Paper:

Cai, Y. & Reeve, D. (2013). Extreme value prediction via a quantile function model. *Coastal Engineering, 77*, 91-98.

http://dx.doi.org/10.1016/j.coastaleng.2013.02.003

Coastal Engineering 77 (2013) 91–98

Contents lists available at SciVerse ScienceDirect

Coastal Engineering

journal homepage: www.elsevier.com/locate/coastaleng

Extreme value prediction via a quantile function model

Yuzhi Cai [*], Dominic E. Reeve

Swansea University, United Kingdom

ARTICLE INFO

Article history:
Received 25 September 2012
Received in revised form 11 February 2013
Accepted 15 February 2013
Available online 21 March 2013

Keywords:
Extreme values
Quantile function models
Bayesian approach
Semi-parametric model
Parametric model
Sea-level

ABSTRACT

Methods for estimating extreme loads are used in design as well as risk assessment. Regression using maximum likelihood or least squares estimation is widely used in a univariate analysis but these methods favour solutions that fit observations in an average sense. Here we describe a new technique for estimating extremes using a quantile function model. A quantile of a distribution is most commonly termed a 'return level' in flood risk analysis. The quantile function of a random variable is the inverse function of its distribution function. Quantile function models are different from the conventional regression models, because a quantile function model estimates the quantiles of a variable conditional on some other variables, while a regression model studies the conditional mean of a variable. So quantile function models allow us to study the whole conditional distribution of a variable via its quantile function, whereas conventional regression models represent the average behaviour of a variable. Little work can be found in the literature about prediction from a quantile function model. This paper proposes a prediction method for quantile function models. We also compare different types of statistical models using sea level observations from Venice. Our study shows that quantile function models can be used to estimate directly the relationships between sea condition variables, and also to predict critical quantiles of a sea condition variable conditional on others. Our results show that the proposed quantile function model and the developed prediction method have the potential to be very useful in practice.

1. Introduction

River and coastal flooding is an acknowledged natural hazard. There are many national examples that can be quoted and that are described in the academic and wider engineering literature. Sea level rise, changes in local climate and development in flood plains have altered the hazard and generally increased the consequences of flooding, and thereby the overall risk. An improved ability to predict, quantify and manage flood probabilities is essential to protecting the public, property and infrastructure, and to maintaining a sustainable economy. More accurate predictions of extreme conditions will improve the assessment, mitigation and management of flood risk.

Flood risk may be measured by the combination of the failure probability of a system such as a sea defence system and a measure of the consequences of the resulting floods. This is commonly represented as the probability of occurrence of an extreme event (or 'hazard') multiplied by the cost of the ensuing damage and hence 'risk' is measured as a rate of expenditure. There are several commonly used approaches to studying the probability of the functional failure of flood defences. That is, the situation in which the natural conditions exceed the severity for which a structure was designed to withstand; leading to the unwanted transmission of water across the line of the defence whilst not necessarily resulting in damage to the structure itself. One of these approaches is based on the concept of a structure variable, which is a known function of flood risk variables, such as sea condition variables. In particular, the structure variable method reduces a multivariate problem to a univariate one by using a formula relating to the structural performance, (for example, the wave overtopping rate), to combine potentially correlated variables such as water level, wave height and wave direction into a single variable which may then be analysed using univariate techniques. The occurrence of failure is then defined in terms of the structure variable exceeding a specified threshold rather than the probability of occurrence of particular combinations of the primary variables. See for example, the studies of Coles and Tawn (1994) and Reeve (1998). Another approach is known as the joint probability method and provides a procedure for estimating the probability that a structure variable exceeds a critical level based on the joint analysis of all flood risk variables. Many papers and reports on the application of these methods can be found in the literature. See, for example, the studies of Owen et al. (1997), Hawkes et al. (2002, 2004) and Meadowcroft et al. (2004).

However, in using the standard methods of fitting distributions to observations, the above two approaches are mainly influenced by observations in the bulk of the distribution which may have little influence on the form of the tail, the most important part of the distribution for estimating failure, and hence may provide poor fits. Methods based on scaled error norms, (e.g., Li et al., 2008; Reeve, 1996), address this to some degree but can be difficult to 'tune' precisely. A further approach

[*] Corresponding author at: College of Business, Economics and Law, Swansea University, Swansea, SA2 8PP, United Kingdom. Tel.: +44 1792 606865.
E-mail address: y.cai@swansea.ac.uk (Y. Cai).

0378-3839 © 2013 Elsevier B.V. Open access under CC BY license.
http://dx.doi.org/10.1016/j.coastaleng.2013.02.003

is to estimate probability distributions by using only extreme values of the flood risk variables, see for example, the studies of Coles (2001), Coles et al. (1999), Tawn (1990, 1992) and Thompson et al. (2009). Such an approach can seem wasteful of data, particularly to those making the observations. Partly in response to this concern, modifications to traditional approaches that use more observations have been developed, (e.g., Smith, 1986).

A common feature of the above approaches is that once a probability model is established for a sea condition variable, the value of the variable corresponding to a failure probability, i.e., the quantile or the return level, needs to be calculated by inverting the estimated conditional distribution function. If the estimated conditional distribution is not a commonly used distribution, then it can be difficult to invert the distribution exactly, and hence adding another layer of inaccuracy to prediction.

Recently, there has been a surge in interest in quantile methods. These methods allow us to directly estimate the conditional quantile, i.e., the value that a sea condition variable takes with a required probability. Hence these models focus on the quantiles at a level of interest (i.e., return levels) instead of the average value of a sea condition variable, because of which, the quantile approaches to statistical modelling have been used in many areas including economics, finance and medical research. See, for example, the studies of Koenker (2005) and Gilchrist (2000).

Generally speaking, there are two types of quantile approaches: one is semi-parametric (see, for example, Koenker, 2005) and another is parametric (see for example, Gilchrist (2000)). Although some work can be found in the literature on prediction with semi-parametric quantile regression models, see, for example, those of Taylor (2005), Cai (2010c) and Cai et al. (2012) and references therein, to the authors' knowledge, little work on prediction with parametric quantile function models can be found in the literature. Therefore, the main contributions of this paper are: (a) to develop a method for extreme value prediction via a parametric quantile function model, and (b) to show, via a real data set, the differences between various statistical models commonly used in extreme value prediction. The advantages of the quantile approach are that it makes full use of all available data including both extreme and non-extreme observations, it allows us to use many non-standard distributions that may provide an improved fit to observations leading to better predictions, and it provides a natural way of dealing with multivariate problems so that predictions on a flood risk variable conditional on a set of other variables can be made. At this point it is perhaps worth emphasising that here we use the words ‘model’, ‘estimation’ to mean ‘a statistical distribution used to model observations’ and ‘determining the best fit values of the distribution parameters’ respectively.

In Section 2 we first give an outline of the two types of quantile approaches. We then introduce a special type of quantile function model and present the new forecasting method in Section 3. The application to the Venice sea-level data and comparisons with other commonly used models are presented in Section 4. Finally, Section 5 provides some further comments and conclusions.

2. An outline of the quantile approaches

2.1. Semi-parametric quantile regression model

Let Y be a continuous flood risk random variable, such as wave overtopping rate or run-up level, and $\mathbf{x} = (x_1,\ldots,x_p)$ be a vector of p covariates. Given a set of observations $\{y_i,x_{1i},\ldots,x_{pi}\}$ $(i = 1,\ldots,n)$, a semi-parametric quantile regression model (Koenker (2005)) for the τth conditional quantile of Y, denoted by $q^{\tau}_{Y|X}$ is given by

$$q^{\tau}_{Y|X} = h\left(\eta^{\tau}, x_1, \ldots, x_p\right), \tag{1}$$

where h is a known function of the covariate $\mathbf{x}$ and model parameter η^{τ} which depends on τ $(0 \le \tau \le 1)$. So, for example, if $\tau = 0.95$, model (1) gives a 95% conditional quantile of Y. Therefore, for a sequence of τ values, model (1) defines a sequence of conditional quantiles of Y. These quantiles provide an estimate of the conditional quantile function of Y. Note that model (1) does not contain an error term, hence, the whole model is semi-parametric. A simple example of (1) is the linear quantile regression model given by $q^{\tau}_{Y|X} = a^{\tau}_0 + a_{\mathrm{T}}$ ${}^{\tau}x_1 + \cdots + a^{\tau}_p x_p$ with $\eta^{\tau} = (a^{\tau}_0,\ldots,a^{\tau}_p)$.

The model parameters can be estimated by using various methods including the method based on solving the minimization problem $\min_{\eta^{\tau}} \sum_{i=1}^{n} \rho_{\tau}(u_i)$, where $\rho_{\tau}(u_i) = u_i(\tau - I_{[u_i<0]})$, in which $I_{[\cdot]}$ is an indicator function and $u_i = y_i - h(\eta^{\tau},x_{1i},\ldots,x_{pi})$, see for example, those of Koenker and D'Orey (1987, 1994) and Koenker (2005). The model parameters can also be estimated by using a Bayesian method. For example, Yu and Moyeed (2001) proposed a Bayesian approach to quantile regression with independent data; Thompson et al. (2010) proposed a Bayesian non-parametric quantile regression method with applications to sea condition variables in coastal engineering; Cai and Stander (2008) studied a Bayesian approach to a nonlinear quantile time series model. Lancaster and Jun (2010) investigated the application of Bayesian exponentially tilted empirical likelihood to inference about quantile regressions. This semi-parametric approach has been used widely. However, it can suffer from some problems. For example, the estimated conditional quantile curves may cross over, leading to a failure of the method. Different methods have been proposed to solve the crossing-over problem. See, for example, the approaches of Bondell et al. (2010) and references therein.

2.2. Parametric quantile function model

Compared with the above semi-parametric approach, much less work can be found in the literature about parametric approaches. Gilchrist (2000) gave an excellent introduction to this parametric approach. A general parametric quantile function model is given by

$$Q_Y(\tau|\xi,\mathbf{x}) = h_1\left(\eta_1, x_1, \ldots, x_p\right) + h_2\left(\eta_2, x_1, \ldots, x_p\right)Q(\tau,\gamma), \tag{2}$$

where $\xi = (\eta_1,\eta_2,\gamma)$ is the model parameter vector; h_i $(i = 1, 2)$ are known functions of $\mathbf{x}$ and η_i, $h_2(\eta_2,x_1,\ldots,x_p) > 0$ and $Q(\tau,\gamma)$ are the quantile functions of the error term with explicit mathematical expression. A special case of model (2) is the linear quantile function model given by $Q_Y(\tau|\xi,\mathbf{x}) = a_0 + a_1x_1 + \cdots + a_px_p + Q(\tau,\gamma)$ with $h_1(\eta_1,x_1,\ldots,x_p) = a_0 + a_1x_1 + \cdots + a_px_p$, $h_2(\eta_2,x_1,\ldots,x_p) = 1$ and $\eta_1 = (a_0,\ldots,a_p)$.

It is seen that in the parametric approach, the number of parameters has been significantly decreased as the model parameters do not depend on τ, and the monotonicity of the conditional quantile function of Y can be guaranteed due to the fact that $Q_Y(\tau|\xi,\mathbf{x})$ is a well defined conditional quantile function. Gilchrist (2000) discussed various methods for estimating ξ. Cai (2009, 2010a, 2010b) proposed Bayesian approaches to estimating parameters of different types of model (2), including polynomial, multivariate and time series quantile function models.

In this paper we focus on a special type of quantile function models with details given in the next section.

3. The prediction method

3.1. Polynomial quantile function model

As mentioned above, different choices of h_1, h_2 and Q in model (2) lead to different quantile function models. If there is only one covariate x, then the dependence between Y and x may be modelled by a polynomial quantile function model studied by Cai (2010b), and is given by:

$$Q_Y(\tau|\beta,\mathbf{x}) = \left(a_0 + a_1x + \cdots + a_{k_1}x^{k_1}\right) + \left(b_0 + b_1x + \cdots + b_{k_2}x^{k_2}\right)Q(\tau,\gamma), \tag{3}$$

where a_i, b_j and γ are model parameters. For model (3) to be fully defined, we need to assign a quantile function for $Q(\tau,\gamma)$. It is worth mentioning that different $Q(\tau,\gamma)$ can be used. So in fact model (3) also defines a class of quantile function models.

Since we will use the model to study the Venice sea-level data, since the measurement of the sea levels is positive, and since we are interested in the extreme behaviour of the sea levels, we propose to use the power-Pareto quantile function defined by

$$Q(\tau,\gamma) = \tau^{\gamma_1}(1-\tau)^{-\gamma_2}, \quad \gamma_1 > 0, \;\; \gamma_2 > 0.$$

The choice of quantile function is to some extent arbitrary. However, although this distribution is rarely used in the literature because the corresponding distribution or density function does not have an explicit mathematical form, and hence it is not easy to use the commonly used maximum likelihood estimation method for parameter estimation, this distribution is very flexible and can deal with different shapes and extremes of a sea condition variable. Fig. 1 shows the density function plots of the power-Pareto distribution for four different pairs of γ_1 and γ_2 values.

Combined with the power-Pareto distribution, the model we will consider further in this paper is defined by

$$Q_Y(\tau|\beta,\mathbf{x}) = \left(a_0 + a_1x + \cdots + a_{k_1}x^{k_1}\right) + \left(b_0 + b_1x + \cdots + b_{k_2}x^{k_2}\right)\tau^{\gamma_1}(1-\tau)^{-\gamma_2}. \tag{4}$$

For any particular application we are faced with the problem of defining suitable values of $a_0, a_1, \ldots, a_{k_1},\ b_0, b_1, \ldots, b_{k_2},\ \gamma_1$ and γ_2. Cai (2010b) proposed a Bayesian method for estimating the parameters of model (4). The basic idea of the Bayesian method is given below. First note that the likelihood function of the observed data can be written as

$$L(y_1,\ldots,y_n|\beta,\mathbf{x}) = \prod_{i=1}^{n} \frac{\tau_i^{1-\gamma_1}(1-\tau_i)^{1+\gamma_2}}{\gamma_1(1-\tau_i) + \gamma_2\tau_i},$$

where τ_i satisfy

$$\frac{y_i - \left(a_0 + a_1x_i + \cdots + a_{k_1}x_i^{k_1}\right)}{b_0 + b_1x_i + \cdots + b_{k_2}x_i^{k_2}} = \tau_i^{\gamma_1}(1-\tau_i)^{-\gamma_2}.$$

So the likelihood function is not an explicit function of the observed data directly, but it is an explicit function of τ_i. In fact, this is why it is difficult to use the conventional maximum likelihood method for the parameter estimation. However, by using a Bayesian method, we can deal with the difficulties easily. Specifically, we need to assign a prior distribution to the model parameters. We let the prior

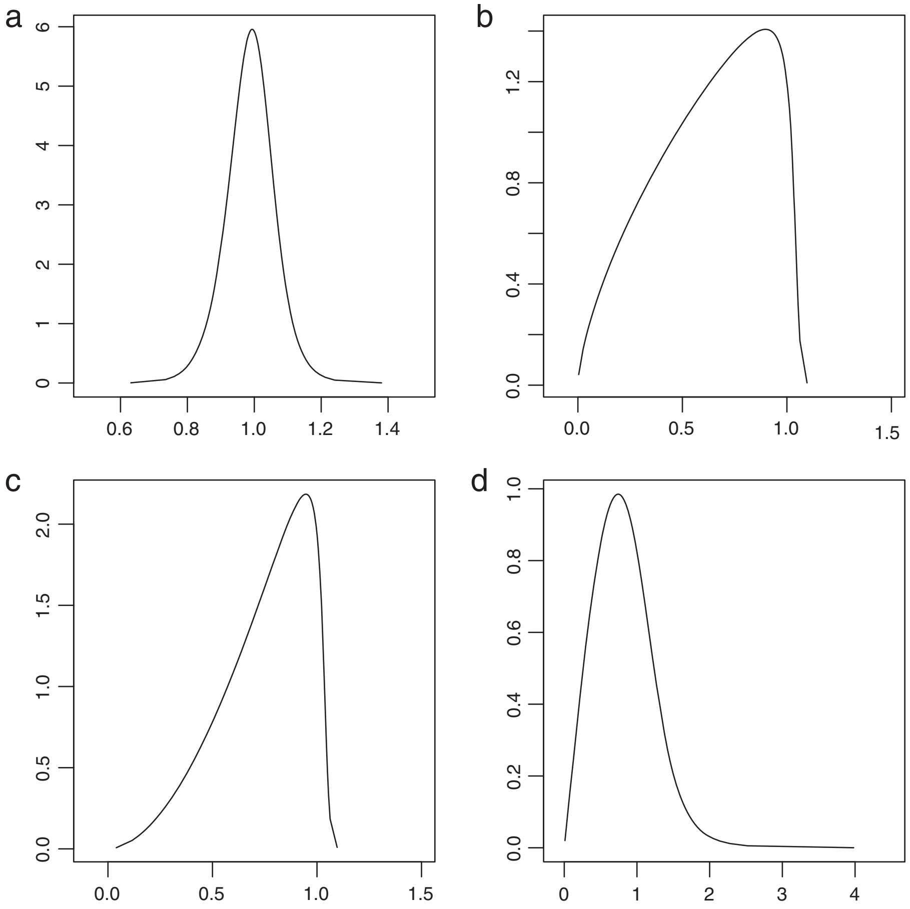

Fig. 1. Density function plots of the power-Pareto distributions for different pairs of γ_1 and γ_2: (a) $\gamma_1 = 0.05, \gamma_2 = 0.035$, (b) $\gamma_1 = 0.6, \gamma_2 = 0.01$, (c) $\gamma_1 = 0.35, \gamma_2 = 0.01$ and (d) $\gamma_1 = 0.5, \gamma_2 = 0.15$.

distribution of the parameters be given by $\pi(\beta) = \pi(\eta_1)\pi(\eta_2)\pi(\gamma)$, where $\eta_1 = (a_0, \ldots, a_{k_1})$, $\eta_2 = (b_0, \ldots, b_{k_2})$, $\gamma = (\gamma_1,\gamma_2)$ and

$$\pi(\eta_1) = \prod_{j_1=0}^{k_1} \pi\left(a_{j_1}\right) = \prod_{j_1=0}^{k_1} \frac{1}{\sqrt{2\pi}\sigma_{j_1}} e^{-\frac{a_{j_1}^2}{2\sigma_{j_1}^2}}, \tag{5}$$
$$\pi(\eta_2) = \prod_{j_2=0}^{k_2} \pi\left(b_{j_2}\right) = \prod_{j_2=0}^{k_2} \frac{1}{\sqrt{2\pi}\sigma_{j_2}} e^{-\frac{b_{j_2}^2}{2\sigma_{j_2}^2}},$$

and

$$\pi(\gamma) = \prod_{\ell=1}^{2} \pi(\gamma_\ell) = \prod_{\ell=1}^{2} \frac{\lambda_\ell}{\gamma_\ell^2} e^{-\lambda_\ell/\gamma_\ell}. \tag{6}$$

Then it can be shown that the posterior distribution of the parameters is given by

$$\pi(\beta|\mathbf{x},\mathbf{y}) \propto \prod_{i=1}^{n} \frac{\tau_i^{1-\gamma_1}(1-\tau_i)^{1+\gamma_2}}{\left(\sum_{j=0}^{k_2} b_j x_i^j\right)[\gamma_1(1-\tau_i)+\gamma_2\tau_i]}$$
$$\times \prod_{j_1=0}^{k_1} \frac{1}{\sqrt{2\pi}\sigma_{j_1}} e^{-\frac{a_{j_1}^2}{2\sigma_{j_1}^2}} \prod_{j_2=0}^{k_2} \frac{1}{\sqrt{2\pi}\sigma_{j_2}} e^{-\frac{b_{j_2}^2}{2\sigma_{j_2}^2}} \prod_{\ell=1}^{2} \frac{\lambda_\ell}{\gamma_\ell^2} e^{-\lambda_\ell/\gamma_\ell}$$

and is well defined on $(\eta_1,\eta_2,\gamma) \in \Omega_1 \times \Omega_2 \times \Omega_3$, where

$$\Omega_1 = \{\left(a_0, \ldots, a_{k_1}\right) \middle| a_0 + a_1 x_i + \cdots + a_{k_1} x_i^{k_1} < y_i, i = 1, \ldots, n\right\},$$

$$\Omega_2 = \left\{\left(b_0, \ldots, b_{k_2}\right) \middle| b_0 + b_1 x_i + \cdots + b_{k_2} x_i^{k_2} > 0, i = 1, \ldots, n\right\},$$

$\Omega_3 = (0,M] \times (0,\infty)$, where M are any fixed positive real numbers.

It is worth mentioning that normal priors (5) can be very useful in practice. These priors say that the values of the model parameters can be positive or negative. The values of σ_js measure the strength of our prior knowledge on the model parameters: large values of them represent that we have little prior knowledge about the model parameters. In this paper we take $\sigma_{j_1} = \sigma_{j_2} = 3$ for all possible values of j_1 and j_2.

The prior (6) is the product of two inverse gamma-distributions. This prior distribution uses the information that $\gamma_1 > 0$ and $\gamma_2 > 0$. In this paper we let $\lambda_{\updownarrow} = 1$, leading to a very flat and right-skewed distribution, which implies the fact that little prior information on these parameters is available in practice.

Once the posterior density function of the parameters is available, we may use the Markov chain Monte Carlo (MCMC) method proposed by Cai (2010b) to estimate the model parameters. It can be shown that the equilibrium distribution of the Markov chain produced by the MCMC method is the posterior distribution of the parameters. Therefore, after a burn-in period, we can collect a posterior sample of the parameters; the average of the samples is the Bayesian estimate of the parameters, denoted by $\hat{\beta}$.

3.2. Forecasting

Now suppose that the parameters of model (4) have been estimated from data. Then conditional on an x value of interest, we want to forecast the τth quantile of Y with an associated prediction interval. Our prediction method consists of the following several steps.

Step 1. Simulate $t_i \sim U(0,1)$, $i = 1,\ldots,m$. That is t_i is a random sample from a uniform distribution between 0 and 1.

Step 2. Calculate y_i:

$$y_i = \left(\hat{a}_0 + \hat{a}_1 x + \cdots + \hat{a}_{k_1} x^{k_1}\right) + \left(\hat{b}_0 + \hat{b}_1 x + \cdots + \hat{b}_{k_2} x^{k_2}\right) t_i^{\hat{\gamma}_1} (1-t_i)^{-\hat{\gamma}_2}.$$

Then $\{y_i, i = 1,\ldots,m\}$ forms a random sample of the conditional distribution of Y.

Step 3. Use the sample obtained in Step 2 to estimate the τth quantile of Y, denoted by q_τ.

Step 4. Repeat the above three steps M times, we have M τth conditional quantiles of Y: $q_\tau^{(j)}$, $j = 1,\ldots,M$.

Step 5. Construct the distribution of the τth quantile of Y using $q_\tau^{(j)}$, $j = 1,\ldots,M$.

Step 6. Use the median and the lower and upper 2.5% quantiles of the distribution obtained in Step 5 as the point forecast of the τth conditional quantile of Y and the associated 95% prediction interval respectively.

It is worth noting that the above prediction method allows us to obtain the whole predictive distribution of a sea condition variable. So the method enables us to forecast any predictive quantities of interest about the sea condition variable.

4. Venice sea-level data

To demonstrate the method we have chosen Venice sea-level data. The application is intended to be illustrative and is unusual in an important respect. That is, the Venice sea-level data set consists of the two highest recorded sea-levels each year in “Venice for the period 1931–1981. The sea level in Venice is conventionally measured relative to the local datum of Punta Salute. The mean sea level is +0.52 m relative to this datum. Sea levels higher than +0.80 m above the datum are termed “acqua alta”, and are associated with travel and transport disruption in the lowest parts of the city (including St. Mark's Square). When the water level is above +1.0 m approximately 5% of public land is liable to be flooded; when the water level reaches +1.10 m about 12% of the city is affected by flooding; this rises to approximately 60% if the water level is +1.40 m. In this application we consider the sea levels relative to the mean sea level. Although no other information is available on the data, for example, how the observations have been de-trended etc., for illustration purposes, we just treated the observations to be the two highest records over all recorded levels each year. The developed method will be used to predict the annual maximum water level conditioned by knowledge of the second highest water level. The methodology can be applied to other data sets with more than one covariate as well.

Let y be the observed highest sea level and x the observed second highest sea level. Fig. 2 shows the scatter plots of x, y and y against x respectively. Clearly, both the second highest and the highest sea-levels increase with time, and there is a positive correlation between x and y. Furthermore, the variation of y also increases as x increases and some extreme sea-levels can also be seen.

Although the upward trend of the sea-levels may be estimated as a function of time by using the highest sea-level data only, we feel that the sea-level may also depend on some other covariates. In this application, a relationship between x and y cannot only reveal the upward trend and non-homogeneous feature of the variations, but also provide us a means of obtaining conditional forecasts on future sea-levels. For example, suppose x_0 is the highest sea level that has been observed in the first six months of a year. Then a well established relationship between x and y will allow us to predict the corresponding y value that may occur in the next six months with a required probability. Hence some necessary procedures may be taken to prevent flooding in the second half of the year.

Note that as (x,y) was collected once a year, it is reasonable to assume that the collected data are independent as we usually do in extreme value analysis (see, for example, Coles (2001)).

It is worth mentioning that a conventional regression model is not appropriate for modelling the relationship between x and y because of the non-homogeneous feature of the data. Fig. 2(c) suggests that model (4)

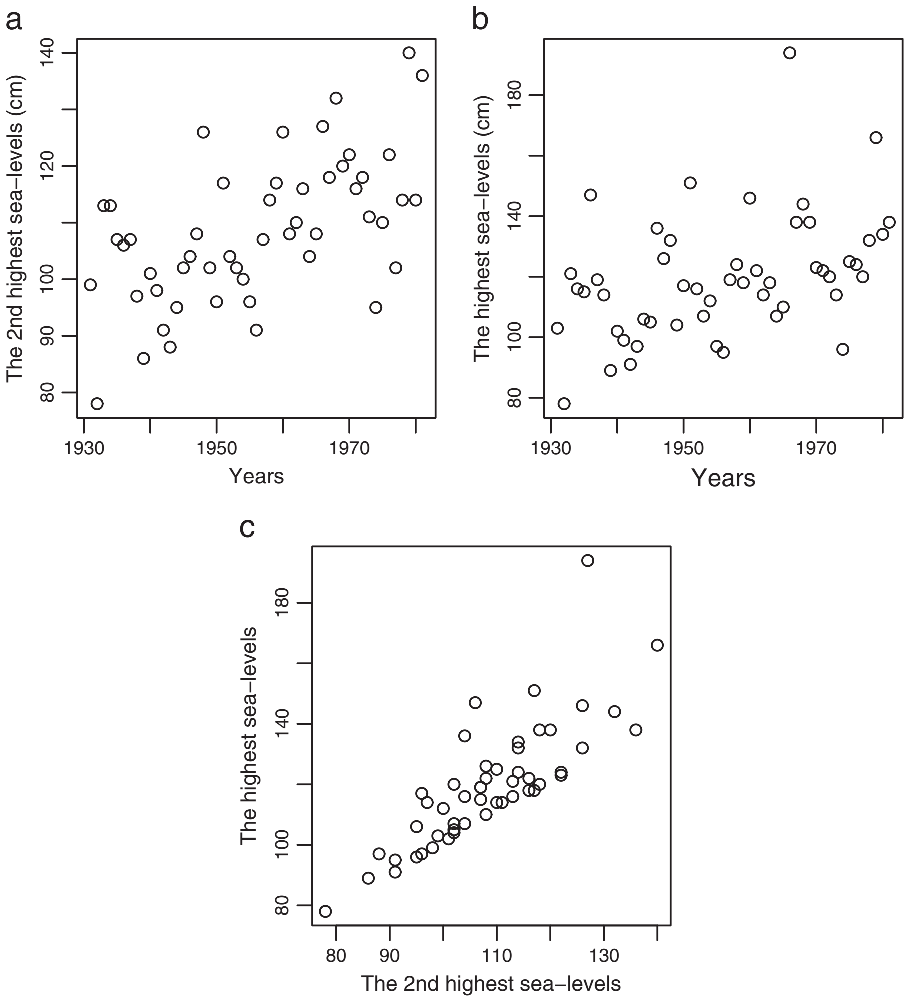

Fig. 2. Scatter plots of Venice sea-level data (relative to Punta Salute datum).

with $k_1 = k_2 = 1$ could be a good candidate for the data, as the strong linearity between the variables is evident. Hence, we applied our method to the model

$$Q_Y(\tau|\beta,\mathbf{x}) = (a_0 + a_1x) + (b_0 + b_1x)\tau^{\gamma_1}(1-\tau)^{-\gamma_2}. \quad (7)$$

Note that model (7) does not include time as a covariate although it is possible to do so. However, x is an implicit function of the time. Hence the long term trend of Y is reflected by the term $a_0 + a_1x$. Similarly, the non-stationarity of Y can also be seen from the two terms $a_0 + a_1x$ and $b_0 + b_1x$ as both the level and the spread of the distribution of Y depend on x and hence on t implicitly. Also it was noted that $\tau^{\gamma_1}(1-\tau)^{-\gamma_2}$ is used to deal with the extremes.

To estimate the parameters of model (7), a Markov chain of length 250,000 was run. A time series plot, (not shown to save space), of the Markov chain output shows that a burn-in of the first 75,000 values would be appropriate. After the burn-in period, we save the parameter values once every 50 steps. Fig. 3 shows the histograms of the collected samples, where the vertical lines correspond to the estimated parameter values which are the sample means of the posterior samples.

So the fitted model is given by

$$Q_Y\left(\tau\middle|\hat{\beta},\mathbf{x}\right) = -3.1794 + 1.0327x + (3.9599 + 0.0992x)\tau^{1.4231}(1-\tau)^{-0.3589}. \quad (8)$$

If the fitted model is fine, then we should expect that the standardized residuals

$$\hat{u}_i = \frac{y_i-(-3.1794 + 1.0327x_i)}{3.9599 + 0.0992x_i}, \quad i = 1,\ldots,n$$

are an independent sample from the distribution defined by

$$Q(\tau,\gamma) = \tau^{1.4231}(1-\tau)^{-0.3589}.$$

The last panel of Fig. 3 provides the plot of the sample quantiles of $\hat{u}_i$ $(i = 1,\ldots,n)$ against the quantiles of $Q(\tau,\gamma)$, which shows no major concerns on the fitted model.

In order to compare our approach with the semi-parametric approach, we also used the statistical software R to fit a sequence of quantile regression models

$$q^{\tau}_{Y|\mathbf{x}} = a^{\tau}_0 + a^{\tau}_1x \quad (9)$$

to the data for $\tau = 0.05, 0.25, 0.5, 0.75, 0.95, 0.99, 0.995$ and 0.999. The estimated parameter values are given in Table 1. Note that estimated parameter values in the last three columns are the same, suggesting that we were not able to estimate extreme quantiles when $\tau > 0.99$ for this data set.

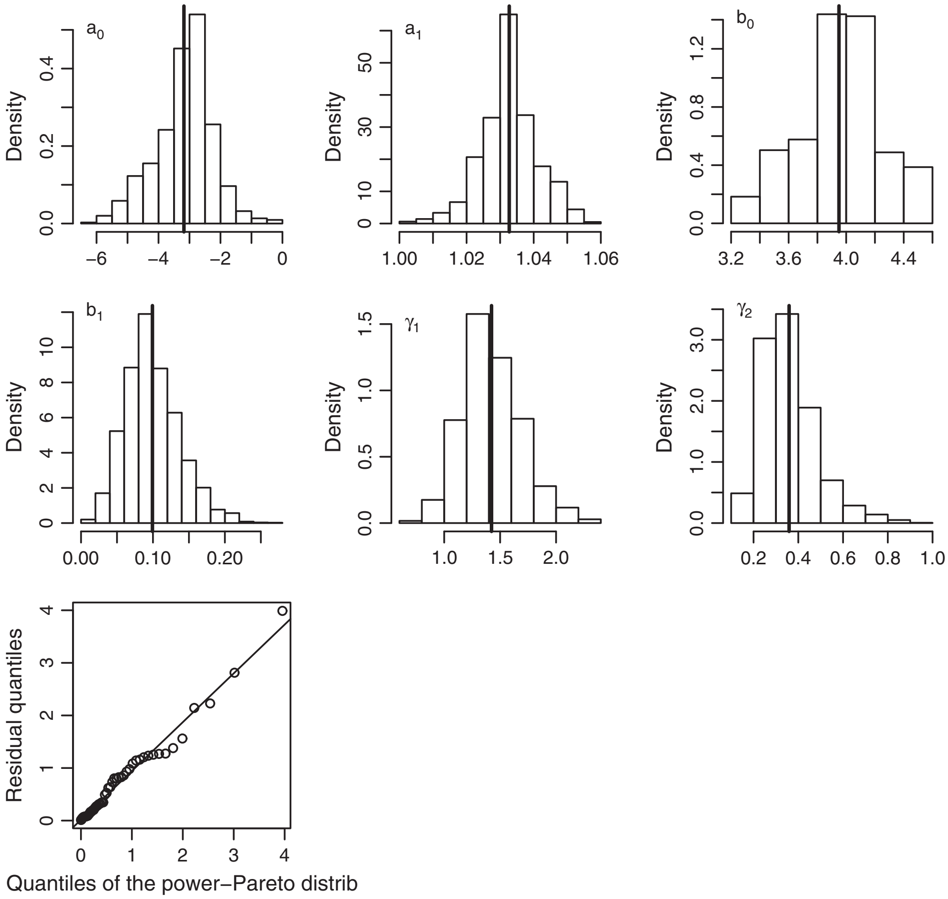

Fig. 3. The histograms of the posterior samples of model parameters and the QQ-plot of the fitted model for Venice sea-level data.

The estimated quantiles can be used to compute conditional quantile functions. Fig. 4 shows the quantile functions conditional on four chosen values: $x = 70$, 100, 120 and 150, where 70 and 150 are outside the range of the observed data. It is seen that the performances of the parametric and the semi-parametric approaches are very similar at lower quantile levels for this data set, but at high quantile levels, the semi-parametric model experiences some difficulties.

Now we consider prediction. Let L_α be the sea level that is expected to be exceeded on average once every $1/\alpha$ years, where $0 < \alpha < 1$. Then L_α is the $1/\alpha$-year return level. A typical value of α in design might be $\alpha = 0.01$, corresponding to the 100-year return level. It is noted that the $1/\alpha$-year return level may depend on some covariates. Furthermore, it is clear that the $1/\alpha$-year return level corresponds to the $(1 - \alpha)$th quantile. So for the 100-year return level, we have $\tau = 1 - \alpha = 0.99$. Our task is to predict the conditional 100-year return level. Note that values for other return levels can be obtained similarly.

Therefore, for our example, if we choose the second highest annual water level to be $x = 194$, (which coincidentally is the maximum observed sea level during the observational period and therefore outside the range of the observed second highest annual maxima), then the distribution of the 100-year return level conditional on $x = 194$ cm is shown in Fig. 5(a), where the darker continuous vertical line corresponds to the median of the 100-year return level, while the two dashed lines form a 95% prediction interval of the 100-year return level. Note that, as we have the whole predicted distribution of the 100-year return level, we could forecast any predictive quantities of this distribution. For example, we may also be interested in obtaining the forecast of the average 100-year return level, which could be easily calculated and is also shown in Fig. 5(a) by the grey vertical line.

Table 1
The estimated parameter values of model (9).

τ	0.05	0.25	0.5	0.75	0.95	0.99	0.995	0.999
a_0^τ	−2.00	−6.50	−13.8	−20.67	−96.00	−90.24	−90.24	−90.24
a_1^τ	1.03	1.08	1.2	1.33	2.23	2.24	2.24	2.24

Similarly, conditional on $x = 108.51$ cm, which is the average of the observed x values, the distribution of the 100-year return level is shown in Fig. 5(b). To quantify the forecasts, Table 2 shows the predicted 100-year return levels (median and mean) with the associated 95% prediction intervals. Note that the predicted mean is larger than the predicted median for this data set because the predictive distribution is skewed to the right.

We also used the semi-parametric model (9) to estimate the conditional 100-year return level when $x = 108.51$ cm, $a_0^{0.99} = -90.24$ and $a_1^{0.99} = 2.24$. We found that the return level is 152.82 cm.

If we consider the annual maximum sea-levels only, then the extreme value theory says that the generalized extreme value (GEV) distribution may be used to model the distribution of these data. Note that the log-normal distribution is also right skewed and can be used to deal with the tail behaviour of a distribution. Also note that from a pure mathematical point of view, the log-normal distribution can provide an excellent approximation to the GEV. Hence we also fitted a log-normal distribution to the annual maximum sea-levels. Fig. 6 shows the probability density function plots of the estimated GEV (dashed curve) and log-normal distribution (continuous curve), obtained by using the statistical software R. It is seen that they are indeed very similar. We found that the 100-year return levels are 178 cm and 172 cm corresponding to the GEV model and the log-normal model respectively.

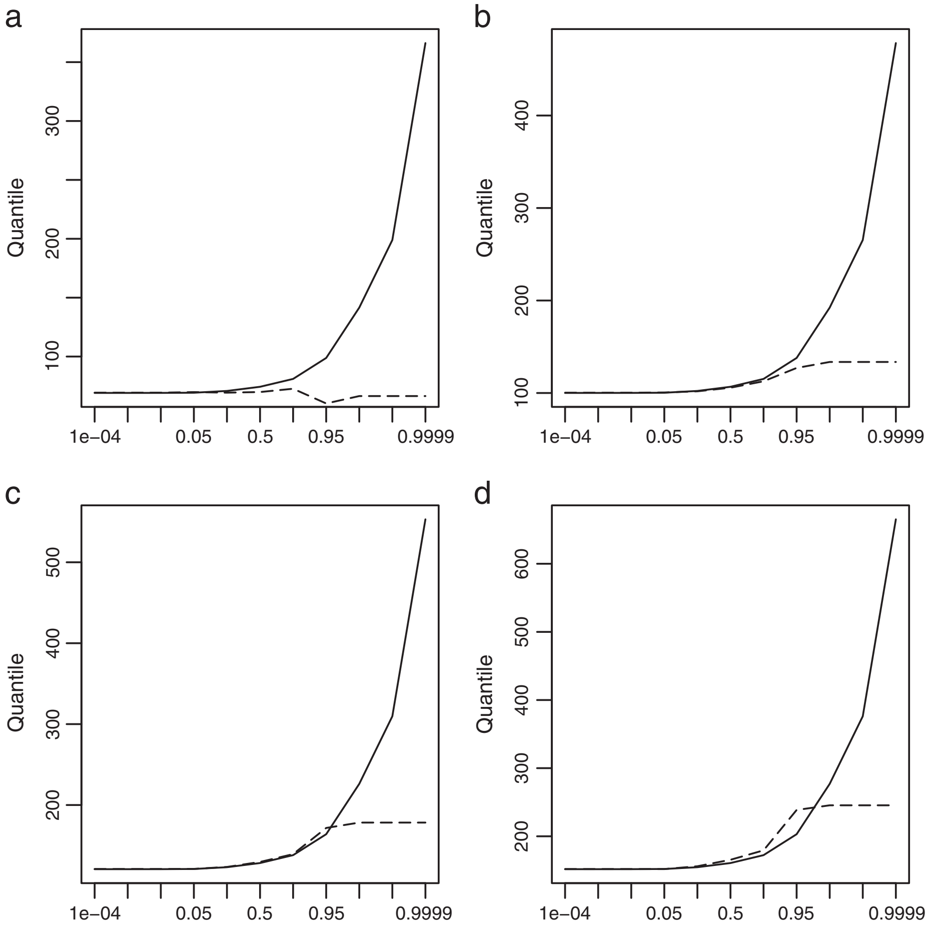

Fig. 4. Conditional quantile function plots conditional on (a) $x = 70$ (b) $x = 100$ (c) $x = 120$ (d) $x = 150$. Continuous curves correspond to the proposed method and dashed curves correspond to the semi-parametric method.

Both estimated return levels are closer to that obtained from our model when $x = 108.51$ cm. The reason for this similarity can be seen from Fig. 7.

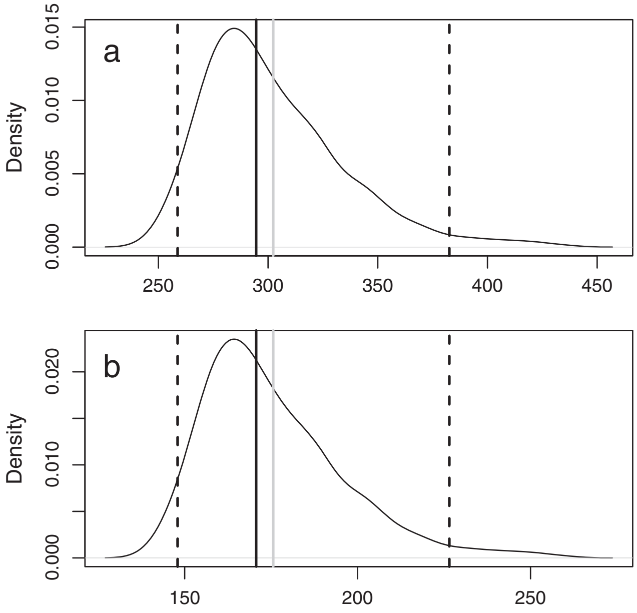

Fig. 5. Density function plots of the 100-year return level conditional on (a) $x = 194$ cm and (b) $x = 108.51$ cm.

Fig. 7 shows the conditional quantile function plots from the above fitted models, where the continuous curve is the quantile function of Y conditional on $x = 108.51$ cm obtained from our model, the dot–dashed curve corresponds to that obtained from the semi-parametric model, the dotted and dashed curves correspond to the unconditional quantile functions of Y obtained from the GEV and log-normal models respectively, and the points (circles) are the sample quantiles at equally spaced quantile levels. In fact the last circle in Fig. 7 corresponds to the sample quantile at the 0.9999 level, while the black square symbol corresponds to the sample quantile at the 0.99 level, indicated by the grey vertical line. It is seen that our model, GEV and log-normal models behave similarly in predicting 100-year return levels conditional on the average value of x.

5. Comments and conclusions

In this paper we have developed a prediction method based on Markov chain simulations for quantile function models. The method allows us to obtain distributional forecasts, and hence enables us to obtain any predictive quantities of a sea condition variable.

We showed that a quantile function model can provide a practical alternative technique for estimating extreme levels, as it allows us to

Table 2
The predicted conditional 100-year return levels for the Venice sea-level data.

x	Median level	Mean level	Lower bound	Upper bound
$x = 194$ cm	294.58	302.35	258.77	382.62
$x = 108.51$ cm	170.68	175.61	147.96	226.52

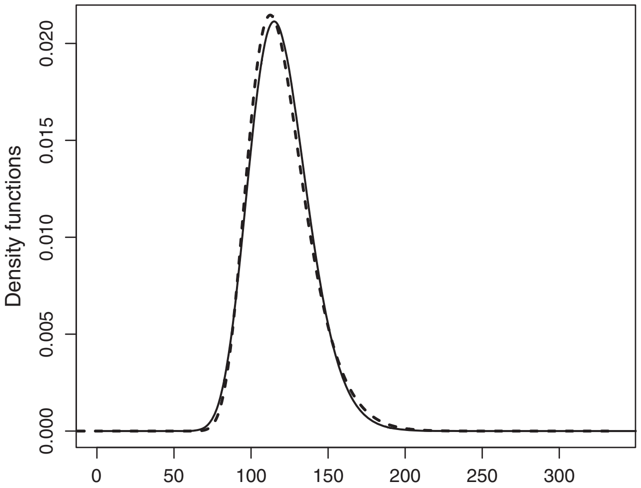

Fig. 6. Probability density function plots of the estimated GEV (dashed curve) and log-normal distribution (continuous curve).

construct a proper statistical model by using many distributions that are non-standard and that are defined only through their quantile functions instead of the probability density functions. Hence, such models can deal with many complicated data structures that may not be dealt with easily by conventional models. The quantile function approach also makes full use of the available data and provides a means of incorporating covariates. It is worth mentioning that quantile function models cannot only be estimated by using our method but also by other methods based on optimization techniques, (see Gilchrist, 2000).

We have also demonstrated the usefulness of the power-Pareto distribution via a polynomial quantile function model and the Venice sea-level data. For extreme or 'tail' probabilities, a parametric model would always give finer results, but only under the condition that it is a proper model. Our results show that it is possible to build up a proper model for a data set by using a quantile function modelling approach.

Our results also show that different models yield a range of performance on the same data set. Our experience also suggests that the performance of a particular model may vary from data set to data set. Therefore, a general recommendation is that in any practical application several different approaches are to be used whenever possible in order to achieve the best model for a given data set.

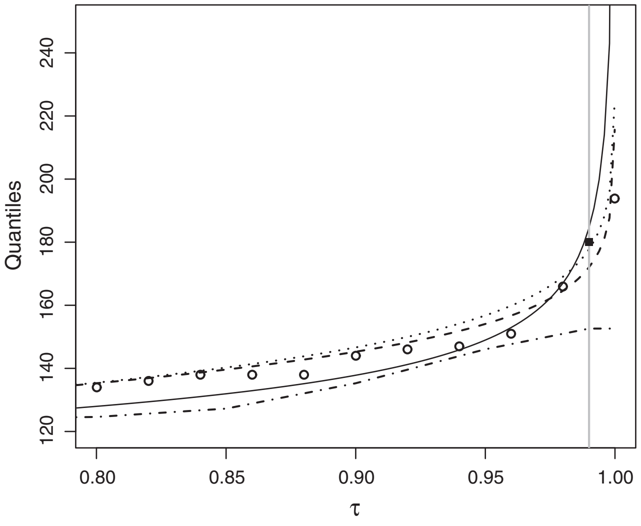

Fig. 7. Quantile function plots of the GEV model (dotted curve), the log-normal model (dashed curve), the power-Pareto quantile function model (continuous curve) and the semi-parametric model (dot–dashed curves). The points are the sample quantiles.

Acknowledgement

This research is supported by the EPSRC Bridging the Gaps Programme. We would like to thank the EPSRC for the financial support. We would also like to thank the referees for their very constructive comments and suggestions which have enhanced the quality and the presentation of the paper.

References

Bondell, H.D., Reich, B.J., Wang, H., 2010. Non-crossing quantile regression curve estimation. Biometrika 97, 825–838.

Cai, Y., 2009. Autoregression with non-Gaussian Innovations. Journal of Time Series Econometrics 1 (2). http://dx.doi.org/10.2202/1941-1928.1016 (Article 2).

Cai, Y., 2010a. Multivariate quantile function models. Statistica Sinica 20, 481–496.

Cai, Y., 2010b. Polynomial power-Pareto quantile function models. Extremes 13, 291–314.

Cai, Y., 2010c. Forecasting for quantile self exciting threshold autoregressive time series models. Biometrika 97, 199–208.

Cai, Y., Stander, J., 2008. Quantile self-exciting threshold autoregressive time series models. Journal of Time Series Analysis 29, 186–202.

Cai, Y., Stander, J., Davies, N., 2012. A new Bayesian approach to quantile autoregressive time series model estimation and forecasting. Journal of Time Series Analysis 33, 684–698.

Coles, S., 2001. An Introduction to Statistical Modelling of Extreme Values. Springer Series in Statistics.

Coles, S., Tawn, J., 1994. Statistical methods for multivariate extremes: an application to structural design (with discussion). Applied Statistics 43, 1–48.

Coles, S., Heffernan, J., Tawn, J., 1999. Dependence measures for extreme value analyses. Extremes 2, 339–365.

Gilchrist, W.G., 2000. Statistical Modelling with Quantile Functions. Chapman & Hall/CRC.

Hawkes, P.J., Gouldby, B.P., Tawn, J.A., Owen, M.W., 2002. The joint probability of waves and water levels in coastal defence design. Journal of Hydraulic Research 40, 241–251.

Hawkes, P.J., Gouldby, B.P., Sun, W., Tawn, J.A., Hames, D., Reeve, D., Blackman, D., Sproson, R., Mavronasos, K., 2004. A comparison of marginal and joint extremes predicted from synthesised wave and water level data. First International Conference on Flood Risk Assessment, University of Bath, Published by the Institute of Mathematics and Its Applications.

Koenker, R., 2005. Quantile Regression. Cambridge University Press.

Koenker, R., D'Orey, V., 1987. Computing regression quantiles. Applied Statistics 36, 383–393.

Koenker, R., D'Orey, V., 1994. A remark on Algorithm AS229: computing dual regression quantiles and regression rank scores. Applied Statistics 43, 410–414.

Lancaster, T., Jun, S.J., 2010. Bayesian quantile regression methods. Journal of Applied Econometrics 25, 287–307.

Li, Y., Simmonds, D.J., Reeve, D.E., 2008. Quantifying uncertainty in extreme values of design parameters with resampling techniques. Ocean Engineering 35, 1029–1038.

Meadowcroft, I.C., Hawkes, P.J., Surendran, S., 2004. Joint probability best practice guide: practical approaches for assessing combined sources of risk for flood and coastal risk managers. Defra Flood and Coastal Management Conference. University of York.

Owen, M.W., Hawkes, P.J., Tawn, J.A., Bortot, P., 1997. The joint probability of waves and water levels: a rigorous but practical new approach. MAFF Conference of River and Coastal Engineers, Keele.

Reeve, D.E., 1996. Estimation of extreme Indian monsoon rainfall. International Journal of Climatology 16, 105–112.

Reeve, D.E., 1998. Coastal flood risk assessment. Journal of Waterway Port, Coastal, and Ocean Engineering, ASCE 124 (5), 219–228.

Smith, R.L., 1986. Extreme value theory based on the r largest annual events. Journal of Hydrology 86, 27–43.

Tawn, J.A., 1990. Modelling multivariate extreme value distributions. Biometrika 77, 245–253.

Tawn, J.A., 1992. Estimating probabilities of extreme sea-levels. Applied Statistics 41, 77–93.

Taylor, J.W., 2005. Generating volatility forecasts from value at risk estimates. Management Science 51, 712–725.

Thompson, P., Cai, Y., Reeve, D.E., Stander, J., 2009. Automated threshold selection methods for extreme wave analysis. Coastal Engineering 56, 1013–1021.

Thompson, P., Cai, Y., Moyeed, R., Reeve, D.E., Stander, J., 2010. Bayesian non-parametric quantile regression using splines. Computational Statistics and Data Analysis 54, 1138–1150.

Yu, K., Moyeed, R.A., 2001. Bayesian quantile regression. Statistics and Probability Letters 54, 437–447.