Ocean Engineering 312 (2024) 119247

Contents lists available at ScienceDirect

Ocean Engineering

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

Research paper

Practical non-stationary extreme value analysis of peaks over threshold using the generalised Pareto distribution: Estimating uncertainties in return values

Stan Tendijck [a], David Randell [a], Graham Feld [b], Philip Jonathan [c,d,*]

[a] Shell Information Technology International B.V., 1031 HW Amsterdam, The Netherlands
[b] Shell UK Ltd., Aberdeen AB12 3FY, United Kingdom
[c] Department of Mathematics and Statistics, Lancaster University, LA1 4YF, United Kingdom
[d] Shell Information Technology International Ltd., London SE1 7NA, United Kingdom

ARTICLE INFO

Keywords:
Extremes
Non-stationary
Generalised Pareto
Return values
Tuning parameters
Bias
Mean-max parameterisation

ABSTRACT

Choice of tuning parameters influences the performance of non-stationary extreme value modelling for peaks over threshold using the generalised Pareto (GP) distribution. We examine the effect of tuning parameter choice on maximum roughness-penalised likelihood estimation of GP models, the shape and scale parameters of which are assumed to vary smoothly on a one-dimensional "directional" covariate domain, under a B-spline representation. We examine the effect of (a) extreme value model parameterisation, (b) relative roughness penalty of GP parameters as a function of covariate, (c) cross-validation strategy for roughness parameter tuning, and (d) estimator for return value, on the estimation of return values corresponding to return periods 1000× the period of a sample of size 1000. Bootstrap resampling is used for thorough uncertainty quantification. We also compare results with those from stationary inference.

Results from a large simulation study of 16 cases broadly representative of North Sea conditions for significant wave height with direction, indicate that (i) multiple two-group cross-validation yields lower return value estimates that ten-group cross-validation (leading to negative bias on average, for the case studies considered), (ii) the quantile of the bootstrap predictive estimator yields larger values than the mean over bootstraps of the quantile estimate (leading to reduced omni-directional bias for the case studies considered). Further, (iii) the use of stationary models for non-stationary tails is only reasonable when a high extreme value threshold is set for the stationary analysis. However, (iv) the relative performance of different modelling strategies is sensitive to the specific characteristics of the case study.

1. Introduction

1.1. Motivation

Recent years have seen important improvements in methodologies to characterise extreme ocean environments. Some of these have emerged naturally on the interface of applied statistical modelling and metocean engineering (e.g. Vanem et al., 2022 and references therein). For example, the effect of directional and seasonal variation of ocean conditions has long been a concern, and various ad-hoc procedures developed to accommodate it. As a result, hierarchical models and software have been introduced allowing rigorous non-stationary extreme value analysis (e.g. Davison and Smith, 1990; Chavez-Demoulin and Davison, 2005; Randell et al., 2015; Youngman, 2022; Wood, 2023; Southworth et al., 2024; Towe et al., 2024). Developments in methodologies for multivariate extremes have facilitated the introduction of more rigorous methods for estimation of joint criteria (e.g. Heffernan and Tawn, 2004; Hansen et al., 2020; Murphy-Barltrop et al., 2024; Mackay et al., 2025), again replacing earlier more ad-hoc approaches (e.g. Haver, 1985; Forristall, 2004; Feld et al., 2019). The need to quantify uncertainties resulting from fitting models to data, and to incorporate uncertainty in decision-making, is acknowledged. There is increasing recognition that the specification of metocean design in terms of return values and similar summary statistics is inadequate, and that the future lies in full probabilistic modelling of the environment and of environment-structure interaction, leading to quantification of risk (e.g. Serinaldi, 2015; Towe et al., 2021; Speers et al., 2024) and hence optimal decision-making.

Adopting more recent methodologies for real-world marine design presents the metocean engineer with different practical challenges. The new methodologies are usually conceptually and computationally more complex, and more cumbersome to use well (e.g. Jones et al., 2018; Hansen et al., 2020; Swan, 2020; Gibson, 2020). We need to

* Corresponding author at: Department of Mathematics and Statistics, Lancaster University, LA1 4YF, United Kingdom.
E-mail address: p.jonathan@lancaster.ac.uk (P. Jonathan).

https://doi.org/10.1016/j.oceaneng.2024.119247
Received 13 June 2024; Received in revised form 4 September 2024; Accepted 10 September 2024
Available online 24 September 2024

think carefully about how to set up a complex inference, and how to confirm that results from the inference are reasonable. Implementations of statistical component models have perhaps not been made sufficiently resistant to the realities of practical application. Recent methodological advances have undoubtedly improved the potential for full and accurate characterisation of extreme ocean environments; but used unwisely without sufficient care and awareness of pitfalls, they can lead the metocean practitioner astray (e.g. Standard Norge, 2022).

1.2. Metocean design using non-stationary extreme value analysis

Section 3 of Randell et al. (2015) outlines a typical implementation of non-stationary extreme value analysis, similar to those of Ross et al. (2017), Hansen et al. (2020), Gibson (2020) and Towe et al. (2024). The modelling procedure requires that high-level modelling choices are made prior to analysis. We will show in this paper that these choices can have a material effect on the quality of inference.

Suppose that the magnitude Y of peaks-over-threshold of storm peak events, assumed conditionally independent given covariates, follows a non-stationary generalised Pareto (GP) distribution, with shape ξ ($\in \mathbb{R}$) and scale σ (>0), conditional on exceeding some covariate-dependent threshold ψ (>0). The support of the generalised Pareto distribution is (ψ, y^+) (where $y^+ = \psi - \sigma/\xi$ for $\xi < 0$ and $= \infty$ otherwise). The corresponding conditional probability density function is

$$f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \frac{1}{\sigma(\boldsymbol{x})}\left(1 + \frac{\xi(\boldsymbol{x})}{\sigma(\boldsymbol{x})}(y - \psi(\boldsymbol{x}))\right)^{-1/\xi(\boldsymbol{x})-1} \tag{1}$$

where all of ψ, σ and ξ in principle are smoothly-varying functions of covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ (including e.g. storm direction, season, climate indices) defined on domain D. In a similar fashion, the annual rate of occurrence of threshold exceedances $\rho(\boldsymbol{x})$ (≥ 0) at $\boldsymbol{X} = \boldsymbol{x}$ can be estimated using a non-stationary Poisson model as explained in Randell et al. (2015); further, the non-stationary extreme value threshold $\psi(\boldsymbol{x})$ (≥ 0) can be estimated for example using a non-stationary quantile regression. Estimating optimally smooth forms for each of ψ, σ and ξ with covariates is the objective of the statistical inference.

The metocean engineer is often ultimately interested in estimating the return value of Y corresponding to a return period of T years, where $T \geq 100$. To achieve this, we use the fitted model first to estimate the cumulative distribution function of the annual maximum (threshold exceedance) event A

$$F_A(y) = \sum_{k=0}^{\infty} \frac{\rho_A^k \exp(-\rho_A)}{k!} F_R^k(y) \tag{2}$$

where ρ_A (>0) is the expected number of threshold exceedances per annum over the whole of D, and F_R is the cumulative distribution function of a random storm (above threshold) given by

$$F_R(y) = \int_D F_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}. \tag{3}$$

where $f_{\boldsymbol{X}}(\boldsymbol{x})$ is the density of threshold exceedances on D, such that

$$\rho(\boldsymbol{x}) = \rho_A f_{\boldsymbol{X}}(\boldsymbol{x}). \tag{4}$$

The return value y_T is then given by the $1 - 1/T$ quantile of F_A

$$F_A(y_T) = 1 - \frac{1}{T}. \tag{5}$$

It can be useful to evaluate the relative performance of different approaches to non-stationary extreme value analysis by assessing the bias and variance of estimates for y_T from different methodologies. In reality, of course, estimates (from the inference) for the quantities in Eqs. (2)–(5) must be used in order to estimate the return value. Specifically, different high-level choices in the modelling procedure may result in different estimates for ψ, σ and ξ, and hence differences in estimates for y_T.

1.3. Objectives, outline and novelty

The objective of this work is to quantify the importance of making apparently arbitrary high-level modelling choices wisely in non-stationary extreme value analysis for metocean design, and specifically to quantify different sources of uncertainty in estimation of return values from non-stationary extreme value analysis of peaks over threshold, using maximum roughness-penalised likelihood methods. As outlined in Section 2, we will focus on four choices, namely that of (a) the extreme value model parameterisation (i.e. the choice of variables with which to define the GP model, e.g. shape ξ and scale σ as in the standard parameterisation of Eq. (1)), (b) the relative roughness of GP parameters as a function of covariate, (c) the cross-validation strategy for hyper-parameter tuning and (d) the estimator for return value subject to uncertainty. We hope to show that the metocean engineer can make these practical choices wisely, leading to improved estimation of return values. Section 3 then provides two motivating applications for the work, namely the estimation of non-stationary extreme value analysis of peaks over threshold of storm peak significant wave height (H_S) as a function of direction for locations in the northern and southern North Sea. Modelling choice (a) involves the adoption of different extreme value model parameterisations, explained in Section 4. Section 5 illustrates the impact of choices (a)–(d) in estimating return values from samples of data simulated from specific underlying models, chosen in part such that the simulated samples have characteristics similar to those observed in the northern and southern North Sea. For comparison, the performance of non-stationary GP models for threshold exceedances is also compared with that of a stationary GP model. A discussion and conclusions are provided in Section 6. A limited number of supporting figures are given in the Appendix A. Derivations of asymptotic covariance matrices for maximum likelihood estimates of generalised Pareto parameters, for standard, orthogonal and (relatively novel) mean-max parameterisations, are given in the Supplementary Material (SM), together with supporting material for the discussion in Sections 5 and 6.

Non-stationary extreme value analysis offers considerable potential to improve metocean design, but the associated statistical methodology is relatively complex, and must be applied carefully and thoughtfully. The novelty of the current work is that it provides a systematic quantitative study of the effects of key high-level non-stationary modelling choices on the quality of estimation of return values corresponding to long return periods. The article also provides a concise introduction to a parameterisation of the GP distribution with negative shape parameter, in terms of its mean and upper end point, which we believe may prove useful in the context of modelling non-stationary metocean extremes.

2. Practically important modelling choices

We choose to quantify the effect of four sources of uncertainty in estimation of return values using maximum roughness-penalised likelihood methods, introduced in Section 1.3. Sections 2.1–2.2 explain the potential influence of choices (a)–(d) from Section 1.3 on the inference. These choices are intended to represent the kinds of high-level modelling decisions necessary in setting up a typical analysis, which are often made on the assumption that their effect on return value estimates will be small.

For definiteness, consider a GP model for a sample $D = \{y_i, x_i\}_{i=1}^n$ of peaks over threshold Y and a single covariate X defined in terms of a set of unknown parameter functions (σ, ξ), where each of σ and ξ is a continuous function $(\sigma(x), \xi(x))$ defined on covariate domain $x \in D \subseteq \mathbb{R}$ for the non-stationary extreme value analysis. We seek to estimate the pair of functions (σ, ξ) from D using maximum penalised likelihood estimation. The variation of a parameter $\eta(x)$, $\eta \in \{\sigma, \xi\}$ for $x \in D$ might itself be represented e.g. as a spline, Gaussian process or Fourier series (see, e.g. Jones et al., 2016; see also Richards and Huser, 2024 for discussion of artificial neural networks). For clarity, and in order to focus on the estimation of uncertainties induces by the extreme value fitting, we assume that the extreme value threshold ψ and the rate of occurrence ρ of threshold exceedances are both known throughout.

2.1. Model parameterisation

The sample negative log likelihood $\ell(\sigma, \xi | D)$ is evaluated using Eq. (1)

$$\ell(\sigma, \xi | D) = -\sum_{i=1}^{n} \log f_{Y|\boldsymbol{X}}(y_i | x_i) \tag{6}$$

and the roughness penalised likelihood by

$$\ell^*(\sigma, \xi, \lambda | D) = \ell(\sigma, \xi | D) + \lambda_\sigma R(\sigma) + \lambda_\xi R(\xi) \tag{7}$$

for roughness coefficient vector $\lambda = (\lambda_\sigma, \lambda_\xi)$, where $R(\eta)$ is the marginal roughness of function $\eta(x)$, for $\eta \in \{\sigma, \xi\}$ on D. For example, we might define $R(\eta)$ using

$$R(\eta) = \int_D (\eta''(x))^2 dx \tag{8}$$

where $\eta''(x)$ is the second derivative of $\eta(x)$ with respect to x on D, with appropriate adjustment for periodic covariates such as direction or season. The objective of maximum penalised likelihood estimation is then to select an optimal value λ° of λ and corresponding values $(\hat{\sigma}_{\lambda^\circ}, \hat{\xi}_{\lambda^\circ})$ of (σ, ξ) which maximise the predictive performance of the model on independent test sample D°, so that

$$(\hat{\sigma}_\lambda, \hat{\xi}_\lambda) = \underset{\sigma, \xi}{\operatorname{argmin}}\, \ell^*(\sigma, \xi, \lambda | D) \tag{9}$$

for a given λ, and then

$$\lambda^\circ = \underset{\lambda}{\operatorname{argmin}}\, \ell(\hat{\sigma}_\lambda, \hat{\xi}_\lambda | D^\circ). \tag{10}$$

Now consider a transformation of variables defined by $(\alpha, \beta) = g(\sigma, \xi)$, for $x \in D$. In terms of transformed variables, the penalised likelihood becomes

$$\ell^{*\prime}(\alpha, \beta, \lambda' | D) = \ell'(\alpha, \beta | D) + \lambda'_\alpha R(\alpha) + \lambda'_\beta R(\beta). \tag{11}$$

for $\lambda' = (\lambda'_\alpha, \lambda'_\beta)$. When $\lambda = \lambda' = \mathbf{0}$, we expect that unconstrained minimisation of $\ell'(\alpha, \beta | D)$ with respect to (α, β) would yield the equivalent solution to unconstrained minimisation of $\ell(\sigma, \xi | D)$ with respect to (σ, ξ). However, there is no reason to expect that solutions to the penalised likelihood problems in Eqs. (7) and (11) for (σ, ξ) and (α, β) with $\lambda, \lambda' \neq \mathbf{0}$ would be equivalent, due to the presence of different marginal roughness penalty terms R. Hence, solutions from maximum penalised likelihood are not invariant to transformation of variables in general. This highlights the importance of considering different parameterisations of the generalised Pareto distribution for non-stationary extreme value inference using penalised likelihoods.

In passing we note from a Bayesian perspective that the combined choice of model parameterisation and roughness penalty amounts to a particular choice of parameter prior for the inference, and that particular choices may physically be more plausible. A procedure for estimating tuning parameters λ is fundamental to estimation of functional forms for the variation of ξ and σ on the covariate domain. In the current work, we use maximum penalised likelihood estimation and cross-validation to achieve this, but an equivalent procedure would be necessary regardless of the estimation method adopted. Below, we evaluate the relative performance of two model parameterisations, the orthogonal parameterisation (e.g. Cox and Reid, 1987), and a less frequently discussed *mean-max* parameterisation. These parameterisations are presented in Section 4. For comparison, we also evaluate the performance of stationary GP models (equivalent to $\lambda = \infty$). We also note that assuming roughness parameters λ to be constant on D is itself a modelling choice. It is possible that for a given application, allowing roughness parameters to vary in some way on D would be more appropriate physically; combining constant roughness with covariate models incorporating non-linear monotonic transformations of the covariate would be one means of achieving this.

2.2. Relative roughness penality for parameter functions

Inspection of Eq. (7) shows that both roughness parameters λ_σ and λ_ξ need to be estimated using the cross-validation scheme, involving multiple model fitting over a 2-D grid of potential values for the roughness parameters; this can be computationally demanding. An alternative approximation, reducing the computational burden considerably, is to fix the ratio $\lambda_\xi / \lambda_\sigma$ to some constant $\kappa > 0$. We might then re-write Equation (7), with λ replacing λ_σ as

$$\ell^*(\sigma, \xi, \lambda, \kappa | D) = \ell(\sigma, \xi | D) + \lambda(R(\sigma) + \kappa R(\xi)). \tag{12}$$

In the standard parameterisation of the GP distribution (see Eq. (1) for the corresponding density), setting $\kappa > 1$ would penalise the GP shape parameter with covariate more than the GP scale parameter; since the sample is generally less informative for the former, this makes intuitive sense. However, the best choice of constant is not clear, and in general problem-specific. Moreover, the choice of constant ratio of roughnesses is also dependent on the GP parameterisation selected. In our numerical study (Section 5) we consider the cases $\kappa = 10$ (found to be reasonable in historical applications) and $\kappa = 50$ for the orthogonal GP parameterisation; the latter choice penalises the roughness of $\xi(x)$ on D even more strongly.

2.3. Cross-validation strategy

In maximum penalised likelihood estimation, a popular procedure to select optimal values of λ and hence of (σ, ξ) (for fixed κ) is cross-validation. Typically in a cross-validation scheme, the full sample D is partitioned in some way into $C > 1$ disjoint sets $D_1, D_2, \ldots, D_C$. Then, in turn for $c = 1, 2, \ldots, C$, subset $D_{-c} = \cup_{\{c'=1, c' \neq c\}}^{C} D_{c'}$ is used as a model training sample (in place of D in Eq. (9)) for parameter estimation for specified λ, and predictive performance assessed using withheld set D_c (in place of D° in Eq. (10)). The value of λ° is then selected to be that which minimises the sum of predictive negative log likelihoods over the set of C withheld sets. In general, different cross-validation strategies (e.g. different partitioning strategies, values of C) will yield different choices of λ° and of $(\hat{\sigma}_{\lambda^\circ}, \hat{\xi}_{\lambda^\circ})$ (see e.g. Joseph, 2022; Risk and James, 2022; Aghbalou et al., 2023; Bates et al., 2023; Lopez et al., 2023; Yates et al., 2023). For extreme value analysis, it might appear advantageous that C be relatively large, so that each choice of D_{-c}, $c = 1, 2, \ldots, C$ is a good basis for estimation of the tail of the distribution from which D is drawn; however, using this strategy, each of the withheld sets D_c would be small, resulting in an uncertain estimate of the penalised likelihood on the withheld set. The "opposite" approach would be to set C as small as possible ($C = 2$), for which the training sample D_{-c} would be relatively small, but of approximately the same size as the withheld set D_c. The full cross-validation "loop" requires the estimation of C models; the computational complexity of the full cross-validation is therefore higher for large C. Because the partitioning of D is typically performed at random, the choice of random partition is a source of uncertainty in inference, particularly for small C. For small C, it is therefore desirable to perform the cross-validation $R \geq 1$ times for different random partitions of D, and choose λ° to maximise average predictive performance over all partitions. In general, we see that it is not clear how to specify the cross-validation strategy (e.g. the choice of C) which leads to the best-performing model in a given application. It may be however that a particular choice of C yields better model performance for an application of a given type. In our numerical study (Section 5) we consider the cases $C = 10, R = 1$ (found to perform reasonably in historical applications) and $C = 2, R = 50$.

2.4. Estimate for return value

Parameter estimates $(\hat{\sigma}_{\lambda^\circ}, \hat{\xi}_{\lambda^\circ})$ from the inference are uncertain. In a frequentist setting, this uncertainty is typically estimated using

bootstrap resampling. That is, the whole inference is repeated for a total of B bootstrap resamples D^b, $b = 1, 2, \dots, B$ of the original sample D, yielding a set of optimal roughness coefficients $\lambda^{\circ,b}$ and corresponding parameter estimates $(\hat{\sigma}^b_{\lambda^\circ}, \hat{\xi}^b_{\lambda^\circ})$ which can be used further to quantify uncertainties in subsequent inferences for quantities such as return values. However, in general, it is not clear how to exploit the set of parameter estimates $\{\hat{\sigma}^b_{\lambda^\circ}, \hat{\xi}^b_{\lambda^\circ}\}_{b=1}^B$ to estimate return values well (e.g. with low bias and variance). Given the cumulative distribution function F_A of the annual maximum event A, the T-year return value is defined using Eq. (5). However, F_A cannot be inferred exactly from data, but we can estimate it using the bootstrap parameter estimates $(\hat{\sigma}^b_{\lambda^\circ}, \hat{\xi}^b_{\lambda^\circ})$, $b = 1, 2, \dots, B$. Writing our estimate for F_A from bootstrap inference with index b as $F_{A|b}$ for brevity, it is nevertheless not clear how to best combine the different estimates $F_{A|b}$, $b = 1, 2, \dots, B$ to provide the best performing estimates of return values. Jonathan et al. (2021) discusses two approaches. The first approach estimates the return value as

$$\mathrm{MQ}(T) = \frac{1}{B}\sum_{b=1}^{B} F^{-1}_{A|b}\left(1 - \frac{1}{T}\right). \tag{13}$$

This estimator is the (predictive) mean over bootstraps of the quantile of the distribution of the annual maximum. For this reason, this estimator will be referred to as the *mean quantile* estimator MQ below. The second estimator first requires estimation of the (predictive) distribution $\tilde{F}_A$ of the annual maximum over bootstraps

$$\tilde{F}_A(y) = \frac{1}{B}\sum_{b=1}^{B} F_{A|b}(y). \tag{14}$$

The return value estimator is then given by the appropriate quantile of $\tilde{F}_A$

$$\mathrm{QM}(T) = \tilde{F}^{-1}_A(1 - \frac{1}{T}). \tag{15}$$

The QM estimator is the quantile of the predictive mean distribution over bootstraps. For this reason, this estimator will be referred to as the *quantile mean* estimator below. Jonathan et al. (2021) showed, using a combination of theoretical arguments and numerical simulations, that MQ and QM estimators have predictable and markedly different bias characteristics for small samples, and that QM estimates can be much larger that those from MQ, for sample characteristics corresponding to typical metocean data. The current work can be thought of in part as extending (Jonathan et al., 2021), which addressed fundamental bias issues in extreme value modelling of stationary peaks over threshold from finite samples, to a more practically important setting, incorporating non-stationarity with respect to covariates.

In passing we note that differences between MQ and QM estimators are not confined to inference using penalised likelihood estimation; similar behaviour is observed in a Bayesian setting. In general, therefore, it is not clear which estimator of return value provides the best estimates in practical application of non-stationary extreme value analysis, in particular using maximum penalised likelihood inference. It is also often unclear which characteristics of return value estimates are most important in a particular application setting (e.g. low bias, low variance, conservatism): what exactly does "best estimate" mean? Probably the best approach would be to evaluate the inference in terms of the expected cost of the decisions made using it (e.g. Berger, 1985; Speers et al., 2024). In our numerical study (Section 5) we consider both MQ and QM estimators.

3. Motivating application

We motivate the current study using data for storm peak significant wave height (H_S) from a locations in the northern and southern North Sea, which show clear directional variation in the characteristics of H_S (Fig. 1). The data are drawn from the NORA10 hindcast of Reistad et al. (2011) for the period 1957–2012, with storm direction indicating the direction from which the storm emanates, given clockwise from North. The left panel of Fig. 1 indicates the sheltering effect of the Norwegian coastline for directions $\in (50, 210)$ and the focussing of Atlantic storms by the coastline around 230°. The presence of large storms approaching from the north from the Norwegian Sea is also clear. The right panel of Fig. 1 indicates broadly similar structure, except that storm severity is reduces markedly. In this case, however, the increased rate of occurrence of events at around 230° is due to storms from the English Channel. The largest events approach from the north, travelling down the North Sea. The general characteristics of Fig. 1 include the presence of covariate intervals on which (a) the rate of occurrence of events is low, and (b) the rate and size of events changes rapidly. We are interested specifically in evaluating the effects of high-level modelling choices on the quality of inference for North Sea applications with these characteristics.

4. Generalised Pareto model parameterisations

As discussed in Section 2.1, model parameterisation can influence inferences from a maximum penalised likelihood analysis of non-stationary extremes. We now summarise three parameterisations of the GP distribution, namely the "standard" scale and shape parameterisation (Section 4.1), the orthogonal parameterisation (Section 4.2) and mean-max parameterisation (Section 4.3). Both the orthogonal and mean-max parameterisations are applied in the numerical study of Section 5. For simplicity of presentation, we suppress dependence on covariates, and assume the extreme value threshold to be zero.

4.1. Standard (σ, ξ) parameterisation

Assume that Y follows a GP distribution with standard parameterisation, for zero threshold, scale $\sigma > 0$, shape $\xi \in \mathbb{R}$. The cumulative distribution function F_s of Y is then

$$F_s(y; \sigma, \xi) := 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi} \tag{16}$$

where the support of Y is $[0, \infty)$ if $\xi \geqslant 0$ and $[0, -\sigma/\xi)$ otherwise. The corresponding density is given in Eq. (1) (with adjustments for threshold and covariates) and sample likelihood in Eq. (6). As shown in Section SM1 of the Supplementary Material, the asymptotic covariance of maximum likelihood estimators for σ and ξ in the standard parameterisation is

$$\mathrm{cov}([\hat{\sigma}, \hat{\xi}]) = \begin{pmatrix} 2\sigma^2(1+\xi) & -\sigma(1+\xi) \\ -\sigma(1+\xi) & (1+\xi)^2 \end{pmatrix} \tag{17}$$

implying that the maximum likelihood estimates show a negative correlation

$$\mathrm{cor}(\hat{\sigma}, \hat{\xi}) = -\frac{1}{\sqrt{2(1+\xi)}}. \tag{18}$$

Asymptotic correlation between parameter estimates is not a desirable feature for numerical inference. For this reason, the orthogonal parameterisation, discussed next, is preferred by many in practice.

4.2. Orthogonal (ν, ξ) parameterisation

Following Cox and Reid (1987) and Davison (2003), we can choose instead to parameterise the generalised Pareto distribution in terms of a pair of orthogonal parameters, namely shape ξ and modified scale $\nu = \sigma(1+\xi)$, where σ and ξ are the parameters for the standard parameterisation above. The form of the corresponding cumulative distribution function is obtained by substituting the expression $\sigma = \nu/(1+\xi)$ for $\xi > -1$ in Eq. (16). As shown in SM2, the asymptotic covariance of maximum likelihood estimators for the standard parameterisation is given by

$$\mathrm{cov}([\hat{\nu}, \hat{\xi}]) = \begin{pmatrix} \nu^2(1+2\xi) & 0 \\ 0 & (1+\xi)^2 \end{pmatrix}$$

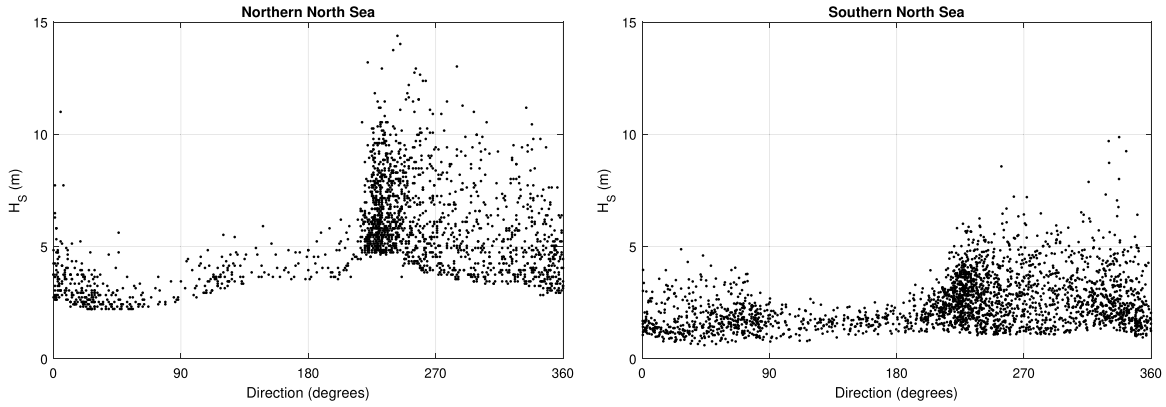

Fig. 1. Illustrations of the variation of storm peak significant wave H_S height with direction for locations in the northern North Sea (left) and southern North Sea (right). Direction is the direction from which the storm emanates, clockwise from North.

from which we note that parameter estimates $\hat{\xi}$ and $\hat{\nu}$ are asymptotically uncorrelated, by construction. This is a desirable feature for numerical inference, since in essence updates for $\hat{\xi}$ and $\hat{\nu}$ can be made independently of each other in an iterative numerical scheme.

4.3. Mean-max (μ, ζ) parameterisation

When $\xi < 0$, random variable Y has an upper end point or maximum ζ given by $\zeta = -\sigma/\xi$, where σ and ξ are the parameters of the standard parameterisation. In metocean analysis, for a variable like H_S, it is usually reasonable to assume that an upper end point ζ exists. Moreover, the upper end point is often a parameter of direct interest, e.g. for analysis of (significant) wave height in shallow water. Indeed, it may be the case that physical considerations allow the value of ζ to be fixed, constrained or elicited more naturally than ξ or σ. It is therefore intuitively appealing to parameterise the generalised Pareto distribution using ζ. A natural partner parameter for ζ would then be the distribution mean $\mu = \sigma/(1-\xi)$, which itself can be interpreted more naturally than either σ or ξ, as the mean excess. The cumulative distribution function F_m of Y is then given by

$$F_m(y;\zeta,\xi) = 1 - \left(1 - \frac{y}{\zeta}\right)^{(\zeta-\mu)/\mu}, \tag{19}$$

for $\zeta > \mu$ and $\zeta \geq y > 0$. As show in Section SM3, for $\zeta > 3\mu$, the asymptotic covariance matrix of the maximum likelihood estimators is

$$\text{cov}([\hat{\mu},\hat{\zeta}]) = \frac{\zeta - 2\mu}{\mu^2\zeta^2(\zeta-\mu)} \begin{pmatrix} \mu^4(\zeta^2 - \mu\zeta + 2\mu^2) & -\mu^3\zeta^2(\zeta - 3\mu) \\ -\mu^3\zeta^2(\zeta-3\mu) & \zeta^4(\zeta-2\mu)(\zeta-3\mu) \end{pmatrix} \tag{20}$$

and their asymptotic correlation is therefore

$$\text{cor}(\hat{\mu},\hat{\zeta}) = -\mu\sqrt{\frac{\zeta-3\mu}{(\zeta-2\mu)\left((\zeta-\mu)(\zeta-2\mu)+2\mu\zeta\right)}} = -\sqrt{\frac{\xi^2(1+2\xi)}{1+\xi^2(1+2\xi)}} \tag{21}$$

in terms of ξ. As shown in Figure SM1 (in Section SM6 of the SM), for $\xi \in (-0.5, 0)$, although the mean-max parameter estimates are correlated asymptotically, the correlation is smaller in magnitude than that of the parameters of the standard parameterisation. This makes the mean-max parameterisation relatively more appealing from a numerical optimisation perspective. We note from Eq. (20) that, as might be expected, the variance of the estimator $\hat{\zeta}$ increases with increasing ζ; indeed $\text{var}(\hat{\zeta}) \sim \zeta^4$, despite the fact that ξ is restricted to be <0.

5. Numerical study

In this section, we perform a numerical study using 16 case studies constructed to mimic the characteristics of the motivating metocean application illustrated in Fig. 1, specifically the strong variation in the rate and size of extreme events with direction. The purpose of the study is to quantify the influence of apparently arbitrary modelling choices on estimates of omni-directional and directional return values corresponding to a return period of 1000 times the period of the sample. Inference is performed using maximum penalised likelihood GP inference for peaks over threshold as outlined in Section 2.1, for different model variants. We quantify the effect of (combinations of) four specific sources of modelling uncertainty on return value estimation, namely (a) the extreme value model parameterisation, (b) the relative roughness of GP parameters as a function of covariate, (c) the cross-validation strategy for hyper-parameter tuning, and (d) the estimator for return value subject to uncertainty, as outlined in Sections 2.1–2.2 For comparison and completeness, we also estimate stationary extreme value models. Section 5.1 introduces the 16 case studies, and outlines their different characteristics. Section 5.2 then outlines the inferences performed. Results are discussed in Section 5.3 in terms of the relative bias in estimation of omni-directional return values, and in Section 5.4 for relative bias in specific directional sectors.

5.1. Case studies

A total of $n_{\text{Cas}} = 16$ case studies were constructed motivated by northern and southern North Sea data discussed in Section 3. Illustrations of typical samples for sample size $n_{\text{Smp}} = 10^3$ are shown in Fig. 2, as a function of a single covariate $x \in D = [0, 360)°$ together with variation of upper end point ζ (blue) and mean excess μ (green) on D. The corresponding sample plot for sample size $n_{\text{Smp}} = 10^6$, shown in Fig. A.2 illustrates for more negative ξ, that upper end point ζ is approached more quickly. Plots of the corresponding variation of σ and ξ on D are given in Fig. A.1. For the current study, we assume that the rate of occurrence of threshold exceedances is known and Poisson-distributed. For cases 1, 2, …, 9 and 16, the density $f_X(x)$ of storm occurrence given by

$$f_X(x) \propto a + b\left(1 + \text{sign}\left(\sin\left(x\frac{\pi}{180}\right)\right)\left|\sin\left(x\frac{\pi}{180}\right)\right|^c\right) \tag{22}$$

where $a = 0.002$, $b = 0.04$ and $c = 0.6$, scaled appropriately so that $\int_D f_X(x)dx = 1$. That is, $f_X(x)$ and hence the directional rate $\rho(x) = \rho_A f_X(x)$ of storm occurrence resembles a "flattened" sine function on D, with maximum at $x = 90°$ and minimum (>0) at $x = 270°$. For

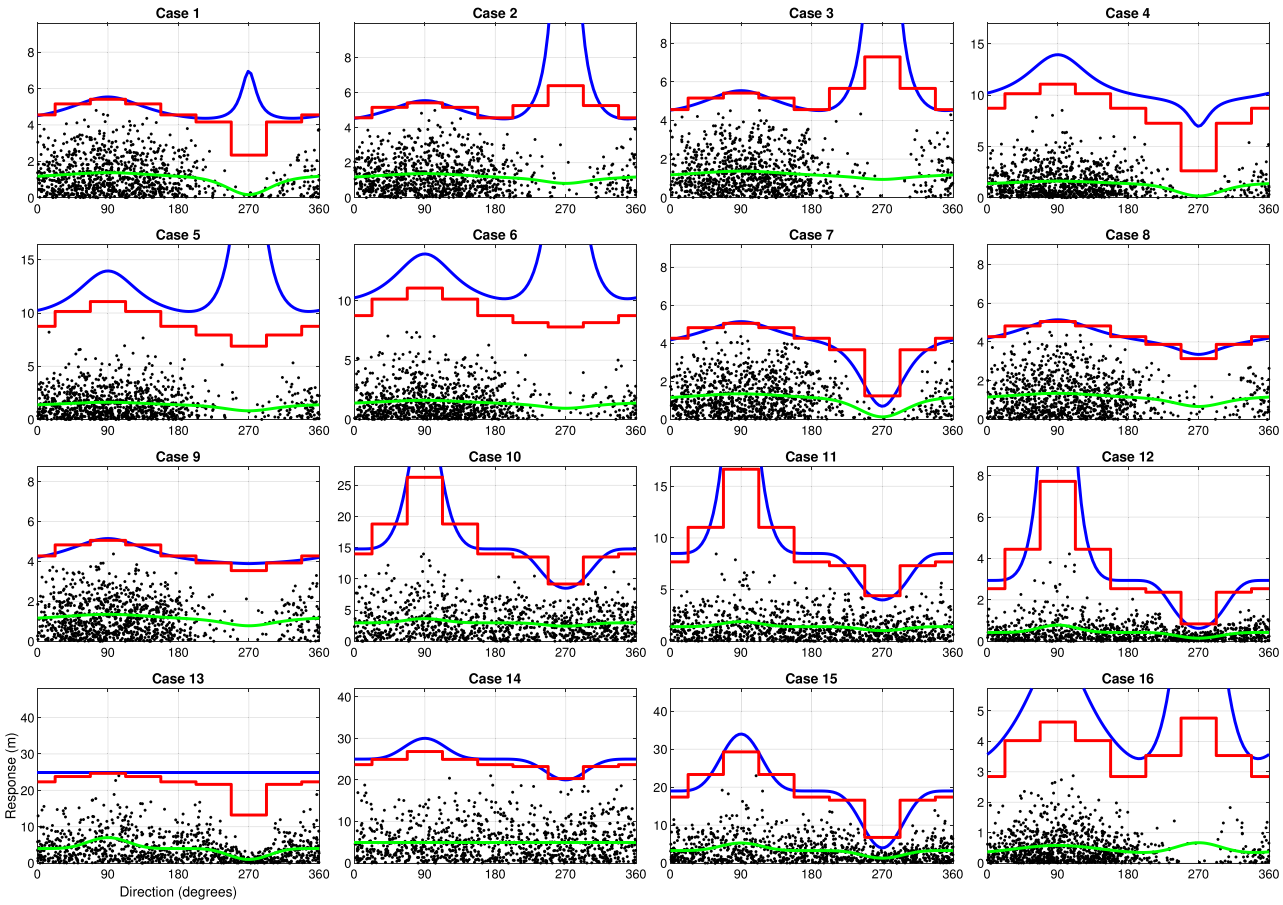

Fig. 2. Illustrative samples (black dots) generated from each of Cases 1–16, for a response as a function of direction, for samples of size 10^3. Also shown per panel is the corresponding conditional distributional mean μ (green) and upper end point ζ (blue). Finally, lines in red indicate the true return values to be estimated, for directional octants corresponding to a return period of 1000× that of the sample.

cases 10, 11, …, 15, the rate is constant on $\mathcal{D}$. From Figs. 2, A.1 and A.2, we observe that the characteristics of the extreme value tail are, by design, quite different at $x \approx 270°$ compared with the remainder of $\mathcal{D}$ for the majority of cases, mimicking the "benign" sector at $\approx [90, 180)°$ observed in the metocean data (Fig. 1) under consideration. We expect therefore that accurate inference for $x \approx 270°$ will be more challenging than elsewhere. Indeed the subset of Cases 1, 4, 8 and 9 are presumed to be "more similar" to the southern North Sea metocean sample than the remaining cases, because these cases exhibit both a dominant directional sector and an effectively "empty" sector, in which observations are relatively more rare and smaller in magnitude. For the current study we also note, given directional covariate, that the conditional distribution of observations is GP. Therefore, we adopt an extreme value threshold of zero on $\mathcal{D}$ for all non-stationary modelling. For fitting of stationary GP models, we consider two extreme value threshold choices discussed further in Section 5.2.

5.2. Outline of inference

For each of the $n_{\text{Cas}} = 16$ case studies, we generate $n_{\text{Rpt}} = 23$ realisations of samples of size $n_{\text{Smp}} = 10^3$, using the specification for $\psi(x)$, $\rho(x)$ and $Y|(X = x)$ given in Section 5.1. For uncertainty quantification, we then generate $n_{Bts} = 50$ bootstrap resamples of the original sample for each case study. We consider a total of n_{Mdl} modelling strategies, which can be characterised as follows. Firstly, we consider three combinations of model parameterisation (orthogonal or mean-max; see Section 4) and relative roughness κ (= 10 and = 50; see Section 2.2). For the mean-max parameterisation, we fix $\kappa = 1$, but acknowledge that this also could be varied. Secondly, we consider two cross-validation procedures ($(C = 10, R = 1)$ and $(C = 2, R = 50)$; see Section 2.3). Thirdly, we consider two return value estimator (MQ and QM; see Section 2.4) of interest. Moreover, we also estimate a stationary extreme value model (using the full sample) for comparison using both MQ and QM estimators. This results in a total of $n_{\text{Mdl}} = 14$ modelling strategies to be executed. Then for each bootstrap resample of each sample realisation of each case study, we estimate a non-stationary GP model, with the forms of $\psi(x)$ (= 0) and $\rho(x)$ (see Eq. (22)) assumed known, focussing on estimation of the GP tail, using each of the n_{Mdl} modelling strategies. In total therefore, the full "kernel" non-stationary extreme value analysis of a sample size n_{Smp} is executed $n_{\text{Cas}} \times n_{\text{Rpt}} \times n_{\text{Bts}} \times n_{\text{Mdl}} \approx 3 \times 10^5$ times. We use the fitted models to estimate the relative bias of return value estimates for a return period of 1000× the period of the original sample.

For clarity and brevity in this section, we use acronyms to refer to the various modelling strategies. The three combinations of parameterisation and relative roughness penalty are referred to as `OK*` (orthogonal parameterisation with relative roughness penality $*$) and `MM` (mean-max), the two combinations of cross-validation procedure as `CV*` (cross-validation with $*$ groups), and the two return value estimators as `MQ` and `QM`. The constant fit is simply referred to as `Constant*` (where $*$ is the auxiliary threshold non-exceedance probability employed).

Inference is performed using maximum penalised likelihood estimation (see Section 2.1), with cross-validation for selection of optimal directional roughness coefficient λ. The variation of GP parameters η

($\in (\nu, \xi)$ for the orthogonal parameterisation, and $\in (\mu, \zeta)$ for the mean-max parameterisation) on covariate domain D is represented by a spline model: that is, for a set of $n_I \geq 1$ index locations uniformly spaced on D, we write the $n_I \times 1$ vector $\boldsymbol{\eta}$ as $\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\beta}_\eta$ giving the values of η at each index location, for pre-specified $n_I \times n_{\text{Spl}}$ B-spline basis matrix $\boldsymbol{B}$ (with $n_{\text{Spl}} \geq 1$ degrees of freedom; see e.g. Zanini et al., 2020; cubic B-splines are used here). Likelihood roughness penalisation is imposed using a second difference roughness penalty $R(\eta) = \boldsymbol{\beta}_\eta' \boldsymbol{G}' \boldsymbol{G} \boldsymbol{\beta}_\eta$, where $n_{\text{Spl}} \times n_{\text{Spl}}$ first difference matrix $\boldsymbol{G}$ has elements $\boldsymbol{G}_{i,j} = 1$ for $i = j$, $= -1$ for $i = j + 1$ and $= 0$ otherwise, with appropriate adjustments for periodicity (see e.g. Zanini et al., 2020).

Since the sample exhibits clear non-stationarity, we might expect it to be unreasonable to estimate a stationary GP model using the full heterogeneous "omni-directional" sample. As mitigation, we define a second "auxiliary" extreme value threshold, corresponding to the 90th "marginal" percentile of the sample in each case, estimated ignoring covariate X. The stationary GP model was then applied to exceedances of this threshold. The expected sample size for estimation of the stationary GP model is therefore $n_{\text{Smp}}/10 = 100$ (noting that the full sample of n_{Smp} points were nevertheless used to estimate the second threshold). Note that cross-validation is unnecessary for estimation of the stationary model, since there is no roughness penalty to tune. We acknowledge that adopting a 90% threshold for stationary analysis is somewhat arbitrary, but nevertheless represents a not-unreasonable choice. Threshold selection for extreme value analysis is itself a challenging topic (e.g. Tancredi et al., 2006; Thompson et al., 2009; Wadsworth and Tawn, 2012; Scarrott and MacDonald, 2012; Northrop et al., 2017; Mackay and Jonathan, 2020; Murphy et al., 2024).

Results are summarised in Sections 5.3 and 5.4 in terms of box-whisker plots for the relative bias of return value estimates, for a return period of 1000× the period of the sample, under the different modelling strategies over all repetitions of the sample. We acknowledge that setting the return period to 1000× the period of the sample is challenging, but not wholly unrealistic: consider estimating a 10,000 year event from 10 years of data.

5.3. Fractional bias in omni-directional return value

Fig. 3 gives the relative error in return value estimate for each combination of case and modelling strategy. In each box-whisker structure, the vertical extent of a box indicates the sample inter-quartile range (q_L, q_H). A sample point s is marked conventionally as an "outlier" if $s - q_H > 1.5(q_H - q_L)$ for $s > q_H$ or $q_L - s > 1.5(q_H - q_L)$ for $s < q_L$. The vertical extent of whiskers then indicates the sample range, considering "non-outlier" points only. The outlier points themselves are then indicated using red crosses. A horizontal line in each panel indicates the optimal outcome of zero bias.

Inspection of Fig. 3 suggests the following: (i) There is systematic negative bias in return value estimates across all modelling strategies in many cases. The negative bias is particularly large for Cases 3, 10, 11 and 12; (ii) quantile mean estimates using cross-validation strategy ($C = 10, R = 1$) (i.e. `OK10_CV10_QM`, `MM_CV10_QM`,`OK50_CV10_QM`) tend to be more positive than all other quantile mean estimates, and all mean quantile estimates, for non-stationary models, for all cases. These estimates also tend to have large inter-quantile ranges. Moreover, there is little difference in performance between `*_CV10_MQ`, `*_CV2_MQ` and `*_CV2_QM` strategies for a given model parameterisation; (iii) The mean-max estimates (blue) tend to be more negatively biased than corresponding estimates using orthogonal parameterisations. There is little difference between outcomes for ratio of roughness penalties $\kappa = 10$ and $= 50$; (iv) Estimates from stationary extreme value fits to exceedances of the 90% sample threshold (labelled "`Constant0.9*`" in the figure, with box plots in magenta) are more variable than those from non-stationary models in general. However, the bias characteristics of stationary estimates are visually no worse than those of estimates from non-stationary models. For further comparison, we also include stationary estimates using the full sample of n_{Smp} points (labelled "`Constant0*`", cyan), in full knowledge of the fact that we are fitting a misspecified stationary model to non-stationary data. A priori we would therefore expect `Constant0*` estimates to show lower variance, but higher bias than those of `Constant0.9*`. Generally speaking, Fig. 3 confirms lower variance but, with the notable exception of Case 13, there seems visually to be little difference in general bias characteristics for the `Constant0*` and `Constant0.9*` inferences.

The bias effect (i) can be explained in part by inspection of estimated parameters in Figs. 4 and 5, where we compare inferences for Cases 5 and 10 under the orthogonal parameterisation with roughness ratio $\kappa = 10$, and the mean-max parameterisation. For the orthogonal parameterisation, estimates for Case 5 on left hand side of Fig. 4 reflect the low rate of occurrence of events at around 270° (see Fig. 2). This makes identification of the specific characteristics of GP shape ξ and scale σ there impossible; the sample is simply not sufficiently informative in this region of the covariate domain. Therefore, the values of estimated parameters in this region tend to be inferred form neighbouring regions of the covariate domain. However, the low rate of occurrence at around 270° also means that estimating this region well in the GP inference is less important from the perspective of estimating the return value, at least for return periods of interest in ocean engineering. For the orthogonal parameterisation and Case 10 (on the right hand side of Fig. 4), the sample is informative (since the rate of occurrence is now constant with direction; see Fig. 2) for the whole covariate domain. The inference fails to recover the full extent of variability in ξ and σ, with the bootstrap 95% uncertainty band for estimated ξ covers the true parameter values everywhere except for around 90° and around 270°. Note further that each sample of size 10^6 illustrated in Fig. A.2 corresponds to a sample 1000× as large as that used for inference in this section. That is, the maxima observed in Fig. A.2 as a function of direction are observations for precisely the return period of interest. For Case 5, this maximum ζ varies relatively smoothly with direction on its domain. For Case 10 however, ζ varies relatively quickly with direction at around 90°, but more smoothly elsewhere. Lack of fit around 90° results in underestimation of omni-directional return values for Case 10. Estimates for the mean μ under the mean-max parameterisation (Fig. 5) are excellent as might be hoped, with 95% bootstrap uncertainty bands covering the true parameter variation on the covariate domain. However, the sample is not at all informative about the directional variation of the upper end point; hence an effectively constant estimate of ζ is inferred, leading (as for inference using the orthogonal parameterisation) to underestimation of return value for Case 10.

Effect (ii) can be explained by the fact that cross-validation strategy ($C = 10, R = 1$) tends to produce more variable parameter estimates than ($C = 2, R = 50$). Coupled with the fact that QM estimates tend to be larger than those using MQ (see Section 2.4), we observed that modelling strategies `*_CV10_QM` tend to produce large estimates. We speculate that effect (iii) is due to the fact that the mean-max parameterisation imposes smoothness directly on the upper end point as a function of direction, and hence tends to produce smoother models than the orthogonal parameterisation. The relative performance of stationary fits (effect (iv)) is perhaps to be expected. If the auxiliary threshold is set high enough, then intuitively inspection of Fig. A.2 suggests that we can effectively eliminate covariate non-stationarity in the sub-sample of auxiliary threshold exceedances, at least for some of the cases, so that the tail of the omni-directional distribution is well-approximated by a GP. Given this, we would expect a stationary model to perform relatively well in terms of fractional bias (see e.g. Mackay et al., 2010). However, the resulting reduced sample size suggests that the variance of return value estimates would be large. Case 13 is interesting because the true upper end point of the GP distribution is constant with direction (see Fig. 2). We might therefore expect that

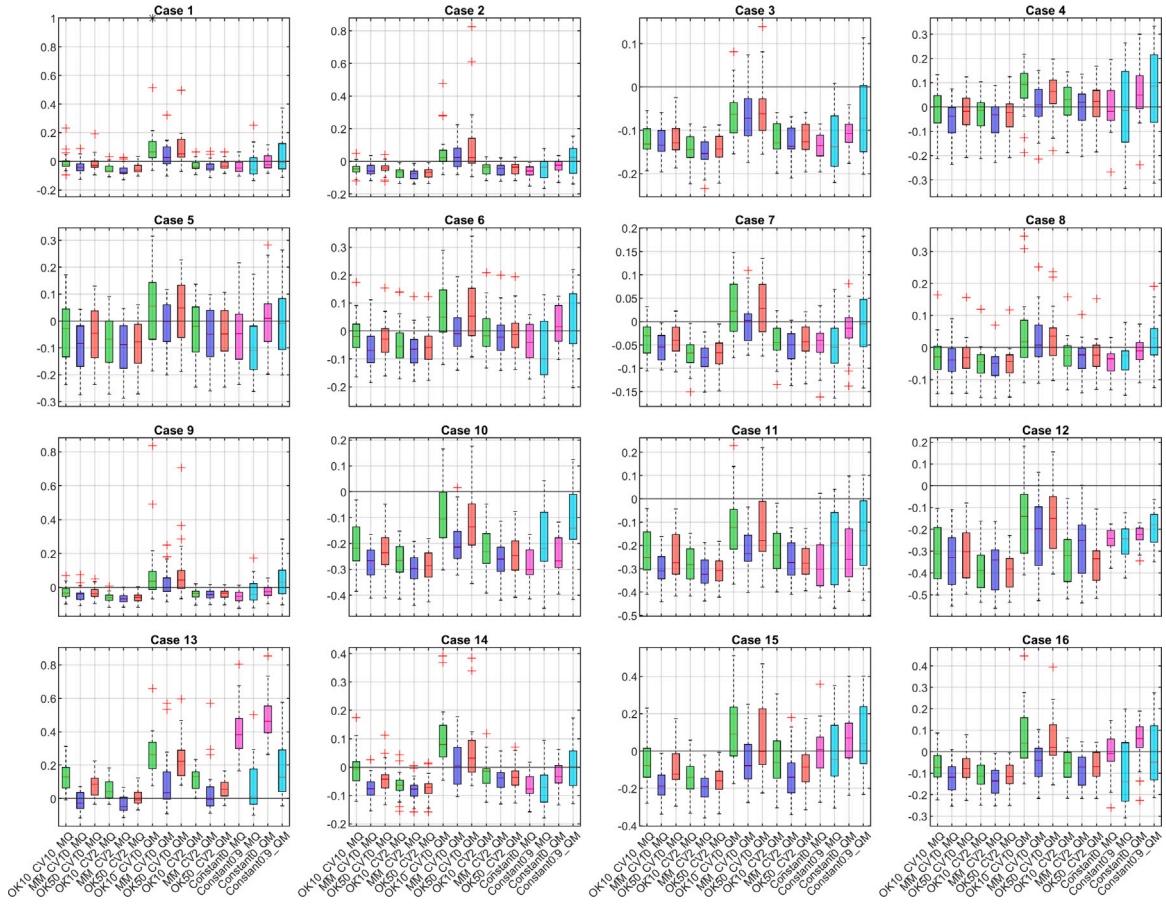

Fig. 3. Fractional bias of return values for each of Cases 1–16, corresponding to a return period of 1000× the period of the data, over $n_{\text{Rpt}} \times n_{\text{Bts}}$ repetitions of the sample. Colours distinguish model fits as follows: green (orthogonal parameterisation, $\kappa = 10$); blue (mean-max parameterisation); red (orthogonal parameterisation, $\kappa = 50$); magenta (stationary fit, full sample); and cyan (stationary fit with 90% auxiliary "marginal" threshold). Models are specified in terms of parameterisation ("`O`", "`MM`" and "`Constant`"), value of κ ("10" or "50" for orthogonal parameterisation only), cross-validation strategy ("`CV10`" and "`CV2`", for all but constant models) and return value estimator ("`MQ`" and "`QM`"). In each panel, the vertical extent of a box indicates the sample inter-quartile range, and whiskers the sample "non-outlier" range. Red crosses indicate "outliers", as defined at the start of Section 5.3. Star symbols at the maximum value of fractional bias for a given model indicate that at least one outlier has been censored at this level. A horizontal black line is included in each panel, corresponding to zero fractional bias.

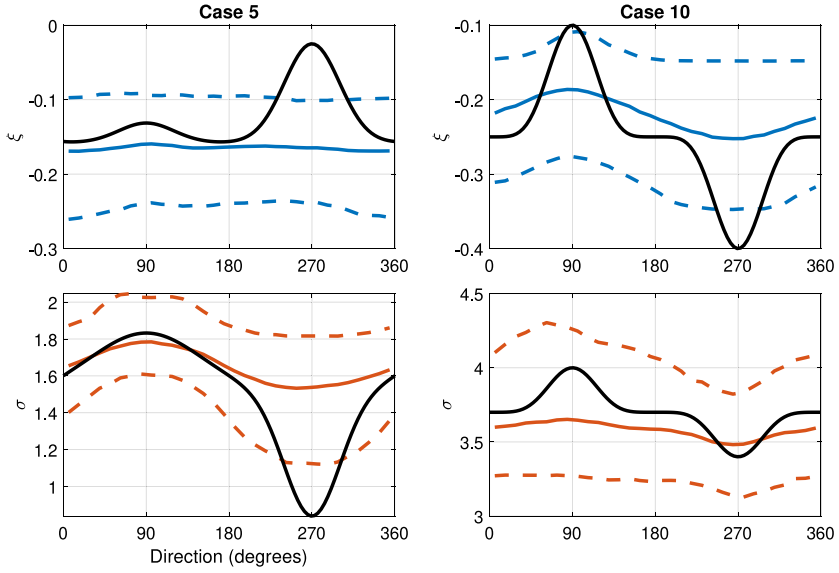

Fig. 4. Parameter estimates for ξ (top) and σ (bottom) for modelling strategy `OK10_CV2_*` aggregated over n_{Bts} samples for each of n_{Rpt} repetitions for Case 5 (left) and Case 10 (right). In each panel, true parameter variation on the covariate domain is shown as a full black line. Estimates from extreme value modelling are summarised in terms of the mean over bootstraps and repetitions (full coloured line), and in terms of a central 95% (bootstrap and repetition) uncertainty band (dashed coloured lines).

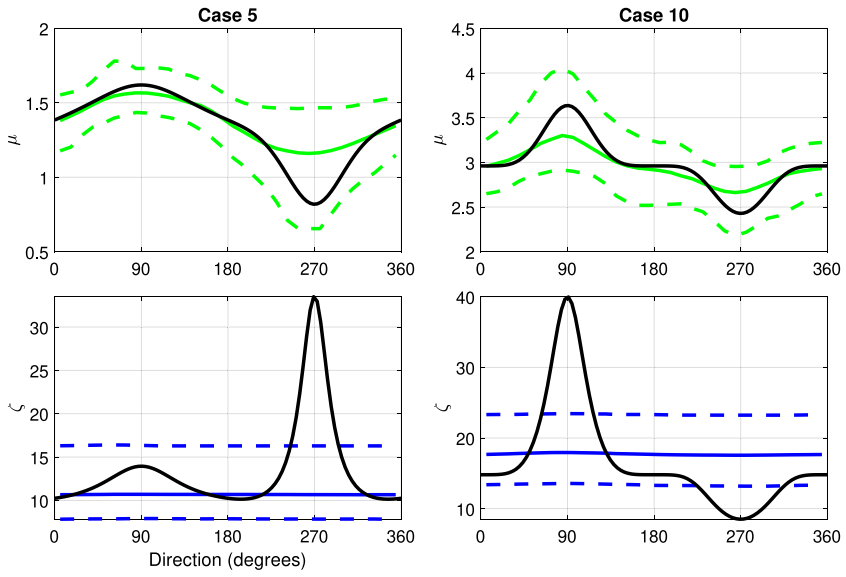

Fig. 5. Parameter estimates for μ (top) and ζ (bottom) for modelling strategies `MM_CV2_*` aggregated over n_{Bts} samples for each of n_{Rpt} repetitions for Case 5 (left) and Case 10 (right). In each panel, true parameter variation on the covariate domain is shown as a full black line. Estimates from extreme value modelling are summarised in terms of the mean over bootstraps and repetitions (full coloured line), and in terms of a central 95% (bootstrap and repetition) uncertainty band (dashed coloured lines).

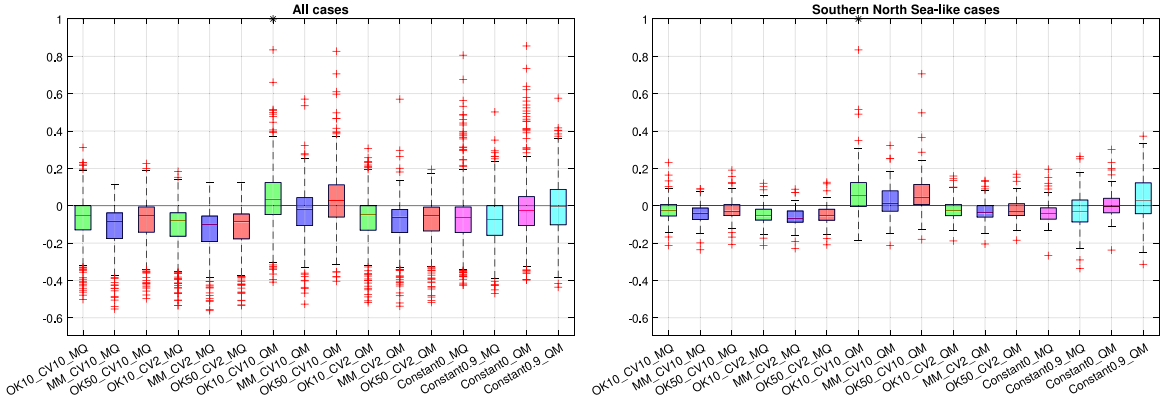

Fig. 6. Fractional bias of return values aggregated over all cases, corresponding to a return period of 1000× the period of the data, for all cases (left) and Cases 1, 4, 8 and 9 (resembling southern North Sea conditions, right). For other details, see Fig. 3.

a mean-max model with heavily penalised max parameter ζ would perform well. Inspection of box-whiskers in Fig. 3 corresponding to `MM*` models (in blue) shows that this is generally the case.

Fig. 6 summarises Fig. 3 by aggregating over multiple cases. Specifically, a box-whisker structure in the left hand panel of Fig. 6 summarises a sample of return value estimates corresponding to all repetitions of all cases, for a given modelling strategy. The right-hand panel aggregates only over Cases 1, 4, 8 and 9 similar to the southern North Sea metocean sample. Comparison of left and right hand panels indicates that different modelling strategies may be appropriate, depending on the characteristics of the sample. Fig. 6 suggest that most if not all the modelling strategies are competitive for estimation of the omni-directional return value. Over all case studies, modelling strategies `*_CV10_QM` are relatively unbiased (for the current choice of case studies) but show increased variance compared to non-stationary competitors. For the southern North Sea-like case studies, variances are generally lower for strategies other than `*_CV10_QM`. Return value estimates from stationary models perform relatively well in terms of bias, but exhibit higher variance compared to the best of the alternatives. However, we stress that the drawing of general conclusions from the figure would be unwise, since the relative performance of different modelling strategies is influenced by the specific characteristics of the case studies considered.

5.4. Fractional bias in directional-octant return value

In this section, we assess the performance of different approaches to non-stationary GP inference for directional sectors of width 45° centred on cardinal directions 0°, 90°, 180° and 270° in terms of fractional bias of return value estimates for a return period 1000× as long as the period of the sample. Results for directional sectors centred at 0° and 270° are given in the main text; results for the remaining sectors are given in Section SM7. Note that, since typical sample size for estimation of the constant model is small in each of the sectors, the constant model is only estimated using all available data in the sector. We consider this a reasonable assumption given that the sample per directional sector will be relatively homogeneous. Further, the sector rate of occurrence of events is assumed known.

For directional sectors centred at 0° and 180°, Fig. 7 and SM4 indicate that estimates from constant models are more negative, and more variable than those from corresponding non-stationary models. When non-stationary inference overestimates return values (e.g. Cases 10–12), this can be advantageous; however when non-stationary inference performs reasonably (e.g. Case 5, 14), estimates from constant models show negative bias. Over all cases, Fig. 8 and SM5 summarise behaviour, suggesting that the range of fractional bias is approximately ±0.5. Fractional biases are somewhat smaller for SNS-like cases. For

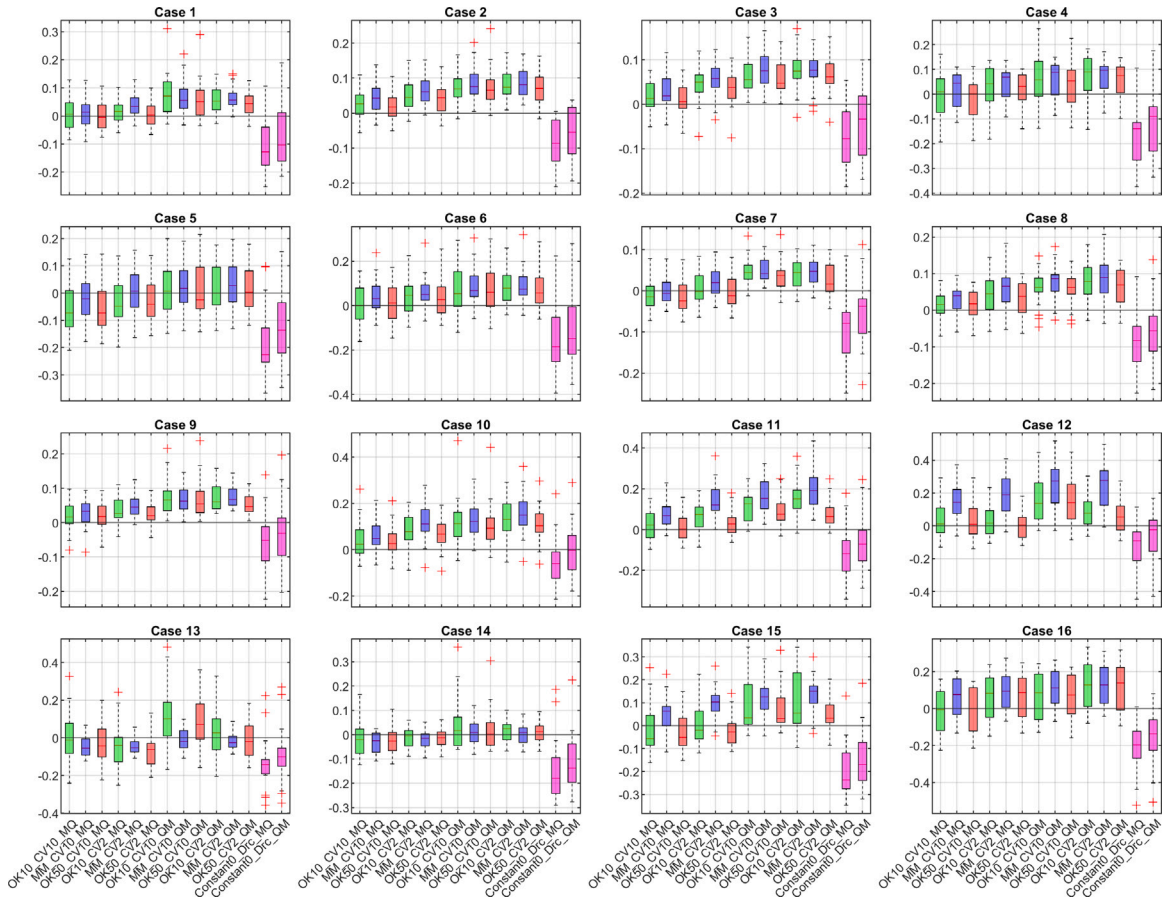

Fig. 7. Fractional bias of return values for each of Cases 1–16, corresponding to a return period of 1000× the period of the data, for directional sector $[-22.5, 22.5)°$. For other details, see Fig. 3.

the directional sector centred at 90°, Figures SM2 and SM3 suggest that estimates from constant models are generally similar to those from non-stationary models, but tend to be more variable. The range of fractional bias is approximately ± 0.6 over all cases, but slightly smaller for SNS-like cases.

For the directional sector centred at 270°, Fig. 9 shows that the fractional bias of estimates from the constant model vary considerably between cases; they are also generally lower than non-stationary estimates. In some cases, estimates from constant models are spectacularly bad (e.g. Cases 3, 5, 6). In other cases, where non-stationary inference overestimates (e.g. Cases 10, 11), estimates from constant models show less bias. The summary plot Fig. 10 shows that the typical range of fractional bias is much greater than ± 1 (and see Figure SM6 and SM7 for comparison, for a truncated y-axis).

Inspection of results corresponding to Case 13 (for which max parameter ζ is constant on the covariate domain) for all four directional sectors (in Figs. 7, 9, SM2 and SM4) again indicates excellent performance from the mean-max parameterisations `MM*` (shown in blue).

6. Discussion and conclusions

This paper addresses the effect of choice of high-level tuning parameters on the performance of non-stationary extreme value modelling for peaks over threshold using the generalised Pareto distribution. Specifically we examine the effect of (a) extreme value model parameterisation, (b) relative roughness of GP parameters as a function of covariate, (c) cross-validation strategy for hyper-parameter tuning, and (d) estimator for return value subject to uncertainty on the estimation of return values corresponding to return periods 1000× the period of the sample. Samples of size 1000 are assumed for inference, using maximum penalised likelihood inference with spline representations for the variation of generalised Pareto parameters on the covariate domain.

We can use regression analysis to quantify the trends observed in the figures summarising results in Sections 5.3 and 5.4. Table 1 and SM2 (in the SM) give the values of t-statistics for regression coefficients from a regression model for the natural logarithm of the fractional absolute bias (FAB) of different modelling strategies and case studies, represented in terms of categorical "treatment" dummy variables. Modelling strategy `OK10_CV10_MQ` on Case 1 is used as the base or reference treatment, with all other strategy-case combinations compared to it. The regression model takes the form

$$\log(\mathrm{FAB}) = \mathrm{Intercept} + \sum_{j=1}^{n_V} \sum_{k=1}^{n_j} b_{jk} x_{jk} + N(0, v^2) \tag{23}$$

for each of n_V categorical variates, each with n_j levels, $j = 1, 2, \ldots, p$, representing the combined effect of case, model parameterisation and roughness penalty ratio, cross-validation strategy and return value estimator. We assume a Gaussian error with mean zero and standard deviation v. The tables report the values of t-statistics $\hat{b}_{jk}/\mathrm{se}_{jk}$, for all combination of j, k, where $\hat{b}_{jk}$ is the estimate of b_{jk} from the regression, and se_{jk} is the standard error of the parameter estimate from the regression. t-statistics with absolute values > 3 are typically considered to indicate significant effects. We caution that the regression analysis is intended only to be indicative of the trends in FAB observed in the results of the simulation study, and that the regression should probably also include some interaction terms to more adequately capture the variability in FAB; we choose not to do this for simplicity of presentation. Negative values of t-statistics in the table indicate effects which reduce the fractional absolute bias of different modelling strategies, and hence indicate more satisfactory strategies. The results in the tables quantify

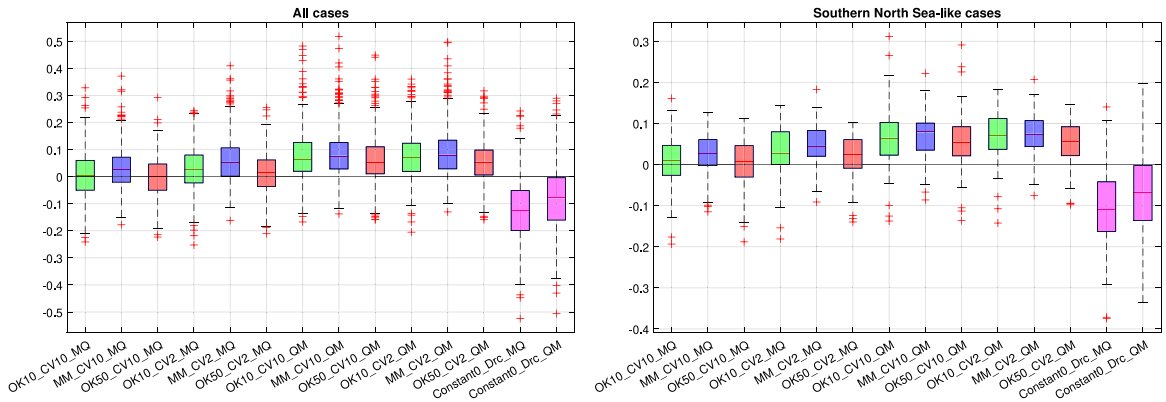

Fig. 8. Fractional bias of return values aggregated over all cases, corresponding to a return period of 1000× the period of the data, for directional sector $[-22.5, 22.5)°$, for northern North Sea (left) and southern North Sea (right). For other details, see Fig. 3.

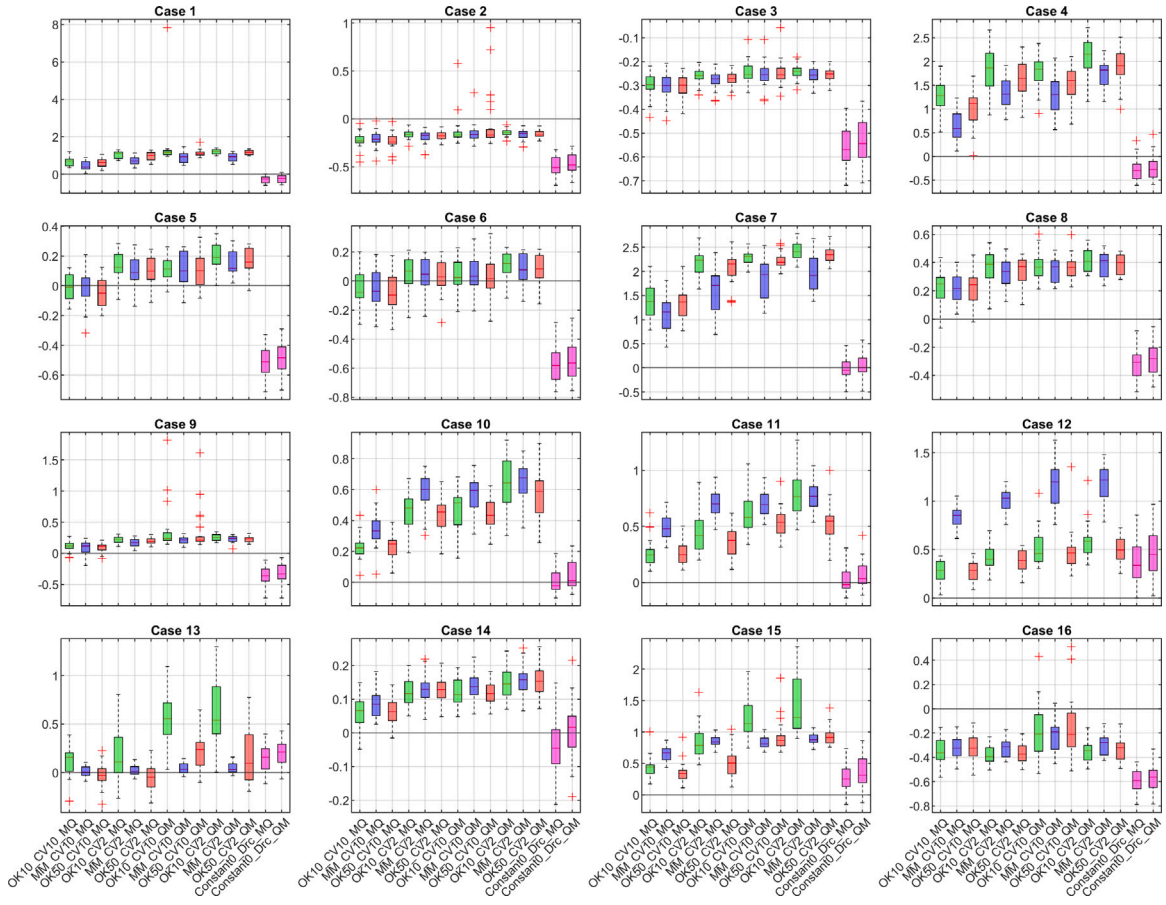

Fig. 9. Fractional bias of return values for each of Cases 1–16, corresponding to a return period of 1000× the period of the data, for directional sector $[247.5, 292.5)°$. For other details, see Fig. 3.

the features noted from inspection of figures in Section 5. Over the full set of case studies, for the omni-directional regression in column 2 of Table 1, we see large effects due to choice of case study; the poor performance of models for Cases 10–12 is clear in particular. We also see a large detrimental effect due to cross-validation strategy CV2 relative to CV10. Constant models with zero threshold (`Constant0*`) also yield poor performance; however with high threshold, the `Constant0.9*` model is competitive with `OK10_CV10_MQ`. For the four directional octants considered, columns 3–6 of the table again indicate two main features: the consistent poor performance of the `Constant0*` strategy, and the large variability in model performance across case studies, especially for directional octants centred on 90° and 270°. Table SM2 provides corresponding results for the southern North Sea - like cases. For estimation of return values using `Constant*` models, note that the generally non-stationary rate of occurrence of events is assumed known; only the conditional GP tail function is assumed not to vary directionally. We do not claim that assuming a stationary GP model on the full covariate domain $\mathcal{D}$ to make inferences for specific directional sectors is necessarily optimal. For this purpose, perhaps estimation of piecewise constant covariate model on $\mathcal{D}$ (e.g. Towe et al., 2024) would be preferable; however, as threshold level increases, decreasing sample size and parsimony typically dictates that the optimal piecewise constant covariate model converges to the stationary GP model.

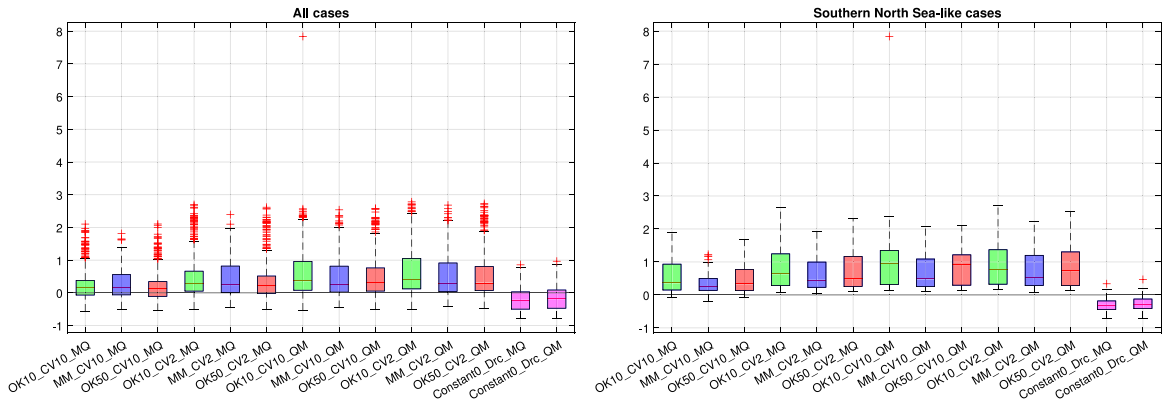

Fig. 10. Fractional bias of return values aggregated over all cases, corresponding to a return period of 1000× the period of the data, for directional sector [247.5, 292.5)°, for northern North Sea (left) and southern North Sea (right). For other details, see Fig. 3.

Table 1
t-statistics of regression coefficients for regression of fractional absolute bias (on log scale) on modelling strategy and case for all 16 case studies, relative to strategy `OK10_CV10_MQ` on Case 1 as reference. When the absolute value of a *t*-statistic is large (>3 for sample size ≫ 20) there is strong evidence that the corresponding value of regression parameter b_{jk} (Eq. (23)) is non-zero.

t-statistic	[0, 360)°	[−22.5, 22.5)°	[67.5, 112.5)°	[157.5, 202.5)°	[247.5, 292.5)°
`QM`	−8.3	2.5	−14.7	3.6	10.1
`CV2`	7.6	0.1	12.3	0.6	8.6
`MM`	2.7	−0.2	5.6	0.3	−7.7
`OK50`	1.1	−3.6	1.7	−4.1	4.3
`Constant0`	6.5	28.8	6.5	28.5	18.6
`Constant0.9`	−0.2	–	–	–	–
Case2	−0.1	−1.5	−0.7	−0.6	−8.4
Case3	7.5	−2.6	−0.5	−0.7	−0.1
Case4	2.5	5.7	4.9	7.7	−1.5
Case5	4.7	6.5	8.1	4.6	−14.4
Case6	1.7	7.2	5.6	6.3	−13.9
Case7	−1.6	−4.1	0.7	−4.3	6.8
Case8	−1.5	−0.9	0.5	−0.3	−10.1
Case9	−1.0	−2.3	−0.2	−1.1	−12.6
Case10	17.3	0.9	20.8	3.4	−14.0
Case11	18.4	2.3	22.1	5.1	−14.5
Case12	20.7	2.6	24.5	5.4	−14.6
Case13	11.2	7.5	2.7	5.8	−15.7
Case14	0.2	1.6	7.9	−0.5	−15.6
Case15	10.2	6.8	13.3	4.7	−8.5
Case16	5.2	8.6	6.6	9.5	1.9

More generally, results from our numerical study reveal trends that we might have anticipated beforehand. For example, compared to using multiple 2-group cross-validation, employing 10-fold cross-validation for roughness parameter estimation tends to produce models with higher variability on the covariate domain, and hence higher return value estimates. Further, return value estimates obtained using the quantile mean (QM) estimator tend to be larger than those obtained using the mean quantile (MQ) estimator (see e.g. Jonathan et al., 2021).

The choice of model parameterisation appears to have little effect on the quality of return value inference, except for cases where it is known that the upper end point ζ can reasonably be expected to be constant on the covariate domain. In this case, there is evidence to favour the mean-max parameterisation. Further, the choice of roughness penalty ratio (see Section 2.2) also appears to have little effect.

In terms of estimation of omni-directional return values, inference using stationary models with an appropriately-chosen high threshold is often competitive in terms of bias if not variance; however this is not always the case. However, we caution that the choice of 16 cases examined in this work may not be representative of the characteristics of tails of distributions of metocean variables with covariate of interest to the practitioner (say, for a given location). Nevertheless, the extent of bias and variance in return values observed is hopefully indicative of what might be anticipated in practice.

For estimation of return values for directional octants, results using non-stationary models mimic those for omni-directional inference. For directional octants which are reasonably populated in the sample, non-stationary models estimates often perform well; estimates from stationary models are generally lower and more variable. Low rate of occurrence of events in the octant of interest is unsurprisingly a key indicator of highly variable performance from all modelling strategies; fractional biases of around +0.3 are typical for non-stationary models, compared to around −0.2 for stationary models. Given the rather pathological behaviour imposed around 270° for some of the cases considered, it is likely that the relative performance of different modelling strategies observed for directional octants centred at 0°, 90° and 180° might be more reflective of the relative behaviour of non-stationary and stationary inference we might expect in practice.

In summary, results from the current numerical study broadly representative of North Sea conditions for significant wave height with direction, indicate that (i) multiple two-group cross-validation yields lower estimates that ten-group cross-validation (leading to negative bias on average, for the case studies considered), (ii) the quantile of the bootstrap predictive estimator yields larger values than the mean over bootstraps of the quantile estimate (leading to reduced omni-directional bias for the case studies considered). Further, (iii) the use of stationary models for non-stationary tails is only reasonable when a high extreme value threshold is set for the stationary analysis. However, (iv) the relative performance of different modelling strategies is very sensitive to the specific characteristics of the case study. The study suggests that, except for replacing the MQ return value estimator with QM, the default strategy `OK10_CV10_MQ` used historically by the authors is generally reasonable.

CRediT authorship contribution statement

Stan Tendijck: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **David Randell:** Project administration, Funding acquisition, Conceptualization. **Graham Feld:** Resources, Investigation, Conceptualization. **Philip Jonathan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization.

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

Acknowledgements

We would like to thank Zak Varty, Paula Cordero-Encinar (both Imperial College, London) and Ross Towe (Shell, London) for discussions.

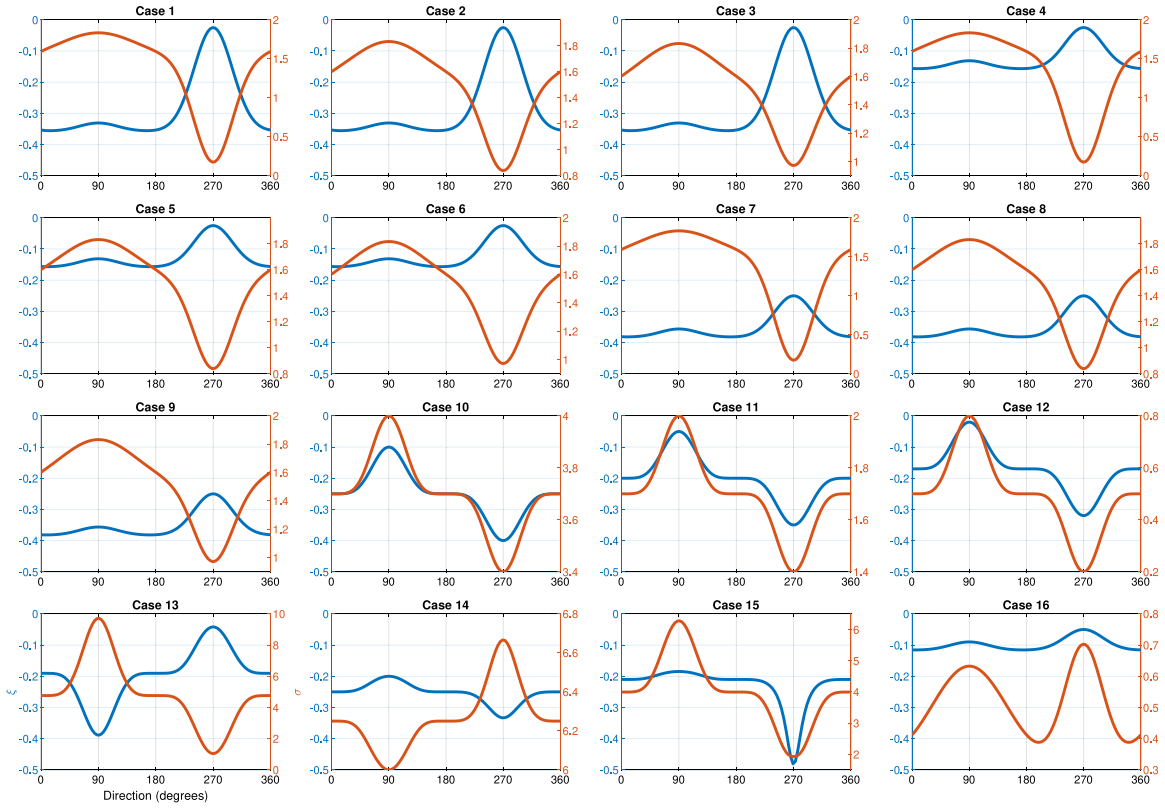

Fig. A.1. Variation of the generalised Pareto shape parameter ξ (blue; left hand y-axis) and scale parameter σ (red; right hand y-axis) with direction for each of the 16 case studies introduced in Section 5.1 of the main text.

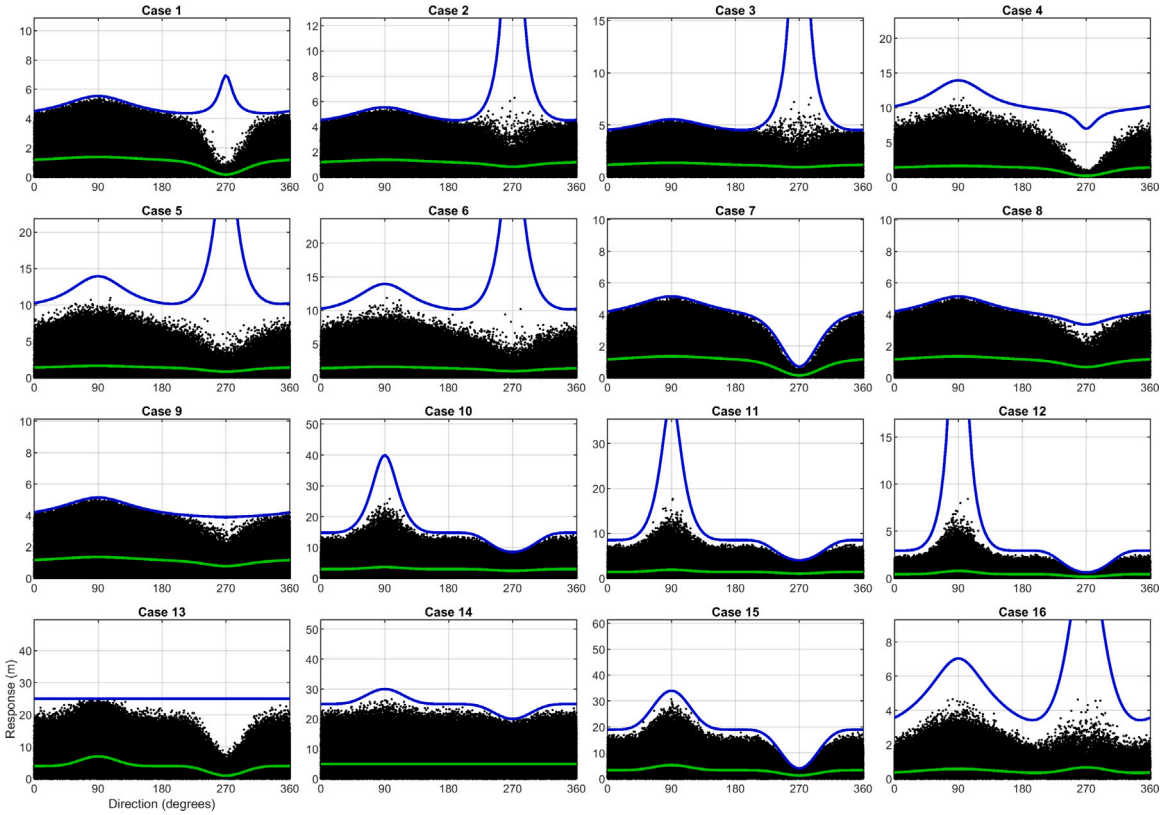

Fig. A.2. Illustrative samples (black dots) generated from each of Cases 1–16, for a response as a function of direction, for samples of size 10^6. Also shown per panel is the corresponding conditional distributional mean μ (green) and upper end point ζ (blue). Notice that the upper end point is approached more rapidly for some combinations of cases and directions. Compare with Fig. 2 of Section 5.1 of the main text.

Appendix A. Illustrations of the 16 case studies considered in the numerical study

A.1. Standard generalised pareto parameter variation for the 16 case studies

Fig. A.1 shows the variation of the generalised Pareto shape parameter ξ (blue) and scale parameter σ (orange) with direction for each of the 16 case studies introduced in Section 5 and illustrated in Figs. 2 and A.2.

A.2. Sample plots for sample size 10^6, with mean-max generalised pareto parameter variation for the 16 case studies

See Fig. A.2.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.oceaneng.2024.119247.

References

Aghbalou, A., Bertail, P., Portier, F., Sabourin, A., 2023. Cross-validation for extreme value analysis. arXiv:2202.00488.

Bates, S., Hastie, T., Tibshirani, R., 2023. Cross-validation: what does it estimate and how well does it do it? J. Am. Statist. Soc. 119, 1434–1445.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, Berlin.

Chavez-Demoulin, V., Davison, A.C., 2005. Generalized additive modelling of sample extremes. J. Roy. Statist. Soc. Series C: Appl. Stat. 54, 207–222.

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

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

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

Feld, G., Randell, D., Jonathan, P., 2019. On the estimation and application of directional design criteria. In: Proc. 38th Int. Conf. on Ocean, Offshore & Arctic Engineering. Scotland.

Forristall, G.Z., 2004. On the use of directional wave criteria. J. Waterw. Port Coast. Ocean Eng. 130, 272–275.

Gibson, R., 2020. Extreme Environmental Loading of Fixed Offshore Structures: Summary Report, Component 2. https://www.hse.gov.uk/offshore/assets/docs/summary-report-component2.pdf.

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.

Haver, S., 1985. Wave climate off northern Norway. Appl. Ocean Res. 7, 85–92.

Heffernan, J.E., Tawn, J.A., 2004. A conditional approach for multivariate extreme values. J. Roy. Statist. Soc. B 66, 497–546.

Jonathan, P., Randell, D., Wadsworth, J., Tawn, J., 2021. Uncertainties in return values from extreme value analysis of peaks over threshold using the generalised Pareto distribution. Ocean Eng. 220, 107725.

Jones, M.J., Hansen, H.F., Zeeberg, A.R., Randell, D., Jonathan, P., 2018. Uncertainty quantification in estimation of ocean environmental return values. Coast. Eng. 141, 36–51.

Jones, M.J., 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.

Joseph, V.R., 2022. Optimal ratio for data splitting. Stat. Anal. Data Min. 15, 531–538.

Lopez, E., Etxebarria-Elezgarai, J., Amigo, J.M., Seifert, A., 2023. The importance of choosing a proper validation strategy in predictive models. A tutorial with real examples. Anal. Chim. Acta 1275, 341532.

Mackay, E.B.L., Challenor, P.G., Bahaj, A.S., 2010. On the use of discrete seasonal and directional models for the estimation of extreme wave conditions. Ocean Eng. 37, 425–442.

Mackay, E., Jonathan, P., 2020. Assessment of return value estimates from stationary and non-stationary extreme value models. Ocean Eng. 207, 107406.

Mackay, E., Murphy-Barltrop, C.J.R., Jonathan, P., 2025. The SPAR model: a new paradigm for multivariate extremes. Application to joint distributions of metocean variables. J. Offshore Mech. Arct. Eng. 147, 011205.

Murphy, C., Tawn, J.A., Varty, Z., 2024. Automated threshold selection and associated inference uncertainty for univariate extremes. arXiv preprint arxiv:2310.17999.

Murphy-Barltrop, C.J.R., Mackay, E., Jonathan, P., 2024. Inference for bivariate extremes via a semi-parametric angular-radial model. arXix preprint arxiv:2401.07259. Accepted by Extremes in July 2024.

Northrop, P., Attalides, N., Jonathan, P., 2017. Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. J. Roy. Statist. Soc. C 66, 93–120.

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, 442–450.

Reistad, M., Breivik, O., Haakenstad, H., Aarnes, O.J., Furevik, B.R., Bidlot, J.-R., 2011. A high-resolution hindcast of wind and waves for the North Sea, the Norwegian Sea, and the Barents Sea. J. Geophys. Res. 116, 1–18.

Richards, J., Huser, R., 2024. Regression modelling of spatiotemporal extreme U.S. wildfires via partially-interpretable neural networks. arXiv preprint arxiv:2208.07581.

Risk, C., James, P.M.A., 2022. Optimal cross-validation strategies for selection of spatial interpolation models for the Canadian forest fire weather index system. Earth Space Sci. 9, e2021EA002019.

Ross, E., Randell, D., Ewans, K., Feld, G., Jonathan, P., 2017. Efficient estimation of return value distributions from non-stationary marginal extreme value models using Bayesian inference. Ocean Eng. 142, 315–328.

Scarrott, C., MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty quantification. Revstat 10, 33–60.

Serinaldi, F., 2015. Dismissing return periods!. Stoch. Environ. Res. Risk A. 29, 1179–1189.

Southworth, H., Heffernan, J.E., Metcalfe, P.D., 2024. texmex: statistical modelling of extreme values. https://cran.r-project.org/package=texmex.

Speers, M., Randell, D., Tawn, J.A., Jonathan, P., 2024. Estimating metocean environments associated with extreme structural response. Ocean Eng. 311, 118754.

Standard Norge, 2022. Shall NORSOK N-0031 and NORSOK N-0062 be updated as a result of findings in LOADS JIP? Conclusions from the evaluation committee. https://standard.no/globalassets/fagomrader-sektorer/petroleum/loads-jip-and-norsok-n_003.pdf.

Swan, C., 2020. Extreme Environmental Loading of Fixed Offshore Structures: Summary Report, Component 1. https://www.hse.gov.uk/offshore/assets/docs/summary-report-component1.pdf.

Tancredi, A., Anderson, C., O'Hagan, A., 2006. Accounting for threshold uncertainty in extreme value estimation. Extremes 9, 87–106.

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

Towe, R., Ross, E., Randell, D., Jonathan, P., 2024. covXtreme: MATLAB software for non-stationary penalised piecewise constant marginal and conditional extreme value models. Environ. Model. Softw. 177, 106035.

Towe, R., Zanini, E., Randell, D., Feld, G., Jonathan, P., 2021. Efficient estimation of distributional properties of extreme seas from a hierarchical description applied to calculation of un-manning and other weather-related operational windows. Ocean Eng. 238, 109642.

Vanem, E., Zhu, T., Babanin, A., 2022. Statistical modelling of the ocean environment: a review of recent developments in theory and applications. Mar. Struct. 86, 103297.

Wadsworth, J.L., Tawn, J.A., 2012. Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. J. Roy. Statist. Soc. B 74, 543–567.

Wood, S., 2023. mgcv: mixed GAM computation vehicle with automatic smoothness estimation. https://cran.r-project.org/package=mgcv.

Yates, L.A., Aandahl, Z., Richards, S.A., Brook, B.W., 2023. Cross validation for model selection: A review with examples from ecology. Ecol. Monogr. 93, e1557.

Youngman, B., 2022. evgam: generalised additive extreme value models. https://cran.r-project.org/package=evgam.

Zanini, E., Eastoe, E., Jones, M.J., Randell, D., Jonathan, P., 2020. Covariate representations for non-stationary extremes. Environmetrics 31, e2624.