1990). Regional estimation from spatially dependent data. (Preprint. http://www.stat.unc.edu/postscript/rs/regest.pdf).

Smith, R. L. (1994). Nonregular regression. *Biometrika 81*, 173–183.

Wadsworth, J. L., J. A. Tawn, and P. Jonathan (2010). Accounting for choice of measurements scale in extreme value modeling. *The Annals of Applied Statistics 4*(3), 1558–1578.

$\mu(\boldsymbol{x}_{ij})$	neg. log-lik	d.f.	ALRS	p-value
constant	22763.20			
linear	22742.59	2	34.23	3.7×10^{-8}
quadratic	22737.09	3	20.50	1.3×10^{-4}
cubic	22737.02	4	2.09	0.72

Table 1: Summary of point process modelling in which the location parameter μ is modelled as a Legendre polynomial function of longitude and latitude and σ and ξ are constant. The likelihood ratio tests compare the model with the model in the row above. d.f. = degrees of freedom; neg. log-lik = negated maximised log-likelihood.

	$\widehat{\mu}_0$	$\widehat{\mu}_1$	$\widehat{\mu}_2$	$\widehat{\mu}_3$	$\widehat{\mu}_4$	$\widehat{\mu}_5$	$\widehat{\sigma}$	$\widehat{\xi}$
MLE	3.652	0.107	−0.153	−0.085	0.104	−0.024	1.885	0.066
adjusted s.e.	(0.132)	(0.045)	(0.025)	(0.024)	(0.024)	(0.012)	(0.185)	(0.069)

Table 2: Maximum likelihood estimates and adjusted standard errors for a point process model in which μ is quadratic in longitude and latitude and σ and ξ are constant. The threshold is quadratic in longitude and latitude.

true value	μ_0=6.8	μ_1=2.0	μ_2=−3.0	σ=1.7	ξ=0.10
p=**0.1**					
bias	−0.0070	0.00096	−0.00025	−0.0082	−0.017
95% CI	(6.3, 7.4)	(2.0, 2.0)	(−3.0, −3.0)	(1.3, 2.1)	(−0.019, 0.18)
st. dev.	0.29	0.016	0.0085	0.20	0.050
adj. se	0.27	0.016	0.0081	0.18	0.044
95% CI	(0.18, 0.38)	(0.012, 0.020)	(0.0060, 0.011)	(0.10, 0.29)	(0.032, 0.061)
p=**0.05**					
bias	−0.012	0.000067	−0.00042	−0.0074	−0.015
95% CI	(6.3, 7.4)	(2.0, 2.0)	(−3.0, −3.0)	(1.3, 2.1)	(−0.059, 0.22)
st. dev.	0.29	0.024	0.013	0.22	0.070
adj. se	0.27	0.024	0.013	0.19	0.060
95% CI	(0.18, 0.38)	(0.017, 0.032)	(0.0091, 0.017)	(0.097, 0.31)	(0.041, 0.086)
p=**0.01**					
bias	0.028	0.00010	−0.0016	0.031	−0.065
95% CI	(6.3, 7.5)	(1.9, 2.1)	(−3.1, −2.9)	(1.3, 2.3)	(−0.25, 0.36)
st. dev.	0.33	0.065	0.034	0.26	0.16
adj. se	0.32	0.065	0.034	0.25	0.11
95% CI	(0.20, 0.49)	(0.040, 0.097)	(0.020, 0.054)	(0.14, 0.40)	(0.063, 0.18)

Table 3: Simulation study results for $\xi = 0.1$, spatial dependence and covariates in μ. For each p: rows 1 – estimated bias; row 2 – 2.5% and 97.5% sample quantiles of the parameter estimates; row 3 – sample standard deviation of the estimates; row 4 – mean of the adjusted standard errors; row 5 – 2.5% and 97.5% sample quantiles of the adjusted standard errors.