On covariate smoothing for non-stationary extremes

Matthew Jones[a], David Randell[a], Kevin Ewans[b], Philip Jonathan[a,*]

[a] *Shell Projects & Technology, Manchester M22 0RR, United Kingdom.*

[b] *Sarawak Shell Bhd., 50450 Kuala Lumpur, Malaysia.*

Abstract

Numerous approaches are proposed in the literature for non-stationarity marginal extreme value inference, including different model parameterisations with respect to covariate, and different inference schemes. The objective of this article is to compare some of these procedures critically. We generate sample realisations from generalised Pareto distributions, the parameters of which are smooth functions of a single smooth periodic covariate, specified to reflect the characteristics of actual samples of significant wave height with direction considered in the literature in the recent past. We estimate extreme values models (1) using Constant, Fourier, B-spline and Gaussian Process parameterisations for the functional forms of generalised Pareto shape and (adjusted) scale with respect to covariate and (2) maximum likelihood and Bayesian inference procedures. We evaluate the relative quality of inferences by estimating return value distributions for the response corresponding to a time period of 10× the (assumed) period of the original sample, and compare estimated return values distributions with the truth using Kullback-Leibler, Cramer-von Mises and Kolmogorov-Smirnov statistics. We find that Spline and Gaussian Process parameterisations estimated by maximum penalised likelihood using the back-fitting algorithm, or Monte Carlo Markov chain inference using the mMALA algorithm, perform equally well in terms of quality of inference and computational efficiency, and generally perform better than alternatives.

Keywords: extreme, covariate, non-stationary, smoothing, non-parametric, spline, Gaussian process, mMALA, Kullback-Leibler

1. Introduction

Accurate estimates of the likely extreme environmental loading on an offshore facility are vital to enable a design that ensures the facility is both structurally reliable and economic.This involves estimating the extreme value behaviour of meteorological and oceanographic (metocean) parameters that quantify the various environmental loading quantities, primarily winds, wave, and currents. Examples of such parameters are the significant wave height, the mean wind speed, and mean current speed, which provide a measure of wave field or sea state intensity, the wind strength, and the ocean current strength respectively. These parameters characterise the short-term variability of the phenomenon involved, which is assumed to be stationary over temporal scales of a few hours.

The long-term variability of these parameters is however non-stationary, in particular with respect to time, space and direction. From a temporal point of view metocean parameters generally have a strong seasonal variation, with an annual periodicity, and longer term variations due to decadal or semi-decadal climate variations. At any given location, the variability of a particular parameter is also dependent on the direction; for example, wind forcing is typically stronger from some directions than others, and fetch and water depth effects can strongly influence the resulting magnitude of the waves. Clearly these effects will vary with location: a more exposed location will be associated with longer fetches, resulting in a more extreme wave climate.

When estimating the long-term variability of parameters, such as significant wave height, the non-stationary effects associated with e.g. direction and season can be incorporated by treating direction and season as covariates. The common practice is to perform extreme value analysis of hindcast data sets, which include many years of metocean parameters, along with their associated covariates. Such data sets have all the information needed for input to covariate analysis.

Numerous authors have reported the essential features of extreme value analysis (e.g. Davison and Smith [9]) and the importance of considering different aspects of covariate effects. Carter and Challenor [3] considers estimation of annual maxima from monthly data, when the distribution functions of monthly extremes are known. Coles and Walshaw [6] describes directional modelling of extreme wind speeds using a Fourier parameterisation. Scotto and Guedes-Soares [23] models the long-term time series of significant wave height with non-linear threshold models. Anderson, Carter, and Cotton [1] reports that estimates for 100-year significant wave height from an extreme

[*] Corresponding author. Email: philip.jonathan@shell.com

value model ignoring seasonality are considerably smaller than those obtained using a number of different seasonal extreme value models. Chavez-Demoulin and Embrechts [5] describes smooth extreme value models in finance and insurance. Chavez-Demoulin and Davison [4] provides a straight-forward description of a nonhomogeneous Poisson model in which occurrence rates and extreme value properties are modelled as functions of covariates. Cooley, Naveau, Jomelli, Rabatel, and Grancher [7] uses a Bayesian hierarchical model to characterise extremes of lichen growth. Renard, Lang, and Bois [22] considers identification of changes in peaks over threshold using Bayesian inference. Fawcett and Walshaw [11] uses a hierarchical model to identify location and seasonal effects in marginal densities of hourly maxima for wind speed. Mendez, Menendez, Luceno, Medina, and Graham [17] considers seasonal non-stationary in extremes of NOAA buoy records. Randell, Feld, Ewans, and Jonathan [19] discusses estimation for return values for significant wave height in the South China Sea using a directional-seasonal extreme value model.

The objective of this article is to evaluate different procedures for estimating non-stationary extreme value models critically. We quantify the extent to which extreme value analysis of samples of peaks over threshold exhibiting clear non-stationarity with respect to covariates, such as those in Figure 1 or simulation case studies in Section 4 below, is influenced by a particular choice of model parameterisation or inference method. We generate sample realisations from generalised Pareto distributions, the parameters of which are smooth functions of a single smooth periodic covariate, specified to reflect the characteristics of actual samples considered in the literature in the recent past. Then we estimate extreme value models (1) using Constant, Fourier, B-spline and Gaussian Process parameterisations for the functional forms of generalised Pareto parameters with respect to covariate and (2) maximum likelihood and Bayesian inference procedures. We evaluate the relative quality of inferences by estimating return value distributions for the response corresponding to a time period of 10× the (assumed) period of the original sample, and compare estimated return values distributions with the truth using Kullback-Leibler (see e.g. Perez-Cruz [18]), Cramer-von Mises (see e.g. Anderson [2]) and Kolmogorov-Smirnov statistics. We cannot hope to compare all possible parameterisations, but choose four parameterisations useful in our experience. Similarly, there are many competing approaches for maximum likelihood and Bayesian inference, and general interest in understanding their relative characteristics. For example, Smith and Naylor [24] compares maximum likelihood and Bayesian inference for the three-parameter Weibull distribution. In this work, we choose to compare frequentist penalised likelihood maximisation with two Monte Carlo Markov Chain (MCMC) methods of different complexities. Non-stationary model estimation is a growing field. There is a huge literature on still further possibilities for parametric (e.g. Chebyshev, Legendre and other polynomial forms) and non-parametric (e.g. Gauss-Markov random fields and radial basis functions) model parameterisations with respect to covariates. Moreover, in extreme value analysis, pre-processing of a response to near stationarity (e.g. using a Box-Cox transformation) is sometimes preferred.

The outline of the paper is as follows. Section 2 illustrates applications of non-stationary extreme value analysis in the recent literature, motivating the model forms adopted subsequently in Section 4.1. Section 3 outlines the different model parameterisations and inference schemes under consideration. Section 4 describes underlying model forms used to generate samples for inference, outlines the procedure for estimation of return value distributions and their comparison, and presents results of those comparisons. Section 5 provides discussion and conclusions.

2. Motivating applications

Randell, Zanini, Vogel, Ewans, and Jonathan [20] explores the directional characteristics of hindcast storm peak significant wave height with direction for locations in the Gulf of Mexico, North-West Shelf of Australia, Northern North Sea, Southern North Sea, South Atlantic Ocean, Alaska, South China Sea and West Africa. Figure 1 illustrates the essential features of samples such as these. The rate and magnitude of occurrences of storm events varies considerably between locations, and with direction at each location. There are directional sectors with effectively no occurrences, there is evidence of rapid changes in characteristics with direction and of local stationarity with direction. Any realistic model for such samples needs to be non-stationary with respect to direction.

The 6 simulation case studies introduced in Section 4 are constructed to reflect the general features of the samples in Figure 1, with the advantage that the statistical characteristics of the case studies are known exactly, allowing objective evaluation and comparison of competing methods of model parameterisation and inference. The results of this study are of course relevant to any application of non-stationary extreme value analysis.

3. Estimating non-stationary extremes

3.1. Generalised Pareto model

We consider estimation of non-stationary marginal extreme value models. We observe a sample $y = \{y_1, \ldots, y_N\}$ of peaks over threshold drawn independently from a generalised Pareto distribution, the parameters of which are functions of corresponding observed covariate

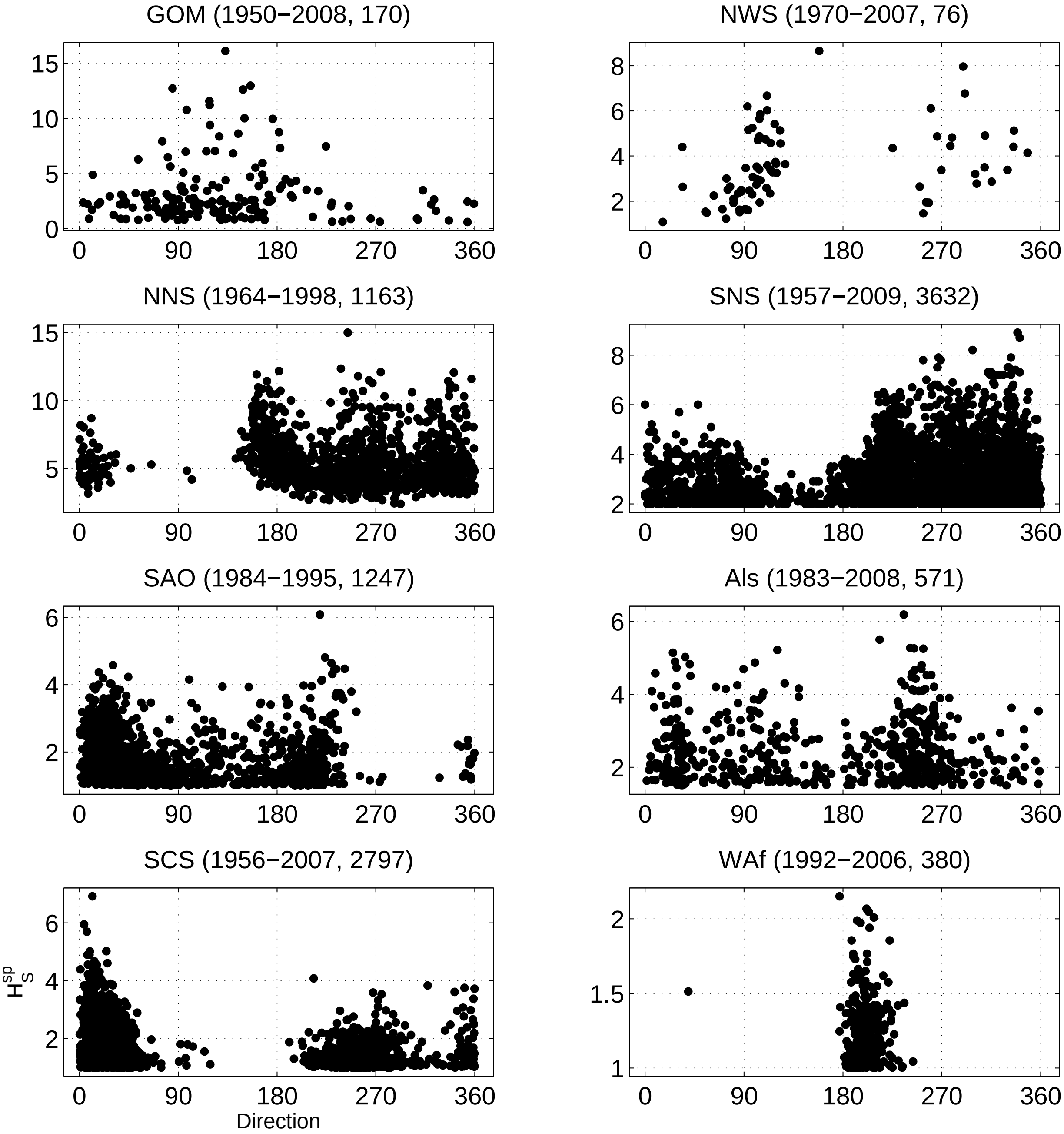


Figure 1: Storm peak significant wave height on direction for 8 locations worldwide. From right to left, top to bottom: Gulf of Mexico (GOM), North-West Shelf of Australia (NWS), Northern North Sea (NNS), Southern North Sea (SNS), South Atlantic Ocean (SAO), Alaska (Als), South China Sea (SCS) and West Africa (WAf). Panel titles give the location, the sample period and storm peak sample size.

values $\Theta = \{\theta_1, \ldots, \theta_N\}$. The sample likelihood is a product of generalised Pareto (GP) likelihoods for each of the observations

$$\begin{aligned} f(y|\Theta, \xi, \sigma, \mu) &= \prod_{i=1}^{N} f(y_i|\xi(\theta_i), \sigma(\theta_i), \mu(\theta_i)) \\ &= \prod_{i=1}^{N} \frac{1}{\sigma(\theta_i)} \left(1 + \xi(\theta_i) \frac{(y_i - \mu(\theta_i))}{\sigma(\theta_i)}\right)^{-(\frac{1}{\xi(\theta_i)}+1)} \end{aligned}$$

where $\xi(\theta)$ and $\sigma(\theta)$ are the shape and scale parameters as functions of covariate. We do not attempt to estimate the threshold function $\mu(\theta)$, assuming it is 0 for all covariate values. We also assume that the rate of occurrence $\rho(\theta)$ of exceedances of μ varies with covariate, but that $\rho(\theta)$ is known. It is computationally advantageous (see e.g. Cox and Reid [8], Chavez-Demoulin and Davison [4]) to transform variables from (ξ, σ) to the asymptotically independent pair (ξ, ν), where $\nu(\theta) = \sigma(\theta)(1 + \xi(\theta))$. Inference therefore amounts to estimating the smooth functions $\xi(\theta)$ and $\nu(\theta)$, although we usually choose to illustrate the analysis in terms of $\xi(\theta)$ and $\sigma(\theta)$.

3.2. Covariate parameterisations

To accommodate non-stationarity, we parameterise ξ and ν as linear combinations of unknown parameters β_ξ and β_ν respectively, where

$$\nu(\theta) = B_\nu(\theta)\beta_\nu \text{ , and } \xi(\theta) = B_\xi(\theta)\beta_\xi$$

and $B_\nu(\theta)$ and $B_\xi(\theta)$ are row vectors of “basis functions” evaluated at θ. We consider four different forms of basis function, corresponding to Constant (stationary), Fourier, Spline and Gaussian Process parameterisations for $\xi(\theta)$ and $\nu(\theta)$, as described below. For each parameterisation, we also specify roughness matrices Q_η (for $\eta = \xi, \nu$) to regulate the roughness of $\eta(\theta)$ with respect to θ during inference. This ensures that the elements of β_η weight the individual basis functions in such a way that the resulting estimate is optimally smooth in some sense. The form of the roughness penalty term R_η is $\frac{1}{2}\lambda_\eta\beta_\eta' Q_\eta \beta_\eta$, for some roughness coefficient λ_η. The penalty is incorporated directly within a penalised likelihood for maximum likelihood inference, and within a prior distribution for β_η in Bayesian inference, as described in Section 3.3.

Constant (stationary) parameterisation

In the Constant parameterisation, the values of ξ and ν do not vary with respect to θ. We therefore adopt a scalar basis function which is constant across all values of covariate, so that $B_\nu(\theta) = B_\xi(\theta) = 1$, and corresponding roughness matrices $Q_\nu = Q_\xi = 1$. We do not expect the return value distributions estimated under this parameterisation to fare well in general in our comparison, since samples are generated from non-stationary distributions. Quality of fit is expected to be poor, at least in some intervals of covariate. However, many practitioners continue to use stationary extreme value models, perhaps with high thresholds to mitigate non-stationarity, in applications; inclusion of a stationary parameterisation provides a useful point of reference for comparison, therefore.

Spline parameterisation

Under a Spline parameterisation, the vector of basis functions for each of ν and ξ is made up of p local polynomial B-spline functions with compact support, joined at a series of knots evenly spaced in the covariate domain (see e.g. Eilers and Marx [10])

$$B_\nu(\theta) = B_\xi(\theta) = \begin{pmatrix} b_1(\theta) & \cdots & b_p(\theta) \end{pmatrix} .$$

We specify roughness matrices $Q_\nu = Q_\xi = D^{\mathrm{T}}D$ which penalise squared differences between adjacent elements of the coefficient vectors, where

$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

is a $(p-1) \times p$ difference matrix.

Fourier parameterisation

We use basis vectors composed of sine and cosine functions of n_p different periods

$$B_\nu(\theta) = B_\xi(\theta) = \begin{pmatrix} 1 & \sin(\theta) & \sin(2\theta) & \cdots & \sin(n_p\theta) & \cos(\theta) & \cdots & \cos(n_p\theta) \end{pmatrix} .$$

The roughness matrix is computed by imposing a condition on the squared second derivative of the resulting parameter function. If we write

$$\eta(\theta) = \sum_{k=1}^{n_p} \left(a_{\eta k} \cos(k\theta) + b_{\eta k} \sin(k\theta) \right)$$

where $\eta = \xi$ or ν, and $a_{\eta k}$ and $b_{\eta k}$ are the parameters from β_η corresponding respectively to the sine and cosine functions of period k. The roughness criterion (from Jonathan, Randell, and Ewans [14]) becomes

$$R_\eta = \int_0^{2\pi} (\eta''(\theta))^2 d\theta = \sum_{k=1}^{n_p} k^4 (a_{\eta k}^2 + b_{\eta k}^2)$$

such that the penalty matrix can be written in matrix form as

$$Q_\eta = \operatorname{diag}\left(0, 1, 2^4, \ldots, k^4, \ldots, {n_p}^4, 1, 2^4, \ldots, k^4, \ldots, {n_p}^4\right)$$

for the $p = 2n_p + 1$ Fourier parameters $(a_{\eta 0}, a_{\eta 1}, \ldots, a_{\eta n_p}, b_{\eta 1}, \ldots, b_{\eta n_p})$.

Gaussian Process parameterisation

We use a Gaussian Process parameterisation [21] for a set of p knots $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p\}$ in the covariate domain, and relate each covariate input to a knot using the following basis vectors

$$B_\nu(\theta) = B_\xi(\theta) = \begin{pmatrix} I_1(\theta) & \cdots & I_p(\theta) \end{pmatrix}$$

where the indicator functions $I_j(.)$ are defined as

$$I_j(\theta) = \begin{cases} 1 & \text{if } |\theta - \hat{\theta}_j| < |\theta - \hat{\theta}_k| \; \forall k \neq j \\ 0 & \text{otherwise} . \end{cases}$$

Roughness matrices Q_η (where $\eta = \nu, \xi$) are defined by the coefficient correlation matrix V_η via $Q_\eta = V_\eta^{-1}$, and the elements of V_η drawn from a periodic squared exponential covariance function [16]

$$V_{\eta jk} = \exp\left(-\frac{2}{r_\eta^2} \sin\left(\frac{\hat{\theta}_j - \hat{\theta}_k}{2} \right)^2 \right)$$

where r_η are correlation lengths for each of the parameters, fixed to likely values by comparison with the covariate functions used to generate the data. Partitioning the covariate domain in this manner, as opposed to fitting a Gaussian Process parameterisation to each of the data inputs, greatly reduces the number of parameters to estimate, and is physically reasonable. Estimating a parameter for each data point would have made the computational burden for the Gaussian Process parameterisation significantly greater than that for any of the other parameterisations.

3.3. Inference procedures

We consider two methods for estimating parameters and return value distributions for the models and parameterisations described above, namely (a) maximum penalised likelihood estimation with bootstrapping to quantify uncertainties, and (b) (two forms of) Bayesian inference using Markov Chain Monte Carlo (MCMC). These are discussed below.

Maximum likelihood estimation

We use an iterative back-fitting optimisation (see Appendix) to minimise the penalised negative log likelihood $-L^*\left(y|\beta_\xi,\beta_\nu;\lambda_\xi,\lambda_\nu\right)$ with respect to β_ξ and β_ν for given roughness coefficients λ_ξ and λ_ν, where

$$\begin{aligned} -L^*\left(y|\beta_\xi,\beta_\nu;\lambda_\xi,\lambda_\nu\right) &= -L\left(y|\beta_\xi,\beta_\nu\right) + R_\xi + R_\nu \\ &= -L\left(y|\beta_\xi,\beta_\nu\right) + \frac{1}{2}\lambda_\xi \beta_\xi' Q_\xi \beta_\xi + \frac{1}{2}\lambda_\nu \beta_\nu' Q_\nu \beta_\nu . \end{aligned}$$

Here, $-L\left(y|\beta_\xi,\beta_\nu\right)$ is the negative log sample GP likelihood from Section 3.1 expressed as a function of ξ and ν, and R_ξ and R_ν are additive roughness penalties. The values of λ_ξ and λ_ν are selected using cross-validation to maximise the predictive performance of the estimated model, and bootstrap resampling is used to quantify the uncertainty of parameter estimates. The original sample is resampled with replacement a large number of times, and inference repeated for each resample. We use the empirical distributions of parameter estimates and return values over resamples as approximate uncertainty distributions.

Bayesian Inference

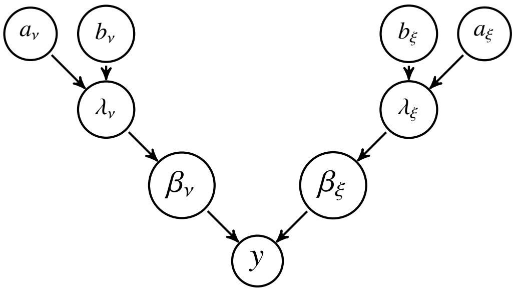
From a Bayesian perspective, all of β_ξ, β_ν, λ_ξ and λ_ν are treated as parameters to be estimated. Their joint posterior distribution given sample responses y and covariates Θ can be written

$$f\left(\beta_\xi, \beta_\nu, \lambda_\xi, \lambda_\nu | y, \Theta\right) \propto f\left(y|\Theta, \xi, \sigma, \mu\right) f\left(\beta_\nu|\lambda_\nu\right) f\left(\beta_\xi|\lambda_\xi\right) f\left(\lambda_\nu|a_\nu, b_\nu\right) f\left(\lambda_\xi|a_\xi, b_\xi\right)$$

where $f\left(y|\Theta, \xi, \sigma, \mu\right)$ is the sample GP likelihood from Section 3.1 and prior distributions $f\left(\beta_\nu|\lambda_\nu\right)$, $f\left(\beta_\xi|\lambda_\xi\right)$, $f\left(\lambda_\nu|a_\nu, b_\nu\right)$ and $f\left(\lambda_\xi|a_\xi, b_\xi\right)$ are specified as follows. Parameter smoothness of GP shape and (modified) scale functions is encoded by adopting Gaussian priors for their vectors β_η of basis coefficients (for $\eta = \xi, \nu$), expressed in terms of parameter roughness R_η

$$f\left(\beta_\eta|\lambda_\eta\right) \propto \lambda_\eta^{1/2} \exp\left(-\frac{\lambda_\eta}{2}\beta_\eta^{\mathrm{T}} Q_\eta \beta_\eta \right) .$$

The roughness coefficient λ_η can be seen, from a Bayesian perspective, as a parameter precision for β_η. It is assigned a Gamma prior distribution, which is conjugate with the prior Gaussian distribution for β_η. The values of hyper-parameters are set such that Gamma priors are relatively uninformative. The Bayesian inference can be illustrated by the directed acyclic graph



Estimates for β_ξ, β_ν, λ_ξ and λ_ν are obtained by sampling the posterior distribution above using MCMC. We choose to adopt a Metropolis-within-Gibbs framework, where each of the four parameters is sampled in turn conditionally on the values of others. The full conditional distributions $\lambda_\xi|\beta_\xi$ and $\lambda_\nu|\beta_\nu$ of precision parameters are Gamma by conjugacy, and are sampled exactly in a Gibbs step. Full conditional distributions for coefficients β_ξ and β_ν are not available in closed form; a more general Metropolis-Hastings (MH) scheme must therefore be used.

There are a number of potential alternative strategies regarding the MH step for β_η ($\eta = \xi, \nu$). We choose to examine two possibilities: (a) a straightforward MH sampling of correlated Gaussian proposals for β_η, and (b) the mMALA algorithm of Girolami and Calderhead [13], exploiting first- and second-derivative information from the log posterior to propose candidate values for the full vector of coefficients in high-probability regions. Implementations are described in the Appendix. Henceforth we refer to these two schemes as "MH" and "mMALA" respectively for brevity. The MH approach is simple to implement, but is likely to generate MCMC chains which mix relatively poorly. The mMALA scheme is expected to explore the posterior with considerably more efficiency; however, its implementation requires knowledge of likelihood derivatives.

4. Evaluation of methods

This section describes evaluation of relative performance of different model parameterisations and inference schemes introduced in Sections 3.2 and 3.3. We assess performance in terms of quality of estimation of distributions of return values corresponding to long return periods, estimated under models for large numbers of replicate samples of data from pre-specified underlying models.

We simulate 100 "sample realisations" each of size 1000 from three different underlying models, described below and referred to henceforth as Cases 1, 2 and 3, and further simulate 100 sample realisations of size 5000 from the same triplet of underlying models, referring to these as Cases 4, 5 and 6 respectively. We next estimate extreme values models for all sample realisations, model parameterisations and inference schemes. We assume that any sample realisation (for any Case) corresponds to a period $\mathcal{T}$ years of observation. We then simulate 1000 replicates of "return period realisations", each replicate consisting of observations of directional extreme values corresponding to a return period of $10 \times \mathcal{T}$, and estimate the distribution of the maximum observed (the $10\mathcal{T}$-year maximum return value) for all model parameterisations and inference methods, by accumulation from the 1000 replicates. Return value distributions are estimated omnidirectionally (that is, including all directions) and for 8 directional octants centred on the cardinal and semi-cardinal direction (by considering only those observations from the return period realisation with the appropriate directional characteristics). For each sample realisation from Cases 4, 5 and 6, we estimate return value distributions for all parameterisations but for only mMALA inference, since as will be discussed in Section

4.3 below, the computational effort associated with any of mMALA, MH and MLE (for large numbers of bootstrap resamples) for these Cases is prohibitively large.

We quantify the quality of return value inference by comparing the empirical cumulative distribution function generated under the fitted model for each sample realisation with that from simulation under the known underlying Case. We quantify the discrepancy between empirical distribution functions by estimating Kullback-Leibler, Cramer-von Mises and Kolmogorov-Smirnov statistics. We visualise relative performance by plotting the empirical cumulative distribution function of the test statistic over the 100 sample realisations, for each combination of Case, model parameterisation and inference method. We also compare performance in terms of prediction of the 37.5th percentile of the $10\mathcal{T}$-year return value distribution, since this sis often used in metocean and coastal design applications. However, we are not only interested in quality of inference, but also in computational efficiency. This is evaluated and illustrated in terms of effective sample size, as discussed below in Section 4.3.

4.1. Case studies considered

First, we describe model Cases 1-6 used to generate sample realisations for extreme value modelling. For each Case, potentially all of Poisson rate ρ of threshold exceedance, GP shape ξ and scale σ of exceedance size vary as a function of covariate θ. The extreme value threshold μ is fixed at zero throughout.

Case 1. : For extreme value threshold $\mu(\theta) = 0$, we simulate 1000 observations with a uniform Poisson rate $\rho(\theta) = 1000/360$ per degree covariate, and a low order Fourier parameterisation of GP shape $\xi(\theta) = \sin(\theta) + \cos(2\theta) + 2$ and scale $\sigma(\theta) = -0.2 + (\sin(\theta - 30))/10$.

Case 2. : For extreme value threshold $\mu(\theta) = 0$ and the same Fourier parameterisation of GP shape and scale as in Case 1, a non-uniform Poisson rate $\rho(\theta) = \max(\sin(\theta) + 1.1, 0) \times 1000/c_\rho$, where $c_\rho = \int_0^{360} \max(\sin(\theta) + 1.1, 0)\, d\theta$ is used to simulate 1000 observations.

Case 3. : For extreme value threshold $\mu(\theta) = 0$, the forms of each of $\rho(\theta)$, $\xi(\theta)$ and $\sigma(\theta)$ are defined by mixtures of between one and five Gaussian densities, as illustrated in Figure 2. Sample size is 1000.

Cases 4, 5 and 6. : These cases are identical to Cases 1, 2 and 3 respectively, except that Poisson rate ρ is increased by a factor of five. Sample size is therefore 5000.

Figure 2 illustrates typical sample realisations of Cases 1, 2 and 3. Parameter variation of GP shape ξ and scale σ with direction θ are identical in Cases 1 and 2. Poisson rate ρ is constant in Case 1 only. In Cases 2 and 3, ρ is very small at $\theta \approx 270°$ leading to a sparsity of corresponding observations. ξ is largest (but negative) at $\theta \approx 120°$ for Cases 1 and 2, leading to larger observations here. For Case 3, ξ is largest (and positive) at $\theta \approx 30°$ leading to the longest tail in any of the Cases considered. Figure 3 shows parameter estimates for ξ and σ, corresponding to the sample realisation of Case 2 shown in Figure 2, for different model parameterisations using mMALA inference. Visual inspection suggests that estimates of similar quality are obtained using all of Spline, Fourier and Gaussian Process parameterisations, but that the Constant parameterisation is poor. It is also apparent that identification of ξ is more difficult than σ. Corresponding plots (not shown) for maximum likelihood and Metropolis-Hastings inference show broadly similar characteristics, as do plots for other realisations of the same Case, and realisations of different Cases. (Posterior) cumulative distribution functions of return values based on models for the sample realisations of Case 2 illustrated in Figures 2 and 3, corresponding to a return period of ten times the period of the original sample, are shown in Figure 4 for different model parameterisations and mMALA inference. It can be seen omnidirectionally that the Constant parameterisation provides best agreement with the known return value distribution, despite the fact that parameter estimates in Figure 3 do not reflect the directional non-stationarity present. Omnidirectionally, and for 8 directional octant sectors, non-stationary model parameterisations perform similarly. However, it is clear that the Constant parameterisation does particularly poorly for the western and north-western sectors, for which the rate of occurrence of events is relatively low, and both ξ and σ are near their minimum values. Figure 5 illustrates uncertainty (over all 100 sample realisations) in the cumulative distribution function of return value for Case 2, corresponding to a return period of ten times the period of the original sample, using the Spline parameterisation and mMALA inference. The median estimate for the return value distribution (over all 100 sample realisations of Case 2) is shown in solid grey, with corresponding point-wise 95% uncertainty band in dashed grey. The true return value distribution is given in solid black. There is good agreement in all sectors. We explore differences in inferences for return value distributions more fully in Section 4.2.

4.2. Assessing quality of inference

The criteria used to compare distributions of return values are now described. Since, for comparison only, we only have access to samples from distributions, where necessary we project empirical distributions onto a linear grid using linear interpolation, and evaluate grid-based approximations to facilitate comparison. Then we compare empirical return value distributions using each of the following three

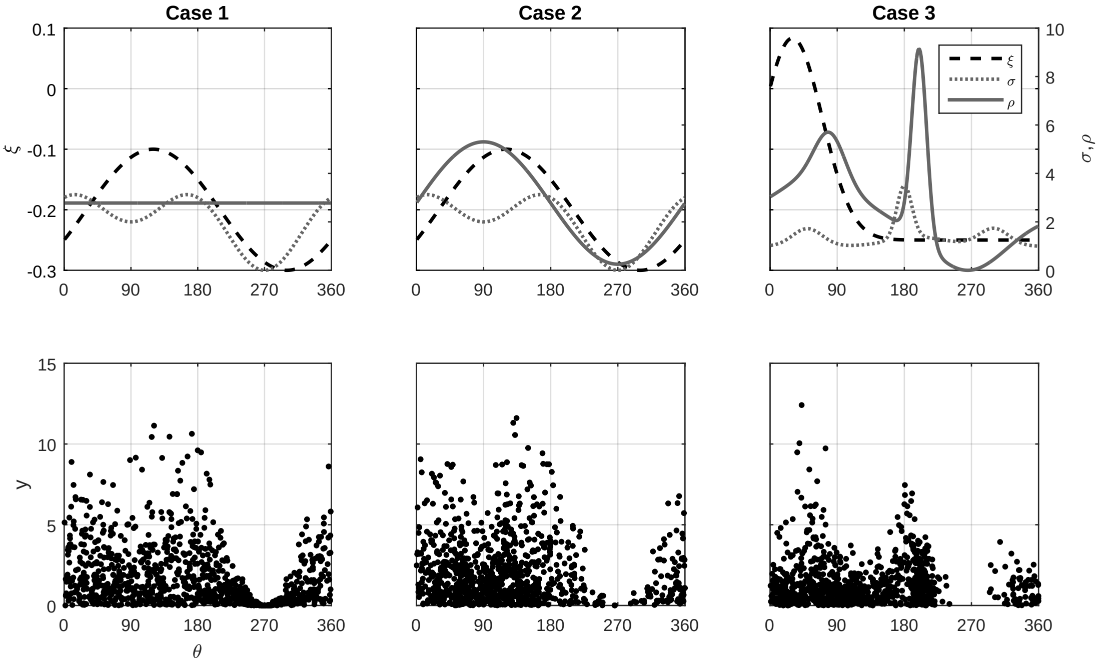


Figure 2: Illustrations of sample realisations from each of Cases 1 (left), 2 (centre) and 3 (right). Upper panels show parameter variation of GP shape ξ, scale σ and Poisson rate ρ with direction θ for each case. Lower panels show the 10th realisation of the corresponding simulated samples. ξ and σ for Cases 4, 5 and 6 are identical to those of Cases 1, 2 and 3 respectively. The value of ρ for Cases 4, 5 and 6 is five times that of Cases 1,2 and 3 respectively.

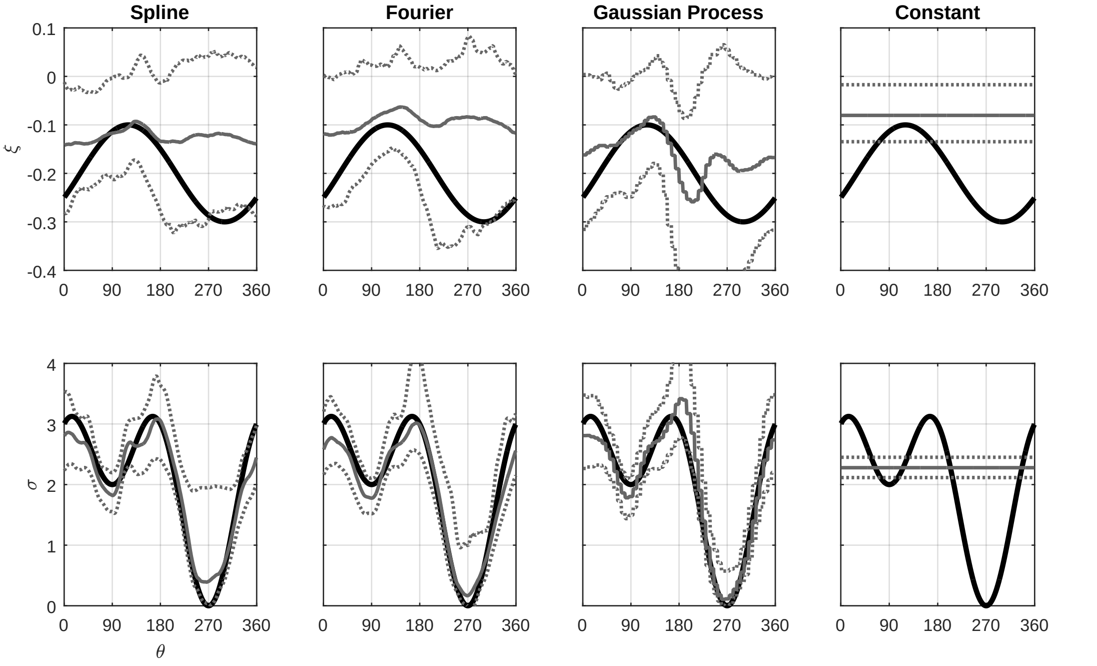


Figure 3: Parameter estimates for GP shape ξ (upper) and scale σ (lower) for the sample realisation of Case 2 shown in Figure 1, for different model parameterisations (left to right: Spline, Fourier, Gaussian Process, Constant) using mMALA inference. Each panel illustrates the true parameter (solid back), posterior median estimate (solid grey) with 95% credible interval (dashed grey).

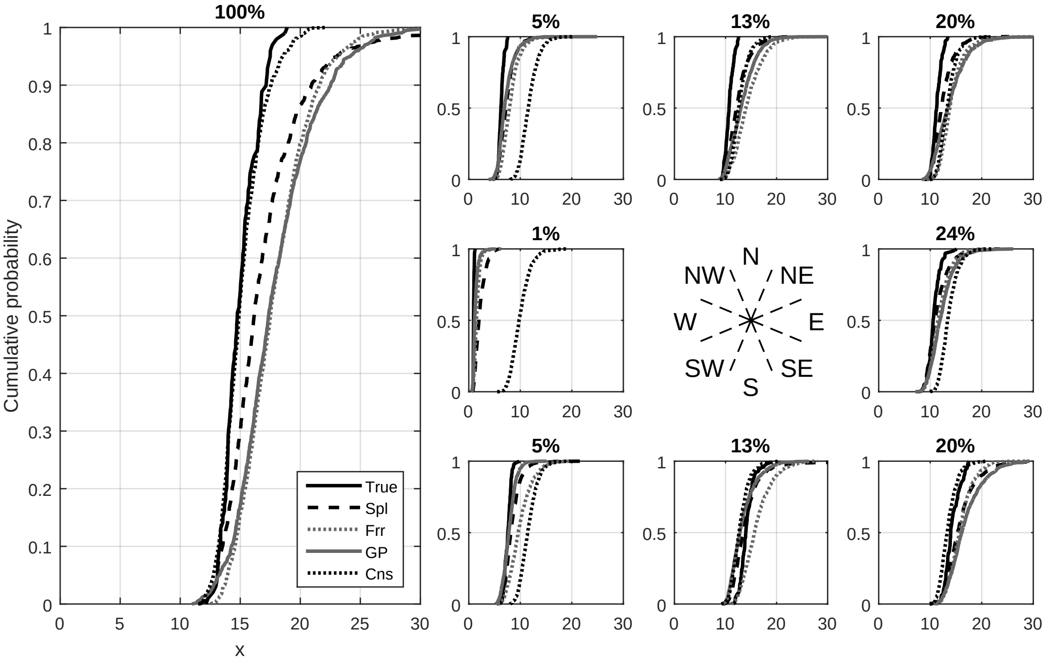


Figure 4: Posterior cumulative distribution functions of return value for the sample realisation of Case 2 shown in Figure 1, corresponding to a return period of ten times the period of the original sample. The left hand panel shows the omnidirectional return value distribution, and right hand panels the corresponding directional estimates. The title for each panel gives the expected percentage of individuals in that directional sector. In each panel, estimates are given for different model parameterisations using mMALA inference. The true return value distribution is given in solid black.

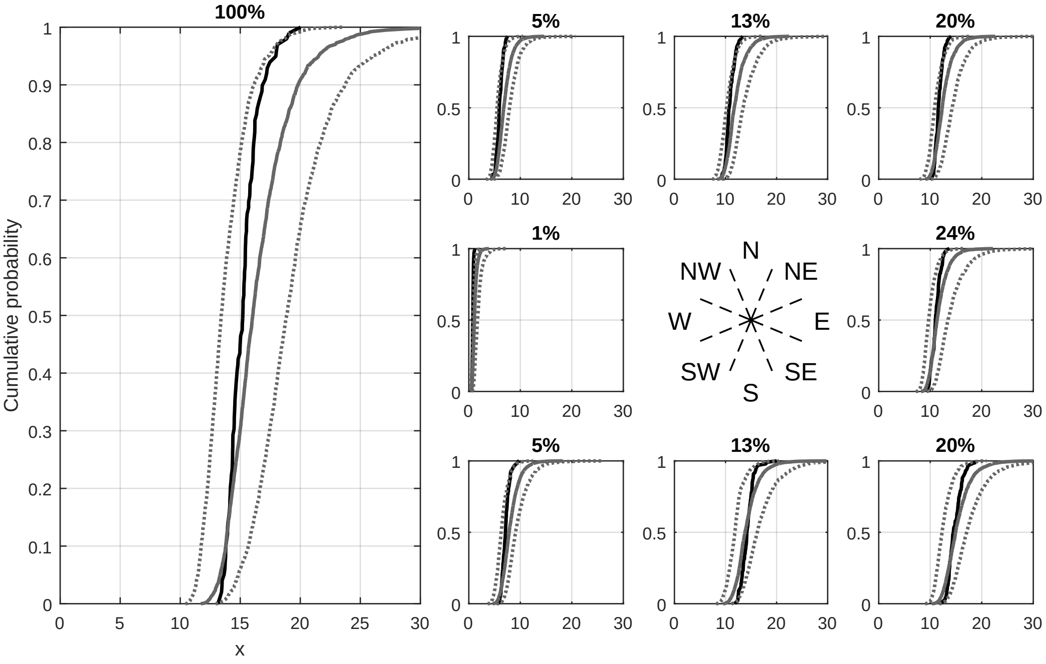


Figure 5: Uncertainty (over all 100 sample realisations) in the cumulative distribution function of return value for Case 2, corresponding to a return period of ten times the period of the original sample. The left hand panel shows the omnidirectional return value distribution, and right hand panels the corresponding directional estimates. The title for each panel gives the expected percentage of individuals in that directional sector. In each panel, the true return value distribution is given in solid black. The median estimate (over realisations) for return value distribution of the Spline model parameterisation using mMALA inference is shown in solid grey, with corresponding point-wise 95% uncertainty band in dashed grey.

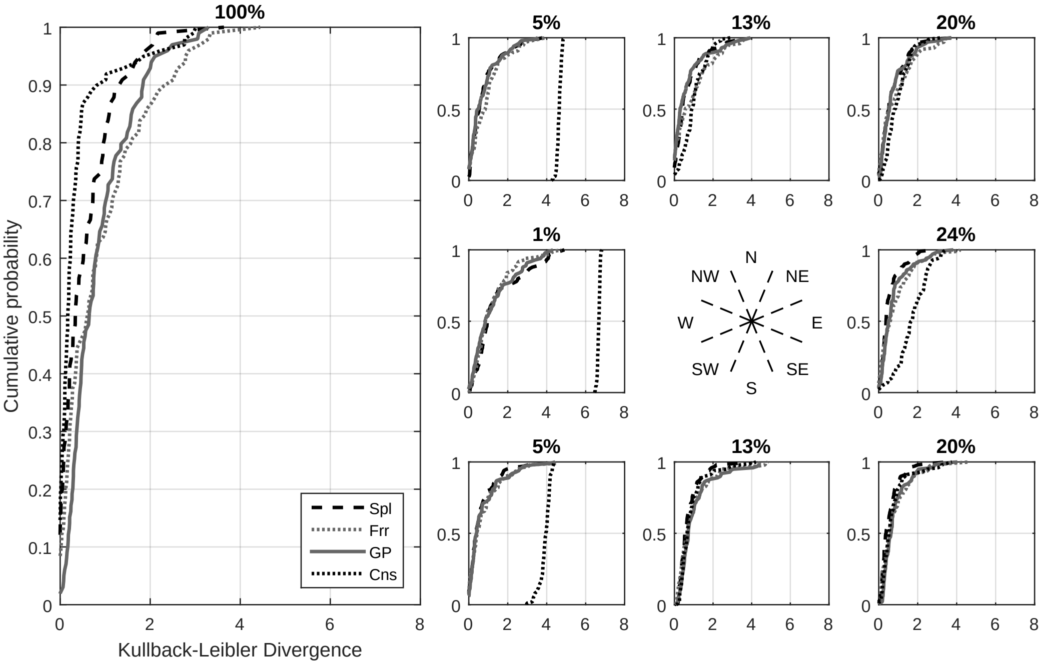


Figure 6: Empirical cumulative distribution functions of the Kullback-Leibler divergence between return value distributions (corresponding to a return period of ten times that the original sample) estimated under the true model and those estimated under models of sample realisations with different parameterisations and mMALA inference for Case 2. The title for each panel gives the expected percentage of individuals in that directional sector.

statistics. The **Kolmogorov-Smirnov criterion** compares two distributions in terms of the maximum "vertical" distance between cumulative distribution functions, as $D_{\text{ks}}(F_0, F_1) = \sup_x |F_1(x) - F_0(x)|$. The **Cramer-von Mises criterion** evaluates the average squared difference of one distribution from a second, reference distribution, using $D_{\text{cm}}(F_0, F_1) = \int_{-\infty}^{\infty} (F_1(x) - F_0(x))^2 f_0(x)\, dx$. The **Kullback-Leibler divergence** compares distributions using the average ratio of density functions $D_{\text{kl}}(F_0, F_1) = \int_{-\infty}^{\infty} \log\left(\frac{f_0(x)}{f_1(x)}\right) f_0(x)\, dx$; in this work, we use the approximation of Perez-Cruz [18]. The general characteristics of differences in return value inference due to model parameterisation and inference method were found to be similar for each of the three statistics. Only comparisons using Kullback-Leibler (KL) divergence are therefore reported here. We note that perfect agreement between $f_1(x)$ and $f_0(x)$ yields a minimum KL divergence of zero.

For illustration, Figure 6 shows empirical cumulative distribution functions for the KL divergence between return value distributions (corresponding to a return period of ten times that the original sample) estimated under the true model and those estimated under models of sample realisations with different parameterisations and mMALA inference for Case 2. The distributions of KL divergence from all non-stationary model parameterisations appear to be very similar, as might be expected from consideration of figures similar to Figure 4. However, the Constant model yields the best performance omnidirectionally in this case (since the corresponding distribution of KL divergence is shifted towards zero). In stark contrast, the Constant model does particularly badly in the eastern, south-western, western and north-western sectors. Figure 7 gives empirical cumulative distribution functions of the KL divergence between return value distributions (corresponding to a return period of ten times that of the original sample) estimated under the true return value distribution and those estimated under models of sample realisations with Spline parameterisations and different inference procedures for the same Case. There appears to be little to choose between the three inference methods for this Case.

Figure 8 summarises the characteristics of distributions for KL divergence corresponding to the omnidirectional return value distribution for all Cases, model parameterisations and inference methods considered in this work. In general, we note that all non-stationary parameterisations perform well with mMALA inference. With MH inference, performance is generally poorer, especially for Fourier parameterisation. MLE does better than MH. We note that the Constant parameterisation generally performs well for the omnidirectional return value, but there is some erratic behaviour, notably for Case 4. Figure 9 is the corresponding plot for the (generally sparsely populated) western directional sector. The Spline parameterisation with mMALA inference performs best. We note that the Fourier parameterisation does less well using MH and MLE, and that the Constant parameterisation behaves very erratically. We also note that, somewhat surprisingly, the Gaussian Process model performs considerably less well than the Spline and Fourier parameterisations. We surmise that this is due to mean-reversion in the absence of observations, compared with the Spline parameterisation in particular which prefers interpolation to reduce parameter roughness. We note that the Constant parameterisation also performs badly for Case 4.

From a practitioner's perspective, it is also interesting to quantify the performance of different model parameterisations and inference methods in estimating some central value (e.g. the mean, median, of 37.5$^{\text{th}}$ percentile) of the distribution of the return value. We choose the 37.5$^{\text{th}}$ percentile, since this is commonly used in the met-ocean community. Figure 10 illustrates this comparison in terms of a box-whisker

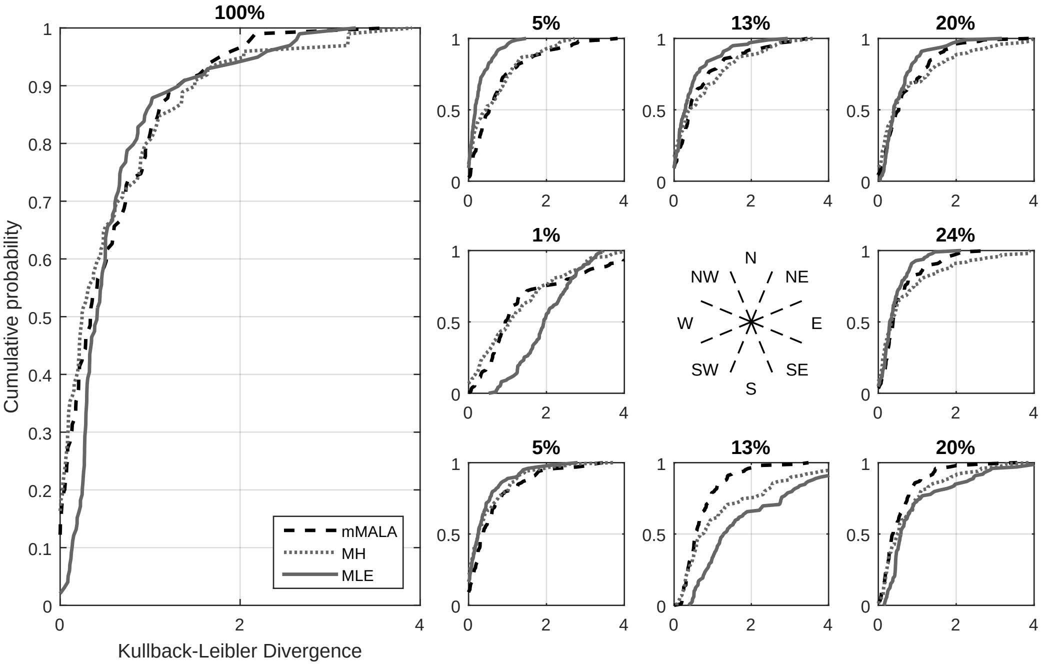


Figure 7: Empirical cumulative distribution functions of the Kullback-Leibler divergence between return value distributions (corresponding to a return period of ten times that the original sample) estimated under the samples from the true return value distribution and those estimated under models of sample realisations with Spline parameterisations and different inference procedures for Case 2. The title for each panel gives the expected percentage of individuals in that directional sector.

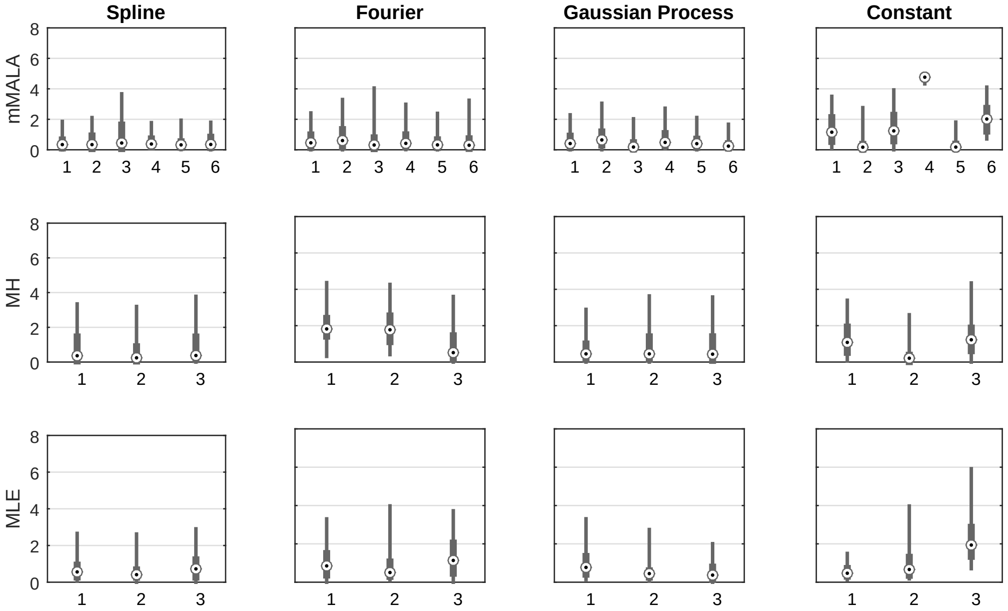


Figure 8: Box-whisker comparison of samples of Kullback-Leibler (KL) divergence between omnidirectional return value distributions (corresponding to a return period of ten times that the original sample) estimated under samples from the true return value distribution and those estimated under models of each of 100 sample realisations. mMALA inference is reported for all six Cases (abscissa labels) and model parameterisations (columns of panels). Metropolis-Hastings (MH) inference and maximum likelihood estimation (MLE) are reported only for the smaller sample sizes. The sample of KL divergence is summarised by the median (white disc with black central dot), the interquartile range (grey rectangular box) and the (2.5%, 97.5%) interval (grey line).

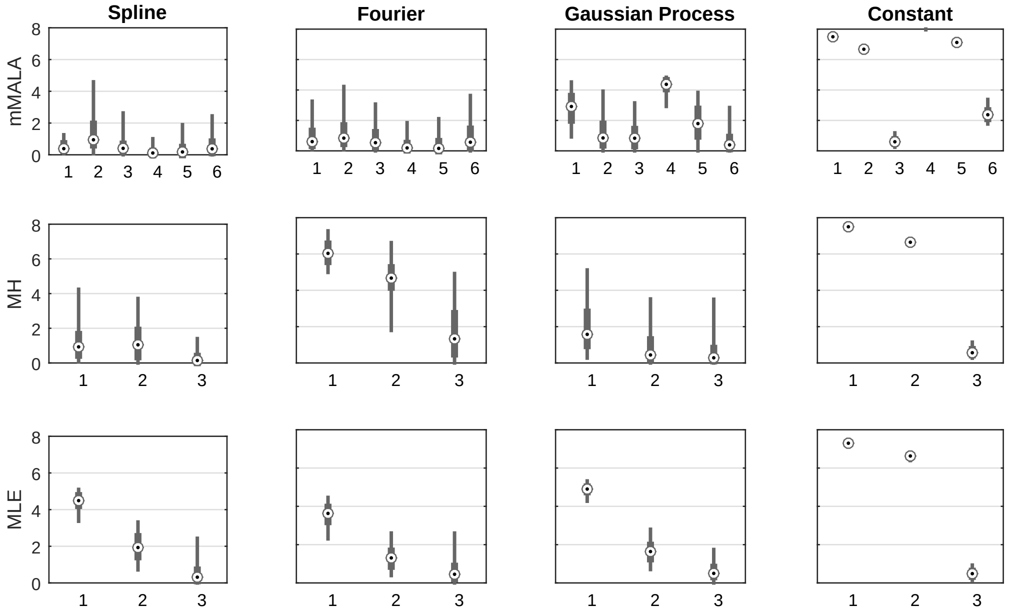


Figure 9: Box-whisker comparison of samples of Kullback-Leibler (KL) divergence between western sector return value distributions (corresponding to a return period of ten times that the original sample) estimated under samples from the true return value distribution and those estimated under models of each of 100 sample realisations. mMALA inference is reported for all six Cases (abscissa labels) and model parameterisations (columns of panels). Metropolis-Hastings (MH) inference and maximum likelihood estimation (MLE) are reported only for the smaller sample sizes. The sample of KL divergence is summarised by the median (white disc with black central dot), the interquartile range (grey rectangular box) and the (2.5%, 97.5%) interval (grey line). The ordinate scale is the same as that of Figure 8 to facilitate comparison. The Constant parameterisation for Case 4 with mMALA inference yields values of KL divergence larger than 8.

plot. In each panel, the "true" value, estimated by simulation under the true model, is shown as a black disc. The distribution of estimates from 100 different sample realisations of each Case is summarised by the median (white disc with black central dot), the interquartile range (grey rectangular box) and the (2.5%, 97.5%) interval (grey line). The performance of mMALA and MLE in all Cases (where comparison is possible) is very similar. In general, MH inference tends to produce greater bias and variability in estimates of the 37.5th percentile. We may surmise that this difference may be due to the fact that both mMALA and MLE exploit knowledge of likelihood gradient and curvature. There is little difference between the performance of the three non-stationary parameterisations; but the Constant model again performs more erratically. Corresponding box-whisker plots for the eastern directional octant (not shown) suggest that MLE and mMALA inference is again very similar, with MH estimates exhibiting greater variability. All non-stationary models yield similar performance, but the Constant model overestimates throughout. True values of the 37.5th percentile for the western sector (see Figure 11) are considerably lower than for the eastern sector, and lower again than the omnidirectional values. In this sector, the rate of occurrences of events is generally lower in all Cases. Nevertheless, Figure 11 has many similar features to Figure 10. However, we note that MH struggles in combination with the Fourier parameterisation, probably since the latter has the whole of the covariate domain as its support; intelligent proposals (like those used here in MLE and mMALA) are necessary. Overall, we note that MLE and mMALA inference for all of Spline, Fourier and Gaussian Process parameterisations perform relatively well, and equally well.

4.3. Assessing efficiency of inference

The effective sample size (m^*, see Geyer [12]) gives an estimate of the equivalent number of independent iterations that a Monte Carlo Markov chain represents, and is defined by $m^* = m/(1 + 2\sum_{k=1}^{\infty} c_k)$, where c_k is the autocorrelation of the MCMC chain at lag k, and m is the actual chain length. The effective sample size per hour is defined by m^*/T, where T is the elapsed computational time (in hours) for m steps of the chain. For maximum likelihood inference with bootstrap uncertainty estimation, since bootstrap resamples are independent of one another, we estimate effective sample size per hour as m_{BS}/T where m_{BS} is the number of bootstrap resamples used and T is now the total elapsed computational time (in hours) to execute analysis of the m_{BS} bootstrap resamples. Comparison of effective sample sizes per hour for different cases, parameterisations and inference methods gives some indication of relative computational efficiency, although objective comparison is difficult. In particular we note that software implementations in MATLAB exploiting common computational structures between different approaches have been used; these are almost certainly to the detriment of computational efficiency for some of the approaches, particulary the Constant parameterisation. For this reason, we do not report effective sample size per hour for the Constant parameterisation. Computational run-times are also of course critically dependent on software and hardware resources used. We note that the focus of this work is primarily quality of inference, rather than its computational efficiency. Specific modelling choices, such as the set-up of the cross-validation strategy adopted for MLE and choice of burn-in length and proposal step-size for MH and mMALA within reasonable

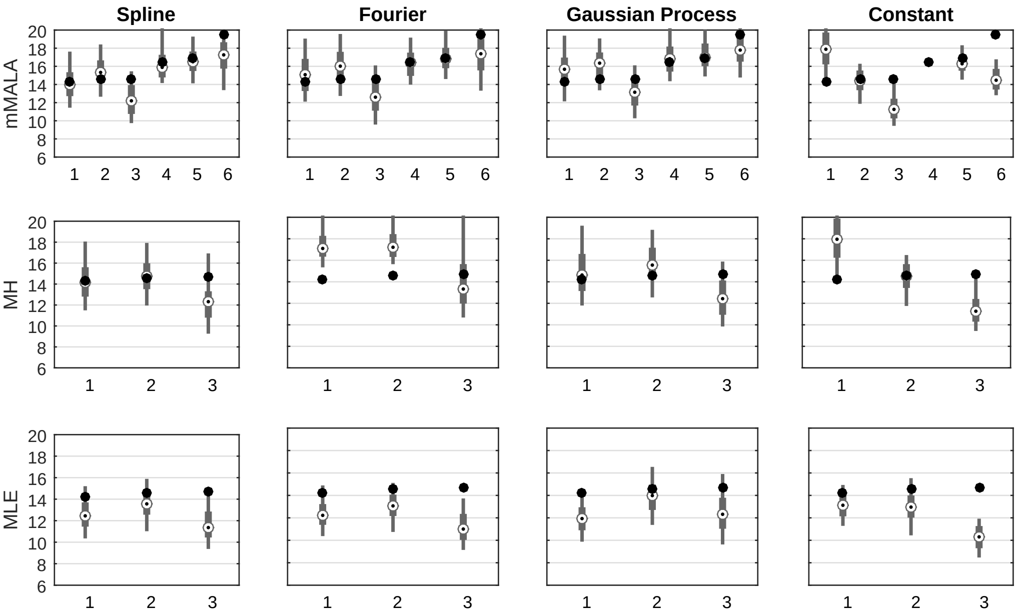


Figure 10: Box-whisker comparison of estimates for the 37.5th percentile of the omnidirectional return value distribution (in metres) for different cases, model parameterisations and inference procedures. mMALA inference is reported for all six cases (abscissa labels) and model parameterisations (columns of panels). Metropolis-Hastings (MH) inference and maximum likelihood estimation (MLE) are reported only for the smaller sample sizes. In each panel, the estimate from simulation under the true model is shown as a black disc. The distribution of estimates from 100 different sample realisations is summarised by the median value (white disc with black central dot), the interquartile range (grey rectangular box) and the (2.5%, 97.5%) interval (grey line). The Constant parameterisation for Case 4 with mMALA inference yields values larger than 20m.

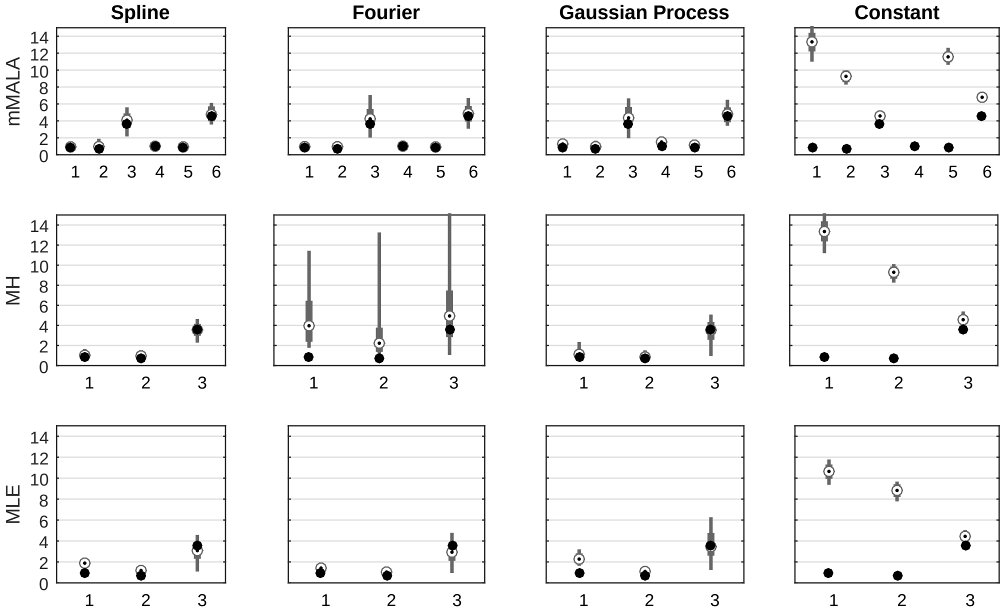


Figure 11: Box-whisker comparison of estimates for the 37.5th percentile of the return value distribution (in metres) for the western directional sector (least populous for cases 2, 3, 5 and 6), for different cases, model parameterisations and inference procedures. mMALA inference is reported for all six cases (abscissa labels) and model parameterisations (columns of panels). Metropolis-Hastings (MH) inference and maximum likelihood estimation (MLE) are reported only for the smaller sample sizes. In each panel, the estimate from simulation under the true model is shown as a black disc. The distribution of estimates from 100 different sample realisations is summarised by the median value (white disc with black central dot), the interquartile range (grey rectangular box) and the (2.5%, 97.5%) interval (grey line). The Constant parameterisation for Case 4 with mMALA inference yields values larger than 15m.

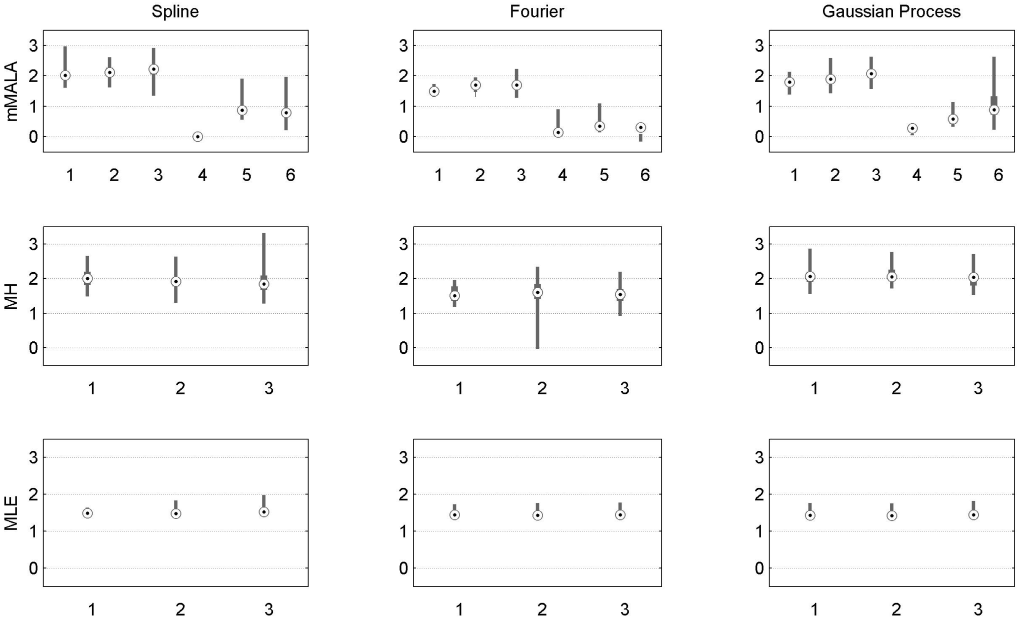


Figure 12: Effective sample size per hour (ESS/hr) estimates (on logarithm base 10 scale, i.e. with $\log_{10}$(ESS/hr) plotted on y-axis) for different cases, non-stationary model parameterisations and inference procedures. mMALA inference is reported for all six cases (abscissa labels) and model parameterisations (columns of panels). Metropolis-Hastings (MH) inference and maximum likelihood estimation (MLE) are reported only for the smaller sample sizes. The distribution of estimates from 100 different realisations of the original sample is summarised by the median value (white disc with black central dot), the interquartile range (grey rectangular box) and the (2.5%, 97.5%) interval (grey line).

bounds may not influence inferences greatly, but will obviously however affect run times. Similarly the Spline, Fourier and Gaussian Process parameterisations used were chosen to be of similar complexity, but small differences may again influence relative computational efficiency of inference. With these caveats in mind, Figure 12 illustrates the distribution of effective sample size per hour estimates for different cases, model parameterisations and inference procedures.

Figure 12 shows that, for mMALA inference, the effective sample size per hour (ESS/hr) is considerably lower for Cases 4, 5 and 6, indicating that inference using large sample sizes is slower. For this reason, in this work, we do not provide results for Cases 4, 5 and 6 using MLE and MH. Overall, comparing non-stationary parameterisations, ESS/hr is larger for Splines and Gaussian Processes than for Fourier. Using Spline and Gaussian Process parameterisations, values of ESS/hr are higher for MH and mMALA inference that for MLE. There is little difference in ESS/hr for different model parameterisations using MLE.

5. Discussion

Adequate allowance for non-stationarity is essential for realistic environmental extreme value inference. Ignoring the effects of covariates leads to unrealistic inference in general. The applied statistics and environmental literature provides various competing possible approaches to modelling non-stationarity. To these authors, adoption of non- or semi-parametric functional forms for generalised Pareto shape and scale in peaks over threshold modelling is far preferable in general in real world applications, than the assumption of a less flexible parameterisation. We find that B-spline and Gaussian Process parameterisations estimated by maximum penalised likelihood (using back-fitting) or Bayesian inference (using mMALA) perform equally well in terms of quality and computational efficiency, and generally outperform alternatives in this work.

The Gaussian Process parameterisation is computationally unwieldy for larger problems, unless covariate “gridding” onto the covariate domain is performed. The Fourier parameterisation, utilising bases with global support (compared to Spline and Gaussian Process basis functions whose support is local on the covariate domain), is generally somewhat more difficult to estimate well in practice, showing greater instability to choices such as starting solution for maximum likelihood estimation. The Constant parameterisation performs surprisingly well in estimating the omnidirectional return values distribution in some cases, but is generally very poor in estimating directional variation.

Various choices of methods of inference are also available. Competing approaches include maximum (penalised) likelihood optimisation and Bayesian inference using Monte Carlo Markov chain sampling. It appears however that the major difference, in terms of practical value of inference, is not between frequentist and Bayesian paradigms but rather the advantage gained by exploiting knowledge of likelihood gradient and curvature. In addition, two aspects give Bayesian inference a slight further advantage: (i) it appears that inference schemes which sample from a likelihood surface randomly, rather that seeking its minimum deterministically, are more stable, and therefore more routinely implementable and useable, and (ii) the computational efficiency of a “gradient-based” Bayesian inference (e.g. mMALA) in quantifying parameter estimates and their variability (evaluated in this work in terms of effective sample size per hour) is higher than that of the corresponding frequentist maximum penalised likelihood with bootstrapping inference. Moreover, Bayesian inference gives a more consistent procedure for statistical learning, and a more intuitive framework for communication of uncertainty, particularly to a non-specialist audience.

Acknowledgement

We are grateful to Kathryn Turnbull of Durham University, UK for useful comments.

Appendix

Maximum likelihood estimation

For maximum likelihood estimation (MLE), we use a back-fitting (or iteratively re-weighted least-squares, IRLS) algorithm to estimate vectors of basis coefficients β_η ($\eta = \xi, \nu$) derived in Jonathan et al. [15]. For fixed value of smoothness parameter λ_η, we initialise coefficients to starting value $\beta_\eta^{(0)}$, and then iterate the following step until convergence

$$\beta_\eta^{(i+1)} = -\left(B_\eta^{\mathrm{T}}\ W\left(\beta_\eta^{(i)}\right)\ B_\eta + Q_\eta\right)^{-1}\left(B_\eta^{\mathrm{T}}\ V\left(\beta_\eta^{(i)}\right) + B_\eta^{\mathrm{T}}\ W\left(\beta_\eta^{(i)}\right)\ B_\eta\beta_\eta^{(i)}\right)$$

where

$$V\left(\beta_\eta\right) = \nabla_\xi L\left(y|\Omega\right) \text{ and } W\left(\beta_\eta\right) = \nabla_{\beta_\eta}\nabla_{\beta_\eta}^{\mathrm{T}} L\left(y|\Omega\right)$$

are derived at the end of this Appendix. This algorithm is similar to the mMALA algorithm (see below) used to generate proposals for the corresponding Metropolis-Hastings step in MCMC, in that both exploit first- and second-derivative information to move towards regions of high probability. The back-fitting iteration is of course deterministic, whereas the mMALA step is stochastic.

MCMC sampling algorithms

Denoting the set of parameters to be estimated by $\Omega = \{\beta_\xi, \beta_\nu, \lambda_\xi, \lambda_\nu\}$, inference proceeds by sampling from the full conditional distributions $f\left(\Omega_k|y, \Theta, \Omega_{\neg k}, \Gamma\right)$ for each parameter in Ω in turn, where $\Gamma = \{a_\xi, b_\xi, a_\nu, b_\nu\}$ is the set of fixed hyper-parameters for prior distributions. The form of the full conditional distribution varies depending on the type of parameter being estimated, as explained below.

Full conditional distributions for basis coefficients

Vector β_η ($\eta = \xi, \nu$) has the following conditional distribution

$$f\left(\beta_\eta|y, \Theta, \Omega_{\neg\beta_\eta}, \Gamma\right) \propto f\left(y|\Theta, \beta_{\neg\eta}\right) f\left(\beta_\eta|\lambda_\eta\right)$$

which is not available in closed form, and therefore cannot be sampled directly in a Gibbs step. Instead, we generate samples using the Metropolis-Hastings algorithm: given current state $\beta_\eta^{(i)}$, we propose new parameter β_η^* from proposal distribution $f\left(\beta_\eta^*|\beta_\eta^{(i)}\right)$ and evaluate the acceptance ratio

$$A\left(\beta_\eta^*, \beta_\eta^{(i)}\right) = \frac{f\left(\beta_\eta^*|y, \Theta, \Omega_{\neg\beta_\eta}\right) f\left(\beta_\eta^{(i)}|\beta_\eta^*\right)}{f\left(\beta_\eta^{(i)}|y, \Theta, \Omega_{\neg\beta_\eta}\right) f\left(\beta_\eta^*|\beta_\eta^{(i)}\right)}$$

accepting the proposal with probability $q = \min\left(1, A\right)$, setting $\beta_\eta^{(i+1)} = \beta_\eta^*$. Otherwise we reject the proposal and set $\beta_\eta^{(i+1)} = \beta_\eta^{(i)}$. As outlined in Section 3.3 and detailed below, we consider two different methods for generating multivariate proposals for β_η. In the first approach (referred to as “MH”), we make an entirely stochastic Gaussian random walk proposal using a fixed covariance matrix; in the second (referred to as “mMALA”), we make a proposal which is partly deterministic and partly stochastic, accounting for local curvature of the likelihood surface.

For Metropolis-Hastings (MH) inference, we generate Gaussian random walk proposals of the form

$$\beta^*_\eta = \beta^{(i)}_\eta + (B'_\eta B_\eta + R_\eta)^{-1} \nu_\eta \epsilon$$

where ϵ is a vector of independent standard Normal random variables, and the value of step size ν_η is adjusted to achieve reasonable acceptance rates of approximately 0.25. For inference using the Riemann manifold Metropolis-adjusted Langevin algorithm (mMALA, as implemented by Girolami and Calderhead [13]) we propose using derivatives of the target distribution at the current sample. This promotes proposals in regions of higher probability, at the additional computational cost of computing necessary derivatives and matrix inverses. At iteration i of the sampling algorithm, where the current sample of the coefficients is $\beta^{(i)}_\eta$, proposals are made as

$$\beta^*_\eta = \beta^{(i)}_\eta + \frac{\nu^2_\eta}{2}\ G^{-1}\left(\beta^{(i)}_\eta\right)\ D\left(\beta^{(i)}_\eta\right) + \nu_\eta \sqrt{G^{-1}\left(\beta^{(i)}_\eta\right)}\ \epsilon$$

where ϵ is a vector of independent standard Normal random variables, ν_η is (adjustable) step size, and

$$D\left(\beta^{(i)}_\eta\right) = -\nabla_{\beta_\eta} L\left(\beta_\eta\right)\Big|_{\beta^{(i)}_\eta} \text{ and } G\left(\beta^{(i)}_\eta\right) = -\nabla_{\beta_\eta} \nabla^{\mathrm{T}}_{\beta_\eta} L\left(\beta_\eta\right)\Big|_{\beta^{(i)}_\eta}$$

are the negative gradient and negative Hessian of the log density, with

$$L\left(\beta_\eta\right) = \log f\left(\beta_\eta | y, \Theta, \Omega_{\neg\beta_\eta}, \Gamma\right) \text{ and } \nabla_{\beta_\eta} = (\partial/\partial\beta_{\eta 1}, \ldots, \partial/\partial\beta_{\eta p})^{\mathrm{T}}\ .$$

Computation of likelihood derivatives is described at the end of this Appendix.

Full conditional distributions for prior precisions

Prior precision parameter λ_η ($\eta = \xi, \nu$) has the following conditional distribution

$$f\left(\lambda_\eta | y, \Theta, \Omega_{\neg\lambda_\eta}, \Gamma\right) \propto f\left(\beta_\eta | \lambda_\eta\right) f\left(\lambda_\eta | a_\eta, b_\eta\right)\ .$$

By construction, since the Gamma distribution is a conjugate prior for the precision of a Gaussian distribution, we know that the full conditional distribution is also Gamma, with updated parameters

$$\hat{a}_\eta = a_\eta + \frac{p_\eta}{2} \text{ and } \hat{b}_\eta = b_\eta + \frac{1}{2}\beta^{\mathrm{T}}_\eta Q_\eta \beta_\eta\ .$$

Derivatives of the posterior distribution

Here we find the derivatives of the log posterior distribution, required for maximum likelihood and mMALA inference. The log likelihood of the observed data under the generalised Pareto distribution is

$$L(y|\Omega) = \begin{cases} \sum_{i=1}^N \left[-\log\left(\frac{\nu_i}{1+\xi_i}\right) - \left(\frac{1}{\xi_i} + 1\right) \log\left(1 + \frac{\xi_i}{\nu_i}(1+\xi_i) y_i\right)\right] & \text{for } \xi_i \neq 0 \\ \sum_{i=1}^N \left[-\log\left(\frac{\nu_i}{1+\xi_i}\right) - \frac{(1+\xi_i) y_i}{\nu_i}\right] & \text{for } \xi_i = 0 \end{cases}$$

The log conditional distribution for the vector of basis coefficients β_η ($\eta = \xi, \nu$) is then the sum of this likelihood plus a contribution from the prior distribution

$$\begin{aligned} L\left(\beta_\eta\right) &= \log\left(f\left(\beta_\eta | y, \Theta, \Omega_{\neg\beta_\eta}, \Gamma\right)\right) \\ &= L(y|\Omega) - \frac{\lambda_\eta}{2}\beta^{\mathrm{T}}_\eta Q_\eta \beta_\eta \end{aligned}$$

We note the obvious similarity between this expression and the penalised (negative log) likelihood used for maximum likelihood inference. The gradient of the log conditional distribution for vector of coefficients β_η ($\eta = \xi, \nu$) is

$$\nabla_{\beta_\eta} L\left(\beta_\eta\right) = \nabla_{\beta_\eta} L(y|\Omega) - Q_\eta \beta_\eta\ .$$

Using the chain rule, the likelihood gradient can be computed as

$$\begin{aligned} \nabla_{\beta_\eta} L\left(\beta_\eta\right) &= \left(D_{\beta_\eta}(B_\eta \beta_\eta)\right)^{\mathrm{T}} \left(\nabla_\eta L(y|\Omega)\right) \\ &= B^{\mathrm{T}}_\eta \left(\nabla_\eta L(y|\Omega)\right)\ . \end{aligned}$$

The components of $\nabla_{\xi} L\left(y|\Omega\right)$ are computed as

$$\frac{\partial}{\partial \xi_i} L\left(y\right) = \begin{cases} -\frac{1}{\xi_i^2 G_i}(1-2\xi_i)(G_i-1) + \frac{1}{1+\xi_i} + \frac{1}{\xi_i}\log\left(G_i\right) & \text{for } \xi_i \neq 0 \\ -\frac{y_i}{\nu_i} + \frac{1}{1+\xi_i} & \text{for } \xi_i = 0 \end{cases}$$

where $G_i = 1 + \frac{\xi_i}{\nu_i}(1+\xi_i)y_i$, and the components of $\nabla_{\nu} L\left(y|\Omega\right)$ are

$$\frac{\partial}{\partial \nu_i} L\left(y\right) = \begin{cases} \frac{1}{\nu_i}\left(1 - (\frac{1}{\xi_i}+1)\frac{G_i-1}{G_i}\right) & \text{for } \xi_i \neq 0 \\ \frac{1}{\nu_i}\left(1 - \frac{G_i-1}{\xi_i}\right) & \text{for } \xi_i = 0\,. \end{cases}$$

Differentiating $\nabla_{\beta_\eta} L\left(\beta_\eta\right)$ $(\eta = \xi, \nu)$ again gives the Hessian matrix

$$\nabla_{\beta_\eta}\nabla_{\beta_\eta}^{\mathrm{T}} L\left(\beta_\eta\right) = \nabla_{\beta_\eta}\nabla_{\beta_\eta}^{\mathrm{T}} L\left(y|\Omega\right) - Q_\eta\,.$$

Applying the chain rule

$$\nabla_{\beta_\eta}\nabla_{\beta_\eta}^{\mathrm{T}} L\left(y|\Omega\right) = B_\eta^{\mathrm{T}}\left(\nabla_\eta\nabla_\eta^{\mathrm{T}} L\left(y|\Omega\right)\right)B_\eta\,.$$

Note that the components of $\nabla_{\xi} L\left(y|\Omega\right)$ and $(\nabla_\eta\nabla_\eta^{\mathrm{T}} L\left(y|\Omega\right))$ are computed separately for $\eta = \xi$ and $\eta = \nu$. Further, the likelihood second derivatives with respect to ξ and ν are

$$\frac{\partial^2}{\partial \xi_i \partial \xi_j} L\left(y|\Omega\right) = \begin{cases} \frac{1}{(1+\xi_i)^2} & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

and

$$\frac{\partial^2}{\partial \nu_i \partial \nu_j} L\left(y|\Omega\right) = \begin{cases} \frac{1}{\nu^2(1+2\xi_i)} & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$$

such that Hessian matrices are diagonal. Moreover, the expectations of all of the cross derivatives $\frac{\partial^2}{\partial \xi_i \partial \nu_j} L\left(y|\Omega\right)$ are zero, since estimates of ξ and ν are asymptotically independent by construction (e.g. Chavez-Demoulin and Davison [4]).

References

[1] Anderson, C., Carter, D., Cotton, P., 2001. Wave climate variability and impact on offshore design extremes. Report commissioned from the University of Sheffield and Satellite Observing Systems for Shell International.

[2] Anderson, T. W., 1962. On the distribution of the two-sample Cramer-von Mises criterion. Ann. Math. Statist. 33, 1148–1159.

[3] Carter, D. J. T., Challenor, P. G., 1981. Estimating return values of environmental parameters. Quart. J. R. Met. Soc. 107, 259.

[4] Chavez-Demoulin, V., Davison, A., 2005. Generalized additive modelling of sample extremes. J. Roy. Statist. Soc. C 54, 207–222.

[5] Chavez-Demoulin, V., Embrechts, P., 2006. Smooth extremal models in finance and insurance. J. Risk. Ins. 71, 183–199.

[6] Coles, S., Walshaw, D., 1994. Directional modelling of extreme wind speeds. Appl. Statist. 43, 139–157.

[7] Cooley, D., Naveau, P., Jomelli, V., Rabatel, A., Grancher, D., 2006. A Bayesian hierarchical extreme value model for lichenometry. Environmetrics 17, 555–574.

[8] Cox, D. R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. J. Roy. Statist. Soc. B 49, 1–39.

[9] Davison, A., Smith, R. L., 1990. Models for exceedances over high thresholds. J. Roy. Statist. Soc. B 52, 393.

[10] Eilers, P. H. C., Marx, B. D., 2010. Splines, knots and penalties. Wiley Interscience Reviews: Computational Statistics 2, 637–653.

[11] Fawcett, L., Walshaw, D., 2006. A hierarchical model for extreme wind speeds. J. Roy. Statist. Soc. C 55, 631–646.

[12] Geyer, C. J., 1992. Practical Markov Chain Monte Carlo. Statist. Sci. 7, 473–483.

[13] Girolami, M., Calderhead, B., 2011. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. Roy. Statist. Soc. B 73, 123–214.

[14] Jonathan, P., Randell, D., Ewans, K., 2013. Joint modelling of extreme ocean environments incorporating covariate effects. Coastal Eng. 79, 22–31.

[15] Jonathan, P., Randell, D., Wu, Y., Ewans, K., 2014. Return level estimation from non-stationary spatial data exhibiting multidimensional covariate effects. Ocean Eng. 88, 520–532.

[16] MacKay, D., 1998. Introduction to Gaussian Processes. In: Bishop, C. (Ed.), Neural Networks and Machine Learning. Springer-Verlag, pp. 84–92.

[17] Mendez, F. J., Menendez, M., Luceno, A., Medina, R., Graham, N. E., 2008. Seasonality and duration in extreme value distributions of significant wave height. Ocean Eng. 35, 131–138.

[18] Perez-Cruz, F., July 2008. Kullback-Leibler divergence estimation of continuous distributions. In: Information Theory, 2008. ISIT 2008. IEEE International Symposium on. pp. 1666–1670.

[19] Randell, D., Feld, G., Ewans, K., Jonathan, P., 2015. Distributions of return values for ocean wave characteristics using directional-seasonal extreme value analysis. (Accepted for publication of Environmetrics in June 2015, draft at www.lancs.ac.uk/~jonathan).

[20] Randell, D., Zanini, E., Vogel, M., Ewans, K., Jonathan, P., 2014. Omnidirectional return values for storm severity from directional extreme value models: the effect of physical environment and sample size. Proceedings of 33nd International Conference on Ocean, Offshore and Arctic Engineering, San Fransisco OMAE2014-23156.

[21] Rasmussen, C., Williams, C., 2006. Gaussian Processes for Machine Learning. MIT Press.
URL `www.gaussianprocess.org/gpml/`

[22] Renard, B., Lang, M., Bois, P., 2006. Statistical analysis of extreme events in a nonstationary context via a bayesian framework. case study with peak-over-threshold data. Stochastic Environmental Research and Risk Assessment 21, 97–112.

[23] Scotto, M., Guedes-Soares, C., 2000. Modelling the long-term time series of significant wave height with non-linear threshold models. Coastal Eng. 40, 313–327.

[24] Smith, R. L., Naylor, J. C., 1987. A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. J. Roy. Statist. Soc. C 36, 358–369.