Ocean Engineering 207 (2020) 107406

Contents lists available at ScienceDirect

Ocean Engineering

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

Assessment of return value estimates from stationary and non-stationary extreme value models

Ed Mackay [a,*], Philip Jonathan [b,c]

[a] College of Engineering, Mathematics and Physical Sciences, University of Exeter, TR10 9FE, United Kingdom
[b] Shell Research Ltd., London SE1 7NA, United Kingdom
[c] Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YW, United Kingdom

ARTICLE INFO

Keywords:
Covariate
Extreme
Generalised Pareto
Metocean
Significant wave height
Non-stationary

ABSTRACT

This article compares the accuracy of return value estimates from stationary and non-stationary extreme value models when the data exhibits covariate dependence. The non-stationary covariate representation used is a penalised piecewise-constant (PPC) model, in which the data are partitioned into bins defined by covariates and the extreme value distribution is assumed to be homogeneous within each bin. A generalised Pareto model is assumed, where the scale parameter can vary between bins but is penalised for the variance across bins, and the shape parameter is assumed constant over all covariate bins. The number and sizes of covariate bins must be defined by the user based on physical considerations. Numerical simulations are conducted to compare the performance of stationary and non-stationary models for various case studies, in terms of quality of estimation of the T-year return value over the full covariate domain. It is shown that a non-stationary model can give improved estimates of return values, provided that model assumptions are consistent with the data. When the data exhibits non-stationarity in the generalised Pareto tail shape, the use of non-stationary model assuming a constant shape parameter can produce biases in return values. In such cases, a stationary model can give a more accurate estimate of return value over the full covariate domain as only the most extreme observations (regardless of covariate) are used to estimate tail shape. In other cases, the assumption of a stationary model will ignore key features of the data and be less reliable than a non-stationary model. For example, if a relatively benign covariate interval exhibits a long (or heavy) tail, extreme values from this interval may influence the T-year return value for very large T. However the sample of peaks over threshold, with high threshold, used to estimate a stationary model in this case may not include sufficient observations from this interval to estimate the return value adequately.

1. Introduction

Accurate estimation of extreme events is important in offshore and coastal engineering. Under-estimation of the magnitude of extreme events can lead to structural failures, whilst over-estimation can lead to overly-conservative and expensive designs and inefficient allocation of limited resources. Return periods of extreme events are usually estimated by fitting a statistical model to observed or modelled data and extrapolating into the tail of the distribution. The accuracy of estimated return values is dependent on numerous factors, including (a) quality of historic data (henceforth "dataset"), (b) length of dataset, (c) characteristics of the actual data-generating distribution, (d) misspecification of the statistical model, and (e) method used to estimate the statistical model.

Bias in metocean data obviously leads to bias in estimates of extremes. Random errors in metocean data lead to positive bias (i.e. a tendency to estimate return values that are higher than the true return values), since the distribution of random errors is convolved with the distribution of the variable (Forristall et al., 1996; Brooker et al., 2004). Shorter datasets lead to higher variance in estimates of extremes, but can also increase bias, since bias in parameter estimators for various distributions can vary with the number of observations. Similarly, the shape of the tail of the distribution affects both the bias and variance of estimates of extreme values, with estimates of longer-tailed distributions having a higher variance for a given sample size. Biases in parameter estimates also vary with the shape of the tail (see e.g.de Zea Bermudez and Kotz (2010), Kang and Song (2017)).

Model misspecification refers to differences between the "true" characteristics of the data (and the data-generating model responsible for it) and the assumptions made in the statistical model. At

* Corresponding author.
E-mail address: e.mackay@exeter.ac.uk (E. Mackay).

https://doi.org/10.1016/j.oceaneng.2020.107406
Received 23 December 2019; Received in revised form 16 March 2020; Accepted 15 April 2020
Available online 27 April 2020

present, the most commonly applied method for estimating extremes of metocean variables is the peaks-over-threshold (POT) method (see e.g. Coles (2001), Jonathan and Ewans (2013)). The POT method makes the following key assumptions about the data: (1) observations are independent and identically distributed (IID) given covariates, and (2) exceedances of a sufficiently high threshold follow a generalised Pareto (GP) distribution. The GP distribution describes the asymptotic behaviour of independent threshold exceedances from a max-stable data-generating distribution. As threshold level increases, theory suggests that the closeness of the conditional distribution of peaks over threshold to the GP form improves. The appropriateness of the GP distribution is therefore based on the threshold being sufficiently high that the asymptotic approximation is valid, with too low a threshold leading to increased bias in the estimated extreme values due to the GP distribution not being an appropriate model. The choice of threshold is a trade-off between increased bias from setting the threshold low and increased variance from setting the threshold high, so that there are fewer observations. The rate of convergence of threshold exceedances from the data-generating distribution to the GP distribution may however be slow. That is, a very large threshold might be required for the GP form to be considered a reasonable approximation, making practical inference difficult for finite samples. For this reason, a number of "pre-asymptotic" parametric distributional forms, or "penultimate approximations", have been proposed (Beirlant et al., 2012; Gomes, 2014); the idea being that the data-generating distribution is in some sense "closer" to the penultimate approximation than to the asymptotic distribution for finite threshold. However, since the GP model is the most widely used at present it has been applied in the present work and the use of penultimate approximations is not pursued further here. In addition, a large literature on non-parametric alternatives for estimation of distributional tails exists (see e.g. Hill (1975), Dekkers et al. (1989)).

Regarding the assumption that observations are IID given covariates, metocean variables typically exhibit serial correlation, so the assumption of independence is not true if a model is fitted to all observations. This is dealt with by declustering the data, where only the largest observation in each storm are considered so that storm maxima can be considered approximately independent. The criterion for what determines independent storms is usually defined in terms of a minimum separation in time. A rigorous treatment of the correlation between successive extreme events can be made by plotting the extremogram (Davis and Mikosch, 2009), an analogue of the autocorrelation function for sequences of extreme events, although care must be taken to first remove the seasonal signal from the data which introduces a longer-range correlation. An example of the use of the extremogram to define a declustering time-scale was presented by Mackay and Johanning (2018), which showed that a time-scale of around 5 days was sufficient for the datasets considered in that study. Alternatively, declustering times can be defined based on more heuristic arguments about the average time scales for weather systems to pass over a site, typically taken to be in the range of 2–5 days. Ewans and Jonathan (2008) discuss a physically-motivated approach to declustering time series of significant wave height, H_s, based on the assumption that the peak severities of different storm events, given covariates, are statistically independent. Storm events are identified from time-series of sea-state H_s. A storm event corresponds to the time interval between the H_s up-crossing of some threshold level and the subsequent down-crossing of the threshold. In addition, storm intervals separated by less than 24 h are merged. The threshold can be defined e.g. in terms of a covariate-dependent quantile of sea-state H_s. The peak value of sea-state H_s during the storm interval then defines the storm peak H_s. Values of storm peak H_s for different storms are taken to be statistically independent.

The distributions of many metocean variables, such as (significant or individual) wave heights, wind speeds and storm surge, exhibit dependence on other variables, referred to as covariates. For example, many studies have considered the dependence of wind speeds or wave heights on the direction of origin of the storm and the time of year (season) (Fawcett and Walshaw, 2006; Méndez et al., 2008; Ewans and Jonathan, 2008; Randell et al., 2015; Jones et al., 2016). Wave heights and wind speeds are also dependent on large-scale climatic indices such as the North Atlantic Oscillation (NAO) (Woolf et al., 2002) or the El Niño Southern Oscillation (ENSO) (Bromirski et al., 2005). Moreover, most studies tacitly assume that the distribution of metocean variables are stationary in time, neglecting the effects of the changing climate which have been observed in some studies (Reguero et al., 2019; Cattrell et al., 2019).

In this study we focus on the effects of periodic covariates such as season and direction and defer consideration of longer-term variations in climate to future work. Specifically, we quantify differences in the performance of models which account for the covariate effects and those that do not (referred to here as constant or stationary models). Obviously, stationary models cannot produce estimates of seasonal or directional extremes, so our interest here is in which model gives the more accurate estimates of return values for the full covariate domain, typically referred to as annual (omniseasonal) or omnidirectional return values; we use the term "omnicovariate" where necessary below for clarity. The quality of historical data is not considered here, but all the other factors listed above influence the comparison between stationary and non-stationary models and therefore need to be considered.

There has been some debate in the literature about the circumstances in which non-stationary models should be applied and whether stationary or non-stationary models produce more accurate estimates of omnicovariate return values. The motivation for using non-stationary models is that their underpinning assumptions better reflect the characteristics of the data and our physical understanding. Non-stationary models assume that the distributions of independent peaks over (covariate-dependent) threshold, conditional on covariates, tend towards a GP distribution. Under this assumption, now consider the highest threshold value (on the covariate domain) for which the GP distribution is a reasonable approximation. For this threshold value, the omnicovariate distribution is a convolution of the GP distributions over the covariate domain and therefore not a GP distribution itself. If we now increase the threshold yet further, we expect to eliminate the influence of all values of covariate except those contributing to the extreme tail of the omnicovariate distribution and that the resulting distribution of threshold exceedances would be "closer" to a GP distribution once more. It may therefore be expected that a high threshold would be required for stationary models to give a similar level of performance as non-stationary models. However, it is also reasonable to expect that the most information about the shape of the tail of the omnicovariate distribution is contained in the largest observations. In applied extreme value analysis there is a maxim that ensuring a good fit to the bulk of the data does not guarantee a good fit to the tail. It is therefore reasonable to ask whether modelling less extreme observations (in a non-stationary model) reduces the bias and variance of return values.

Model complexity is another consideration. Stationary models are simpler to implement and have fewer parameters to estimate. Whilst the complexity of non-stationary models is not an argument against their use on its own, practical considerations aside, it may be expected that the larger number of parameters that need to be estimated for non-stationary models would increase the variance of those estimates. The need to estimate more parameters is traded off against two effects. Firstly, due to the larger number of parameters, non-stationary models offer a more flexible (hence potentially more accurate) fit to the data. Secondly, non-stationary models are typically fitted to a larger proportion of the observations, increasing the sample size. From this discussion it is apparent that theoretical arguments alone cannot justify the use of a stationary or non-stationary model exclusively. From the practitioner's perspective, the challenge is knowing which type of model gives the most accurate estimates of extreme values in a given situation.

Many of the earlier studies on the use of non-stationary models (e.g. Carter and Challenor (1981), Morton et al. (1997), Anderson et al. (2001)) compared their performance to stationary models in situations where the true return values were not known. In these studies it is not possible to conclude which type of model is more accurate, only that results differ. Jonathan et al. (2008) presented a comparison of stationary and non-stationary models using simulated data where the observations are drawn from two distinct distributions, representing storms from two directions. They demonstrate that in the cases they consider the non-stationary models give lower bias in estimates of return values. Mackay et al. (2010) argued that these results were not representative of real situations, where the distribution of a variable will vary smoothly with direction, season or other covariate, rather than changing sharply at the boundary of two sectors. Mackay et al. (2010) presented the results of simulations where the distribution of storm peak H_S conditioned on season varied continuously through the year. Piecewise-constant models were fitted, where the data were divided into a number of discrete bins and independent fits were made in each bin. It was shown that the piecewise-constant models performed worse in estimating omniseasonal return values than the stationary models, with higher bias and variance in all cases considered, and with both bias and variance increasing with the number of bins used. It was explained that the reason the non-stationary models performed worse in these case studies was due to the independent estimates of the GP shape parameter in each bin. As the number of bins increases, the sample size in each bin decreases and the variance of the parameter estimates increases. A high estimate of the GP shape parameter in one bin is not compensated for by a low estimate in another bin and therefore leads to a positive bias in the annual return values. Jonathan and Ewans (2011) argued that the results in Mackay et al. (2010) were due to a fortuitous choice of extreme value threshold for the stationary model and that there was no way of knowing in practice where the correct threshold should be set. Jones et al. (2016) extended the study of Jonathan et al. (2008) using more sophisticated covariate representations (splines, Fourier series and Gaussian processes), and suggested that the performance of stationary models in estimating omnicovariate models is in general more variable than the performance of a non-stationary model.

The purpose of the present study is to extend the results of Jonathan et al. (2008) and Mackay et al. (2010) in an attempt to provide further guidance on the relative performance of stationary and non-stationary models in realistic situations. We extend the results from Mackay et al. (2010) in two main ways. Firstly, case studies are constructed where the threshold for both the stationary and non-stationary models can be varied, so that the effect of threshold choice can be examined. Secondly, we consider a penalised piecewise-constant (PPC) non-stationary model (Ross et al., 2018, 2019). In this model the data are partitioned into bins defined by covariates, and the GP scale parameter is allowed to vary between bins but the shape parameter is constant over all bins. The likelihood function used to estimate the parameters is penalised for the variance in the scale parameter over all the bins, with the roughness penalty selected using cross-validation to maximise predictive likelihood.

More advanced non-stationary models than the PPC model have been proposed, which have the objective of providing optimally flexible descriptions of the systematic variability of extreme values with covariate (e.g. Zanini et al. (2020)). Typically, a regression approach underpins these models (e.g. Northrop et al. (2016)). A suitable set of basis functions for the covariate domain is defined, and the value of each of the extreme value model parameters (on the covariate domain) is then defined as a linear combination of basis functions; the basis coefficient vector is estimated statistically. Suitable bases for one-dimensional covariate domains include splines and Fourier series. Basis functions with compact support, such as B-splines, are advantageous computationally; PPC exploits a piecewise-constant basis in one-dimension. There are numerous variants of spline parameterisations. These include P-splines (penalised B-splines, Eilers and Marx (2010)), for which squared differences of neighbouring basis coefficients are penalised to increase the smoothness of the representation, and adaptive regression splines (e.g. Biller (2000)), for which locations of spline basis knots are also estimated to optimise model fit. Useful bases for higher-dimensional covariates include thin-plate splines (e.g. Wood (2003)), suitable kernels (e.g. radial basis functions), and Voronoi tessellations (e.g. Bodin and Sambridge (2009)); bases for higher-dimensional covariates can also be formed from tensor products of lower-dimensional bases (e.g. Raghupathi et al. (2016)). Higher-dimensional bases formed from tensor products of penalised B-splines admit efficient computation using generalised linear additive models (Currie et al., 2016).

The motivation for using the PPC model over more advanced forms of non-stationary model is that is represents a good compromise between simplicity, robustness and flexibility. The PPC model represents a step up in complexity from binning the data and fitting independent models in each bin (the non-stationary model considered by Mackay et al. (2010)), where the additional complexity of the roughness penalisation makes the model more robust to increasing uncertainties from dividing the data into bins. The complexity of the PPC model is determined by the number of bins used, rather than the number of covariates. It can therefore be used for multidimensional covariate problems without modification, making it very flexible.

As with previous studies, the scope of the current study is necessarily limited to a relatively small number of case studies. Hence the conclusions drawn here may not be applicable universally. The results presented apply to the PPC model and similar types of non-stationary model. However, we have also attempted to draw more general conclusions that extend to other types of non-stationary model. In particular, the conclusions about the effects of binning the data and assuming a piecewise-constant distribution apply to other types of model that take this approach, and the conclusions about the effects of assuming a stationary shape parameter are likely to be applicable to any non-stationary model that makes this assumption. Moreover, as discussed further in Section 3, since the PPC model only considers the total level of variability between bins, the particular choice of patterns of covariate dependence are not restrictive.

The paper is organised as follows. A brief overview of the theory and model assumptions is presented in Section 2. The design of the simulation case studies is described in Section 3. In Section 4 we examine the effect of non-stationarity in the data on the shape of the tail of the omnicovariate distribution, and the effect this has on quality of estimation from return values from a stationary fitted model. The effect of partitioning the data into bins and fitting a piecewise-constant non-stationary model is considered in Section 5. Section 6 summarises the results of the simulation studies. Finally, conclusions are given in Section 7.

2. Theory and assumptions

2.1. Return values from a non-stationary distribution

In the present study we consider estimation of the distribution of an arbitrary variable, X, where X could be interpreted as storm-peak H_s or another environmental variable, showing dependence on covariates. It is assumed that storm peaks are sufficiently separated in time that adjacent observations are independent. Further, it is assumed that X follows some arbitrary distribution, with parameters dependent on one or more covariates. In the current study we consider the influence of a single covariate, denoted ι, which could be interpreted as the time of year (season) or mean wave direction at the storm peak.

Denote the cumulative distribution function (CDF) of X conditional on a particular choice of ι as $P_S(X \leq x|\iota)$. For simplicity, it is assumed that $\iota \in \mathcal{T} = [0, 360)$. It is further assumed that the occurrence rate of storm peaks, $\rho_\iota(\iota)$, is dependent on ι, where the rate is defined as the

number of storms per year per unit covariate. The probability that a storm, selected at random, has associated covariate t is

$$p_t(t) = \frac{\rho(t)}{M}, \tag{1}$$

where

$$M = \int_0^{360} \rho(t)\mathrm{d}t \tag{2}$$

is the expected number of storms per year.

The unconditional CDF of X for a storm selected at random, denoted P_{RS}, is obtained by integrating the conditional CDF over the covariate domain, weighted by occurrence

$$P_{RS}(X \leq x) = \int_0^{360} P_S(X \leq x|t)\ p_t(t)\ \mathrm{d}t. \tag{3}$$

The T-year return value, x_T, is then the solution of

$$P_{RS}(X > x_T) = \frac{1}{TM}. \tag{4}$$

2.2. Penalised piecewise-constant (PPC) model

Consider a sample $D = \{x_i\}_{i=1}^n$ of n values of peaks over threshold for a random variable X. Further, let $\{t_i\}_{i=1}^n$ be the corresponding values of a covariate t on some domain $\mathcal{T}$. We assume a single covariate, but extension to more complex covariate domains is straightforward as explained in Ross et al. (2018). We make inferences about extreme values of X given t, for $t \in \mathcal{T}$.

The piecewise-constant model uses a particularly simple description of non-stationarity with respect to covariates. For each observation in the sample, the value of covariate t_i is used to allocate the observation to one and only one of N_{bin} covariate intervals (or bins) $\{C_k\}_{k=1}^{N_{bin}}$ by means of an allocation vector A such that $k = A(i)$ and $\mathcal{T} = \bigcup C_k$. For each k, all observations in the set $\{x_i\}_{A(i)=k}$ with the same covariate interval C_k are assumed to have common extreme value characteristics.

A non-stationary GP model is then estimated using cross-validated roughness-penalised maximum likelihood estimation. For covariate interval C_k, the extreme value threshold $u_k > 0$ is assumed to be a quantile of the empirical distribution of X in that interval, with specified non-exceedance probability $\psi \in (0, 1)$, with ψ constant across intervals, and estimated by counting. Threshold exceedances are assumed to follow the GP distribution with shape $\xi \in [-0.5, \infty)$ and scale $\sigma_k > 0$, with CDF

$$F_{GP}(x|\xi, \sigma_k, u_k) = 1 - z_k, \tag{5}$$

where

$$z_k = \begin{cases} \left(1 + \xi(x - u_k)/\sigma_k\right)^{-1/\xi}, & \xi \neq 0, \\ \exp\left(-(x - u_k)/\sigma_k\right), & \xi = 0 \end{cases} \tag{6}$$

per covariate interval C_k. F_{GP} is defined on $x \in (u_k, x_k^+)$ with $x_k^+ = u_k - \sigma_k/\xi$ when $\xi < 0$ and ∞ otherwise. The parameters u_k, σ_k and ξ are the threshold, scale and shape parameters, respectively. Since estimation of the shape parameter is particularly problematic, ξ is assumed constant (but unknown) across covariate intervals, and the reasonableness of the assumption assessed by inspection of diagnostic plots. Parameters $\xi, \{\sigma_k\}$ are estimated by maximising the predictive performance of a roughness-penalised model, optimally regulating the extent to which $\{\sigma_k\}$ varies across intervals, using a cross-validation procedure.

The sample GP likelihood $\mathcal{L}$ under the piecewise stationary model is

$$\mathcal{L} = \prod_{k=1}^{N_{bin}} \prod_{\substack{i:\, A(i)=k;\\ x_i > u_k}} \frac{1}{\sigma_k}\left[1 + \frac{\xi}{\sigma_k}\left[x_i - u_k\right]\right]^{-1/\xi - 1}, \tag{7}$$

where $\mathcal{L}, \{u_k\}, \{\sigma_k\}$ and ξ are functions of marginal extreme value threshold non-exceedance probability ψ, and ξ is constant across the N_{bin} intervals $\{C_k\}$. The negative log likelihood, penalised for the roughness of $\{\sigma_k\}$ across intervals, is then

$$\ell^* = -\log \mathcal{L} + \lambda_\sigma \left(\frac{1}{N_{bin}} \sum_{k=1}^{N_{bin}} \left[\sigma_k - \bar{\sigma}\right]^2\right), \tag{8}$$

where ℓ^* is a function of both ψ and roughness coefficient λ_σ and $\bar{\sigma}$ is the mean value of σ_k over the bins:

$$\bar{\sigma} = \frac{1}{N_{bin}} \sum_{j=1}^{N_{bin}} \sigma_j. \tag{9}$$

For given ψ and λ_σ, estimates for ξ and $\{\sigma_k\}$ are found by minimising ℓ^*. The minimisation is conducted using a simplex search method (Lagarias et al., 1998). The search is initialised using first guess of $\hat{\xi} = 0$ (where the caret $\hat{}$ denotes an estimate of a parameter) and the moment estimates of σ in each interval. The optimisation is constrained to give $\hat{\xi} \geq -0.5$ and $\max\left\{x_i | A(i) = k\right\} \leq \hat{u_k} - \hat{\sigma_k}/\hat{\xi}$ when $\hat{\xi} < 0$. A random 10-fold cross-validation is then used to select the value $\hat{\lambda}_\sigma$ of λ_σ and corresponding $\hat{\xi}$, $\{\hat{\sigma}_k\}$ which, for each ψ, maximises predictive performance. In the PPC model, if $\lambda_\sigma = \infty$ then the model has only one degree of freedom for σ, whereas if $\lambda_\sigma = 0$ then the fitted model has N_{Bin} degrees of freedom for σ. For intermediate values, the "effective" degrees of freedom for σ is at some intermediate value.

In a typical application, the complete PPC modelling procedure is repeated for a number of bootstrap resamples of the original sample to capture sampling uncertainty. Moreover, for each sample, the extreme value model is evaluated for multiple thresholds with non-exceedance probability ψ drawn at random from the interval $\mathcal{I}_\psi \subseteq (0, 1)$ on which model performance is deemed reasonable from inspection of diagnostics. However, in the current study, where the model is applied in a large number of Monte Carlo simulated datasets, only the original sample is used. Moreover, as we wish to study the effect of threshold level on the estimates, the PPC model is fitted for several values of ψ and the results compared directly.

The method used to fit the PPC model is relatively simple. It is conceivable that other methods such as Markov Chain Monte Carlo (MCMC) could potentially improve results. However, this would represent a significant step up in terms of complexity. As mentioned above, the motivation for using the PPC model is for its balance between simplicity and flexibility. Examples of non-stationary models using MCMC can be found in e.g. Hansen et al. (2020), Zanini et al. (2020).

Once the PPC model has been estimated the omnicovariate distribution is obtained using the discretised form of (3):

$$\hat{P}_{RS}(X \leq x) = \frac{1}{n_T} \sum_{k=1}^{N_{bin}} n_k F_{GP}(x|\hat{\xi}, \hat{\sigma_k}, \hat{u_k}), \tag{10}$$

where n_k is the number of observations in interval C_k and

$$n_T = \sum_{k=1}^{N_{bin}} n_k. \tag{11}$$

Return values can then be estimated using (4). For consistency, the stationary model used in this work is a special case of the PPC model with a single covariate bin and no roughness penalisation.

2.3. Assessment criteria

The performance of the stationary and non-stationary models are assessed in terms of the bias, standard deviation (STD) and root-mean-square error (RMSE) of estimated model parameters and return values over N realisations of Monte Carlo simulated datasets. Let $\hat{\theta}$ denote an estimator of either a model parameter or return value, θ. The expected value, bias, STD and RMSE of the estimator are defined as

$$E(\hat{\theta}) = \frac{1}{N} \sum_{j=1}^{N} \hat{\theta}_j, \tag{12}$$

$$\mathrm{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta, \tag{13}$$

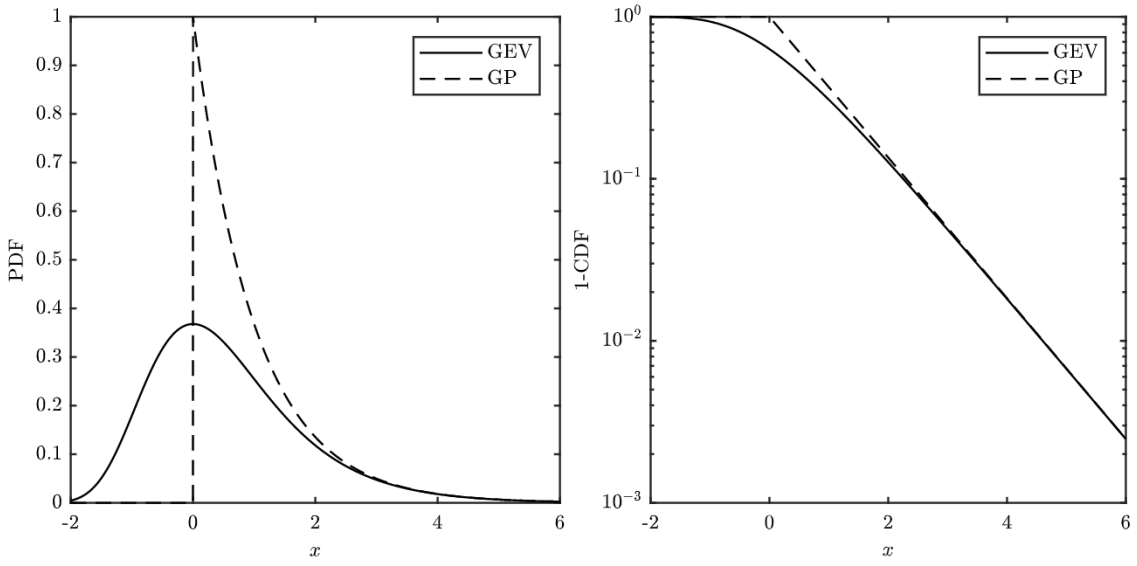

Fig. 1. Comparison of GEV and GP probability density functions (PDFs) and cumulative distribution functions (CDFs, shown as the tail) for the case $u = \mu = 0$, $\sigma = 1$, $\xi = 0$.

Table 1
Model parameters used in the second set of case studies. See Eqs. (20)–(22) for functional forms of data-generating GEV distribution model parameterisations.

	α	β	γ
Case 1	0, 1, 2, 3	0	0
Case 2	0	0.25, 0.5	0
Case 3	1	0.25, 0.5	0
Case 4	1	0.5	±0.1, ±0.2

$$\mathrm{STD}^2(\hat{\theta}) = \frac{1}{N}\sum_{j=1}^{N}\left(\hat{\theta}_j - E(\hat{\theta})\right)^2, \tag{14}$$

$$\mathrm{RMSE}^2(\hat{\theta}) = \frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_j - \theta)^2 = \mathrm{bias}^2(\hat{\theta}) + \mathrm{STD}^2(\hat{\theta}), \tag{15}$$

where $\hat{\theta}_j$ is the estimate corresponding to the jth Monte Carlo simulated dataset (henceforth referred to as a "trial" for brevity).

3. Design of case studies

Previous simulation studies comparing stationary and non-stationary models have generated data from a GP distribution, where the parameters depend on covariate values. The limitation of this type of study is that the minimum threshold for which the stationary model can be applied is the maximum threshold value over all covariates, since below this level the distribution is not defined for all covariate values. To overcome this limitation, a model is required for the distribution of all storm-peak data, not just the tails. This could be achieved by using a two-part model with a parametric distribution for the body of the distribution and a GP model for the tail. The problem with this approach is that the choice of distribution for the body is arbitrary and it is difficult to ensure continuity of the density function on the boundary between body and tail.

In our simulations we have opted to simulate from the generalised extreme value distribution (GEV) rather than the GP distribution, avoiding the need for a two-part model. Previous investigations (details available from the authors on request) with measured data also show that the GEV distribution is a reasonable model for storm-peak H_S. The GEV is the asymptotic distribution of "block maxima" of fixed block size (e.g hourly, daily or weekly maxima). Storm peak data can be considered block maxima in a sense, where the block size is related to the method used for identifying storm peaks, although the block size is not strictly constant. However, we are not using the GEV to generate data to conduct a block-maxima analysis. Instead, we are using the GEV to generate data for a non-stationary POT analysis (using the PPC model). A POT analysis can be applied to data generated from any distribution. The motivation for using the GEV as the data-generating distribution in the current study, is that it has the convenient property that the tail converges to the GP distribution with the same shape parameter, in the sense illustrated below. The CDF of the GEV can be written as

$$F_{GEV}(x) = \exp(-z), \tag{16}$$

where z is defined in the same way as the for the GP distribution in (6). In the tail of the distribution z is small. As $z \to 0$ we have $\exp(-z) \to 1-z$ and $F_{GEV}(x|\mu,\sigma,\xi) \to F_{GP}(x|\mu,\sigma,\xi)$. That is, the GEV and GP CDFs converge, with common scale and shape parameters, and GEV location parameter μ equal to the GP threshold u, as illustrated in Fig. 1 for the case $\mu = 0$, $\sigma = 1$, $\xi = 0$.

It is well known that there is a relation between block-maxima modelled using the GEV distribution and threshold exceedances modelled using the GP distribution (see e.g. Coles (2001)). However, the argument above merely relates to the similarity of the functional forms of the GEV and GP tails, and is not the same as the argument associating the GP distribution for peaks over threshold when the GEV is used for block maxima.

Fitting a GP model to a dataset with GEV as the data-generating model will introduce some bias at lower threshold values, due to the mismatch between the fitted model and data-generating model (see Fig. 1). The resulting bias and STD of parameter and quantile estimates when fitting the GP distribution to GEV data is examined in the Appendix. The bias in parameter and quantile estimates are slightly higher when a GP model is fitted to GEV data than when a GP model is fitted to GP data. However the STD is slightly lower, resulting in an RMSE that is comparable. The use of the GEV distribution as the data-generating model rather than the GP distribution will therefore not significantly influence the results.

For the PPC model, the likelihood is penalised on the variance of the scale parameter. The difference between the estimates of the scale parameters in adjacent bins is not considered explicitly, only the total variance over all bins. The complexity of the PPC model is therefore determined by the total number of covariate bins only and not the number of covariates used. Therefore, the case studies considered here focus on a single covariate and the results can be expected to apply to cases with multiple covariates.

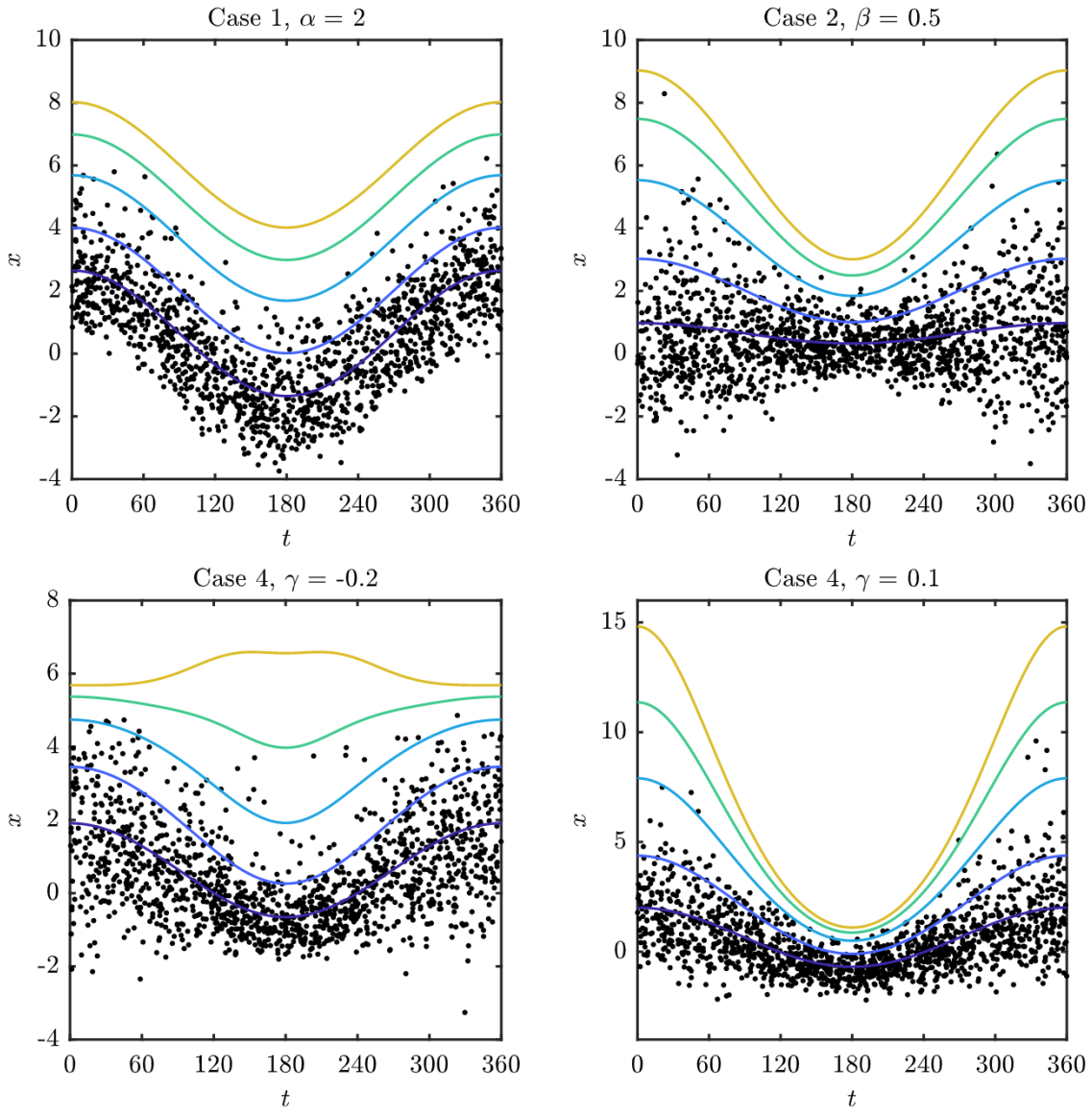

Fig. 2. Illustration of some of the case studies considered. Black dots: Simulated 20 year datasets. Coloured lines: Quantiles at non-exceedance probabilities of $\psi = 0.6$, 0.9, 0.99, 0.999 and 0.9999. See Eqs. (20)–(22) and Table 1.

We now consider two sets of case studies. In the first, the GEV parameters are assumed to vary linearly with covariate t, and in the second the parameters are assumed to vary sinusoidally with t. The parameters in the first set of case studies are defined as

$$\mu = at/360, \qquad a \in [-3, 3], \tag{17}$$

$$\sigma = 1 + bt/360, \qquad b \in [0, 2], \tag{18}$$

$$\xi = -0.1. \tag{19}$$

The first set of case studies is designed to illustrate the effect of fitting a stationary extreme value model to data from a non-stationarity data-generating distribution, and is similar to the PPC fit in a specific covariate bin (see Section 4). The parameters in the second set of case studies are defined as

$$\mu = \alpha \cos\left(\frac{2\pi t}{360}\right), \tag{20}$$

$$\sigma = 1 + \beta \cos\left(\frac{2\pi t}{360}\right), \tag{21}$$

$$\xi = -0.1 + \gamma \cos\left(\frac{2\pi t}{360}\right), \tag{22}$$

where different choices of α, β and γ are also considered. As the PPC model does not directly account for the difference in parameter estimates between adjacent bins on the covariate domain, it is mainly the level of variation between bins that influences model fit and not the pattern of variation. The assumption of sinusoidal variation in model parameters is therefore not particularly restrictive. However, it will be shown in Section 6.2 that the level of non-stationarity of the data within a bin does influence model fit.

The second set of case studies is designed to be more representative of a real situation and are used to compare the performance of the stationary and non-stationary models. The GEV parameters for each case are listed in Table 1. The first case with $\alpha = \beta = \gamma = 0$ is included to illustrate the effect of increasing the number of bins on the estimated omnicovariate return values in absence of covariate effects and is discussed in Section 5. The subsequent cases illustrate the effect of different patterns in the variation of the data-generating distribution parameters and are discussed in Section 6.

For each case, simulations are conducted as follows. The sample size is fixed at 1440 observations (this corresponds to a mean time between storm peaks of 5 days and a dataset length of 20 years, if a year is assumed to last 360 days, as defined in Section 2). The occurrence

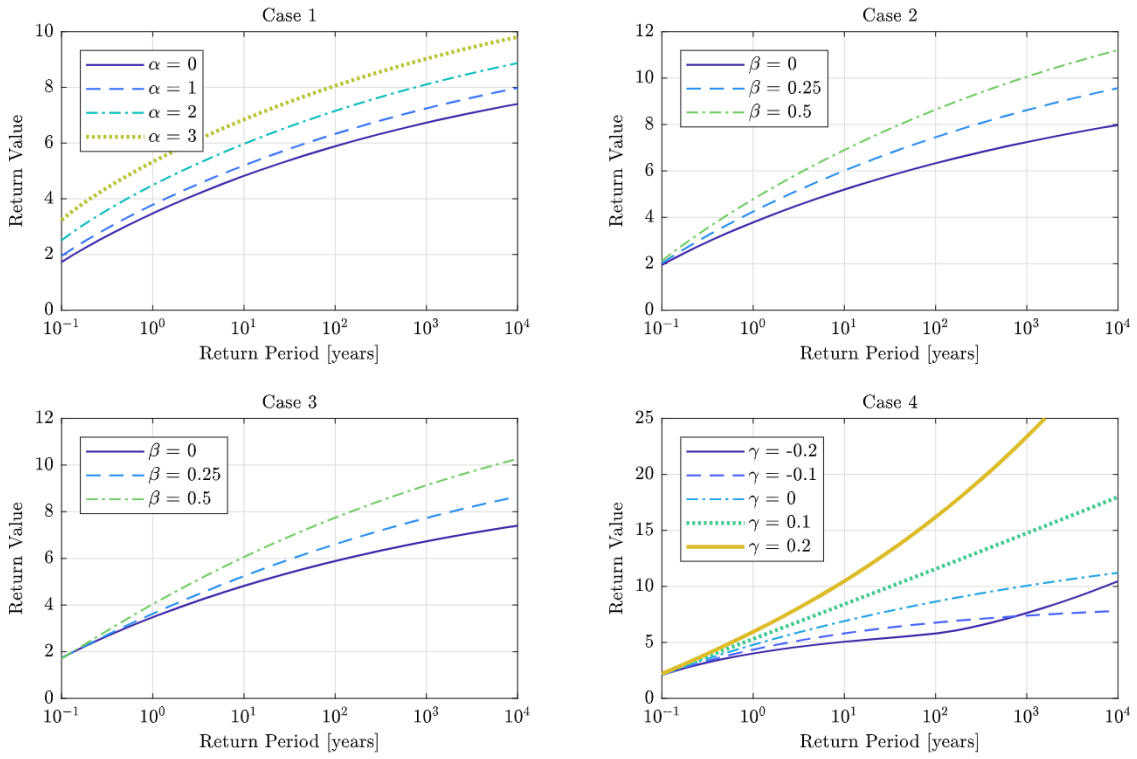

Fig. 3. True return values for the four cases described in Table 1.

rate is assumed to be constant with covariate, so the values of the covariate t are simulated as uniformly distributed numbers in $[0, 360)$. After the values of t have been simulated, GEV parameters are defined for each storm peak, conditional on t and a random value of X is generated. The PPC model is fitted to the data using between 1 and 8 (or sometimes 12) bins, with bin edges spaced evenly over the covariate domain, with the first bin centred at $t = 0$. The threshold for the GP model in each bin is defined as the empirical quantile corresponding to a fixed non-exceedance probability, ψ, where the levels are set at $\psi = 0.6, 0.7, 0.8, 0.9$. For the cases where the observations are partitioned into 8 bins this gives approximately $n = 72, 54, 36, 18$ threshold exceedances per bin respectively.

For each case 10,000 trials were performed. The estimation of the optimal penalty via cross-validation is the most time consuming step in estimating the PPC model. To reduce computation, the optimal penalty is estimated for only the first 100 trials; subsequent trials use the median penalty from the first 100 trials. The estimated optimal penalty showed very little variation over the first 100 trials, justifying the use of the median value in the remaining trials.

Examples of simulated datasets for four of the cases (see Eqs. (20)–(22) and Table 1) are shown in Fig. 2, together with the theoretical quantiles at non-exceedance levels of $\psi = 0.6$, 0.9, 0.99, 0.999 and 0.9999. Since we have defined the location parameter to be sinusoidally varying about zero, there are some observations that are negative, which is not representative of some environmental variables such as storm peak wave heights or wind speeds. This could be rectified by adding an offset to μ, which would have the effect of offsetting all the observations. However, the choice of offset would be arbitrary, so has been left as zero. The return values for each case, calculated from (3) and (4), are shown in Fig. 3.

In Case 1, the location parameter varies with t whilst other parameters remain constant. The result is an offset in the return value curves, which grows with α. The offset in the return values does not change much with return period. In Case 2, the scale parameter varies with t while other parameters are held constant. The resulting distribution, shown in Fig. 2 is 'pinched' in the middle. This pattern of variation is less representative of real situations, but is included for illustration. The resulting effect on return values grows with return period. In Case 3, both the scale and location parameters vary with t, which is more representative of real situations. Finally, in Case 4 all the parameters are non-stationary. When $\gamma = -0.2$ there is a change in the gradient of the return value curve that occurs at a return period of approximately 100 years. For other values of γ the return value varies smoothly with return period. In Case 4 both the stationary and PPC models are misspecified, since the PPC model assumes a constant shape parameter. The assumption of a stationary shape parameter is commonly used in oceanographic applications (e.g. Davison and Smith (1990), Anderson et al. (2001)). It is therefore interesting to assess how well the PPC model performs in this situation. The cases with $\gamma = \pm 0.2$ may be less realistic for metocean variables, due to the positive shape parameter in some sectors. However, they are instructive to include as they illustrate some potentially important effects.

The performance of the stationary and non-stationary models is assessed in terms of bias, STD and RMSE in the 100-year and 1000-year return values. As the size of the return values differs between cases, we need to compare the relative size of the uncertainties in estimates. To achieve this, we have normalised the bias, STD and RMSE by the size of the true return values. This means that the normalised results are influenced by the arbitrary choice of the mean value of the location parameter μ. For a larger mean value of μ the true return values would increase and the normalised bias, STD and RMSE would be reduced. However, since any normalisation is somewhat arbitrary, we have opted to use this convention. The choice of normalisation used here does not influence the conclusions of the study in terms of which model performs better, the choice of normalisation only influences the relative magnitude of the effects.

4. Fitting a stationary model to data from a non-stationary data-generating distribution

Here we examine the effect of non-stationarity in the data-generating model on the tail of the estimated omnicovariate

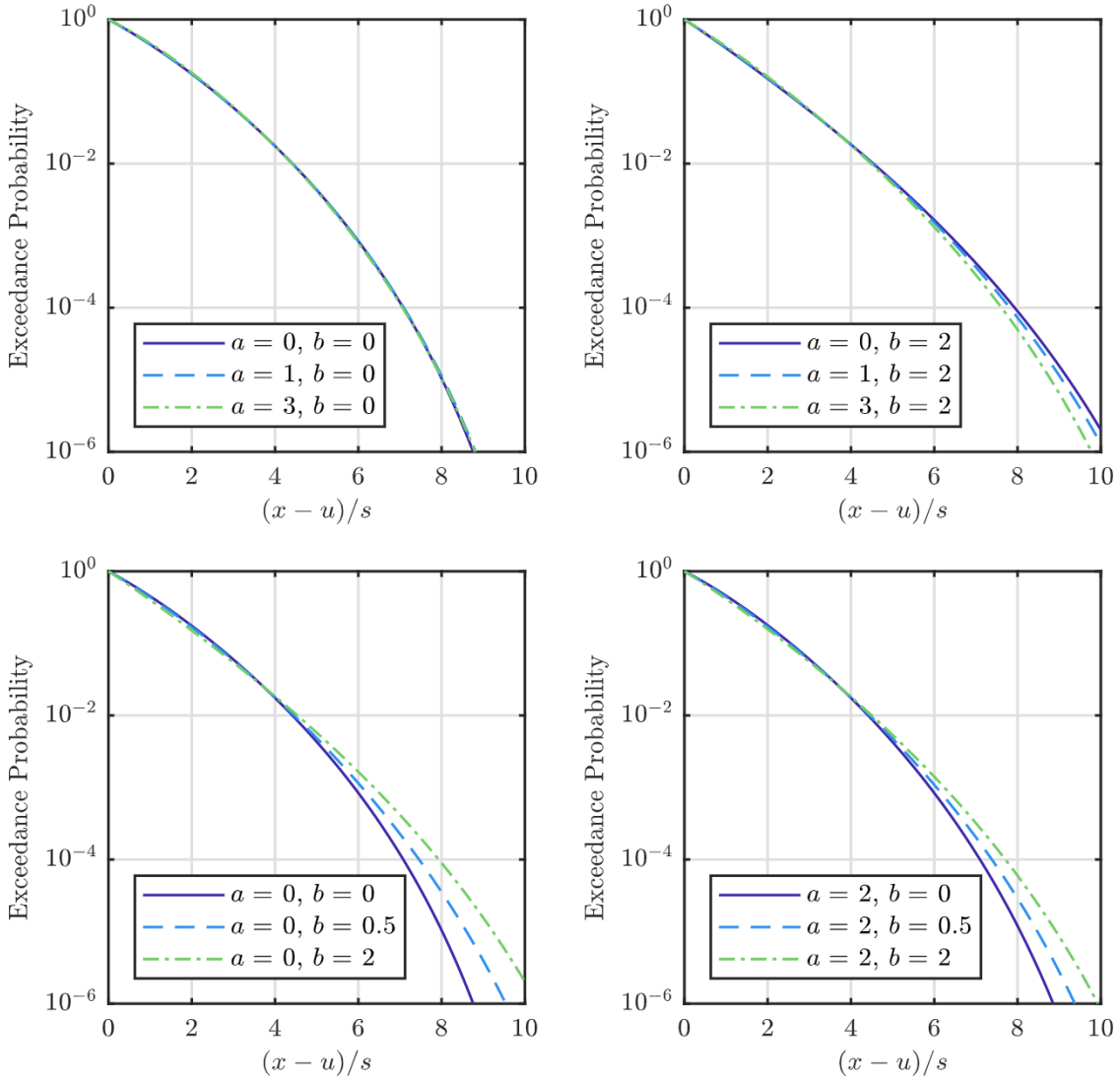

Fig. 4. Normalised shape of upper tail of distributions with linear variation in parameters. Upper tail defined as exceedances of threshold with non-exceedance probability $\psi = 0.7$.

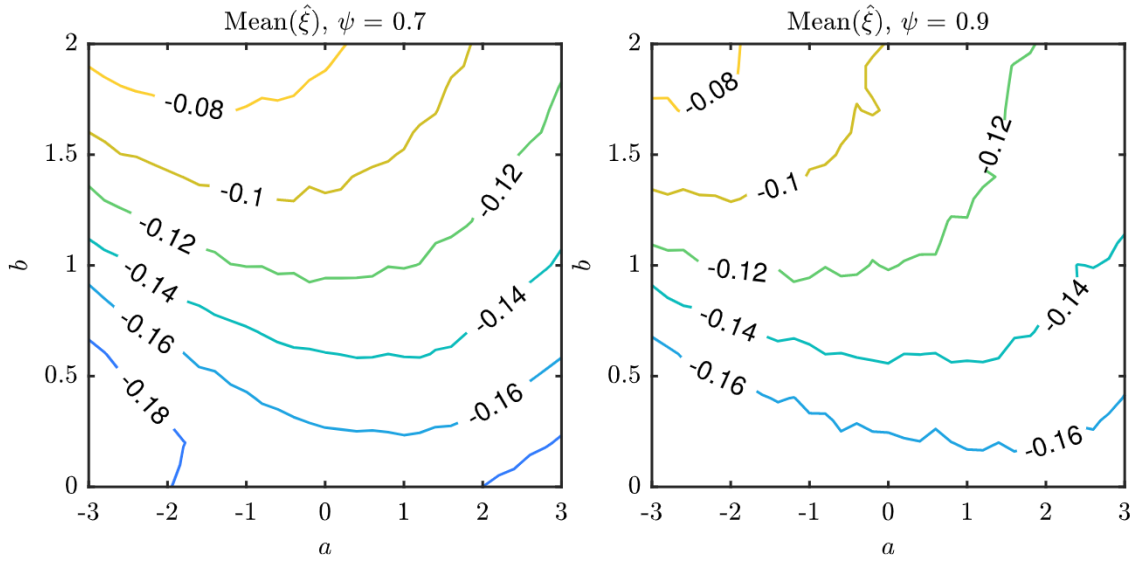

Fig. 5. Estimated shape parameter for stationary model as a function of a and b for thresholds at $\psi = 0.7$ and 0.9.

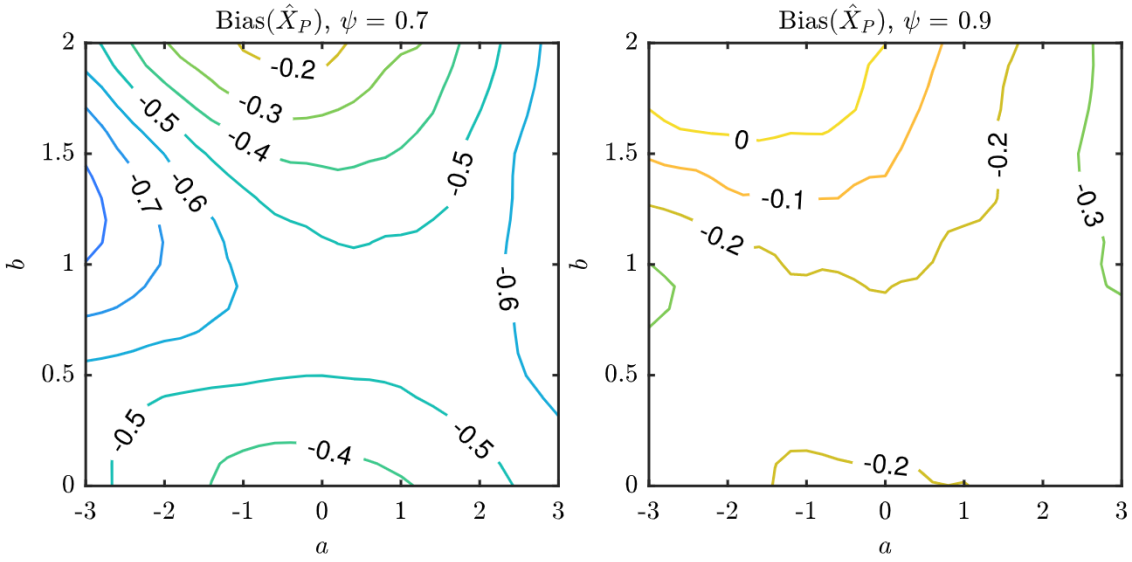

Fig. 6. Bias in estimated omnicovariate return value, $\hat{X}_P$ from stationary model as a function of a and b for thresholds at $\psi = 0.7$ and 0.9. Return value defined as quantile of distribution with non-exceedance probability 0.9999.

distribution using a stationary fitted model, so that the effect of non-stationarity can be assessed in isolation, before considering the effect of partitioning the data by covariate. The true data-generating model, with a linear variation in parameters, is described in (17)–(19).

To illustrate the influence of non-stationarity on the shape of tail of the distribution, we apply a normalisation, so that the tail shape can be considered without the influence of varying location and scale parameters. Suppose we wish to examine the shape of the tail above threshold level u, corresponding to non-exceedance probability ψ. Define threshold exceedances as $Y = X - u$ for $X > u$. The conditional distribution of threshold exceedances is

$$F_Y(y) = P(Y \leq y | X > u) = 1 - \frac{1 - P_{RS}(X \leq y + u)}{1 - \psi}. \tag{23}$$

The mean, m, and STD, s, of the conditional distribution are given by

$$m = \int_0^\infty y f_Y(y) \mathrm{d}y, \tag{24}$$

$$s^2 = \int_0^\infty (y - m)^2 f_Y(y) \mathrm{d}y, \tag{25}$$

where $f_Y(y) = \mathrm{d}F_Y(y)/\mathrm{d}y$ is the probability density function of threshold exceedances. Fig. 4, shows the tail distribution $1 - F_Y(y)$ against the normalised quantity $(x - u)/s$, where u is defined to correspond to a non-exceedance probability of $\psi = 0.7$, for various values of a and b. From the upper left plot, it is evident that for these values of ξ and ψ there is almost no change in the shape of the tail of the distribution for a linear variation in location. However, for the upper left plot where $b = 2$, when the scale is also non-stationary, increase in a makes the distribution appear marginally shorter-tailed. The lower plots show that a non-stationary scale parameter has a more significant effect, making the distribution longer-tailed with increasing b. However, the effect is reduced slightly when there is also an increase in location parameter a. Similar trends (not shown) were observed for other choices of ξ and ψ.

Now we consider how non-stationarity affects bias in estimates when fitting a stationary model. For each value of a and b a simulation study was conducted as follows. The sample size was fixed at $n = 500$ observations. For each simulation, covariate values were generated as uniform random variables $t \in [0, 360)$. The parameters of the GEV conditional on t were defined according to (17) and (18) and a stationary model (the PPC model with one covariate bin) was fitted at threshold levels corresponding to $\psi = 0.7$ and 0.9. For each value of a and b, 10,000 random trials were conducted. Fig. 5 shows the mean of the estimated shape parameter, $\hat{\xi}$, as a function of a and b for thresholds at $\psi = 0.7$ and 0.9. Note that only the mean $\hat{\xi}$ can be shown, rather than bias, since there is no 'true' shape parameter when $a, b \neq 0$ since the data-generating distribution is in fact a non-stationary GEV integrated over t and is not GEV itself. In the case $a = b = 0$, the true shape parameter is $\xi = -0.1$. We observe a negative bias in $\hat{\xi}$ due to two effects: the known bias in maximum-likelihood estimators (Hosking and Wallis, 1987) and the fact that we are fitting a GP distribution to GEV data (see discussion in the Appendix). The trends in the mean $\hat{\xi}$ with a and b are similar to the results indicated in Fig. 4. When $b = 0$, non-stationarity in the location has little influence on the estimated shape parameter, but when there is non-stationarity in the scale then the estimated shape becomes more negative with increasing a. It is also clear that non-stationarity in the scale has more influence on the shape that non-stationarity in the location.

Fig. 6 shows the bias in the estimated omnicovariate return value, X_P, where the return value is defined to be the quantile at a non-exceedance probability of 0.9999, corresponding to a return period around 20 times the length of the observations. For the threshold at $\psi = 0.9$ the non-stationarity has relatively little effect on the bias in the return value, since much of the non-stationarity is removed by the high threshold and the GP model is a good fit for the tail of the distribution. In fact, for a less than approximately 1.5, the bias actually reduces with increasing b, since the positive bias introduced by the non-stationary scale is compensated by the negative bias which results from the parameter estimation method. For the lower threshold, the bias initially increases with b (becomes more negative) then decreases again. The effect of non-stationarity in the location parameter has a smaller effect. Overall, the change in the bias with a and b is relatively small compared to the bias in the case of a stationary distribution at $a = b = 0$.

5. Fitting a non-stationary model to data from a stationary data-generating distribution

If we are confident there is no non-stationarity in the data, then there is obviously no need to apply a non-stationary model. It is, however, instructive to consider the application of a non-stationary model in this situation. PPC and similar models are designed so that, in application to stationary data, a large roughness parameter λ_σ would be estimated, and the variability of estimated σ with covariate t would consequently be small, corresponding to an approximately stationary

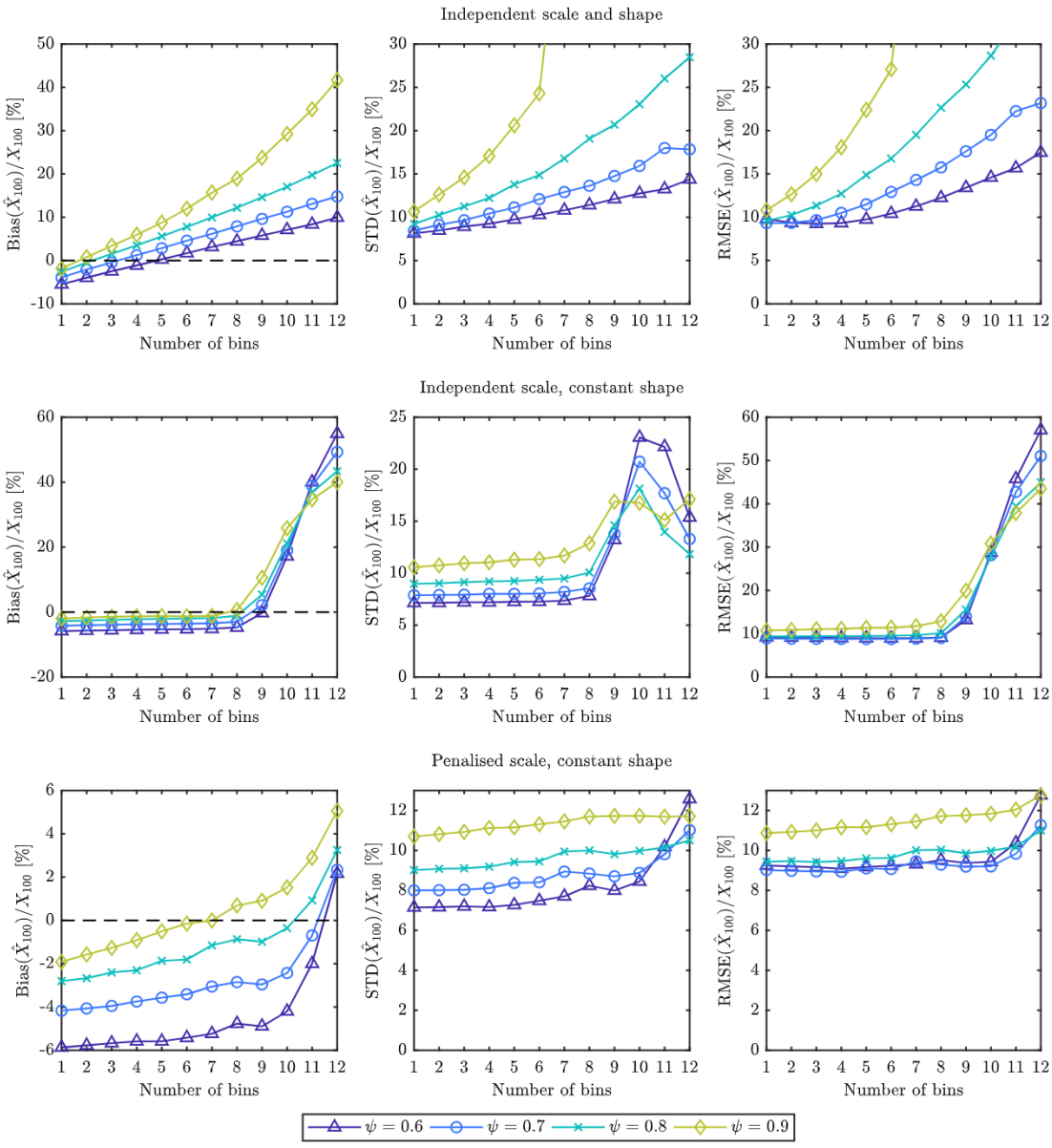

Fig. 7. Bias, STD and RMSE in 100-year omnicovariate return value as a function of number of bins used for fits to data from a stationary distribution (Table 1, Case 1, $\alpha = 0$) at various threshold non-exceedance levels, ψ. Top row: independent fits in each bin. Middle row: Constant shape parameter across all bins, with independent scale parameter (i.e. PPC with $\lambda_\sigma = 0$). Bottom row: Constant shape parameter across all bins, scale parameter penalised for variance (full PPC model with optimal λ_σ).

GP fit. However, it is interesting to study the practical performance of PPC in this setting, in particular the effect of choice of number of covariate bins and other characteristics of covariate binning. Since we know the data-generating distribution is stationary, any effects observed cannot be due to non-stationary in the dataset. We consider the simplest case of a fit to data from a constant distribution (Case 1 of Table 1 with $\alpha = 0$).

We consider three model types with increasing complexity. In the first model, independent fits to the data in each covariate bin are performed. In the second model, the shape parameter is assumed to be constant over all bins but the scale parameter fit is unconstrained (i.e. PPC fit with $\lambda_\sigma = 0$). The third model is full PPC, where the shape parameter is constant over all bins and the scale parameter per bin is chosen to maximise predictive performance. Note that in the case of a single covariate bin, all models are equivalent.

The bias, STD and RMSE in the 100-year omnicovariate return value, X_{100} are shown in Fig. 7 for fits using between 1 and 12 bins. The results for the 1000-year return value are similar and are not shown here. For the one-bin (stationary) fitted model there is a negative bias in the estimate of X_{100}, which is a result of previously-mentioned bias in the maximum likelihood estimators and fitting a GP model to GEV data. In the case of independent fits to each bin, the bias increases with the number of bins used. This effect was reported by Mackay et al. (2010) and is caused by the increased uncertainty in the shape parameter, with a high estimate of ξ in one bin not being compensated for by a low estimate in another. The STD of estimates also increases with both the threshold non-exceedance probability, ψ, and the number of bins used, since sample size per bin reduces with both ψ and the number of bins. The RMSE for the fits with $\psi = 0.6$ decreases slightly from its value in the 1 bin case to a minimum in the 3 bin case. We attribute this to a balancing between the negative bias from parameter estimation and positive bias from increased binning, resulting in a slightly lower RMSE. For all other threshold levels, the RMSE increases monotonically with the number of bins used.

For the PPC ($\lambda_\sigma = 0$) case, the performance of the fitted model is much more stable as a function of the number of bins used, up to 8

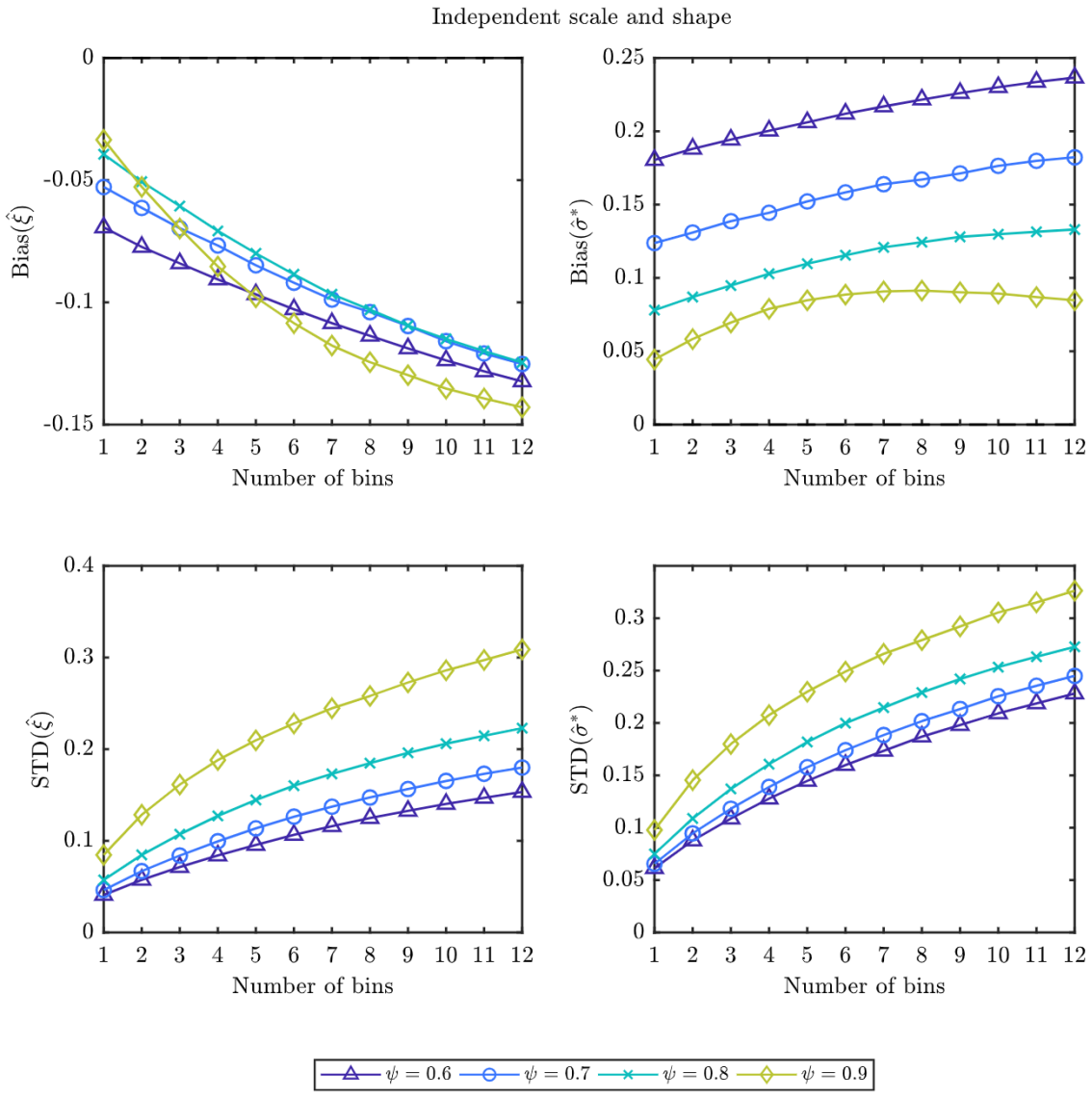

Fig. 8. Bias and STD in parameter estimates against number of bins for fits to data from a stationary distribution (Table 1, Case 1, $\alpha = 0$) using independent fits per bin at various threshold levels.

bins. For more than 8 bins there is a large increase in both the bias and STD of the estimates. For PPC (optimal λ_σ) model, the performance is very similar to the PPC ($\lambda_\sigma = 0$) model up to 8 bins. However, when using more than 8 bins, the performance of the full PPC model is much more stable due to the influence of the roughness penalty on σ. For more than 10 bins there is some increase in the bias from the PPC model. It is thought that this bias results from lack of convergence of the simple simplex-type optimisation algorithm used for maximum likelihood inference. Nevertheless, the bias is still very small compared with the other approaches even for 12 covariate bins.

The bias and STD in parameter estimates from the independent fits-per-bin model are shown in Fig. 8, with the corresponding plots for full PPC model (with optimal λ_σ) in Fig. 9. For the independent fits the bias and STD increases with the number of bins used, due to the reduced sample size in each bin. For the full PPC model, the results are again considerably more stable as a function of number of covariate bins. There is a small reduction in the bias with increasing number of bins used. This is likely due to the increased influence of the σ-roughness penalty, which acts to optimise the performance of the model. The STD in the estimates remains fairly constant with the number of bins used in the full PPC model. For fits with 11 and 12 bins, the STD increases when using a threshold at $\psi = 0.6$, but reduces for the threshold at $\psi = 0.9$. Again, we attribute this effect at least in part to lack of convergence, for large numbers of covariate bins, of the simplex optimisation algorithm used in PPC.

We conclude from this study that the full PPC model provides a good representation of stationary data-generating distributions (with parameters considered), at least when the number of covariate bins does not exceed 10. Therefore, for the studies reported in Section 3, we focus on the fits using up to 8 bins. We note that more sophisticated optimisation schemes (exploiting likelihood slope and curvature characteristics, Davison, 2003; Raghupathi et al., 2016) are available for more challenging applications.

6. Fitting a non-stationary model to data from a non-stationary data-generating distribution

In this section we consider the performance of stationary and non-stationary fitted PPC models for the four cases with sinusoidal parameter variation described in Section 3 (Eqs. (20)–(22) and Table 1), samples of which are illustrated in Fig. 2. For these case studies the true omnicovariate data-generating distribution is an integral of GEV distributions over the covariate domain and therefore not itself a GEV distribution. Hence it is not possible to assess performance in terms of

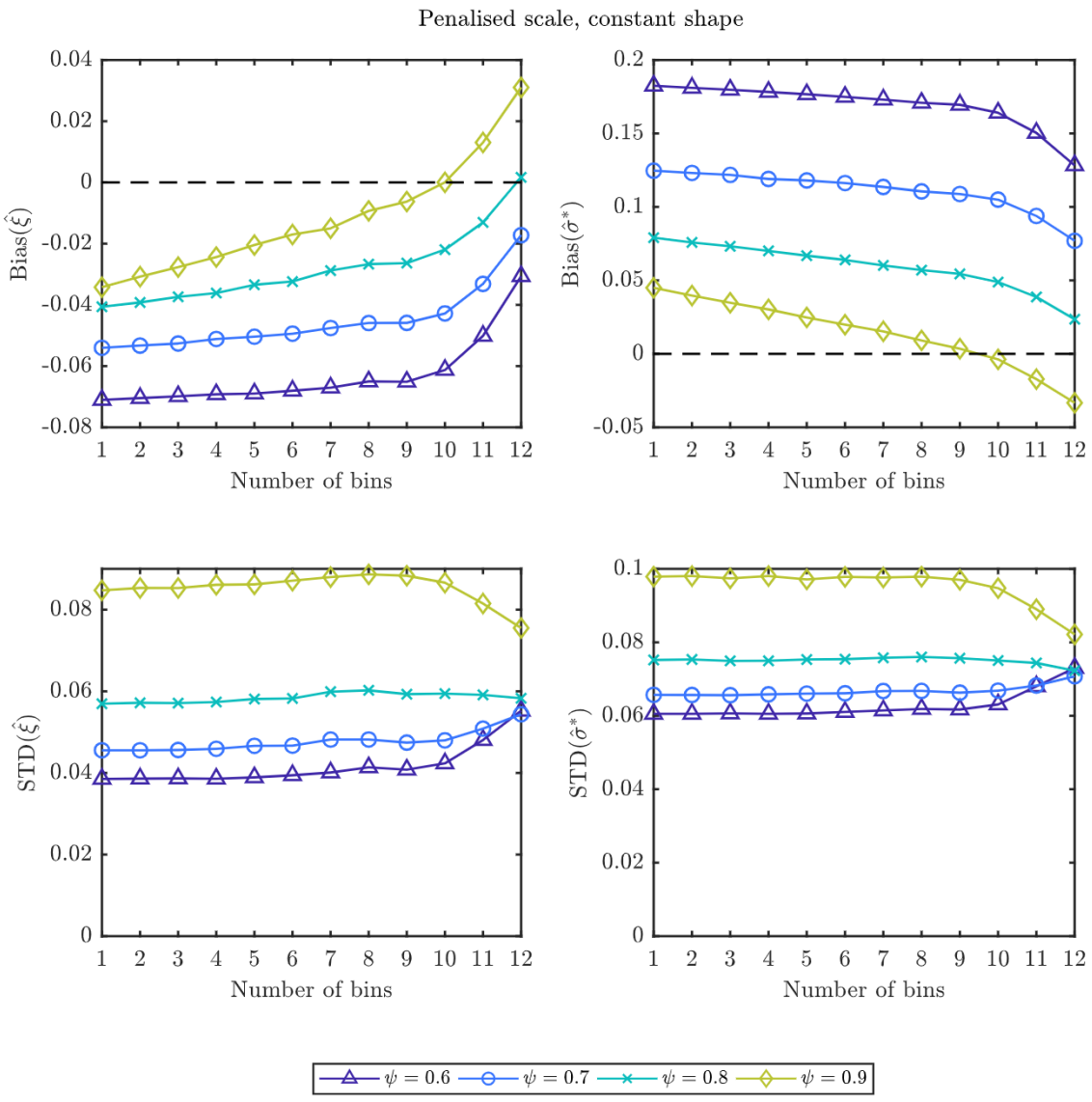

Fig. 9. Bias and STD in parameter estimates against number of bins for fits to data from a stationary distribution (Table 1, Case 1, $\alpha = 0$) using the full PPC model with optimal λ_σ at various threshold levels.

the parameter estimates from the fitted GP models. Instead we focus on the estimating omnicovariate return values for which the true values can be calculated from (4), as illustrated in Fig. 3. We consider the four cases from Table 1 in turn.

6.1. Case 1: Data from distribution with non-stationary location parameter

Fig. 10 shows the bias, STD and RMSE of the estimated 100-year omnicovariate return value as a function of the number of covariate bins and threshold levels used in the (full) PPC model. Results for the 1000-year omnicovariate return value display similar trends and are not shown here. The number of observations used for modelling is dependent only on the threshold non-exceedance probability and not on the number of bins used. However, the observations used for fitting change depending on how the data are binned.

The fitted (one-bin) stationary model shows a negative bias, which reduces with increasing threshold level, consistent with the results shown in Fig. 6. For the lower threshold levels, the bias becomes slightly more negative as α increases and the amplitude of variation in location parameter grows. In contrast, at $\psi = 0.9$, the bias and STD does not vary much with α. The reduction in bias with increasing threshold is due to two effects. First, as threshold increases, there is less covariate variation in sample of threshold exceedances form modelling. Secondly, the GP distribution provides a better fit to the GEV distribution in the tail, as discussed in the Appendix. For non-stationary fits, bias and STD reduce initially as a function of increasing number of covariate bins, up to 3 bins. Performance thereafter stabilises, with STD and RMSE remaining approximately constant up to 8 bins. The stability in performance of the PPC model with the number of bins used is due to the use of the σ-roughness penalty; as the number of bins used increases, the optimal penalty also increases, so that the model does not over-fit. The trend in bias with increasing number of bins for $\alpha = 2$ and 3 is somewhat more complex than might be anticipated. This is due to the effect of the location of bin edges, which is discussed further in Section 6.2. In general, the PPC model fitted using 5–8 bins using a threshold at $\psi = 0.6$ or 0.7 gives the best performance in this case.

6.2. Case 2: Data from distribution with non-stationary scale parameter

In a similar fashion to Section 6.1, bias, STD and RMSE in the 100-year omnicovariate return values for Case 2 (Table 1) are shown in Fig. 11. For the fitted (one-bin) stationary model, bias is negative for

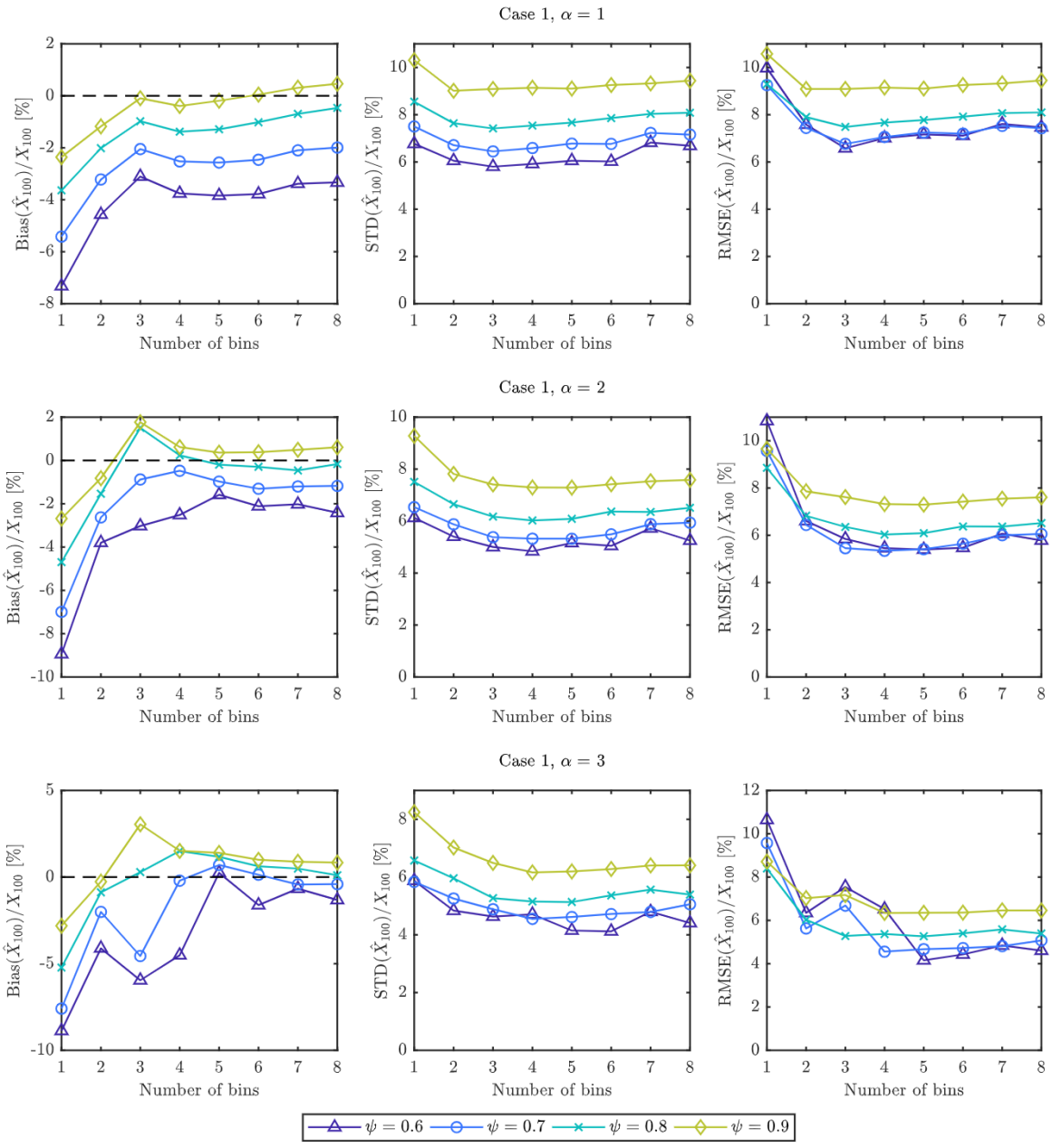

Fig. 10. Bias, STD and RMSE in 100-year omnicovariate return value for Case 1 (Table 1), as a function of the number of covariate bins and extreme value thresholds used. ψ is the non-exceedance probability for the extreme value threshold.

$\beta = 0.25$, but slightly positive for $\beta = 0.5$. Bias becomes more negative as the number of bins increases in general, but there is an excursion in the bias for the 3-bin model, most pronounced for $\beta = 0.5$ and $\psi = 0.6, 0.7$. Despite the increasingly negative bias with the number of bins used in the model, the STD and RMSE remains approximately constant for more than 2 bins, due to σ-roughness penalisation. For $\beta = 0.25$, the performance of the stationary (one-bin) and non-stationary models are similar in terms of RMSE. For the case with $\beta = 0.5$, the PPC model with $N_{bin} \geq 4$ gives a small improvement in performance over the stationary model.

The excursion for three-bin fits is due to the placement of the bin edges. Fig. 12 shows the true tail distributions in each covariate bin for a threshold level at $\psi = 0.6$, when the data is partitioned into 2, 3, 4 or 5 bins, together with the omnicovariate distribution as reference. The distributions in each bin have been normalised using the procedure described in Section 4. In each case, the first bin is centred at $t = 0$ and all bins are of equal width. For the two bin case, the distribution in bin 1 is shorter-tailed than the distribution in bin 2. As the PPC model assumes a constant GP shape parameter across bins, the value of $\hat{\xi}$ will be an average over the shape for each bin. For the three bin case, the distribution in bins 2 and 3 has a longer tail than that in bin 1, since there is a larger change in the scale parameter in bins 2 and 3 than in bin 1 (see Fig. 2). As discussed in Section 4, the non-stationarity in the shape parameter in bins 2 and 3 results in the distribution being longer-tailed in these bins. The estimated shape parameter over the three bins will be more influenced by the two longer-tailed distributions in the lower sectors than the shorter-tailed distribution in the higher sector in bin 1. For the cases with four and five bins, there is less difference between the shapes of the distributions in each bin, since the bins are smaller and the distribution in each bin is more homogeneous. Examination of the distribution of $\hat{\xi}$ showed that the estimates are indeed more positive for three-bin than for other cases. For higher threshold levels, the size of the excursion is reduced since the sample of threshold exceedances is smaller an

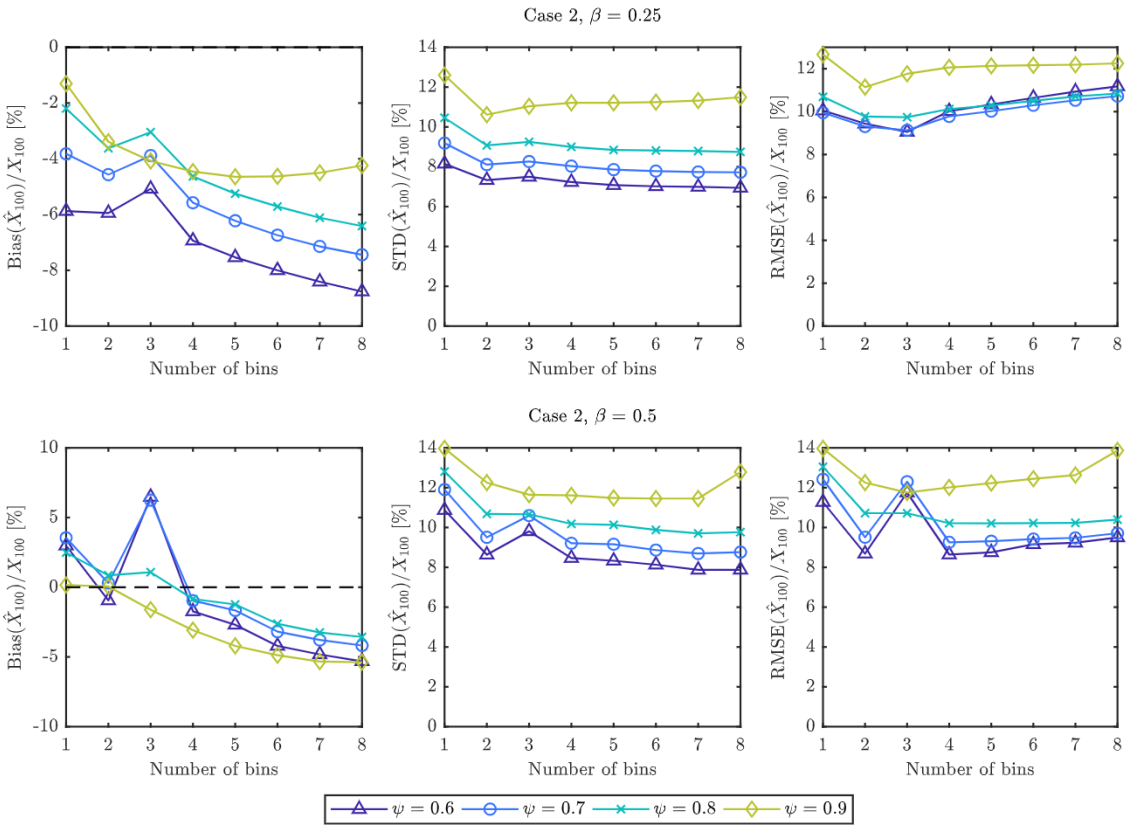

Fig. 11. Bias, STD and RMSE in 100-year omnicovariate return value for Case 2 (Table 1), as a function of number of covariate bins and extreme value thresholds used. ψ is the non-exceedance probability for the extreme value threshold.

more homogeneous. In practice, where smooth variation of the data with covariate is expected, it is not possible to define bins within which there is a homogeneous population. This means that some bin placement effects are unavoidable. However, increasing the number of bins means that a piecewise-constant covariate model is a better approximation to the true data-generating distribution, which should improve the performance of the PPC model. Optimisation of bins widths and locations for directional analysis of extreme conditions is discussed in Ewans and Jonathan (2008).

To investigate the effect of bin placement further, additional simulations were conducted with random placement of the first bin edge on $[0, 360)$, whilst keeping bin widths constant. This procedure effectively eliminated the excursion discussed above, but otherwise the characteristics of the results (not shown) are similar to those shown in Fig. 11. Cases 3 and 4 discussed below utilise random bin placement for this reason. It is possible to optimise the number of bins used and the placement of bin edges (see e.g. Zanini et al. (2020)). However, this represents a significant step up from the PPC model in terms of complexity and has not been pursued further here.

6.3. Case 3: Data from distribution with non-stationary location and scale parameters

The corresponding STD and RMSE in omnicovariate return value estimates for Case 3 (Table 1) using random bin placement (as described in Section 6.2) are similar to those for Case 2 with random bin placement, and are therefore not shown here. The bias for $\beta = 0.5$ was found to be somewhat more negative than that for Case 2, but of a similar magnitude between 0 and -10%. The similarity in performance of PPC models for Cases 2 and 3 agrees with results from Section 4; the effect of non-stationary scale is similar regardless of whether the location parameter is stationary or non-stationary.

6.4. Case 4: Data from distribution with non-stationary location, scale and shape parameters

Fig. 13 shows the bias, STD and RMSE in the 100-year omnicovariate return value estimates for the cases with $\gamma = -0.1$ and -0.2. The data-generating distribution in the "benign" covariate interval has a longer tail in than elsewhere (see Fig. 3). Results for $\gamma = -0.1$ show a small negative bias (of 2%–4%) for the one-bin case and a small positive bias (2%–4%) for the PPC model with 3 or more bins. Bias for the two bin case is close to zero. STD is also relatively stable as a function of the number of bins used. Since STD is larger than bias, RMSE is also relatively stable. There is little difference between stationary and non-stationary models in this case. Results for $\gamma = -0.2$ show a small negative bias for the stationary model. For non-stationary models, bias increases with both the number of bins and threshold level. This behaviour is related to the model misspecification: the PPC model estimates a constant shape parameter by maximising predictive likelihood over all bins. The shape parameter estimate will therefore be influenced by the long tail for the benign sector, resulting in a positive bias overall. STD is relatively stable with increasing number of bins used. Due to the large bias effect, RMSE is lowest for the stationary model and increases with the number of bins used.

Corresponding results for the 1000-year omnicovariate return value are shown in Fig. 14. Now the effect of the long tail in the benign sector is more pronounced (see Fig. 3). Results for $\gamma = -0.1$ are similar to those in Fig. 13, but with slightly larger biases and STDs. For $\gamma = -0.2$ the stationary model displays a large negative bias, since it does not account for the effect of the longer tail in the benign sector, which has a stronger influence on the 1000-year return value than the 100-year return value. Bias reduces with increasing number of bins used, up to approximately four bins. Since PPC assumes a constant ξ, this reduction in bias can only be explained by compensating optimal choices for bin scale parameters. STD is slightly lower for the stationary model than

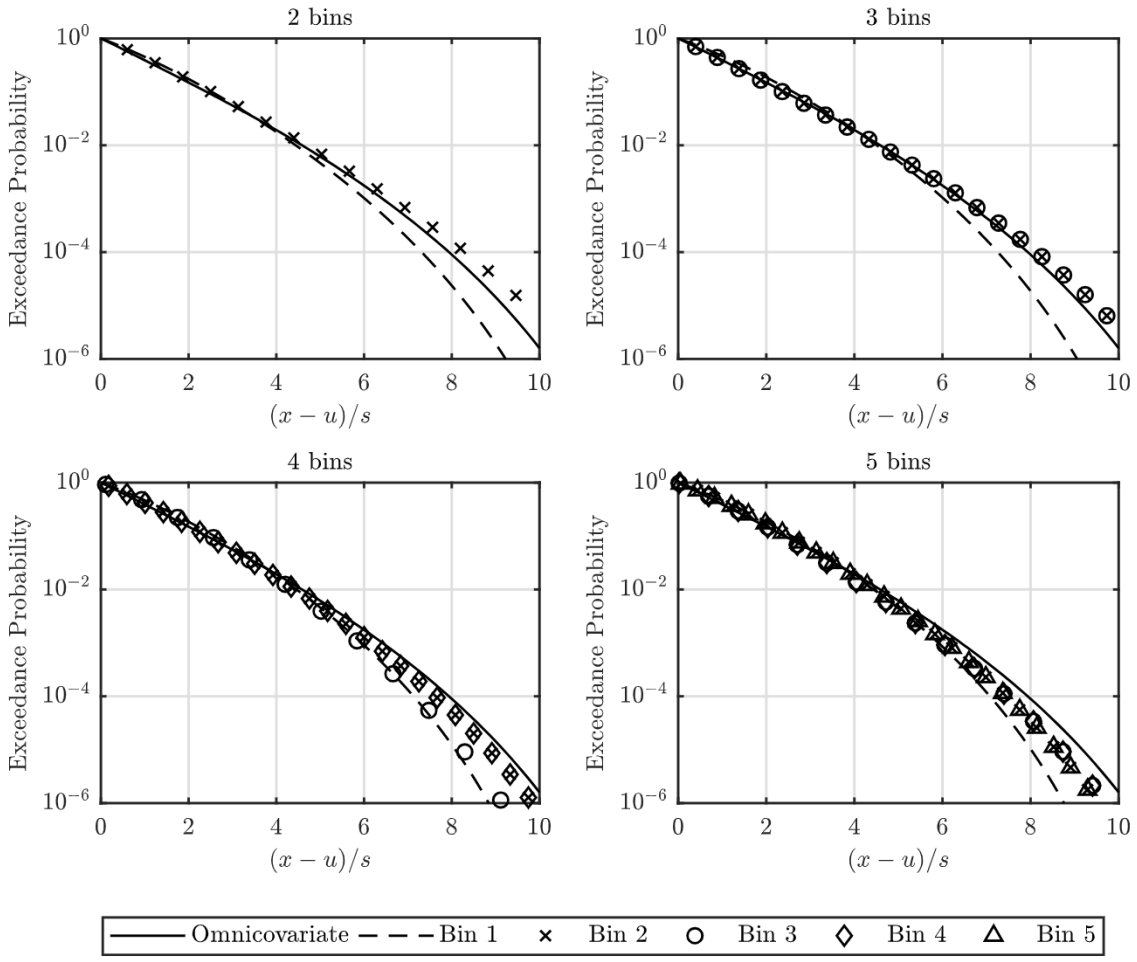

Fig. 12. Normalised distributions of threshold exceedances at non-exceedance probability $\psi = 0.6$ for Case 2 (Table 1), $\beta = 0.5$.

the non-stationary models, but due to the large negative bias in the stationary model, RMSE is lowest for the non-stationary models using four or more bins. RMSE for $\psi = 0.9$ is higher than the fits using the lower thresholds. It is likely that this is because of lack of evidence in the sample of threshold exceedances to justify a large variation in the scale parameter to account for the longer tails in the benign sector.

Fig. 15 shows the bias, STD and RMSE in the estimated omnicovariate 100-year return values for the cases with $\gamma = 0.1$ and 0.2. In these cases the distribution in the benign covariate sector has a shorter tail. Trends in results are similar for both cases. Results from the fitted (one-bin) stationary model indicate a negative bias, between −2 and -12% depending on threshold level, with the highest threshold giving the least biased results, as expected. Bias becomes more negative with increasing number of bins used. This effect is the opposite to that observed for the cases with negative γ. The estimated shape parameter is lower for the non-stationary models, due to the influence of the shorter tails in the more benign sectors that do not contribute to the overall return values. In contrast, the stationary model is not influenced by the distribution in these benign sectors. RMSE is similar between the stationary and non-stationary models and relatively constant as a function of the number of bins. Overall, the performance of both models is poor, due in part to model misspecification and in part to difficulty of estimating data-generating distributions with positive shape parameters. The results for the 1000-year return value (not shown) are similar, but with larger bias and RMSE.

In practice, non-stationarity in the tail shape can be assessed by examining diagnostic plots in each bin, comparing the model to the data. This can be assessed in terms of the fit of the model to the tail of the distribution, or by plotting empirical and modelled return values. A systematic variation in the fit of the model between bins, with the model over-predicting in some bins and under-predicting in other bins, can indicate that there is non-stationarity in the tail shape. In this case the use of more advanced non-stationary models discussed in the introduction may be appropriate.

7. Conclusions

This study compared the performance of stationary and non-stationary extreme value models in estimating omnicovariate return values in the presence of covariate effects, for samples of peaks over threshold. The non-stationary model considered was a penalised piecewise-constant (PPC) GP model, assuming a constant shape parameter, but covariate dependence of scale and extreme value threshold.

The effects of linear trends in the location and scale parameters of the data-generating GEV model on the shape of the omnicovariate tail distribution were examined. For the cases considered, linear variation of the location parameter has only a small effect on the tail. Linear variation in the scale parameter of the data-generating model results in the omnicovariate distribution having a longer tail. Further, we examined the performance of a stationary GP fit to non-stationary data-generating distribution. For the cases considered, the change in bias due to a linear variation in location or scale was small relative to the bias for the case of a stationary data-generating distribution.

The effect of fitting a non-stationary piecewise-constant model to data from a stationary data-generating distribution was also investigated. It was found that when independent GP models are fitted per covariate bin, bias and variance of estimated return values increase with the number of bins used. When the shape parameter is constrained

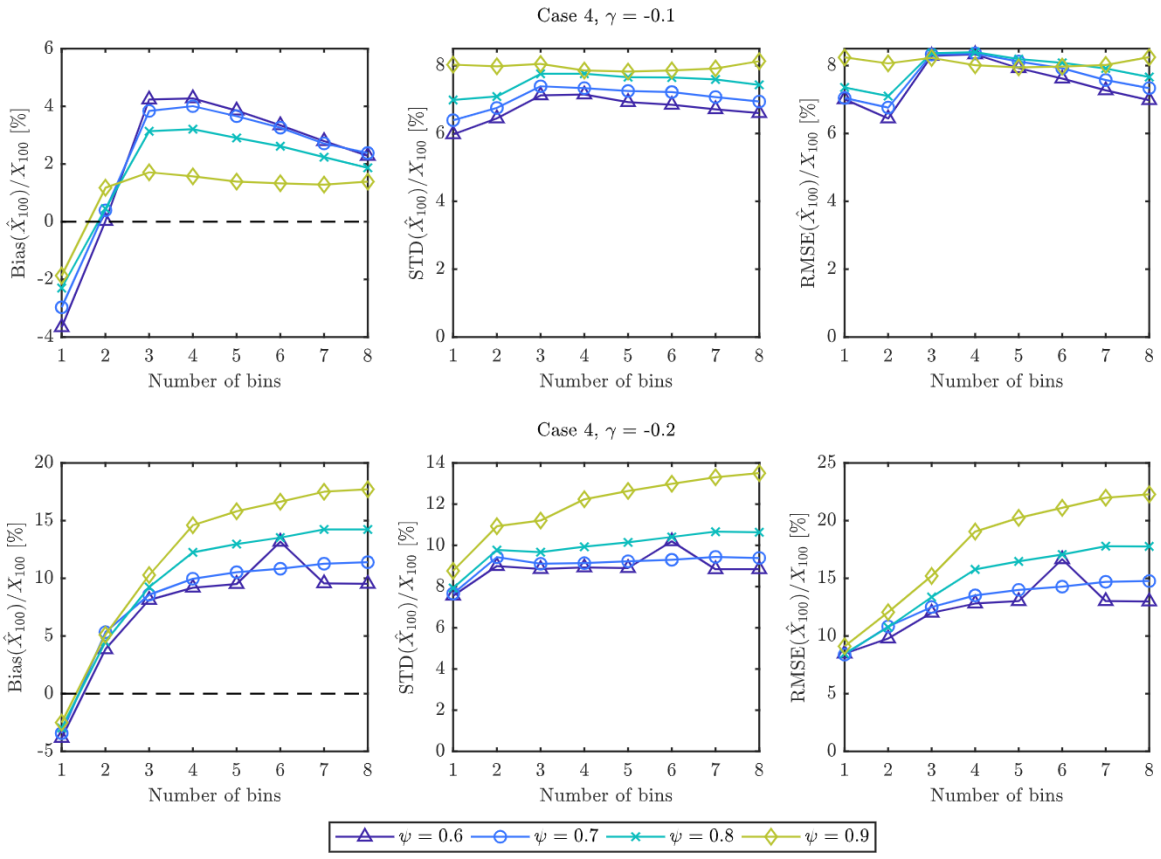

Fig. 13. Bias, STD and RMSE in 100-year omnicovariate return value for Case 4 (Table 1) with negative γ, as a function of number of covariate bins and extreme value thresholds used. ψ is the non-exceedance probability for the extreme value threshold.

to be constant across all bins, and the values of scale per bin estimated freely, it was shown that both bias and variance of estimates stabilise as a function of number of covariate bins. For fits using more than 8 covariate bins, bias and variance increased significantly with the number of bins used. This effect was greatly reduced in the PPC model, where the likelihood maximised to estimate the model parameters is penalised for the variance of estimated scale parameters over covariate bins, with the roughness penalty estimated for optimal out-of-sample predictive performance.

Further case studies involved datasets from data-generating distributions with sinusoidal parameter variation, estimated using a full PPC model for threshold exceedances. For the cases considered, results suggest that the PPC model performs better than a stationary model in estimating return values given non-stationary location parameter in the data-generating model and gives some improvement in performance given non-stationary scale parameter in the data-generating model. Care must be taken over the choice of the width and placement of covariate bins to ensure that the data is as homogeneous as possible within-bin. The choice of the number of bins and location of bin edges can influence model performance when the within-bin data-generating distribution is particularly inhomogeneous, in violation of PPC model assumptions. Case studies with non-stationary shape parameter in the data-generating model showed mixed results. Here both stationary and non-stationary PPC fitting models were misspecified, and hence there was less expectation that the non-stationary model would perform better. Clearly additional case studies need to be considered, for which the non-stationary model incorporates a non-stationary representation for shape parameter should be made, for more useful comparison with fits using a stationary GP.

In summary, a non-stationary extreme value model can give improved estimates of omnicovariate return values compared with stationary models, provided that the characteristics of the data-generating model, and the model to be estimated are consistent. However, the relative performance of stationary and non-stationary extreme value models in estimating an omnicovariate return value is problem specific; either approach works reasonably well when the analysis is performed carefully. When all that is needed from the analysis is an estimate of an omnicovariate return value, a stationary fitted model may be sufficient. However, when a set of return values corresponding to multiple different partitions of the covariate domain is required, in addition to the omnicovariate return value, the non-stationary model exploiting suitable covariate representations is likely to provide a more consistent and statistically efficient framework for inference.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ed Mackay: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. **Philip Jonathan:** Conceptualization, Methodology, Software, Writing - review & editing.

Acknowledgements

The authors acknowledge useful discussions with Kevin Ewans, and colleagues at Shell and Lancaster University, UK. EM was funded under EPSRC, United Kingdom grant EP/R007519/1. The PPC software, developed during the part EU-funded project ECSADES (Ross et al., 2019), is freely available from the authors, and from https://github.com/ECSADES/ecsades-matlab.

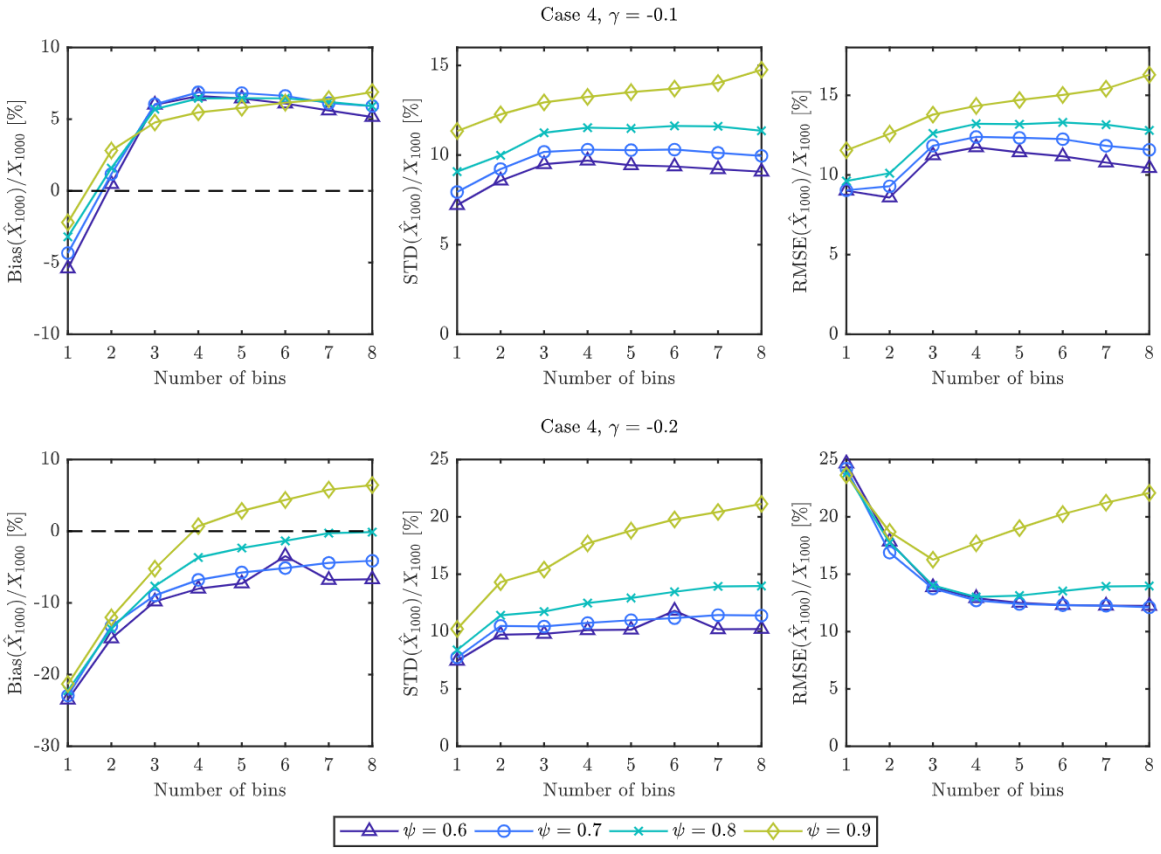

Fig. 14. As Fig. 13, but for 1000-year return value.

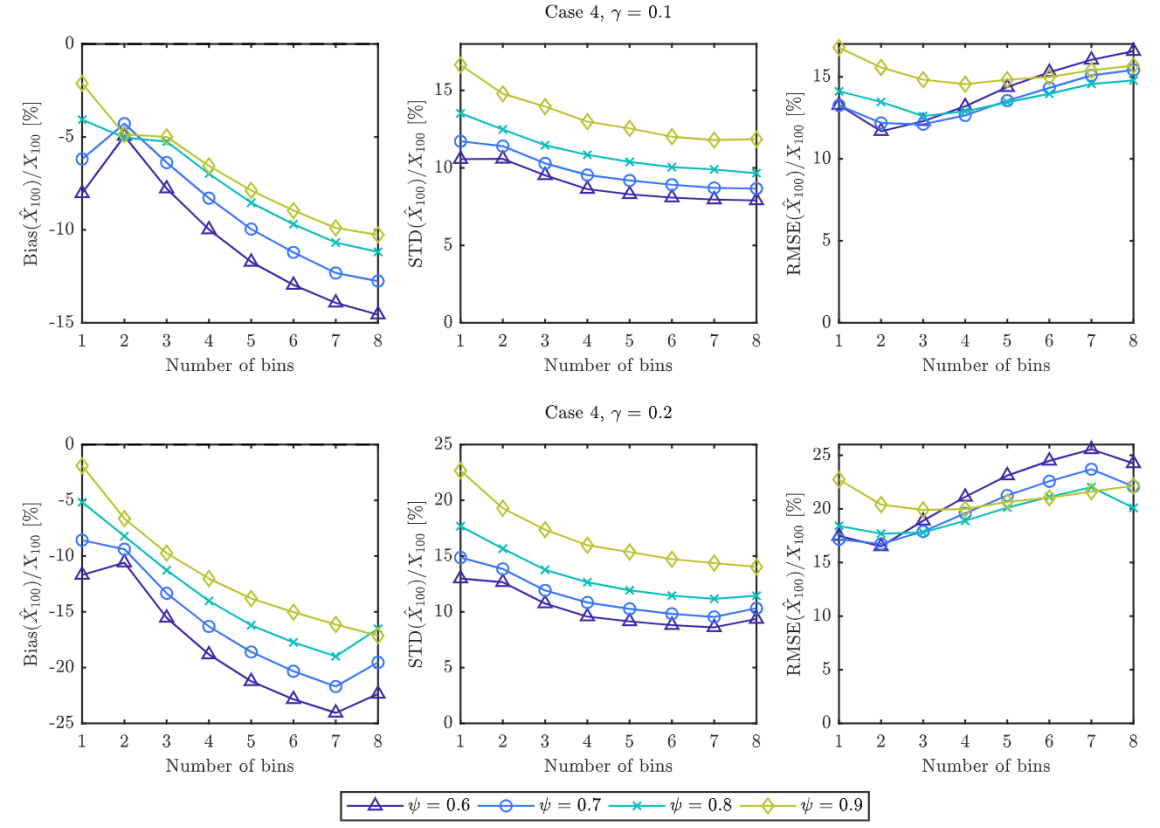

Fig. 15. Bias, STD and RMSE in 100-year omnicovariate return value for Case 4 (Table 1) with positive γ, as a function of number of covariate bins and extreme value thresholds used. ψ is the non-exceedance probability for the extreme value threshold.

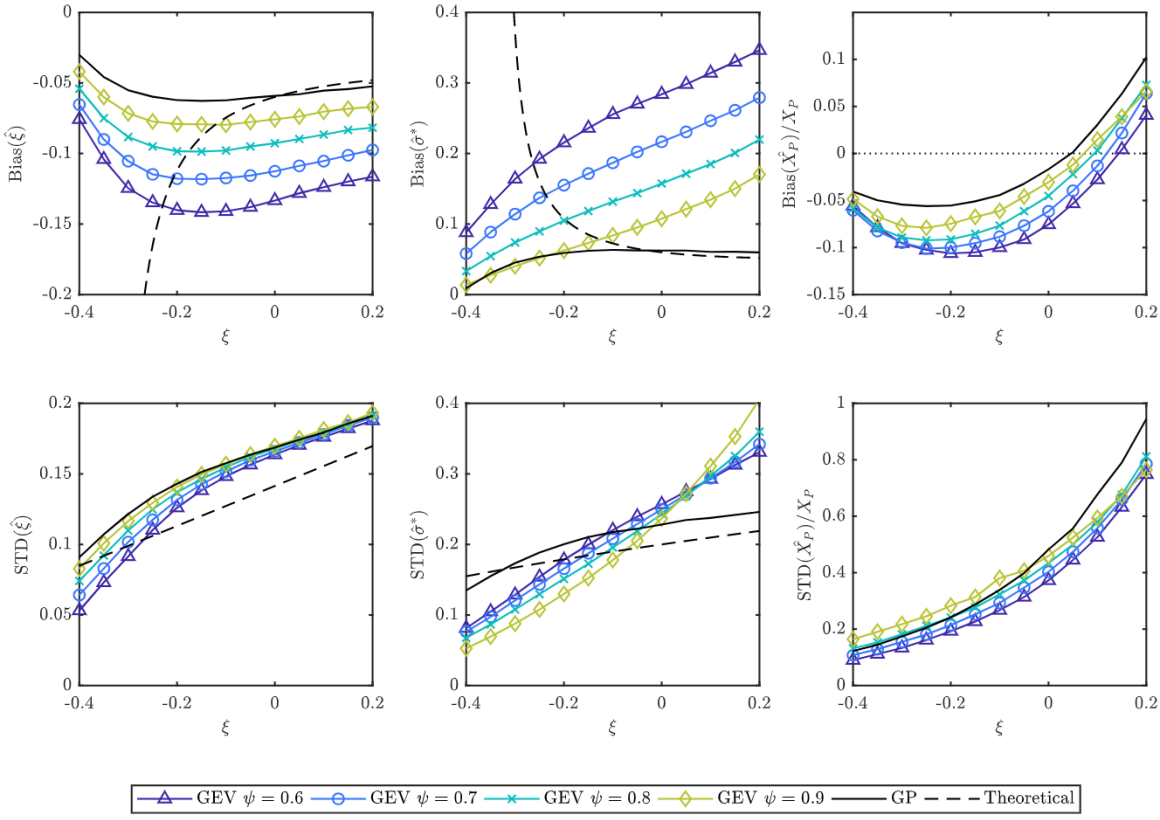

Fig. 16. Bias and STD of estimators of GP shape and scale parameters and return values for a sample size of $n = 50$. The return value X_P is the quantile corresponding to non-exceedance probability 0.999. Coloured lines are for GP fit to GEV data at different threshold levels with non-exceedance probability ψ. Black lines are for GP fit to GP data. Dashed black lines are theoretical bias and STD given by (26) and (27).

Appendix. Errors when fitting the gp distribution to gev data

This study addresses the relative performance of stationary and non-stationary extreme value models, in the presence of covariate effects. Maximum likelihood estimation, potentially penalised to ensure optimal parameter smoothness, is used as discussed in the main text, in conjunction with a GP distribution for exceedances of a high threshold. It is instructive, in addition, to consider the performance of the maximum likelihood estimators for the GP parameters and extreme quantiles in a stationary case. Moreover, in the current work, the data-generating distribution is the GEV. It is important also therefore to assess the bias and variance in parameter and quantile estimates for a GP fit to data generated from a GP distribution, with a GP fit to data generated from a GEV distribution.

There is a wide range of methods for estimating the parameters of the GP distribution, differing in bias and variance characteristics, with the performance depending on sample size and the value of the GP shape parameter (see e.g. Mackay et al. (2011), Kang and Song (2017)). The maximum likelihood (ML) estimators are asymptotically unbiased and efficient (as the sample size tends to infinity the ML estimators achieve the Cramer–Rao lower bound for the variance of an unbiased estimator). However, the ML estimates do not achieve this asymptotic property for small sample sizes and other methods can produce lower bias and variance.

A key step in the estimation of the PPC model is the penalisation of the likelihood function for the "roughness" of the GP scale parameter estimates, which makes ML the most suitable computational framework for inference. We therefore focus on the properties of ML estimators. Various methods have been proposed for calculating the ML estimators for the GP distribution (e.g. Grimshaw (1993), Chaouche and Bacro (2006)) and the performance depends somewhat on the numerical algorithm used. Convergence of the algorithm is sometimes problematic and some methods can give results inconsistent with data, in the sense that $\hat{\xi} < 0$ and $\max(x) > \hat{u} - \hat{\sigma}/\hat{\xi}$. In the PPC model, parameter estimates are forced to be consistent with the data and the shape parameter is constrained to be $\hat{\xi} > -0.5$, as described in Section 2.2. The asymptotic covariance matrix for the ML estimators of GP parameters is (Smith, 1984)

$$\mathrm{var}\begin{bmatrix}\hat{\sigma}\\ \hat{\xi}\end{bmatrix} \approx \frac{1}{n}\begin{bmatrix}2\sigma^2(1+\xi) & \sigma(1+\xi)\\ \sigma(1+\xi) & (1+\xi)^2\end{bmatrix}, \quad \xi > -\frac{1}{2}, \tag{26}$$

where n is the sample size. This provides a lower bound for the variance of unbiased parameter estimates for the stationary model. The second-order bias in the ML estimators was derived by Giles et al. (2016)

$$n\,\mathrm{bias}\,(\hat{\sigma}) = \sigma\frac{3+5\xi+4\xi^2}{(1+3\xi)} + O\left(n^{-1}\right), \quad \xi > -\frac{1}{3} \tag{27}$$

$$n\,\mathrm{bias}\left(\hat{\xi}\right) = -\frac{3+4\xi+\xi^2}{(1+3\xi)} + O\left(n^{-1}\right), \quad \xi > -\frac{1}{3} \tag{28}$$

A simulation study was conducted to compare the bias and variance of GP fits to GP and GEV data and to the theoretical values given above. The aim was to investigate the influence of the threshold level at which the GP distribution is fitted to the GEV. To make a meaningful comparison, the sample size must be consistent between the different threshold levels, which requires generating more extreme GEV values for fits using higher threshold values. The approach taken is summarised as (a) set shape parameter ξ, (b) set GEV threshold non-exceedance probability ψ, (c) set number of GP samples n_{GP}, (d) define number of GEV observations to generate as $n_{GEV} = \lfloor \frac{n_{GP}}{1-\psi} + \frac{1}{2} \rfloor$, where $\lfloor\cdot\rfloor$ is the floor function, (e) generate n_{GP} samples from GP distribution with $u = 0$ and $\sigma = 1$ and fit GP distribution to all samples, and (f) generate n_{GEV} samples from GEV distribution with $\mu = 0$ and $\sigma = 1$ and fit GP distribution to largest $(1-\psi)n_{GEV} \approx n_{GP}$ samples. For each

value of ξ and ψ, 100,000 trials were conducted. As the GP distribution is fit to the GEV data at different threshold levels the estimated scale parameter must be adjusted to allow consistent comparison. A feature of the GP distribution is that if exceedances of threshold u_0 follow a GP distribution with parameters σ_{u_0} and ξ, then for threshold $u > u_0$, the exceedances are GP distributed with same ξ and scale parameter $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$ (see e.g. Coles (2001)). The parameter $\sigma^* = \sigma_u - \xi u$ is therefore threshold-independent. We therefore compare estimates of σ^* rather than σ. Note that since $u = 0$ in this example the true value is $\sigma^* = 1$.

Fig. 16 shows the results of the simulation study for a sample size of $n = 50$. Results for $n = 200$ yields similar results, and are not reproduced here. In this example the return value X_P is defined as the quantile at a non-exceedance probability of $P = 0.999$ for the GP data. The bias of parameter estimates for fits to GP data agree reasonably well with the theoretical values from (27) and (28), when $\xi > -0.1$, but the theoretical values depart significantly from the simulations when $\xi < -0.1$ due to the influence of singularity in the theoretical expressions when $\xi = -1/3$. For the fits to the GEV data, the bias is larger than that for the fit to the GP data. The bias reduces as the threshold increases and the tail of the GEV converges to a GP distribution. The STD of estimates (lower panel) for the fit to GP data is slightly above that predicted by the asymptotic result. For lower values of ξ, STD is closer to the asymptotic values. This is because the estimated shape parameter is constrained to be greater than -0.5, restricting the range of values that the estimates can take. The STD of $\hat{\xi}$ for the fits to the GEV data is slightly lower than that for the fit to the GP data. The STD for $\hat{\sigma}^*$ is lower for the fit to the GEV data for ξ less than approximately -0.1 and higher than that for the fit to the GP data for larger values of ξ. For the estimated return values, there is an increase in absolute bias for fits to GEV data. There is a slight reduction in STD for the fits to the GEV data, except for higher threshold case with $\psi = 0.9$ and negative shape parameter.

In summary, fitting a GP distribution to threshold exceedances from a GEV data-generating distribution results in a slight increase in the bias of parameter and quantile estimates relative to fitting to GP data, with the bias decreasing as the threshold increases. STD of estimates in fits to GEV data is generally slightly lower, meaning that RMSE in return value estimates is comparable to that for fits to GP data. It therefore seems reasonable to use a GEV model in the case studies in this work, especially considering that the appropriate model for environmental data is not known beforehand.

References

Anderson, C.W., Carter, D.J.T., Cotton, P.D., 2001. Wave Climate Variability and Impact on Offshore Design Extremes. Tech. Rep. Report for Shell International and the Organization of Oil & Gas Producers, pp. 1–100.

Beirlant, J., Caeiro, F., Gomes, M., 2012. An overview and open research topics in statistics of univariate extremes. REVSTAT – Stat. J. 10 (1), 1–31.

Biller, C., 2000. Adaptive Bayesian regression splines in semiparametric generalized linear models. J. Comput. Graph. Statist. 9, 122–140.

Bodin, T., Sambridge, M., 2009. Seismic tomography with the reversible jump algorithm. Geophys. J. Int. 178, 1411–1436.

Bromirski, P.D., Cayan, D.R., Flick, R.E., 2005. Wave spectral energy variability in the northeast Pacific. J. Geophys. Res. C 110 (3), 1–15. http://dx.doi.org/10.1029/2004JC002398.

Brooker, D., Cole, G., McConochie, J., 2004. The influence of hindcast modelling uncertainty on the prediction of high return period wave conditions. In: Proc. 23th Int. Conf. Offshore Mech & Arctic Eng., Vancouver, Canada.

Carter, D., Challenor, P., 1981. Estimating return values of environmental parameters. Q. J. R. Meteorol. Soc. 1 107, 259–266.

Cattrell, A.D., Srokosz, M., Moat, B.I., Marsh, R., 2019. Seasonal intensification and trends of rogue wave events on the US western seaboard. Nature 9 (1), http://dx.doi.org/10.1038/s41598-019-41099-z.

Chaouche, A., Bacro, J.N., 2006. Statistical inference for the generalized pareto distribution: Maximum likelihood revisited. Comm. Statist. Theory Methods 35 (5), 785–802. http://dx.doi.org/10.1080/03610920500501429.

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer, http://dx.doi.org/10.1198/tech.2002.s73.

Currie, I.D., Durban, M., Eilers, P.H.C., 2016. Generalized linear array models with applications to multidimensional smoothing. J. R. Stat. Soc. Ser. B Stat. Methodol. 68, 259–280.

Davis, R.A., Mikosch, T., 2009. The extremogram: A correlogram for extreme events. Bernoulli 15 (4), 977–1009. http://dx.doi.org/10.3150/09-BEJ213.

Davison, A.C., 2003. Statistical Models. Cambridge University Press, Cambridge, UK.

Davison, A., Smith, R., 1990. Models for exceedances over high thresholds. J. R. Stat. Soc. Ser. B Stat. Methodol. 52, 393.

de Zea Bermudez, P., Kotz, S., 2010. Parameter estimation of the generalized pareto distribution-Part I. J. Statist. Plann. Inference 140 (6), 1353–1373. http://dx.doi.org/10.1016/j.jspi.2008.11.019.

Dekkers, A.L.M., Einmahl, J.H.J., Haan, L.D., 1989. A moment estimator for the index of an extreme-value distribution. Ann. Statist. 17, 1833–1855.

Eilers, P., Marx, B., 2010. Splines, knots and penalties. Wiley Intersci. Rev.: Comput. Stat. 2, 637–653.

Ewans, K., Jonathan, P., 2008. The effect of directionality on northern north sea extreme wave design criteria. In: Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering - OMAE, vol. 2. pp. 479–488. http://dx.doi.org/10.1115/OMAE2007-29657.

Fawcett, L., Walshaw, D., 2006. A hierarchical model for extreme wind speeds. J. R. Stat. Soc. Ser. C. Appl. Stat. 55 (5), 631–646. http://dx.doi.org/10.1111/j.1467-9876.2006.00557.x.

Forristall, G., Heideman, J., Leggett, I., Roskam, B., Vanderschuren, L., 1996. Effect of sampling variability on hindcast and measured wave heights. J. Waterway Port Coast. Ocean Eng. 122 (5), 216–225.

Giles, D.E., Feng, H., Godwin, R.T., 2016. Bias-corrected maximum likelihood estimation of the parameters of the generalized Pareto distribution. Comm. Statist. Theory Methods 45 (8), 2465–2483. http://dx.doi.org/10.1080/03610926.2014.887104.

Gomes, M.I., 2014. Statistics of extremes and applications : An introduction. In: ASA 2014: The 56th Annual Conference of the South African Statistical Association. Rhodes University, Grahamstown, South Africa, Octo.

Grimshaw, S.D., 1993. Computing maximum likelihood estimates for the generalized Pareto distribution. Technometrics 35 (2), 185–191.

Hansen, H.F., Randell, D., Zeeberg, A.R., Jonathan, P., 2020. Directional–seasonal extreme value analysis of North Sea storm conditions. Ocean Eng. 195, 106665. http://dx.doi.org/10.1016/j.oceaneng.2019.106665.

Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. Ann. Statist. 3, 1163–1174.

Hosking, J.R.M., Wallis, J.R., 1987. Parameter and quantile estimation for the generalized pareto distribution. Technometrics 29 (3), 339–349.

Jonathan, P., Ewans, K., 2011. Discussion of 'On the use of discrete seasonal and directional models for the estimation of extreme wave conditions' by Edward B.L. Mackay, Peter G. Challenor, AbuBakr S. Bahaj [Ocean Engineering 37(56), April 2010, pp. 425442]. Ocean Eng. 38 (1), 205. http://dx.doi.org/10.1016/j.oceaneng.2010.10.013.

Jonathan, P., Ewans, K., 2013. Statistical modelling of extreme ocean environments for marine design : A review. Ocean Eng. 62, 91–109. http://dx.doi.org/10.1016/j.oceaneng.2013.01.004.

Jonathan, P., Ewans, K., Forristall, G., 2008. Statistical estimation of extreme ocean environments: The requirement for modelling directionality and other covariate effects. Ocean Eng. 35 (11–12), 1211–1225. http://dx.doi.org/10.1016/j.oceaneng.2008.04.002.

Jones, M., Randell, D., Ewans, K., Jonathan, P., 2016. Statistics of extreme ocean environments: Non-stationary inference for directionality and other covariate effects. Ocean Eng. 119, 30–46. http://dx.doi.org/10.1016/j.oceaneng.2016.04.010,.

Kang, S., Song, J., 2017. Parameter and quantile estimation for the generalized pareto distribution in peaks over threshold framework. J. Korean Stat. Soc. 46 (4), 487–501. http://dx.doi.org/10.1016/j.jkss.2017.02.003.

Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E., 1998. Convergence properties of the nelder-mead simplex method in low dimensions. SIAM J. Optim. 9, 112–147.

Mackay, E.B., Challenor, P.G., Bahaj, A.B.S., 2010. On the use of discrete seasonal and directional models for the estimation of extreme wave conditions. Ocean Eng. 37, 425–442. http://dx.doi.org/10.1016/j.oceaneng.2010.01.017.

Mackay, E.B., Challenor, P.G., Bahaj, A.S., 2011. A comparison of estimators for the generalised Pareto distribution. Ocean Eng. 38 (11–12), 1338–1346. http://dx.doi.org/10.1016/j.oceaneng.2011.06.005.

Mackay, E., Johanning, L., 2018. Long-term distributions of individual wave and crest heights. Ocean Eng. 165 (May), 164–183. http://dx.doi.org/10.1016/j.oceaneng.2018.07.047.

Méndez, F.J., Menéndez, M., Luceño, A., Medina, R., Graham, N.E., 2008. Seasonality and duration in extreme value distributions of significant wave height. Ocean Eng. 35 (1), 131–138. http://dx.doi.org/10.1016/j.oceaneng.2007.07.012.

Morton, I.D., Bowers, J., Mould, G., 1997. Estimating return period wave heights and wind speeds using a seasonal point process model. Coast. Eng. 31 (1–4), 305–326. http://dx.doi.org/10.1016/S0378-3839(97)00016-1.

Northrop, P., Jonathan, P., Randell, D., 2016. Threshold modeling of nonstationary extremes. In: Dey, D., Yan, J. (Eds.), Extreme Value Modeling and Risk Analysis: Methods and Applications. CRC Press, Boca Raton, USA, pp. 87–108.

Raghupathi, L., Randell, D., Ewans, K., Jonathan, P., 2016. Fast computation of large scale marginal extremes with multi-dimensional covariates. Comput. Stat. Data Anal. 95, 243–258.

Randell, D., Feld, G., Ewans, K., Jonathan, P., 2015. Distributions of return values for ocean wave characteristics in the south China Sea using directional-seasonal extreme value analysis. Environmetrics 26 (6), 442–450. http://dx.doi.org/10.1002/env.2350.

Reguero, B.G., Losada, I.J., Méndez, F.J., 2019. A recent increase in global wave power as a consequence of oceanic warming. Nature Commun. 10 (1), http://dx.doi.org/10.1038/s41467-018-08066-0.

Ross, E., Astrup, O.C., Bitner-Gregersen, E., Bunn, N., Feld, G., Gouldby, B., Huseby, A., Liu, Y., Randell, D., Vanem, E., Jonathan, P., 2019. On environmental contours for marine and coastal design. Ocean Eng. 195, 106194. http://dx.doi.org/10.1016/j.oceaneng.2019.106194.

Ross, E., Sam, S., Randell, D., Feld, G., Jonathan, P., 2018. Estimating surge in extreme North Sea storms. Ocean Eng. 154, 430–444. http://dx.doi.org/10.1016/j.oceaneng.2018.01.078.

Smith, R.L., 1984. Threshold methods for sample extremes. In: de Oliveira, J. (Ed.), Statistical Extremes and Applications. Springer, Dordrecht, pp. 621–638.

Wood, S.N., 2003. Thin plate regression splines. J. R. Stat. Soc. Ser. B Stat. Methodol. 65 (1), 95–114.

Woolf, D.K., Challenor, P.G., Cotton, P.D., 2002. Variability and predictability of the north atlantic wave climate. J. Geophys. Res. C 107 (10), 9–1..

Zanini, E., Eastoe, E., Jones, M., Randell, D., Jonathan, P., 2020. Flexible covariate representations for extremes. Environmetrics http://dx.doi.org/10.1002/env.2624.