ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)

Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity

Paul J. Northrop and Nicolas Attalides

University College London, UK

and Philip Jonathan

Shell Projects and Technology, Manchester, UK

[Received April 2015. Revised April 2016]

Summary. Design conditions for marine structures are typically informed by threshold-based extreme value analyses of oceanographic variables, in which excesses of a high threshold are modelled by a generalized Pareto distribution. Too low a threshold leads to bias from model misspecification, and raising the threshold increases the variance of estimators: a bias–variance trade-off. Many existing threshold selection methods do not address this trade-off directly but rather aim to select the lowest threshold above which the generalized Pareto model is judged to hold approximately. In the paper Bayesian cross-validation is used to address the trade-off by comparing thresholds based on predictive ability at extreme levels. Extremal inferences can be sensitive to the choice of a single threshold. We use Bayesian model averaging to combine inferences from many thresholds, thereby reducing sensitivity to the choice of a single threshold. The methodology is applied to significant wave height data sets from the northern North Sea and the Gulf of Mexico.

Keywords: Cross-validation; Extreme value theory; Generalized Pareto distribution; Predictive inference; Threshold

1. Introduction

Ocean and coastal structures, including breakwaters, ships and oil and gas producing facilities, are designed to withstand extreme environmental conditions. Marine engineering design codes stipulate that estimated failure probabilities of offshore structures, associated with one or more return periods, should not exceed specified values. To characterize the environmental loading on an offshore structure, return values for winds, waves and ocean currents corresponding to a return period of typically 100 years, but sometimes to 1000 and 10000 years, are required. The severity of waves in a storm is quantified by using significant wave height. Extreme value analyses of measured and hindcast samples of significant wave height are undertaken to derive environmental design criteria, typically by fitting a generalized Pareto (GP) distribution to excesses of a high threshold. The selection of appropriate threshold(s) is important because inferences can be sensitive to threshold.

Address for correspondence: Paul Northrop, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK.
E-mail: p.northrop@ucl.ac.uk

 0035–9254/17/66000

1.1. Storm peak significant wave height data sets

The focus of this paper is the analysis of two sequences of hindcasts of storm peak significant wave height, which are shown in Fig. 1. Significant wave height H_s is a measure of sea surface roughness. It is defined as four times the standard deviation of the surface elevation of a *sea state*, which is the ocean surface observed for a certain period of time (3 h for our data sets). Using global satellite observations, Cardone *et al.* (2015) reported multiple estimates of very extreme sea state significant wave height in excess of 18 m, and some estimates in excess of 20 m. Hindcasts are samples from physical models of the ocean environment, calibrated to observations of pressure, wind and wave fields.

For each of the data sets raw data have been declustered, using a procedure described in Ewans and Jonathan (2008), to isolate cluster maxima (storm peaks) that can reasonably be treated as being mutually independent. We also assume that storm peaks are sampled from a common distribution. Even in this simplest of situations practitioners have difficulty in selecting appropriate thresholds. The first data set (Oceanweather, 1995), from an unnamed location in the northern North Sea, contains 628 storm peaks from October 1964 to March 1995, but restricted to the period from October to March within each year. The other data set (Oceanweather, 2005) contains 315 storm peaks from September 1900 to September 2005. Most (213) of the peaks occur during August–October, but there is no obvious seasonality in their magnitudes. For the

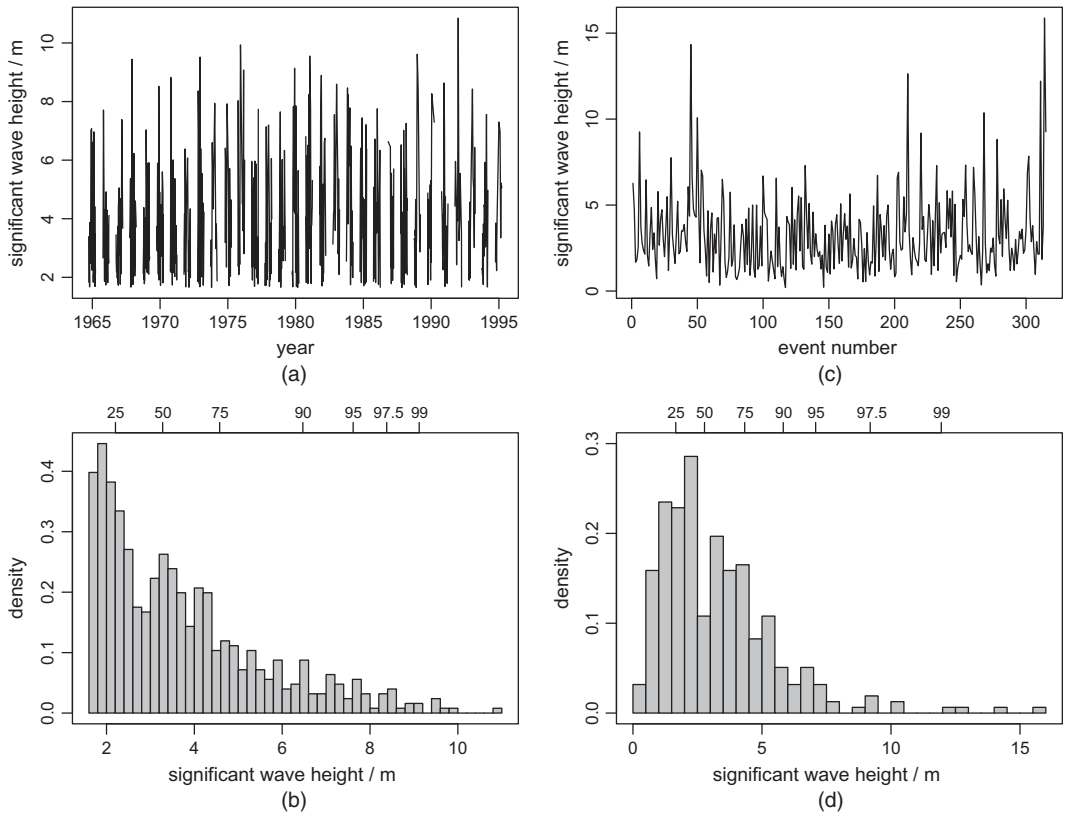

Fig. 1. Storm peak significant wave height hindcast data sets: (a), (b) North Sea data (628 observations); (c), (d) Gulf-of-Mexico data (315 observations); (a), (c) time series plots (in (a) distinct October–March periods are separated); (b), (d) histograms (the upper axis scales give the sample quantiles)

North Sea data there is evidence of some seasonality within the October–March window: storm peaks tend to be slightly larger near the middle of this window than near the ends. Ignoring such seasonality amounts to an analysis of the extremes from a distribution mixing random deviations across seasonal effects. This omniseasonal analysis is of practical relevance but may result in some loss of efficiency. In on-going work we are extending the methodology to incorporate seasonal, or other, covariate effects.

In the northern North Sea the main fetches are the Norwegian Sea to the north, the Atlantic Ocean to the west and the North Sea to the south. Extreme sea states from the directions of Scandinavia to the east and the British Isles to the south-west are not possible, owing to the shielding effects of these land masses. At the location under consideration, the most extreme sea states are associated with storms from the Atlantic Ocean (Ewans and Jonathan, 2008). With up to several tens of storms impacting the North Sea each winter, the number of events for analysis is typically larger than for locations in regions such as the Gulf of Mexico, where hurricanes produce the most severe sea states. Most hurricanes originate in the Atlantic Ocean between June and November and propagate west to north-west into the Gulf, producing the largest sea states with dominant directions from the south-east to east directions. It is expected that there is greater potential for very stormy sea conditions in the Gulf of Mexico than in the northern North Sea and therefore that the extremal behaviour is different in these two locations.

1.2. Extreme value threshold selection

Extreme value theory provides asymptotic justification for a particular family of models for excesses of a high threshold. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables, with common distribution function H, and u_n be a threshold, increasing with n. Pickands (1975) showed that if there is a non-degenerate limiting distribution for appropriately linearly rescaled excesses of u_n then this limit is a GP distribution. In practice, a suitably high threshold u is chosen empirically. Given that there is an exceedance of u, the excess $Y = X - u$ is modelled by a GP(σ_u, ξ) distribution, with positive threshold-dependent scale parameter σ_u, shape parameter ξ and distribution function

$$G(y; \sigma_u, \xi) = \begin{cases} 1 - (1 + \xi y/\sigma_u)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y/\sigma_u), & \xi = 0, \end{cases} \tag{1}$$

where $y > 0$ and $x_+ = \max(x, 0)$. The $\xi = 0$ case is defined in the limit as $\xi \to 0$. When $\xi < 0$ the distribution of X has a finite upper end point of $u - \sigma_u/\xi$; otherwise, X is unbounded above. The frequency with which the threshold is exceeded also matters. Under the current assumptions the number of exceedances over the threshold u has a binomial (n, p_u) distribution, where $p_u = P(X > u)$, giving a binomial–GP BGP(p_u, σ_u, ξ) model (Coles (2001), chapter 4).

Many threshold diagnostic procedures have been proposed: Scarrott and MacDonald (2012) provides a review. Broad categories of methods include assessing stability of model parameter estimates with threshold (Drees *et al.*, 2000; Wadsworth, 2016), goodness-of-fit tests (Davison and Smith, 1990; Dupuis, 1999), approaches that minimize the asymptotic mean-squared error of estimators of ξ or of extreme quantiles, under particular assumptions about the form of the upper tail of H (Hall and Welsh, 1985; Hall, 1990; Ferreira *et al.*, 2003; Beirlant *et al.*, 2004; Caeiro and Gomes, 2016) and specifying a model for (some or all) data below the threshold (Wong and Li, 2010; MacDonald *et al.*, 2011; Wadsworth and Tawn, 2012). In the last category, the threshold above which the GP model is assumed to hold is treated as a model parameter and *threshold uncertainty* is incorporated by averaging inferences over a posterior distribution of model parameters. In contrast, in a *single-threshold approach* threshold level is viewed as a

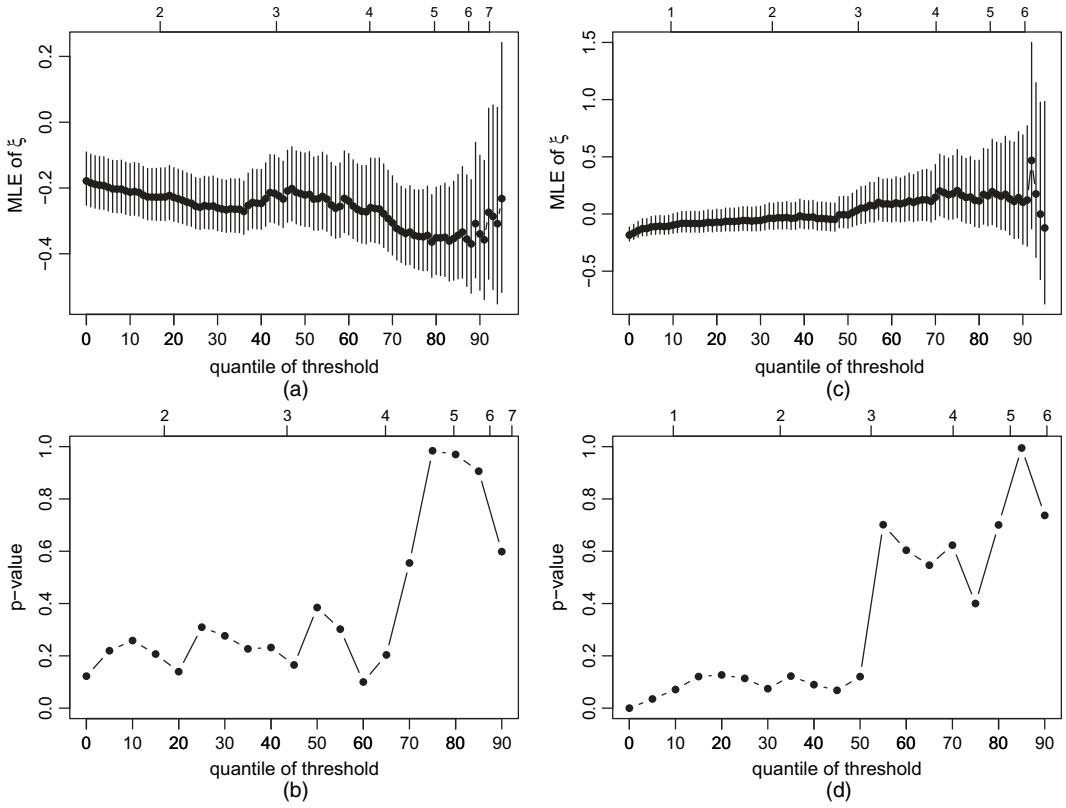

Fig. 2. Threshold diagnostic plots for the storm peak significant wave height hindcast data sets: (a), (b) North Sea data; (c), (d) Gulf-of-Mexico data; (a), (c) parameter stability plots for ML estimates of ξ, with 95% pointwise profile-likelihood-based confidence intervals; (b), (d) p-values associated with a test of constant shape parameter against the lowest threshold considered (the upper axis scales give the level of the threshold in metres)

tuning parameter, whose value is selected before the main analysis and is treated as fixed and known when subsequent inferences are made.

Single-threshold selection involves a bias–variance trade-off (Scarrott and MacDonald, 2012): the lower the threshold the greater the estimation bias due to model misspecification; the higher the threshold the greater the estimation uncertainty. Many existing approaches do not address the trade-off directly but rather examine sensitivity of inferences to threshold and/or aim to select the lowest threshold above which the GP model appears to hold approximately. We seek to deal with the bias–variance trade-off on the basis of the main purpose of the modelling, i.e. prediction of future extremal behaviour. We make use of a data-driven method that is commonly used for such purposes: cross-validation (Stone, 1974). We consider only the simplest of modelling situations, i.e. where observations are treated as independent and identically distributed. However, selecting the threshold level is a fundamental issue for all threshold-based extreme value analyses and we expect that our general approach will have much wider applicability.

We illustrate some of the issues that are involved in threshold selection by applying to the significant wave height data sets two approaches that assess parameter stability. In Figs 2(a) and 2(b) maximum likelihood (ML) estimates $\hat{\xi}$ of ξ are plotted against threshold. The aim is to choose the lowest threshold above which $\hat{\xi}$ is approximately a constant function of thresh-

old, taking into account sampling variability summarized by the confidence intervals. It is not possible to make a definitive choice and different viewers may choose rather different thresholds. In both of these plots *our* eyes are drawn to the approximate stability of the estimates at around the 70% sample quantile. One could argue for lower thresholds, to incur some bias in return for reduced variance, but it is not possible to assess this objectively from these plots. In practice, it is common not to consider thresholds below the apparent mode of the data because the GP distribution has its mode at the origin. For example, on the basis of the histogram of the Gulf-of-Mexico data in Fig. 1 one might judge a threshold below the 25% sample quantile to be unrealistic. However, we *shall* consider such thresholds because it is interesting to see to what extent the bias expected is offset by a gain in precision.

The inherent subjectivity of this approach, and other issues such as the strong dependence between estimates of ξ based on different thresholds, motivated more formal assessments of parameter stability (Wadsworth and Tawn, 2012; Northrop and Coleman, 2014; Wadsworth, 2016). Figs 2(b) and 2(d) are based on Northrop and Coleman (2014). A subasymptotic piecewise constant model (Wadsworth and Tawn, 2012) is used in which the value of ξ may change at each of a set of thresholds, here set at the 0%, 5%,. . ., 95% sample quantiles. For a given threshold the null hypothesis that the shape parameter is constant from this threshold upwards is tested. In the plots p-values from this test are plotted against the threshold. Although these plots address many of the inadequacies of the parameter estimate stability plots subjectivity remains because one must decide how to make use of the p-values. One could prespecify a size, e.g. 0.05, for the tests or take a more informal approach by looking for a sharp increase in p-value. For the North Sea data the former would suggest a very low threshold and the latter a threshold in the region of the 70% sample quantile. For the Gulf-of-Mexico data respective thresholds close to the 10% and 55% sample quantiles are indicated.

An argument against selecting a single threshold is that this ignores uncertainty concerning the choice of this threshold. As mentioned above, one way to account for this uncertainty is to embed a threshold parameter within a model. We use an approach based on Bayesian model averaging, on which Hoeting *et al*. (1999) provide a review. Sabourin *et al*. (2013) have recently used a similar approach to combine inferences from different multivariate extreme value models. We treat different thresholds as providing competing models for the data. Predictions of extremal behaviour are averaged over these models, with individual models weighted in proportion to the extent to which they are supported by the data. There is empirical and theoretical evidence (Hoeting *et al*. (1999), section 7) that averaging inferences in this way results in better average predictive ability than provided by any single model.

For the most part we work in a Bayesian framework because prediction is handled naturally and regularity conditions that are required for making inferences by using ML (Smith, 1985) and probability-weighted moments (Hosking and Wallis, 1987), namely $\xi > -\frac{1}{2}$ and $\xi < \frac{1}{2}$ respectively, can be relaxed. This requires a prior distribution to be specified for the parameters of the BGP model. Initially, we consider three 'reference' prior distributions, in the general sense of priors constructed by using formal rules (Kass and Wasserman, 1996). Such priors can be useful when information that is provided by the data is much stronger than prior information from other sources. This is more likely to be so for a low threshold than for a high threshold. We use simulation to assess the utility of these priors for our purpose, i.e. making predictive statements about future extreme observations, and use the results to formulate an improved prior. For high thresholds, when the data are likely to provide only weak information, it may be important to incorporate at least some basic prior information to avoid making physically unrealistic inferences.

In Section 2 we use cross-validation to estimate a measure of threshold performance to select

a single threshold. Sections 2.1 and 2.2 describe the cross-validation procedure and its role in selecting a single threshold. In Section 2.3 we discuss two related formulations of the objective of an extreme value analysis and, in Section 2.4, we use one of these formulations in a simulation study to inform the choice of prior distribution for GP parameters. In Section 2.5 we use our methodology to make inferences about extreme significant wave heights in the North Sea and in the Gulf of Mexico. In Section 3 we use the measure of threshold performance to combine inferences over many thresholds. Another simulation study, in Section 3.1, compares choosing a single 'best' threshold and averaging inferences over many thresholds and in Section 3.2 we apply the latter to the significant wave height data sets. In Section 3.3 we incorporate prior information to avoid physically unrealistic inferences.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Single-threshold selection

We use a Bayesian implementation of leave-one-out cross-validation to compare the predictive ability of BGP inferences based on different thresholds. We take a *predictive* approach, averaging inferences over the posterior distribution of parameters, to reflect differing parameter uncertainties across thresholds: uncertainty in GP parameters will tend to increase as the threshold is raised. In contrast, under an *estimative*, or *plug-in*, approach, predictions use point estimates of parameters, acting as if these are the true values, with no account made for parameter uncertainty. A point estimate of GP model parameters can give a zero likelihood for a validation observation: this occurs if $\hat{\xi} < 0$ and this observation is greater than the estimated upper end point $u - \hat{\sigma}_u/\hat{\xi}$. In this event an estimative approach would effectively rule out the threshold u. Accounting for parameter uncertainty alleviates this problem by giving weight to parameter values other than a particular point estimate.

A naive implementation of leave-one-out cross-validation is computationally intensive. To avoid excessive computation we use importance sampling to estimate cross-validation predictive densities based on Bayesian inferences from the entire data set. One could use a similar strategy in a frequentist approximation to predictive inference based on large sample theory or bootstrapping (Young and Smith (2005), chapter 10). However, large sample results may provide poor approximations for high thresholds (small numbers of excesses) and the GP observed information is known to have poor finite sample properties (Süveges and Davison, 2010). Bootstrapping, of ML or probability-weighted moments estimates, increases computation time further and is subject to the regularity conditions that were mentioned in Section 1.

2.1. Assessing threshold performance by using cross-validation

Suppose that $\mathbf{x} = (x_1, \ldots, x_n)$ is a random sample of raw (unthresholded) data from H. Without loss of generality we assume that $x_1 < \ldots < x_n$. Consider a *training threshold* u. A BGP(p_u, σ_u, ξ) model is used at threshold u, where $p_u = P(X > u)$ and (σ_u, ξ) are the parameters of the GP model for excesses of u. Let $\boldsymbol{\theta} = (p_u, \sigma_u, \xi)$ and $\pi(\boldsymbol{\theta})$ be a prior density for $\boldsymbol{\theta}$. Let $\mathbf{x}^s$ denote a subset of $\mathbf{x}$, possibly equal to $\mathbf{x}$. The posterior density $\pi_u(\boldsymbol{\theta}|\mathbf{x}^s) \propto L(\boldsymbol{\theta}; \mathbf{x}^s, u)\,\pi(\boldsymbol{\theta})$, where

$$L(\boldsymbol{\theta}; \mathbf{x}^s, u) = \prod_{i: x_i \in \mathbf{x}^s} f_u(x_i|\boldsymbol{\theta}),$$

$$f_u(x_i|\boldsymbol{\theta}) = (1 - p_u)^{I(x_i \leqslant u)} \{p_u\, g(x_i - u; \sigma_u, \xi)\}^{I(x_i > u)},$$

$I(x)=1$ if x is true and $I(x)=0$ otherwise, and $g(x;\sigma_u,\xi)=\sigma_u^{-1}(1+\xi x/\sigma_u)_+^{-(1+1/\xi)}$ is a GP (σ_u,ξ) density. Note that $f_u(x_i|\boldsymbol{\theta})$ is not a probability density but the contribution to $L(\boldsymbol{\theta};\mathbf{x}^s,u)$ from a mixed indicator continuous variable that depends on whether x_i is above or below u.

We quantify the ability of BGP inferences based on threshold u to predict (out of sample) at extreme levels. For this we introduce a *validation threshold* $v \geqslant u$. If $1+\xi(v-u)/\sigma_u > 0$ then a BGP(p_u,σ_u,ξ) model at threshold u implies a BGP(p_v,σ_v,ξ) model at threshold v, where $\sigma_v=\sigma_u+\xi(v-u)$ and $p_v=P(X>v)=\{1+\xi(v-u)/\sigma_u\}^{-1/\xi}p_u$. Otherwise, $p_v=0$ and excesses of v are impossible. For a particular value of v we wish to compare the predictive ability of the implied BGP(p_v,σ_v,ξ) model across a range of values of u. We use a fixed validation threshold v for different values of u so that the performances of the training thresholds are compared by using exactly the same validation data.

We employ a leave-one-out cross-validation scheme in which $\mathbf{x}_{(r)}=\{x_i, i\neq r\}$ forms the training data and x_r the validation data. The *cross-validation predictive densities* at validation threshold v, based on a training threshold u, are given by

$$f_v(x_r|\mathbf{x}_{(r)},u)=\int f_v(x_r|\boldsymbol{\theta},\mathbf{x}_{(r)})\,\pi_u(\boldsymbol{\theta}|\mathbf{x}_{(r)})\,\mathrm{d}\boldsymbol{\theta}, \qquad r=1,\ldots,n, \tag{2}$$

although 'density' is an abuse of terminology owing to the presence of the indicator variables. The conditioning in $\pi_u(\boldsymbol{\theta}|\mathbf{x}_{(r)})$, and hence $f_v(x_r|\mathbf{x}_{(r)},u)$, is on those values in $\mathbf{x}_{(r)}$ above u and below-threshold indicators of the remaining components, not all the $n-1$ numerical values in $\mathbf{x}_{(r)}$.

Suppose that the $\{x_i\}$ are conditionally independent given $\boldsymbol{\theta}$. If $p_v>0$ then

$$f_v(x_r|\boldsymbol{\theta},\mathbf{x}_{(r)})=f_v(x_r|\boldsymbol{\theta})=(1-p_v)^{I(x_r\leqslant v)}\{p_v\,g(x_r-v;\sigma_v,\xi)\}^{I(x_r>v)}. \tag{3}$$

If $p_v=0$ then $f_v(x_r|\boldsymbol{\theta},\mathbf{x}_{(r)})=I(x_r\leqslant v)$. Suppose that we have a sample $\boldsymbol{\theta}_j^{(r)}$, $j=1,\ldots,m$, from the posterior $\pi_u(\boldsymbol{\theta}|\mathbf{x}_{(r)})$. Then a Monte Carlo estimator of $f_v(x_r|\mathbf{x}_{(r)},u)$ based on expression (2) is given by

$$\hat{f}_v(x_r|\mathbf{x}_{(r)},u)=\frac{1}{m}\sum_{j=1}^{m} f_v(x_r|\boldsymbol{\theta}_j^{(r)},\mathbf{x}_{(r)}). \tag{4}$$

Evaluation of estimator (4), for $r=1,\ldots,n$, is computationally intensive because it involves generating samples from n different posterior distributions. To reduce computation time we use an importance sampling estimator (Gelfand, 1996; Gelfand and Dey, 1994) that enables estimation of $\hat{f}_v(x_r|\mathbf{x}_{(r)},u)$, for $r=1,\ldots,n-1$, using a single sample only. We rewrite expression (2) as

$$f_v(x_r|\mathbf{x}_{(r)},u)=\int f_v(x_r|\boldsymbol{\theta},\mathbf{x}_{(r)})\,q_r(\boldsymbol{\theta})\,h(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}, \qquad r=1,\ldots,n, \tag{5}$$

where $q_r(\boldsymbol{\theta})=\pi_u(\boldsymbol{\theta}|\mathbf{x}_{(r)})/h(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta})$ is a density whose support must include that of $\pi_u(\boldsymbol{\theta}|\mathbf{x}_{(r)})$. In the current context a common choice is $\pi_u(\boldsymbol{\theta}|\mathbf{x})$ (Gelfand and Dey (1994), page 511). However, the support of $\pi_u(\boldsymbol{\theta}|\mathbf{x})$: $\xi>-\sigma_u/(x_n-u)$ does not contain that of $\pi_u(\boldsymbol{\theta}|\mathbf{x}_{(n)})$: $\xi>-\sigma_u/(x_{n-1}-u)$, since $x_n>x_{n-1}$. Therefore we use $h(\boldsymbol{\theta})=\pi_u(\boldsymbol{\theta}|\mathbf{x})$ only for $r\neq n$.

Suppose that we have a sample $\boldsymbol{\theta}_j$, $j=1,\ldots,m$, from the posterior $\pi_u(\boldsymbol{\theta}|\mathbf{x})$. For $r=1,\ldots,n-1$ we use the importance sampling ratio estimator

$$\hat{f}_v(x_r|\mathbf{x}_{(r)},u)=\frac{\sum_{j=1}^{m} f_v(x_r|\boldsymbol{\theta}_j)\,q_r(\boldsymbol{\theta}_j)}{\sum_{j=1}^{m} q_r(\boldsymbol{\theta}_j)}=\frac{\sum_{j=1}^{m} f_v(x_r|\boldsymbol{\theta}_j)/f_u(x_r|\boldsymbol{\theta}_j)}{\sum_{j=1}^{m} 1/f_u(x_r|\boldsymbol{\theta}_j)}, \tag{6}$$

where $q_r(\boldsymbol{\theta}) = \pi_u(\boldsymbol{\theta}|\mathbf{x}_{(r)})/\pi_u(\boldsymbol{\theta}|\mathbf{x}) \propto 1/f_u(x_r|\boldsymbol{\theta})$. If we also have a sample $\boldsymbol{\theta}_j^{(n)}$, $j = 1, \ldots, m$, from the posterior $\pi_u(\boldsymbol{\theta}|\mathbf{x}_{(n)})$ then $\hat{f}_v(x_n|\mathbf{x}_{(n)}, u) = (1/m)\,\Sigma_{j=1}^{m} f_v(x_n|\boldsymbol{\theta}_j^{(n)})$. We use

$$\hat{T}_v(u) = \sum_{r=1}^{n} \log\{\hat{f}_v(x_r|\mathbf{x}_{(r)}, u)\} \tag{7}$$

as a measure of predictive performance at validation threshold v when using training threshold u.

2.2. Comparing training thresholds

Consider k training thresholds $u_1 < \ldots < u_k$, resulting in estimates $\hat{T}_v(u_1), \ldots, \hat{T}_v(u_k)$. Up to an additive constant, $\hat{T}_v(u)$ provides an estimate of the negated Kullback–Leibler divergence between the BGP model at validation threshold v and the true density (see Silverman (1986), page 53). Thus, $u^* = \arg\max_u \hat{T}_v(u)$ has the property that, of the thresholds considered, it has the smallest estimated Kullback–Leibler divergence. Some inputs are required: $\mathbf{u} = (u_1, \ldots, u_k)$, v and $\pi(\boldsymbol{\theta})$.

2.2.1. Training thresholds $\mathbf{u}$

Choosing a set $\mathbf{u}$ of thresholds for analysis, or an interval $(u_{\min}, u_{\max})$, is the starting point for all the threshold selection methods listed in Section 1.2, apart from those based on the minimization of the mean-squared error of estimators. The thresholds should span the range over which the bias–variance trade-off is occurring. An initial graphical diagnostic, such as a parameter stability plot, can assist this choice.

The highest threshold u_k (or $u_{\max}$) should not be so high that little information is provided about GP parameters. There is no definitive rule for limiting u_k but Jonathan and Ewans (2013) suggested that there should be no fewer than 50 threshold excesses. Applying this rule would restrict u_k to be no higher than the 84% and 92% sample quantiles for the Gulf-of-Mexico and North Sea data sets respectively, but later we shall use u_k that break the rule and examine the consequences.

We shall also set u_1 lower than is typical, to illustrate the effect of the bias–variance trade-off on predictive performance at extreme levels. When selecting a single threshold there is no problem in considering low thresholds that we expect to perform badly: only the best-performing threshold is used and inferences from other thresholds do not affect inferences about extremes.

2.2.2. Validation threshold v

The main additional requirement of our method is the choice of v. We need $v \geqslant u_k$, but the larger v is the fewer excesses of v there are and the smaller the information from data thresholded at v. Consider two validation thresholds: $v_1 = u_k$ and $v_2 > u_k$. If we use v_2 we lose validation information: if $v_1 < x_r \leqslant v_2$ then in equation (3) x_r is censored rather than enters the GP part of the predictive density, and we gain nothing. The prediction of $x_r > v_2$ is unaffected by the choice of v_1 or v_2 because $p_{v_1}\, g(x - v_1; \sigma_{v_1}, \xi) = p_{v_2}\, g(x - v_2; \sigma_{v_2}, \xi)$. Therefore, we should use $v = u_k$, so that v is determined by the highest threshold in $\mathbf{u}$. In some applications results may be sensitive to the choice of u_k. All threshold selection methods involve tuning parameters or assumptions that can have non-negligible effects on results. See the references in Section 1.2 for details.

2.2.3. Generalized Pareto prior distribution $\pi(\theta)$

In Section 2.4 we compare predictive properties of three 'reference' priors for GP parameters. Such priors are intended for use when substantial prior information is not available and it is ex-

pected that information provided by the data will dominate the posterior distribution (O'Hagan, 2006). The general issue of quantifying the relative contributions to a posterior distribution of information from the prior and from the data is an area of current research; see, for example, Reimherr *et al.* (2014). Here we judge the extent to which the data dominate the posterior distribution by using graphical summaries. In Fig. 8 in Section 2.5 we assess sensitivity of the posterior for (σ_u, ξ) to the choice of reference prior distribution and compare the marginal prior and posterior densities of ξ. If the data dominate then a posterior should not be sensitive to the choice of reference prior and the prior density for ξ should be almost flat over the range of ξ for which the posterior density is non-negligible.

For high thresholds it may be that the data do not dominate. Then the use of a reference prior will tend to result in high uncertainty about model parameters and about extrapolations to long future time horizons. If such time horizons are important and the lack of precision is unacceptable then we may wish to incorporate more information, particularly if physically unrealistic extrapolations have resulted. A more considered prior distribution or a model that better represents the physics of the data-generating process could be used. We consider the former strategy in Section 3.3.

2.3. Prediction of extreme observations

In an extreme value analysis the main focus is often the estimation of extreme quantiles called *return levels*. Let M_N denote the largest value observed over a time horizon of N years. The N-year return level $z(N)$ is defined as the value that is exceeded by an annual maximum M_1 with probability $1/N$. In offshore engineering, design criteria are usually expressed in terms of return levels, for values of N such as 100, 1000 or 10000. A related approach defines the quantity of interest as the random variable M_N, rather than particular quantiles of M_1. Under a BGP(p_u, σ_u, ξ) model, for $z > u$,

$$F(z;\boldsymbol{\theta}) = P(X \leqslant z) = 1 - p_u \left\{ 1 + \xi\left(\frac{z-u}{\sigma_u} \right) \right\}^{-1/\xi}.$$

Then $z(N) = z(N;\boldsymbol{\theta})$ satisfies $F\{z(N);\boldsymbol{\theta}\}^{n_y} = 1 - 1/N$, where n_y is the mean number of observations per year. Similarly, for $z > u$, $P(M_N \leqslant z) = F(z;\boldsymbol{\theta})^{n_y N}$. For large N ($N = 100$ is sufficient), $z(N)$ is approximately equal to the 37% quantile of the distribution of M_N (Cox *et al.*, 2002). In an estimative approach, based on a point estimate of $\boldsymbol{\theta}$, the value of $z(N)$ is below the median of M_N. A common interpretation of $z(N)$ is the level that is exceeded on average once every N years. However, for large N (again $N = 100$ is sufficient) and under an assumption of independence at extreme levels, $z(N)$ is exceeded 0, 1, 2, 3, 4 times with respective approximate probabilities of 37%, 37%, 18%, 6% and 1.5%. It may be more instructive to examine directly the distribution of M_N, rather than very extreme quantiles of the annual maximum M_1.

The relationship between these two approaches is less clear under a predictive approach, in which posterior uncertainty about $\boldsymbol{\theta}$ is incorporated in the calculations. The N-year *(posterior) predictive return level* $z_{\mathrm{P}}(N)$ is the solution of

$$P(M_1 \leqslant z_{\mathrm{P}}(N) | \mathbf{x}) = \int F\{z_{\mathrm{P}}(N);\boldsymbol{\theta}\}^{n_y} \pi(\boldsymbol{\theta}|\mathbf{x})\, \mathrm{d}\boldsymbol{\theta} = 1 - 1/N,$$

and the predictive distribution function of M_N is given by

$$P(M_N \leqslant z | \mathbf{x}) = \int F(z;\boldsymbol{\theta})^{n_y N} \pi(\boldsymbol{\theta}|\mathbf{x})\, \mathrm{d}\boldsymbol{\theta}. \tag{8}$$

As noted by Smith (2003), section 1.3, accounting for parameter uncertainty tends to lead to larger estimated probabilities of extreme events, i.e. $z_{\mathrm{P}}(N)$ tends to be greater than an estimate $\hat{z}(N)$ based on, for example, the ML estimator $\hat{\boldsymbol{\theta}}$. The strong non-linearity of $F(z;\boldsymbol{\theta})^{n_y}$ for large z, and the fact that it is bounded above by 1, means that averages of $F(z;\boldsymbol{\theta})^{n_y}$ over areas of the parameter space relating to the extreme upper tail of M_1 tend to be smaller than point values near the centre of such areas. This is less critical when working with the distribution of M_N because now central quantiles of M_N also have relevance, not just particular extreme tail probabilities. Numerical results in Section 2.5 (Fig. 7) show that $z_{\mathrm{P}}(N)$ can be rather greater than the median of the predictive distribution of M_N, particularly when posterior uncertainty about $\boldsymbol{\theta}$ is large.

For a given value of N, we estimate $P(M_N \leqslant z|\mathbf{x})$ by using the sample $\boldsymbol{\theta}_j$, $j = 1, \ldots, m$, from the posterior density $\pi(\boldsymbol{\theta}|\mathbf{x})$ to give

$$\hat{P}(M_N \leqslant z|\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m} F(z;\boldsymbol{\theta}_j)^{n_y N}. \tag{9}$$

The solution $\hat{z}_{\mathrm{P}}(N)$ of $\hat{P}\{M_1 \leqslant \hat{z}_{\mathrm{P}}(N)|\mathbf{x}\} = 1 - 1/N$ provides an estimate of $z_{\mathrm{P}}(N)$.

2.4. Simulation study 1: priors for generalized Pareto parameters

We compare approaches for predicting future extreme observations: a predictive approach using different prior distributions and an estimative approach using the ML estimator. We use Jeffreys's prior $p_u \sim \mathrm{beta}(\frac{1}{2}, \frac{1}{2})$ for p_u, so that $p_u|\mathbf{x} \sim \mathrm{beta}(n_u + \frac{1}{2}, n - n_u + \frac{1}{2})$, where n_u is the number of threshold excesses. Initially we consider three prior distributions for GP parameters: a Jeffreys prior

$$\pi_{\mathrm{J}}(\sigma_u, \xi) \propto \sigma_u^{-1}(1+\xi)^{-1}(1+2\xi)^{-1/2}, \qquad \sigma_u > 0, \quad \xi > -\tfrac{1}{2}, \tag{10}$$

a maximal data information (MDI) prior (Zellner, 1998; Beirlant *et al.*, 2004)

$$\pi_{\mathrm{M}}(\sigma_u, \xi) \propto \sigma_u^{-1}\exp\{-(\xi+1)\}, \qquad \sigma_u > 0, \quad \xi \geqslant -1, \tag{11}$$

truncated from $\xi \in \mathbb{R}$ to $\xi \geqslant -1$, and a flat prior (Pickands, 1994)

$$\pi_{\mathrm{F}}(\sigma_u, \xi) \propto \sigma_u^{-1}, \qquad \sigma_u > 0, \quad \xi \in \mathbb{R}, \tag{12}$$

which is equivalent to placing independent uniform priors on $\log(\sigma_u)$ and ξ. Motivated by findings presented later in this section we generalize prior (11) to an MDI(a) prior:

$$\pi_{\mathrm{M}a}(\sigma_u, \xi; a) \propto \sigma_u^{-1} a \exp\{-a(\xi+1)\}, \qquad \sigma_u > 0, \quad \xi \geqslant -1, \quad a > 0. \tag{13}$$

These priors are improper. Let n_u be the number of threshold excesses. Castellanos and Cabras (2007) showed that the Jeffreys prior yields a proper posterior for $n_u \geqslant 1$ and Northrop and Attalides (2016) showed that under the flat prior a sufficient condition for posterior propriety is $n_u \geqslant 3$. Northrop and Attalides (2016) also showed that, for any sample size, if, and only if, ξ is bounded below *a priori*, the MDI prior, and the generalized MDI prior, yields a proper posterior. The way in which MDI priors are constructed (Zellner (1998), section 2.2) means that the functional form of prior (11) is invariant to the particular lower bound that is chosen. At the particular bound of -1 used in prior (11) the GP distribution reduces to a uniform distribution on $(0, \sigma)$ and corresponds to a change in the behaviour of the GP density: for $\xi < -1$, this density increases without limit as it approaches its mode at the upper end point $-\sigma_u/\xi$, which is behaviour that is not expected in extreme value analyses. The constraint $\xi \geqslant -1$ is also imposed

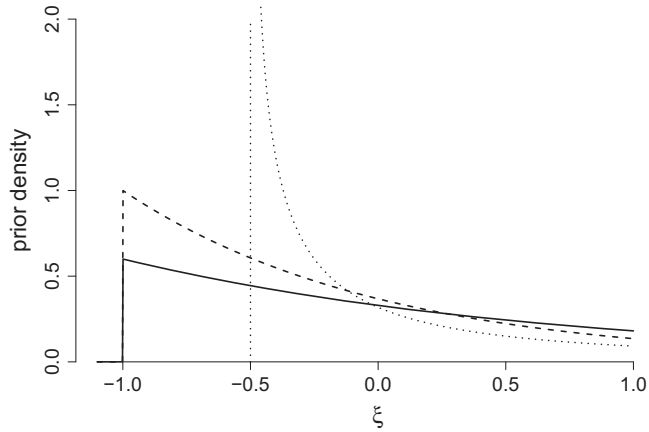

Fig. 3. Jeffreys (· · · · · ·), truncated MDI (- - - - - -) and generalized MDI (———) priors as functions of ξ

in ML estimation for the GP distribution because for $\xi < -1$ the likelihood increases without limit as $-\sigma_u/\xi$ approaches $x_n - u$ (Hosking and Wallis, 1987).

Fig. 3 compares the Jeffreys, MDI and generalized MDI prior (for $a = 0.6$) as functions of ξ. The Jeffreys prior (10) is unbounded as $\xi \downarrow -\frac{1}{2}$. If there are small numbers of threshold excesses this can result in a bimodal posterior distribution, with one mode at $\xi = -\frac{1}{2}$. In this simulation study we also find that the Jeffreys prior results in poorer predictive performance than the truncated MDI and flat priors.

Let Z_{new} be a future N-year maximum, sampled from a distribution with distribution function $F(z;\boldsymbol{\theta})^{n_y N}$. If the predictive distribution function (8) is the same as that of Z_{new} then $P(M_N \leqslant Z_{\text{new}}|\mathbf{x})$ has a $U(0,1)$ distribution. In practice this can only hold approximately: the closeness of the approximation under repeated sampling provides a basis for comparing different prior distributions. Performance of an estimative approach based on the ML estimator $\hat{\boldsymbol{\theta}}$ can be assessed by using $F(Z_{\text{new}};\hat{\boldsymbol{\theta}})^{n_y N}$. For a given prior distribution and given values of N, n_y and n, the simulation scheme is as follows.

Step 1: simulate a data set $\mathbf{x}_{\text{sim}}$ of n independent observations from a BGP(p_u, σ_u, ξ) model and then a sample $\boldsymbol{\theta}_j$, $j = 1, \ldots, m$, from the posterior $\pi(\boldsymbol{\theta}|\mathbf{x}_{\text{sim}})$.
Step 2: simulate an observation z_{new} from the distribution of M_N, i.e. $\max(X_1, \ldots, X_{N_u})$, where $N_u \sim \text{bin}(n_y N, p_u)$ and $X_i \sim^{\text{IID}} \text{GP}(\sigma_u, \xi)$, $i = 1, \ldots, N_u$.
Step 3: use equation (9) to evaluate $\hat{P}(M_N \leqslant z_{\text{new}}|\mathbf{x})$.

Steps 1–3 are repeated 10 000 times, providing a putative sample of size 10 000 from a $U(0, 1)$ distribution. In the estimative approach step 3 is replaced by evaluation of $F(z_{\text{new}};\hat{\boldsymbol{\theta}})^{n_y N}$. Here, and throughout this paper, we produce samples of size m from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$ by using the generalized ratio-of-uniforms method of Wakefield *et al.* (1991), following their suggested strategy of relocating the mode of $\pi(\boldsymbol{\theta}|\mathbf{x})$ to the origin and setting a tuning parameter r to $\frac{1}{2}$. In the simulation studies we use $m = 1000$ and when analysing real data we use $m = 10\,000$.

We assess the closeness of the $U(0, 1)$ approximation graphically (Geweke and Amisano, 2010), comparing the proportion of simulated values in each $U(0, 1)$ decile with the null value of 0.1. To aid the assessment of departures from this value we superimpose approximate pointwise 95% tolerance intervals based on the number of points within each decile having a bin(10 000, 0.1) distribution, i.e. $0.1 \pm 1.96(0.1 \times 0.9/10\,000)^{1/2} = 0.1 \pm 0.006$. We use $p_u \in \{0.1, 0.5\}$, $\sigma_u = 1$ and

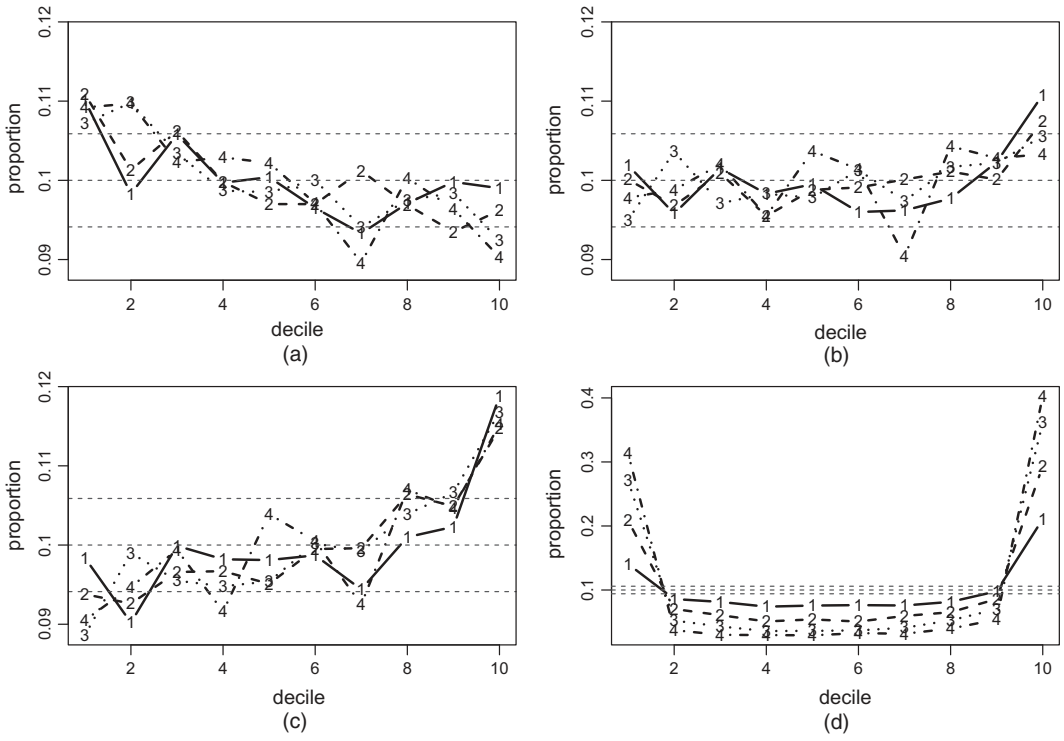

Fig. 4. Proportions of simulated values of $\hat{P}(M_N \leqslant z_{\text{new}}|\mathbf{x})$ falling in $U(0,1)$ deciles for the case $\xi = 0.1$ and $p_u = 0.5$ (1, $N = 100$; 2, $N = 1000$; 3, $N = 10000$; 4, $N = 100000$) (95% limits are superimposed): (a) flat prior; (b) MDI prior; (c) Jeffreys prior; (d) ML estimation

values of ξ suggested approximately in Section 2.5 by the analyses of the Gulf-of-Mexico data ($\xi \approx 0.1$) and the North Sea data ($\xi \approx -0.2$).

The plots in Figs 4 ($\xi = 0.1$; $p_u = 0.5$) and 5 ($\xi = -0.2$; $p_u = 0.1$) are based on simulated data sets of length $n = 500$ and $n_y = 10$, i.e. 50 years of data with a mean of 10 observations per year, for $N = 100, 1000, 10\,000, 100\,000$ years. Note that Figs 4(d) and 5(d) have much wider vertical axis scales than the other plots. It is evident that the estimative approach based on the ML estimator produces too few values in deciles 2–9 and too many in deciles 1 and 10. When the true BGP distribution of M_N (from which z_{new} is simulated in step 2) is wider than that inferred from data (in step 1 and using equation (9)) we expect a surplus of values in the first and last deciles. The estimative approach fails to take account of parameter uncertainty, producing distributions that tend to be too concentrated and resulting in underprediction of large values of z_{new} and overprediction of small values of z_{new}.

The predictive approaches perform much better. Although departures from desired performance are relatively small, and vary with N in some cases, some general patterns appear. In Fig. 4 the flat prior tends to overpredict large values and small values. The MDI prior tends to result in underprediction of large values. The Jeffreys prior underpredicts large values, to a greater extent than the MDI prior, and also tends to underpredict small values. All these tendencies are slightly more pronounced for $\xi = 0.1$ and $p_u = 0.1$ (not shown).

Fig. 5 gives similar findings, although the $N = 100$ case behaves a little differently from those for the larger values of N. The Jeffreys prior is replaced by a control plot based on values sampled from a $U(0, 1)$ distribution. For $\xi = -0.2$ and with small numbers of threshold excesses

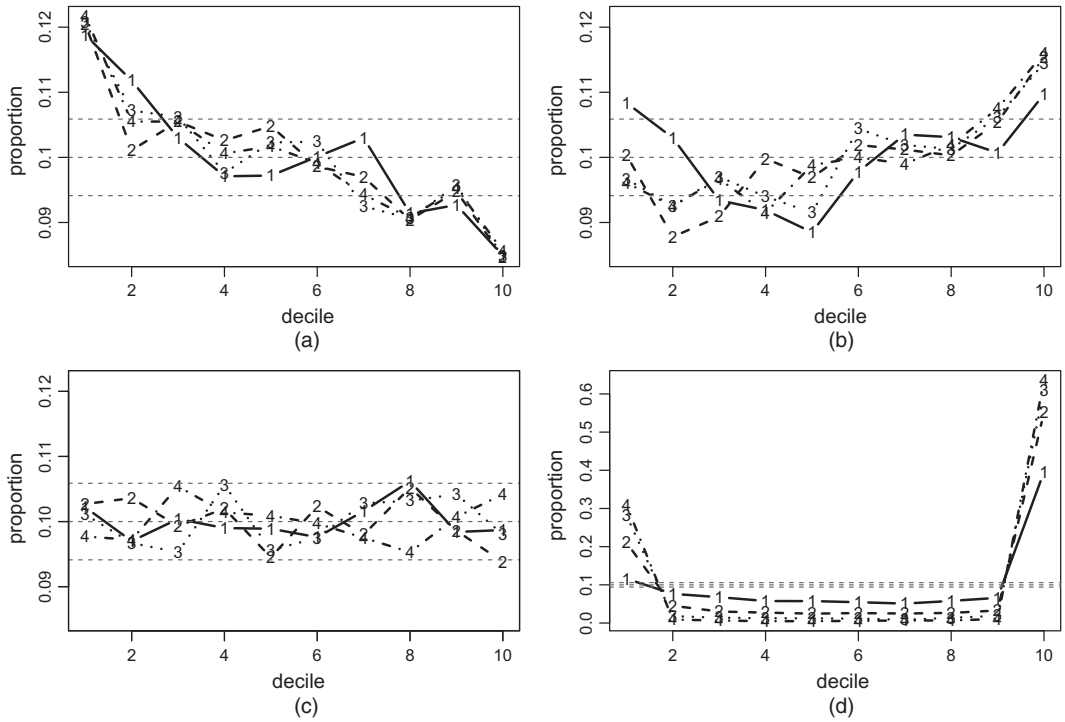

Fig. 5. Proportions of simulated values of $\hat{P}(M_N \leqslant z_{\text{new}}|\mathbf{x})$ falling in $U(0,1)$ deciles for the case $\xi = -0.2$ and $p_u = 0.1$ (1, $N = 100$; 2, $N = 1000$; 3, $N = 10000$; 4, $N = 100000$) (95% limits are superimposed): (a) flat prior; (b) MDI; (c) control plot based on random $U(0,1)$ samples (replacing the Jeffreys prior); (d) ML estimation

the Jeffreys prior occasionally produces a posterior that is also unbounded as $\xi \downarrow -\frac{1}{2}$, making sampling from the posterior difficult.

These results suggest that, in terms of predicting M_N for large N, the MDI prior performs better than the flat prior and the Jeffreys prior. However, a prior for ξ that is in some sense intermediate between the flat prior and the MDI prior could have better properties. To explore this we consider prior (13) for $0 < a \leqslant 1$. Letting $a \to 0$ produces a flat prior for ξ on the interval $[-1, \infty)$. To explore quickly a range of values for a we reuse the posterior samples based on the priors $\pi_{\text{F}}(\sigma, \xi)$ and $\pi_{\text{M}}(\sigma, \xi)$. We use the importance sampling ratio estimator (6) to estimate $P(M_N \leqslant Z_{\text{new}}|\mathbf{x})$ twice: once using $\pi_{\text{F}}(\boldsymbol{\theta}|\mathbf{x})$ as the importance sampling density $h(\boldsymbol{\theta})$ and once using $\pi_{\text{M}}(\boldsymbol{\theta}|\mathbf{x})$. We calculate an overall estimate of $P(M_N \leqslant Z_{\text{new}}|\mathbf{x})$ by using a weighted mean of the two estimates, with weights equal to the reciprocal of the estimated variances of the estimators (Davison (2003), page 603).

Fig. 6 shows plots based on the MDI(0.6) prior. This value of a has been selected on the basis of plots for $a \in \{0.1, 0.2, \ldots, 0.9\}$. We make no claim that this is optimal: just that it is a reasonable compromise between the flat and MDI priors, providing relatively good predictive properties for the cases that we have considered.

2.5. Significant wave height data: single thresholds

We analyse the North Sea and Gulf-of-Mexico storm peak significant wave heights by using the MDI(0.6) prior that was suggested by the simulation study in Section 2.4. We use the methodology that was proposed in Section 2.1 to quantify the performance of various training

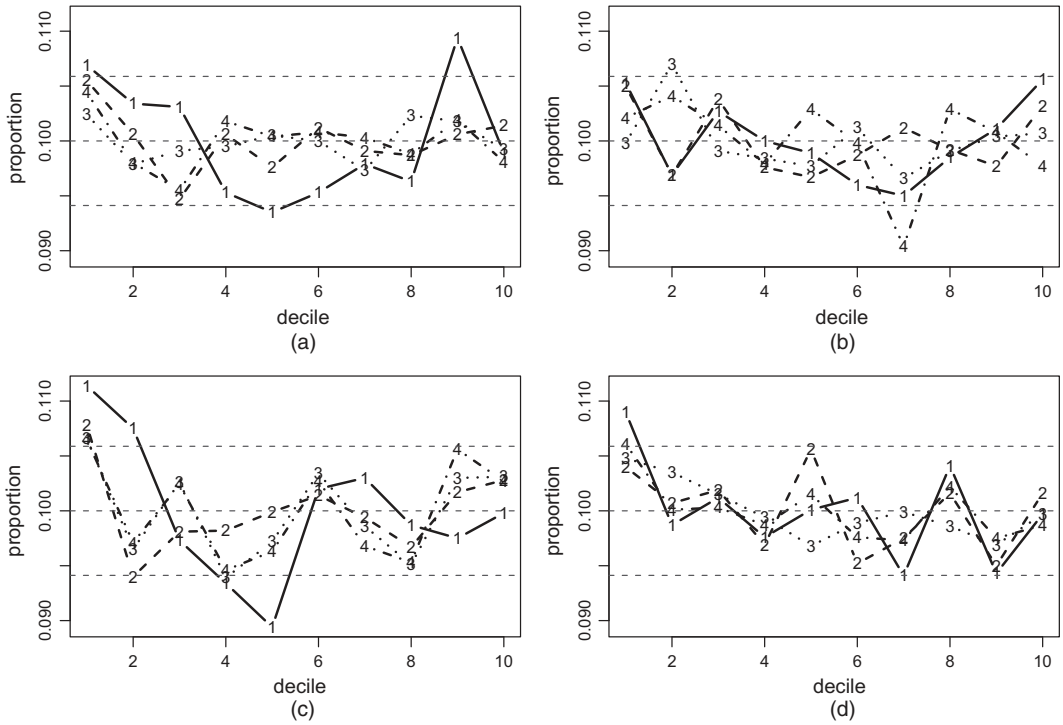

Fig. 6. Proportions of simulated values of $\hat{P}(M_N \leqslant z_{\text{new}}|\mathbf{x})$ falling in $U(0,1)$ deciles for various combinations of ξ and p_u under the MDI(0.6) prior (1, $N = 100$; 2, $N = 1000$; 3, $N = 10000$; 4, $N = 100000$) (95% tolerance limits are superimposed): (a) $\xi = 0.1$, $p_u = 0.1$; (b) $\xi = 0.1$, $p_u = 0.5$; (c) $\xi = -0.2$, $p_u = 0.1$; (d) $\xi = -0.2$, $p_u = 0.5$

thresholds. We use *training thresholds* set at the 0%, 5%, ..., u_k% sample quantiles, for various u_k. We define the estimated *threshold weight* that is associated with training threshold u_i, assessed at *validation threshold* $v\,(=u_k)$, by

$$w_i(v) = \exp\{\hat{T}_v(u_i)\} \Big/ \sum_{j=1}^{k} \exp\{\hat{T}_v(u_j)\}, \tag{14}$$

where $\hat{T}_v(u)$ is defined in equation (7). The ratio $w_2(v)/w_1(v)$, which is an estimate of a *pseudo-Bayes factor* (Geisser and Eddy, 1979), is a measure of the relative performance of threshold u_2 compared with threshold u_1. In Section 3 these weights will be used to combine inferences from different training thresholds.

Figs 7(a) and 7(b) show plots of the estimated training weights against training threshold for various u_k. For the North Sea data, training thresholds in the region of the sample 25–35% quantiles (for which the ML estimate of ξ is approximately –0.2) have relatively large threshold weight and there is little sensitivity to u_k. For the Gulf-of-Mexico data, training thresholds in the region of the 60–70% sample quantiles (for which the ML estimate of ξ is approximately 0.1) are suggested, and the threshold at which the largest weight is attained is more sensitive to u_k. As expected from the histogram of the Gulf-of-Mexico data in Fig. 1, training thresholds below the 25% sample quantile have low threshold weight. Note that for the Gulf-of-Mexico data the 90% and 95% thresholds have far fewer excesses (32 and 16) than the suggested 50.

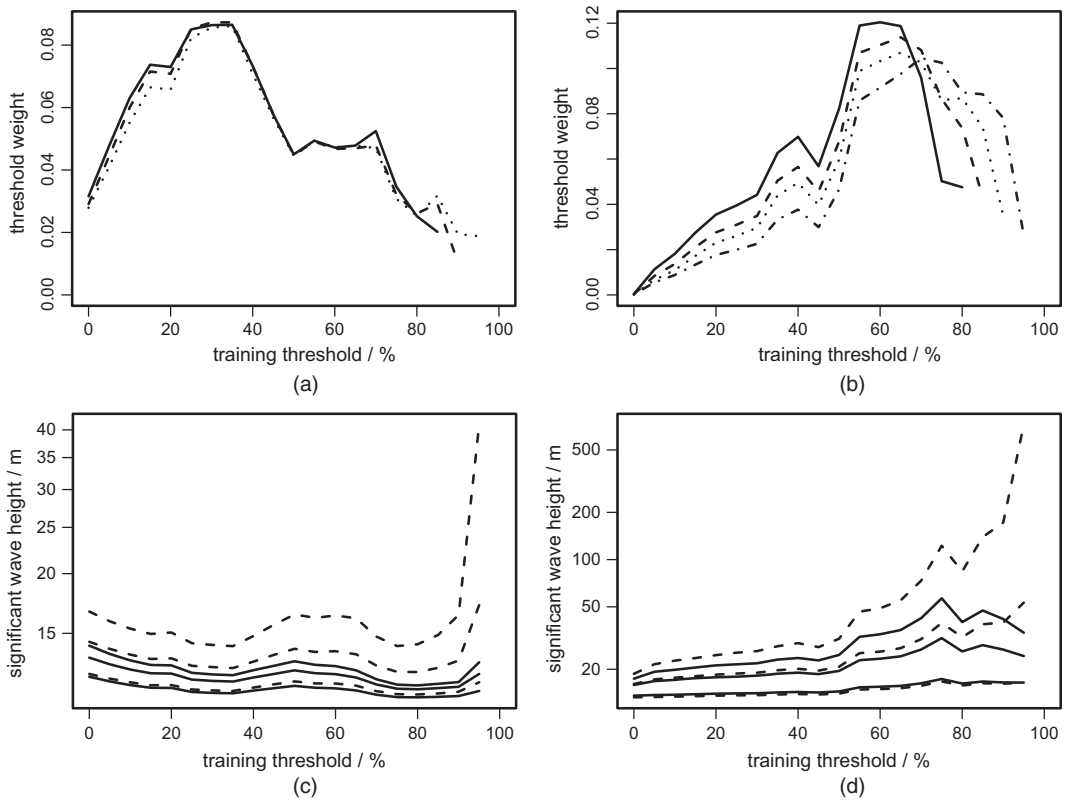

Fig. 7. Analyses of significant wave data by training threshold u: (a) estimated threshold weights by the highest training threshold considered (———, highest u 85%; – – –, highest u 90%; · · · · · ·, highest u 95%); (b) estimated threshold weights by the highest training threshold considered (———, highest u 80%; – – –, highest u 85%; · · · · · ·, highest u 90%; · – · – ·, highest u 95%); (c), (d) N-year predictive return levels (– – –) and medians of the predictive distribution of M_N (———) for $N = 100, 1000, 10000$; (a), (c) North Sea data; (b), (d) Gulf-of-Mexico data

Figs 7(c) and 7(d) show that the N-year predictive return levels and the medians of the predictive distribution of N-year maxima M_N are close for $N = 100$, where little or no extrapolation is required, but for $N = 1000$ and $N = 10\,000$ the former is much greater than the latter. For the North Sea data the results appear sensible and broadly consistent with estimates from elsewhere. From the 55% training threshold upwards, which includes thresholds that have high estimated training weights, estimates of the median of M_{1000} and M_{10000} from the Gulf-of-Mexico data are, in the opinion of experts, implausibly large, e.g. 31.6 m and 56.7 m for the 75% threshold. The corresponding estimates of the predictive return levels are even less credible. High posterior probability of large positive values of ξ, caused by high posterior parameter uncertainty, translates into large predictive estimates of extreme quantiles. That the estimated medians M_{1000} and M_{10000} from the Gulf-of-Mexico data are considered implausible suggests that there is expert prior information that could be included.

Fig. 8 gives examples of the posterior samples of σ_u and ξ underlying the plots in Fig. 7. As we would expect from the fact that quantiles of a GP distribution increase in both σ_u and ξ, these parameters are negatively associated *a posteriori*. The conditional posterior distributions of ξ given σ_u are positively skewed, particularly so for the 95% training thresholds, mainly

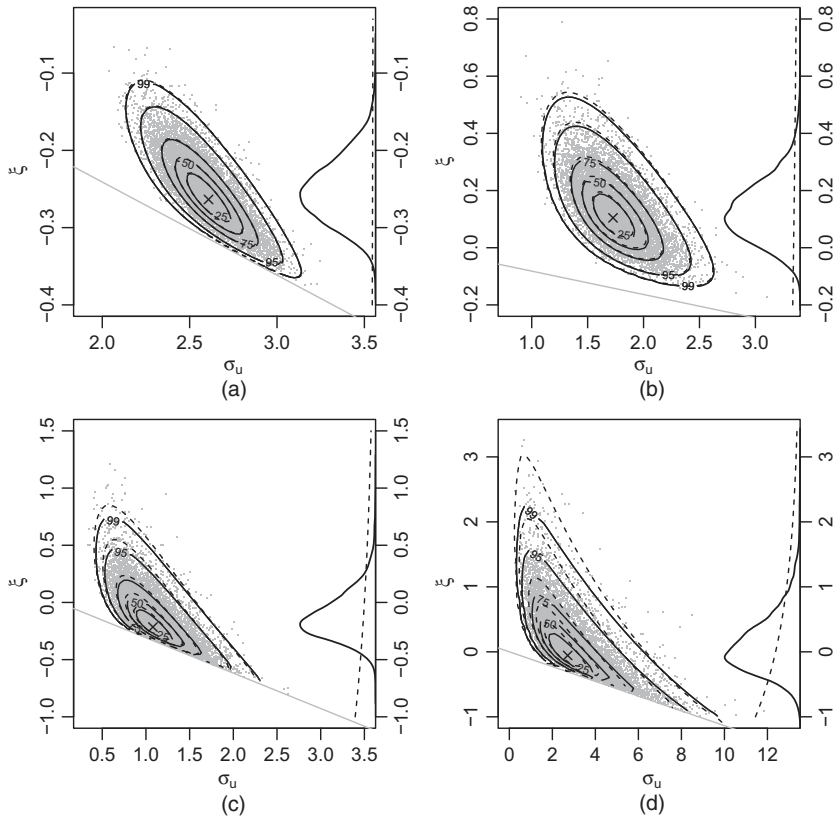

Fig. 8. Samples from $\pi(\sigma_u, \xi|\mathbf{x})$, with 25%, 50%, 75%, 95% and 99% highest posterior density (———) contours and the corresponding contours under the flat prior $\pi_F(\sigma_u, \xi)$ (– – –) (x, posterior mode; ╲, support of the posterior distribution) (on the right-hand axes are plotted the prior (– – –) and posterior (———) marginal densities for ξ): (a) 35% (best) and (b) 65% (best) training threshold (for u_k set at the 85% sample quantile); (c), (d) 95% training threshold; (a), (c) North Sea data; (b), (d) Gulf-of-Mexico data

because, for fixed σ_u, ξ is bounded below by $\sigma_u/(x_n - u)$. The higher the threshold the larger the posterior uncertainty and the greater the skewness towards values of ξ that correspond to a heavy-tailed distribution. For the Gulf-of-Mexico data at the 95% threshold $\hat{P}(\xi > \frac{1}{2} \mid \mathbf{x}) \approx 0.20$ and $\hat{P}(\xi > 1|\mathbf{x}) \approx 0.05$. This issue is not peculiar to a Bayesian analysis: frequentist confidence intervals for ξ and for extreme quantiles are also unrealistically wide.

Fig. 8 also contains posterior contours under the flat prior $\pi_F(\sigma_u, \xi)$. We could also think of these as contours of the likelihood for $(\log(\sigma_u), \xi)$. This change of prior has virtually no effect for the 'best' training thresholds, and little effect when a 95% threshold is used for the North Sea data. When a 95% threshold is used for the Gulf-of-Mexico data the posterior under the flat prior places much greater probability on large positive values of ξ than that under the MDI prior, indicating that at this threshold the data do not dominate the prior. This is also suggested by the marginal posterior density of ξ being far from flat over the effective posterior support of ξ.

Physical considerations suggest that there is a finite upper limit to storm peak H_s (Jonathan and Ewans, 2013), but if there is positive posterior probability on $\xi \geqslant 0$ then the implied distribution of H_s is unbounded above and on extrapolation to a sufficiently long time horizon, N_1 say, unrealistically large values will be implied, i.e., in the absence of information external

to the data, high uncertainty about long extrapolations is to be expected. This may not be a problem if N_1 is greater than the time horizon of practical interest, i.e. the information in the data is sufficient to allow extrapolation over this time horizon. Otherwise, we could incorporate supplementary data (perhaps by pooling data over space as in Northrop and Jonathan (2011)), prior information or a model that better accounts for the physics of the process. A physical characterization of the limiting behaviour of H_s for a given wave environment and bathymetry is not available, but a model based on a mixture of distributions with different tail behaviours (Süveges and Davison, 2012) may provide a useful generalization of a single BGP model. Some practitioners assume that $\xi < 0$ *a priori*, to ensure a finite upper limit, but such a strategy may sacrifice performance at time horizons of importance and produce unrealistically small estimates for the magnitudes of rare events. In Section 3.3 we consider how one might incorporate expert prior information to avoid unrealistic inferences.

3. Accounting for uncertainty in threshold

We use Bayesian model averaging (Hoeting *et al.*, 1999; Gelfand and Dey, 1994) to combine inferences based on different thresholds. Consider a set of k training thresholds $u_1, \ldots, u_k$ and a particular validation threshold v. We view the k BGP models that are associated with these thresholds as competing models. There is evidence that one tends to obtain better predictive performance by interpolating smoothly between all models entertained as plausible *a priori*, than by choosing a single model (Hoeting *et al.* (1999), section 7). Suppose that we specify prior probabilities $P(u_i)$, $i = 1, \ldots, k$, for these models. In the absence of more specific prior information, and in common with Wadsworth and Tawn (2012), we use a discrete uniform prior $P(u_i) = 1/k$, $i = 1, \ldots, k$. We suppose that the thresholds occur at quantiles that are equally spaced on the probability scale. We prefer this to equal spacing on the data scale because it seems more natural than an equal spacing on the data scale and retains its property of equal spacing under data transformation.

Let $\boldsymbol{\theta}_i = (p_i, \sigma_i, \xi_i)$ be the BGP parameter vector under model u_i, under which the prior is $\pi(\boldsymbol{\theta}_i | u_i)$. By Bayes's theorem, the *posterior threshold weights* are given by

$$P_v(u_i|\mathbf{x}) = \frac{f_v(\mathbf{x}|u_i)\, P(u_i)}{\sum_{j=1}^{k} f_v(\mathbf{x}|u_j)\, P(u_j)},$$

where $f_v(\mathbf{x}|u_i) = \int f_v(\mathbf{x}|\boldsymbol{\theta}_i, u_i)\, \pi(\boldsymbol{\theta}_i|u_i)\, \mathrm{d}\boldsymbol{\theta}_i$ is the prior predictive density of $\mathbf{x}$ based on validation threshold v under model u_i. However, $f_v(\mathbf{x}|u_i)$ is difficult to estimate and is improper if $\pi(\boldsymbol{\theta}_i|u_i)$ is improper. Following Geisser and Eddy (1979) we use $\Pi_{r=1}^{n} f_v(x_r|\mathbf{x}_{(r)}, u_i) = \exp\{\hat{T}_v(u_i)\}$ as a surrogate for $f_v(\mathbf{x}|u_i)$ to give

$$\hat{P}_v(u_i|\mathbf{x}) = \frac{\exp\{\hat{T}_v(u_i)\}\, P(u_i)}{\sum_{j=1}^{k} \exp\{\hat{T}_v(u_j)\}\, P(u_j)}. \tag{15}$$

Let $\boldsymbol{\theta}_{ij}$, $j = 1, \ldots, m$, be a sample from $\pi(\boldsymbol{\theta}_i|\mathbf{x})$, the posterior distribution of the GP parameters based on threshold u_i. We calculate a threshold-averaged estimate of the predictive distribution function of M_N by using

$$\hat{P}_v(M_N \leqslant z|\mathbf{x}) = \sum_{i=1}^{k} \hat{P}(M_N \leqslant z|\mathbf{x}, u_i)\, \hat{P}_v(u_i|\mathbf{x}), \tag{16}$$

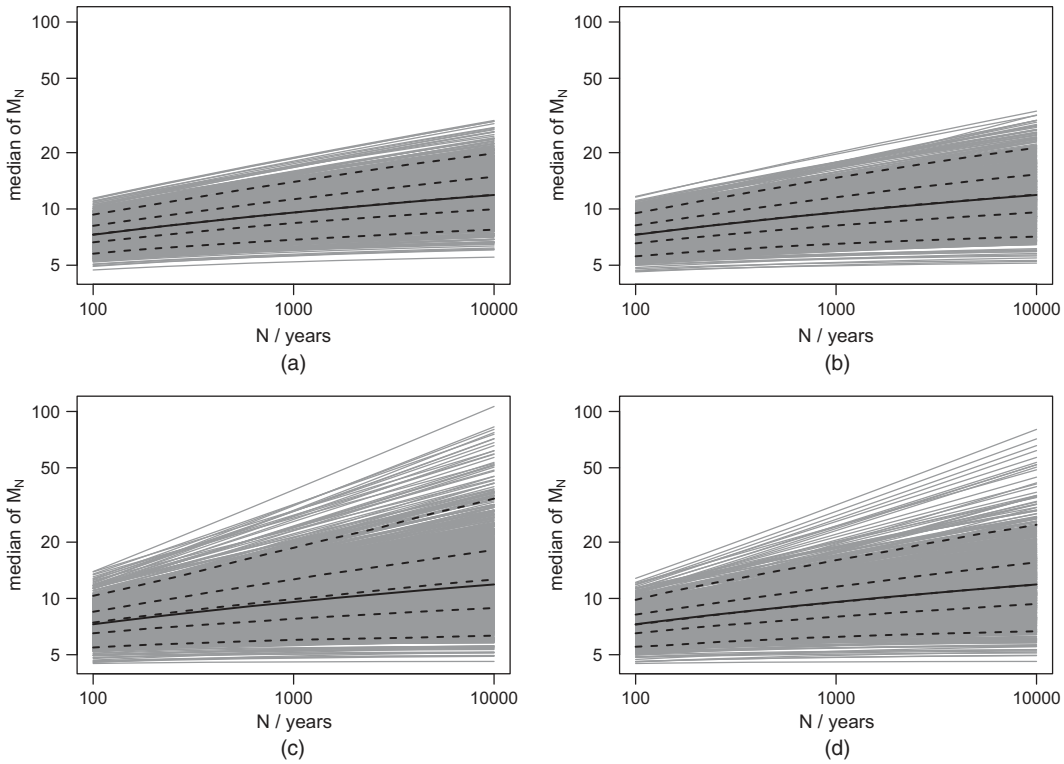

Fig. 9. Predictive medians of M_N by N for the exponential example (———, individual data sets; – – –, N-specific 5%, 25%, 50%, 75% and 95% sample quantiles; ———, true median): (a) median threshold strategy; (b) threshold-averaged threshold strategy; (c) 90% quantile threshold strategy; (d) 'best' single-threshold strategy

where, by analogy with equation (9), $\hat{P}(M_N \leqslant z|\mathbf{x}, u_i) = (1/m)\Sigma_{j=1}^{m} F(z;\boldsymbol{\theta}_{ij})^{n_y N}$. The solution $\hat{z}_{\mathrm{PM}}(N)$ of

$$\hat{P}_v\{M_1 \leqslant \hat{z}_{\mathrm{PM}}(N)|\mathbf{x}\} = 1 - 1/N \tag{17}$$

provides a threshold-averaged estimate of the N-year predictive return level, based on validation threshold v. All training thresholds with non-zero prior probability contribute to inferences, with thresholds producing relatively good predictive performance at extreme levels having greater influence than those with weaker performance.

3.1. Simulation study 2: single and multiple thresholds

We compare inferences from a single threshold with those from averaging over many thresholds, based on random samples simulated from three distributions, chosen to represent qualitatively different behaviours. With knowledge of the simulation model we should be able to choose a suitable single threshold, at least approximately. In practice this would not be so and so it is interesting to see how well the strategies of choosing the 'best' threshold u^* (Section 2), and of averaging inferences over different thresholds (Section 3), compare with this choice and how the estimated weights $\hat{P}_v(u_i|\mathbf{x})$ in equation (15) vary over u_i.

The three distributions are now described. A (unit) exponential distribution has the property that a GP(1,0) model holds above any threshold. Therefore, choosing the lowest available

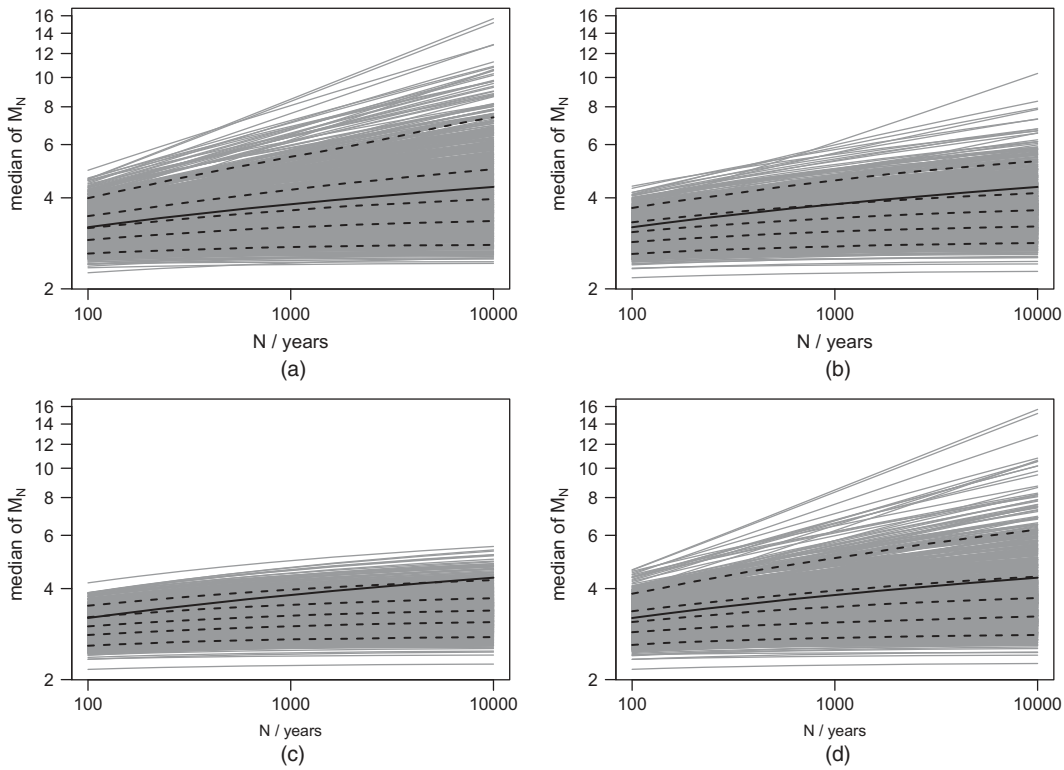

Fig. 10. Predictive medians of M_N by N for the normal distribution example (———, individual data sets; — — —, N-specific 5%, 25%, 50%, 75% and 95% sample quantiles; ———, true median): (a) 90% quantile threshold strategy; (b) threshold-averaged strategy; (c) median threshold strategy; (d) 'best' single-threshold strategy

threshold is optimal. For a (standard) normal distribution the GP model does not hold for any finite threshold, the quality of a GP approximation improving slowly as the threshold increases. In the limit $\xi = 0$, but at finite levels the effective shape parameter is negative (Wadsworth and Tawn, 2012) and we expect a relatively high threshold to be indicated. A uniform–GP hybrid has a constant density up to its 75% quantile and a GP density (here with $\xi = 0.1$) for excesses of the 75% quantile. Thus, a GP distribution holds only above the 75% threshold.

In each case we simulated 1000 samples each of size 500, representing 50 years of data with an average of 10 observations per year. We set training thresholds at the 50%, 55%, ..., 90% sample quantiles, so that there are 50 excesses of the (90%) validation threshold. For each sample, and for values of N between 100 and 10000, we solved $\hat{P}_v(M_N \leqslant z|\mathbf{x}) = \frac{1}{2}$ for z (see equation (9)) to give estimates of the median of M_N. We show results (in Figs 9–11) for three single thresholds: the threshold that we might choose based on knowledge of the simulation model (Figs 9(a), 10(a) and 11(a)), the 'best' threshold u^* (Figs 9(d), 10(d) and 11(d); see Section 2) and another (clearly suboptimal) threshold chosen to facilitate further comparisons (Figs 9(c), 10(c) and 11(c)). We compare these estimates, and a threshold-averaged estimate based on equation (16) with the true median of M_N, $H^{-1}\{(\frac{1}{2})^{10N}\}$, where H is the distribution function of the underlying simulation model.

The results for the exponential distribution are summarized in Fig. 9. As expected, all strategies have negligible bias. The threshold-averaged estimates match closely the behaviour of the

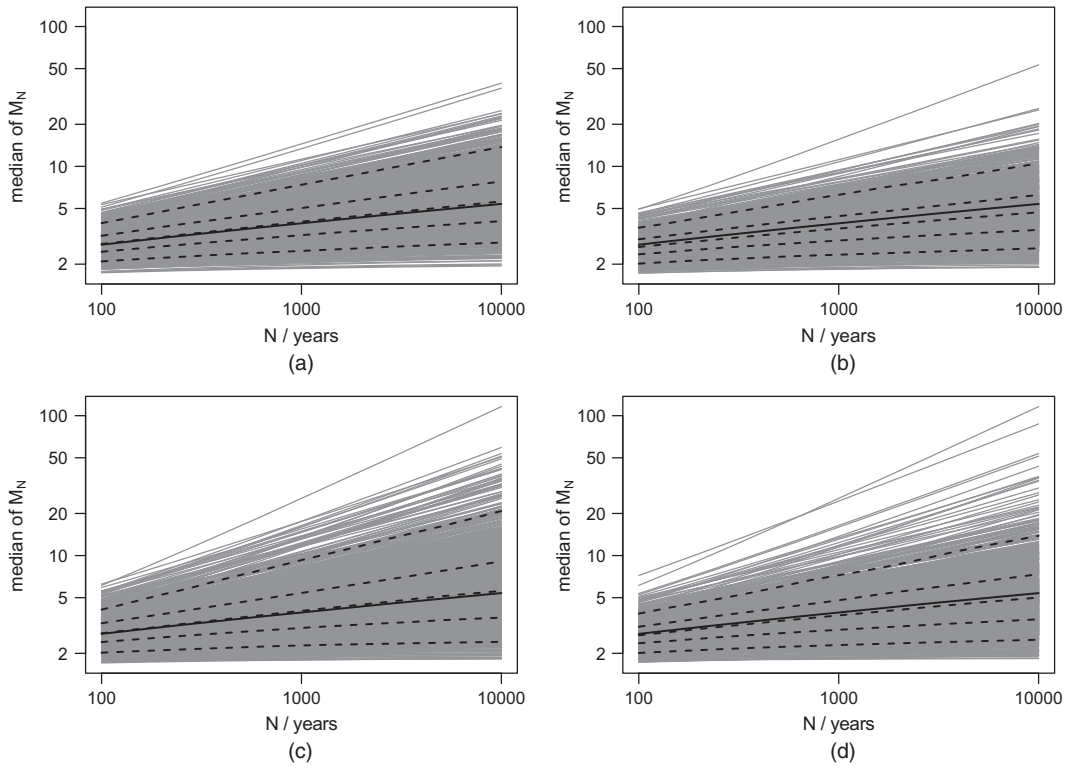

Fig. 11. Predictive median of M_N by N for the uniform–GP hybrid example (———, individual data sets; — — —, N-specific 5%, 25%, 50%, 75% and 95% sample quantiles; ———, true median): (a) 75% quantile threshold strategy; (b) threshold-averaged threshold strategy; (c) 90% quantile threshold strategy; (d) 'best' single-threshold strategy

optimal strategy (the 50% threshold). The best single threshold results in slightly greater variability, offering less protection than threshold averaging against estimates that are far from the truth. In the normal case (Fig. 10) the expected underestimation is evident for large N: this is substantial for a 50% threshold but small for a 90% threshold. The cross-validation-based strategies have greater bias than those based on a 90% threshold, because inferences from lower thresholds contribute, but have much smaller variability. Similar findings are evident in Fig. 11 for the uniform–GP hybrid distribution: contributions from thresholds lower than the 75% quantile produce negative bias but threshold averaging achieves lower variability than the optimal 75% threshold.

In all these examples the cross-validation-based strategies seem preferable to a poor choice of a single threshold, and, in a simple visual comparison of bias and variability, are not dominated clearly by a (practically unobtainable) optimal threshold. Using threshold averaging to account for threshold uncertainty is conceptually attractive but, the exponential example aside, compared with the best-threshold strategy its reduction in variability is at the expense of slightly greater bias. A more definitive comparison would depend on problem-dependent losses associated with overestimation and underestimation.

Fig. 12 summarizes how the posterior threshold weights vary with training threshold. For a few data sets the 90% training threshold receives highest weight. This occurs when inferences about ξ by using a 90% threshold differ from those by using each lower threshold. This effect diminishes if the number of excesses in the validation set is increased. In the exponential and

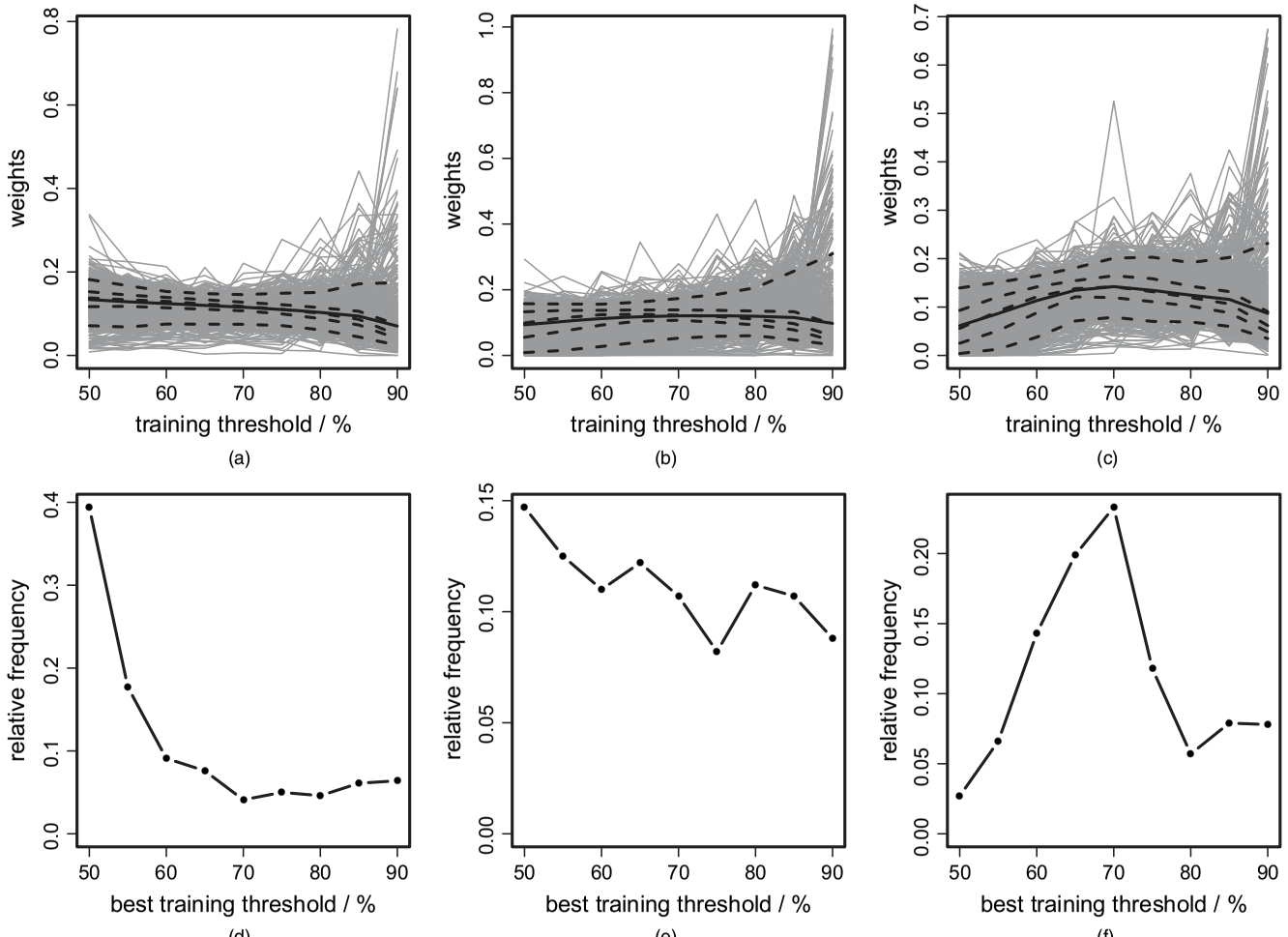

Fig. 12. Threshold weights by training threshold: (a)–(c) individual data sets (——) with threshold-specific sample means (———) and 5%, 25%, 50%, 75% and 95% sample quantiles (– – –); (d)–(f) relative frequency with which each threshold has the largest weight; (a), (d) exponential distribution; (b), (e) normal distribution; (c), (f) uniform–GP hybrid

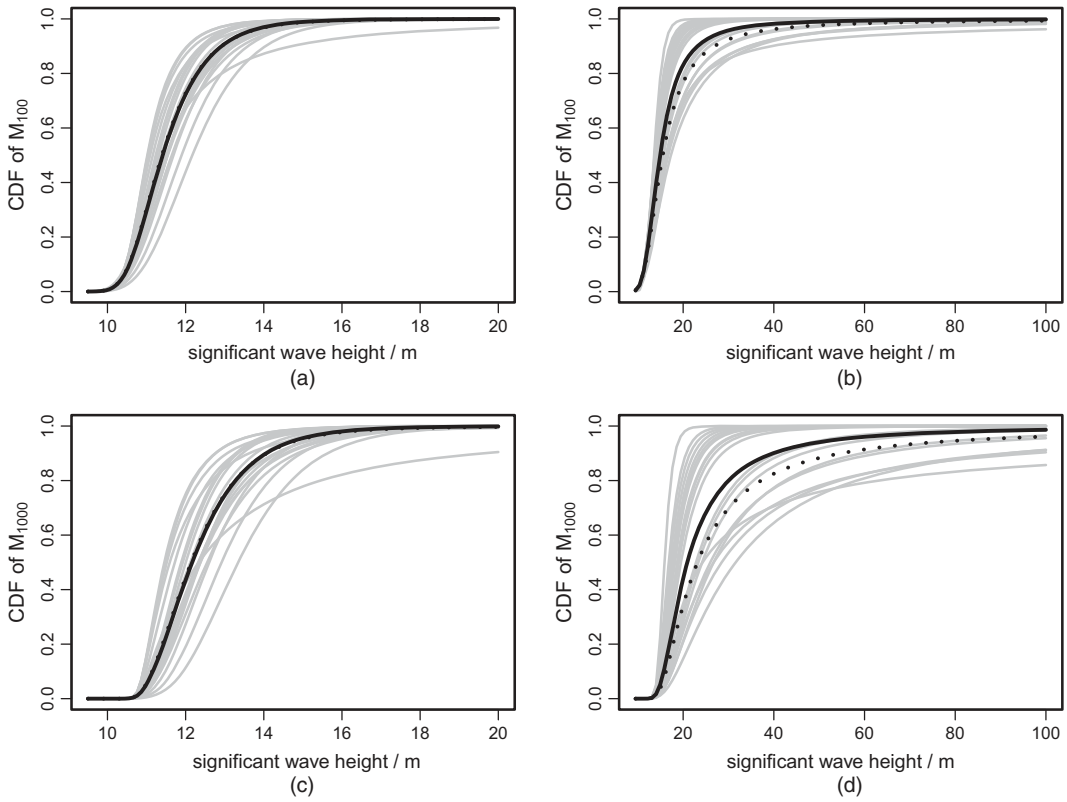

Fig. 13. Threshold-specific (——) and threshold-averaged (———, · · · · · ·) predictive distribution functions of (a), (b) M_{100} and (c), (d) M_{1000}: (a), (c) North Sea data (——, highest u 85%; · · · · · ·, highest u 95%); (b), (d) Gulf-of-Mexico data (———, highest u 80%; · · · · · ·, highest u 95%)

hybrid cases the average weights behave as expected: decreasing in u in the exponential case, and peaking at approximately the 70% quantile (i.e. lower than the 75% quantile) in the uniform–GP case. In the exponential example the best available threshold (the 50% quantile) receives the highest weight with relatively high probability. In the hybrid example the 70% quantile receives the highest weight most often. The 75% quantile is the lowest threshold at which the GP model for threshold excesses is correct. The 70% quantile performs better than the 75% quantile by trading some model misspecification bias for increased precision resulting from larger numbers of threshold excesses. In the normal distribution case there is no clear-cut optimal threshold. This is reflected in the relative flatness of the graphs, with the average weights peaking at approximately the 70–80% quantile and the 50% threshold being the best slightly more often than higher thresholds. Given the slow convergence in this case it may be that much higher thresholds should be explored, requiring much larger simulated sample sizes, such as those used by Wadsworth and Tawn (2012).

3.2. Significant wave height data: threshold uncertainty

We return to the significant wave height data sets, using the methodology of Section 3 to average extreme value inferences obtained from different thresholds. We use the full set of training thresholds given in Section 2.5, but the influence on inferences of particular thresholds, e.g. the

very lowest thresholds, could be eliminated completely by setting to 0 their prior probabilities. Fig. 13 shows the estimated threshold-specific predictive distribution functions of M_{100} and M_{1000}. Also plotted are estimates from the weighted average (16) over thresholds, for various choices of the highest threshold u_k. For the North Sea data there is so little sensitivity to u_k that the black curves are indistinguishable. For the Gulf-of-Mexico data there is greater sensitivity to u_k, although on the basis of the discussion in Section 2.2 setting u_k at the 95% sample quantile is probably inadvisable with only 315 observations. However, for both choices of u_k, averaging inferences over thresholds has provided some protection against the high probability of unrealistically large values of H_s estimated under some individual thresholds.

3.3. An informative prior

We have used prior distributions for model parameters that are constructed without reference to the particular problem in hand. This strategy is inadvisable when the data contain insufficient information to dominate such priors, because inferences are then influenced strongly by a generically chosen prior. In the analysis of the Gulf-of-Mexico data in Section 2.5 we saw that for the highest thresholds unrealistic extreme value extrapolations were produced at long time horizons. Sensitivity of posterior inferences to the choice of reference prior suggests that this is at least partly caused by a lack of information in the data. If small sample sizes cannot be avoided and long time horizons are important then unrealistic inferences can be avoided by providing application-specific prior information. This prior could be elicited from an expert (Coles and Tawn, 1996; Stephenson, 2016), or specified to reflect general experience of the quantity under study, such as the beta-type prior for ξ on $-\frac{1}{2} \leqslant \xi \leqslant \frac{1}{2}$ that was used by Martins and Stedinger (2001) for river flows and rainfall totals.

We illustrate the effects on the Gulf-of-Mexico analysis of providing expert information, with the aim of preventing unrealistic inferences. Let m_N be the median of M_N. Oceanographers with knowledge of the hurricane-induced storms in the Gulf of Mexico suggest approximate values of 15 m for m_{100} and 1.5 for the ratio m_{10000}/m_{100}, i.e. a value of 22.5 m for m_{10000}. Assuming independence of distinct annual maxima $P(M_{10000} \leqslant 22.5) = \frac{1}{2}$ implies that $P(M_{100} \leqslant 22.5) = (\frac{1}{2})^{1/100} \approx 0.993$. The experts also assert that M_{100} is unlikely to exceed 20 m, so we take $P(M_{100} \leqslant 20) = 0.9$.

Let $r_q = P(M_{100} \leqslant q)$. We use Crowder (1992) to specify a prior distribution for $(r_{q_1}, r_{q_2}, r_{q_3})$, for quantiles $q_1 < q_2 < q_3$. Here $(q_1, q_2, q_3) = (15, 20, 22.5)$ m. A Dirichlet($\boldsymbol{\alpha}$) distribution (Kotz *et al.*, 2000), where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, is placed on $(r_{q_1}, r_{q_2} - r_{q_1}, r_{q_3} - r_{q_2}, 1 - r_{q_3})$, from which it follows that, marginally, $r_{q_i} \sim \text{beta}(\Sigma_{j=1}^{i} \alpha_j, \Sigma_{j=i+1}^{4} \alpha_j)$, $i = 1, 2, 3$. We set $\boldsymbol{\alpha}$ so that the prior modes of $(r_{q_1}, r_{q_2}, r_{q_3})$ are $(0.5, 0.9, 0.993)$ and r_{q_1} lies in $(0.25, 0.75)$ with probability 0.99. This gives $\boldsymbol{\alpha} = (11.77, 8.62, 2.01, 1.15)$.

Following (Stephenson, 2016), and on the basis of theory concerning the limiting behaviour of the maximum of independent and identically distributed random variables (Coles (2001), chapter 3), we suppose that M_N (with $N = 100$ here) has a generalized extreme value GEV(μ, σ, ξ) distribution, so

$$r_q = \exp[-\{1 + \xi(q - \mu)/\sigma\}_+^{-1/\xi}] = F_{\text{GEV}}(q). \tag{18}$$

The prior for $(r_{q_1}, r_{q_2}, r_{q_3})$ implies a prior for $\boldsymbol{\phi} = (\mu, \sigma, \xi)$. For a given threshold u we require a prior distribution for the BGP parameters $\boldsymbol{\theta} = (p_u, \sigma_u, \xi)$, where $p_u = 1 - F_{\text{GEV}}(u)^{1/n_y N}$ and $\sigma_u = \sigma + \xi(u - \mu)$. Further transformation from $\boldsymbol{\phi}$ to $\boldsymbol{\theta}$ gives this prior as

$$\pi(\boldsymbol{\theta}) \propto J_1(\boldsymbol{\phi})\, J_2(\boldsymbol{\theta}) \prod_{i=1}^{4} (r_{q_i} - r_{q_{i-1}})^{\alpha_i - 1}, \qquad 0 < p_u < 1, \quad \sigma_u > 0, \quad \xi > -\sigma_u/(q_3 - u), \tag{19}$$

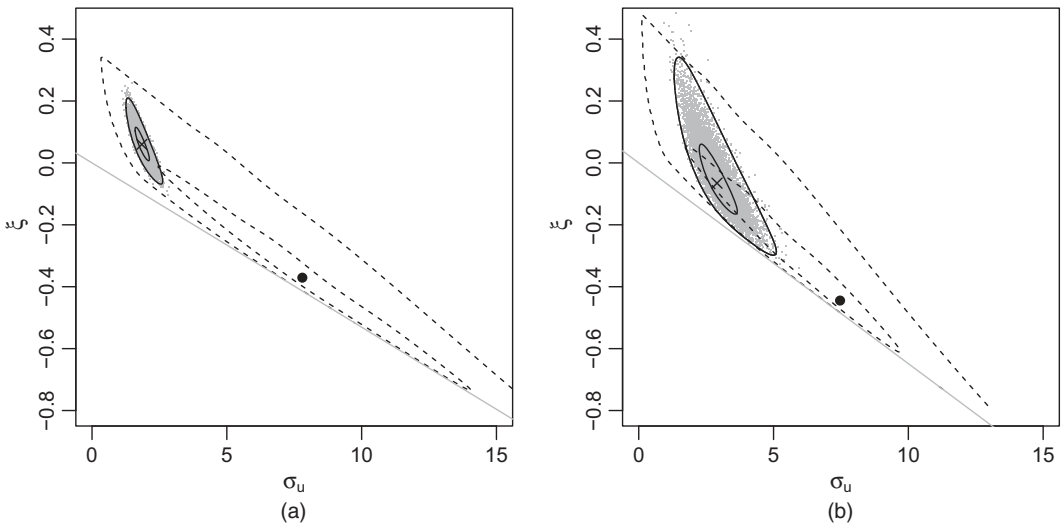

Fig. 14. Samples from the marginal posterior density of (σ_u, ξ) with 50% and 95% highest posterior density contours (———) and prior density contours (------) (×, posterior mode; •, prior mode; ╲, support of the posterior distribution): (a) 65% threshold; (b) 95% threshold

where $r_{q_0} = 0$ and $r_{q_4} = 1$, and $J_1(\phi)$ and $J_2(\boldsymbol{\theta})$ are the respective Jacobians of the transformations from $(r_{q_1}, r_{q_2}, r_{q_3})$ to ϕ and from ϕ to $\boldsymbol{\theta}$. It can be shown that

$$J_1(\phi) = \sigma\xi^{-2}\left\{\prod_{i=1}^{3} f_{\mathrm{GEV}}(q_i)\right\}\left|\sum_{i,j\in\{1,2,3,\},i<j}(-1)^{i+j+1}(t_i t_j)^{-\xi}\log(t_j/t_i)\right|, \tag{20}$$

$$J_2(\boldsymbol{\theta}) = \sigma_u(n_y N)^{\xi}(1-p_u)^{-1}\{-\ln(1-p_u)\}^{\xi-1} \tag{21}$$

where $t_i = -\log(r_{q_i})$ and $f_{\mathrm{GEV}}(q_i) = t_i^{1+\xi}\exp(-t_i)/\sigma$ is the density function of a GEV(μ, σ, ξ) distribution.

This construction results in priors for BGP parameters whose marginal distributions and dependence structures reflect the expert probabilistic statements regarding extreme quantiles. The (broken) contours of marginal prior densities for (σ_u, ξ) in Fig. 14 show that the marginal prior distributions of σ_u and ξ are quite diffuse but the prior information induces negative association between σ_u and ξ. Relatively to the MDI(0.6) prior that was used in Section 2.5 the informative prior downweights parameter combinations corresponding to extrapolations that are substantially larger or smaller than expected by the experts.

Fig. 14 also contains graphical summaries of the marginal posterior distribution of σ_u and ξ under the informative prior for analyses of the Gulf-of-Mexico data by using 65% and 95% training thresholds, for comparison with the plots under the MDI(0.6) prior in Figs 8(b) and 8(d). The posterior distributions under the informative prior are less diffuse, with lower posterior probability on large positive values of ξ, and exhibit stronger negative association between σ_u and ξ. As expected, the change of prior has had a greater effect at the higher of these two thresholds, but for both thresholds the disparity between the prior and the posterior suggests that the data have a meaningful influence on the inferences.

Figs 15(a) and 15(b) show, by comparison with Figs 7(b) and 7(d), the effect of the change of prior on the threshold weights and the threshold-specific predictive extreme value inferences.

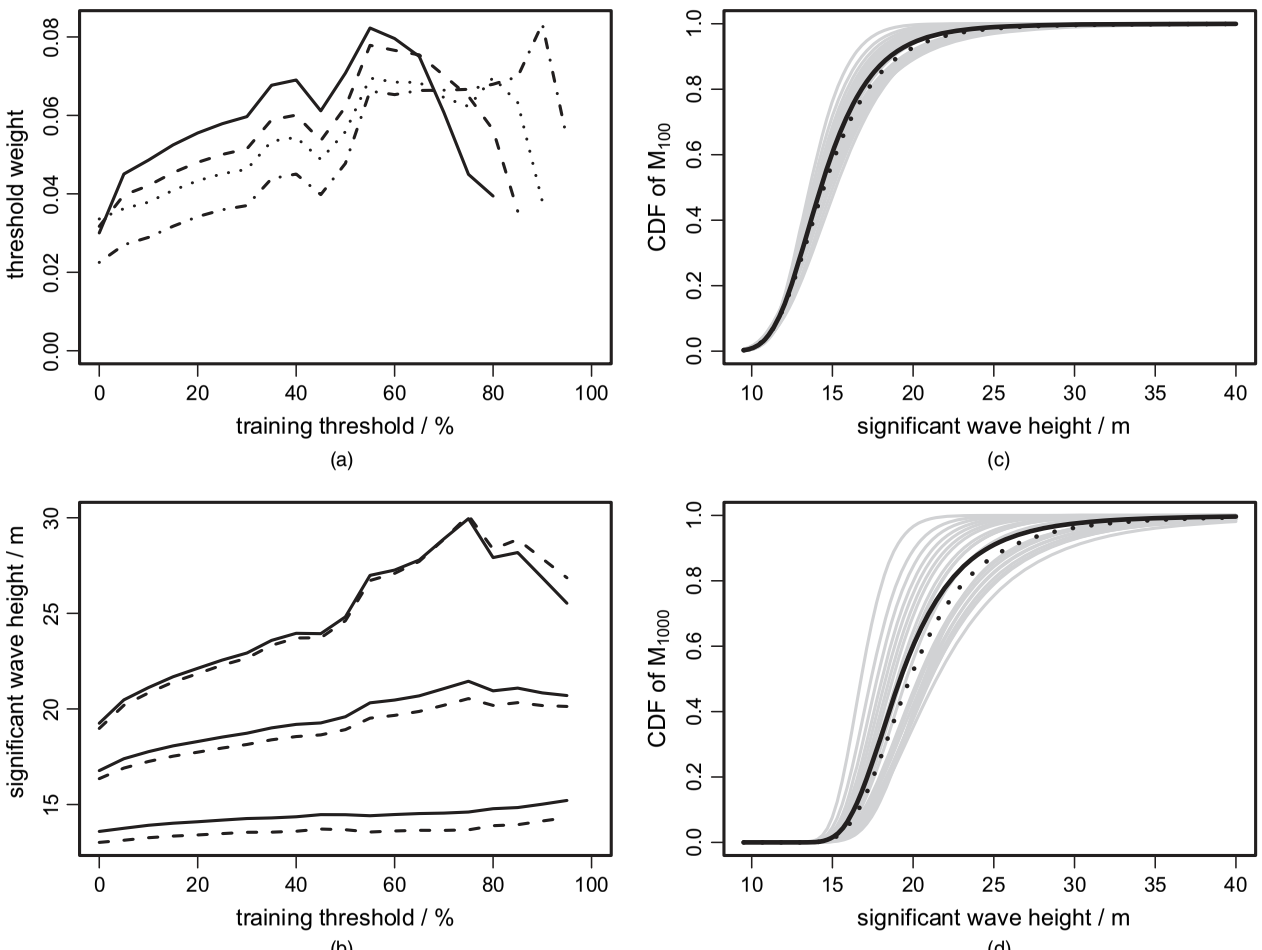

Fig. 15. Extreme value inferences for the Gulf-of-Mexico data using an informative prior: (a) estimated threshold weights by the highest training threshold u (———, 80%; – – –, 85%; · · · · · ·, 90%; · – · –, 95%); (b) N-year predictive return levels (– – –) and medians (———) of the predictive distribution of M_N for $N = 100, 1000, 10000$; (c), (d) threshold-specific (———) and threshold-averaged (———, highest u 80%; · · · · · ·, highest u 95%) predictive distribution functions of M_{100} and M_{1000} respectively

The general pattern of the weights is similar under both priors but the relative performance of the lowest thresholds has improved. With little prior information the posterior distributions that are produced by these thresholds are relatively precise but have locations for ξ that are rather smaller than those at the best-performing thresholds. This behaviour can be seen in Fig. 2(b). The use of the informative prior increases these posterior locations sufficiently to improve performance of the lowest thresholds. In fact, the use of the informative prior has improved predictive performance, as measured by $\hat{T}_v(u)$ in equation (7), for all u and v. Another change relative to Fig. 7 is the anomalously high weight at the 90% training threshold if validation is performed at the 95% sample quantile. As discussed earlier, in practice we would not use such a high validation threshold as it produces only 16 excesses. Fig. 15(b) shows that for the highest thresholds the informative prior has prevented the very unrealistic estimates that were obtained under the MDI(0.6) prior. This can also be seen by comparing Figs 13(b) and 13(d) and 15(c) and 15(d): the grey curves corresponding to high thresholds have shifted to the left, i.e. towards giving higher density to smaller values of M_N, with a similar knock-on effect on the threshold-averaged black curves.

4. Discussion

We have proposed new methodology for extreme value threshold selection based on a GP model for threshold excesses. It can be used either to inform the choice of a best single threshold or to reduce sensitivity to a particular choice of threshold by averaging extremal inferences from several thresholds, weighting thresholds with better cross-validatory predictive performance more heavily than those with poorer performance. The simulation study in Section 3.1 shows that the estimated threshold weights behave as expected in cases where the GP model holds exactly above some threshold and illustrates the potential benefit of averaging different estimated tail behaviours to perform extreme value extrapolation.

The methodology has been applied to significant wave height data sets from the northern North Sea and the Gulf of Mexico. For the latter data set the highest thresholds result in physically unrealistic extrapolation to long future time horizons. Averaging inferences over different thresholds avoids basing inferences solely on one of these thresholds, but we also explored how the incorporation of basic prior information can be used to address this problem. Stronger prior information about GP model parameters, or indeed prior information about the threshold level itself, could also be used.

In common with all existing threshold selection methods some subjective input is required. These inputs are discussed in detail in Section 2.2. The main requirement of our methodology is the choice of the highest training threshold to be considered, as this is also the validation threshold at which extreme value predictions from different training thresholds are compared.

The fact that our methodology is based on inferences from standard unmodified extreme value models makes it relatively amenable to generalization. In significant wave height examples that are considered in this paper it is standard to extract event maxima from raw data, thereby producing observations that are treated as approximately independent. Otherwise, data may exhibit short-term temporal dependence at extreme levels, leading to clusters of extremes. In on-going work we are extending our general approach to this situation and to deal with other important issues: the presence of covariate effects, the choice of measurement scale and inference for multivariate extremes. Another possibility is to work with the GEV parameterization of the point process approach of Smith (1989) so that the rate of threshold exceedance is modelled jointly with the tail characteristics.

Acknowledgements

We thank Richard Chandler, Kevin Ewans, Tom Fearn and Steve Jewson for helpful comments. Nicolas Attalides was funded by an Engineering and Physical Sciences Research Council studentship while carrying out this work. We are grateful to the Joint Editor and two reviewers for comments that led to improvements in the paper.

References

Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004) *Statistics of Extremes: Theory and Applications*. Oxford: Oxford University Press.

Caeiro, F. and Gomes, M. I. (2016) Threshold selection in extreme value analysis. In *Extreme Value Modeling and Risk Analysis: Methods and Applications* (eds D. K. Dey and J. Yan), pp. 69–86. London: Chapman and Hall.

Cardone, V. J., Callahan, B. T., Chen, H., Cox, A. T., Morrone, M. A. and Swail, V. R. (2015) Global distribution and risk to shipping of very extreme sea states (VESS). *Int. J. Clim.*, **35**, 69–84.

Castellanos, E. M. and Cabras, S. (2007) A default Bayesian procedure for the generalized Pareto distribution. *J. Statist. Planng Inf.*, **137**, 473–483.

Coles, S. G. (2001) *An Introduction to Statistical Modelling of Extreme Values*. London: Springer.

Coles, S. G. and Tawn, J. A. (1996) A Bayesian analysis of extreme rainfall data. *Appl. Statist.*, **45**, 463–478.

Cox, D. R., Isham, V. S. and Northrop, P. J. (2002) Floods: some probabilistic and statistical approaches. *Philos. Trans. R. Soc. Lond.* A, **360**, 1389–1408.

Crowder, M. (1992) Bayesian priors based on parameter transformation using the distribution function. *Ann. Inst. Statist. Math.*, **44**, 405–416.

Davison, A. C. (2003) *Statistical Models*. Cambridge: Cambridge University Press.

Davison, A. C. and Smith, R. L. (1990) Models for exceedances over high thresholds (with discussion). *J. R. Statist. Soc*. B, **52**, 393–442.

Drees, H., de Haan, L. and Resnick, S. (2000) How to make a Hill plot. *Ann. Statist.*, **28**, 254–274.

Dupuis, D. J. (1999) Exceedances over high thresholds: a guide to threshold selection. *Extremes*, **1**, 251–261.

Ewans, K. and Jonathan, P. (2008) The effect of directionality on northern North Sea extreme wave design criteria. *J. Offshr. Mech. Arct. Engng*, **130**.

Ferreira, A., de Haan, L. and Peng, L. (2003) On optimising the estimation of high quantiles of a probability distribution. *Statistics*, **37**, 401–434.

Geisser, S. and Eddy, W. F. (1979) A predictive approach to model selection. *J. Am. Statist. Ass.*, **74**, 153–160.

Gelfand, A. E. (1996) Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 144–161. London: Chapman and Hall.

Gelfand, A. E. and Dey, D. K. (1994) Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc*. B, **56**, 501–514.

Geweke, J. and Amisano, G. (2010) Comparing and evaluating Bayesian predictive distributions of asset returns. *Int. J. Forecast.*, **26**, 216–230.

Hall, P. (1990) Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multiv. Anal.*, **32**, 177–203.

Hall, P. and Welsh, A. H. (1985) Adaptive estimates of parameters of regular variation. *Ann. Statist.*, **13**, 331–341.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999) Bayesian model averaging: a tutorial. *Statist. Sci.*, **14**, 382–417.

Hosking, J. R. M. and Wallis, J. R. (1987) Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, **29**, 339–349.

Jonathan, P. and Ewans, K. (2013) Statistical modelling of extreme ocean environments for marine design: a review. *Ocean Engng*, **62**, 91–109.

Kass, R. E. and Wasserman, L. (1996) The selection of prior distributions by formal rules. *J. Am. Statist. Ass.*, **91**, 1343–1370.

Kotz, S., Balakrishnan, N. and Johnson, N. L. (2000) *Continuous Multivariate Distributions*, vol. 1, *Models and Applications*, 2nd edn, ch. 49. New York: Wiley.

MacDonald, A., Scarrott, C., Lee, D., Darlow, B., Reale, M. and Russell, G. (2011) A flexible extreme value mixture model. *Computnl Statist. Data Anal.*, **55**, 2137–2157.

Martins, E. S. and Stedinger, J. R. (2001) Generalized maximum likelihood Pareto-Poisson estimators for partial duration series. *Wat. Resour. Res.*, **37**, 2551–2557.

Northrop, P. J. and Attalides, N. (2016) Posterior propriety in Bayesian extreme value analyses using reference priors. *Statist. Sin.*, **26**, 721–743.

Northrop, P. J. and Coleman, C. L. (2014) Improved threshold diagnostic plots for extreme value analyses. *Extremes*, **17**, 289–303.

Northrop, P. J. and Jonathan, P. (2011) Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics*, **22**, 799–809.
Oceanweather (1995) NEXT—North Sea hindcast study. Oceanweather, Cos Cob.
Oceanweather (2005) GOMOS—Gulf of Mexico hindcast study. Oceanweather, Cos Cob.
O'Hagan, A. (2006) Science, subjectivity and software (comment on articles by Berger and by Goldstein). *Baysn Anal.*, **1**, 445–450.
Pickands, J. (1975) Statistical inference using extreme order statistics. *Ann. Statist.*, **3**, 119–131.
Pickands, J. (1994) Bayes quantile estimation and threshold selection for the generalized Pareto family. In *Extreme Value Theory and Applications* (eds J. Galambos, J. Lechner and E. Simiu), pp. 123–138. New York: Springer.
Reimherr, M., Meng, X.-L. and Nicolae, D. L. (2014) Being an informed Bayesian: assessing prior informativeness and prior likelihood conflict. *Preprint*. (Available from `arxiv.org/abs/1406.5958`.)
Sabourin, A., Naveau, P. and Fougres, A.-L. (2013) Bayesian model averaging for multivariate extremes. *Extremes*, **16**, 325–350.
Scarrott, C. and MacDonald, A. (2012) A review of extreme value threshold estimation and uncertainty quantification. *Revstat*, **10**, 33–60.
Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
Smith, R. L. (1985) Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, **72**, 67–92.
Smith, R. L. (1989) Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statist. Sci.*, **4**, 367–377.
Smith, R. L. (2003) Statistics of extremes, with applications in environment, insurance and finance. In *Extreme Values in Finance, Telecommunications and the Environment* (eds B. Finkenstädt and H. Rootzén), pp. 1–78. Boca Raton: Chapman and Hall–CRC.
Stephenson, A. (2016) Bayesian inference for extreme value modelling. In *Extreme Value Modeling and Risk Analysis: Methods and Applications* (eds D. K. Dey and J. Yan), pp. 257–280. Boca Raton: Chapman and Hall.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc*. B, **36**, 111–147.
Süveges, M. and Davison, A. C. (2010) Model misspecification in peaks over threshold analysis. *Ann. Appl. Statist.*, **4**, 203–221.
Süveges, M. and Davison, A. C. (2012) A case study of a "Dragon-King": the 1999 Venezuelan catastrophe. *Eur. Phys. J. Specl Top.*, **205**, 131–146.
Wadsworth, J. L. (2016) Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics*, **58**, 116–126.
Wadsworth, J. L. and Tawn, J. A. (2012) Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *J. R. Statist. Soc.* B, **74**, 543–567.
Wakefield, J. C., Gelfand, A. E. and Smith, A. F. M. (1991) Efficient generation of random variates via the ratio-of-uniforms method. *Statist. Comput.*, **1**, 129–133.
Wong, T. S. T. and Li, W. K. (2010) A threshold approach for peaks-over-threshold modelling using maximum product of spacings. *Statist. Sin.*, **20**, 1257–1272.
Young, G. A. and Smith, R. L. (2005) *Essentials of Statistical Inference*. Cambridge: Cambridge University Press.
Zellner, A. (1998) Past and recent results on maximal data information priors. *J. Statist. Res.*, **32**, 1–22.