

# A penalised piecewise-linear model for non-stationary extreme value analysis of peaks over threshold

Anna Maria Barlow<sup>a</sup>, Ed Mackay<sup>b,\*</sup>, Emma Eastoe<sup>a</sup>, Philip Jonathan<sup>a,c</sup>

<sup>a</sup>Department of Mathematics and Statistics, Lancaster University LA1 4YF, United Kingdom.

<sup>b</sup>College of Engineering, Mathematics and Physical Sciences, University of Exeter, Penryn TR10 9FE, United Kingdom.

<sup>c</sup>Shell Research Limited, London SE1 7NA, United Kingdom.

---

## Abstract

Metocean extremes often vary systematically with covariates such as direction and season. In this work, we present non-stationary models for the size and rate of occurrence of peaks over threshold of metocean variables with respect to one- or two-dimensional covariates. The variation of model parameters with covariate is described using a piecewise-linear function in one or two dimensions, defined with respect to pre-specified node locations on the covariate domain. Parameter roughness is regulated to provide optimal predictive performance, assessed using cross-validation, within a penalised likelihood framework for inference. Parameter uncertainty is quantified using bootstrap resampling. The models are used to estimate extremes of storm-peak significant wave height with respect to direction and season for a site in the northern North Sea. A covariate representation based on a triangulation of the direction-season domain with six nodes gives good predictive performance. The penalised piecewise-linear framework provides a flexible representation of covariate effects at reasonable computational cost.

*Keywords:* Extreme, Non-stationary, Covariate, Penalised likelihood, Significant wave height

---

## 1. Introduction

### 1.1. Background

Metocean variables such as significant wave height, wind speed and current speed usually show dependence on covariates such as direction, season and water depth (e.g. Randell et al. 2015), as well as having complex inter-relationships: e.g. the relationship between wind speed and significant wave height (e.g. Towe et al. 2017; Mackay and Jonathan 2020b), or significant wave height and spectral peak period (e.g. Haver 1987; Jonathan et al. 2010), or significant wave height and surge (e.g. Ross et al. 2018). Good characterisation of the metocean environment and its extremes demands statistical models which accommodate dependencies of variables on covariates, and dependencies between extremes of variables, in a careful manner.

In this work, we consider a univariate response  $Y$  (e.g. storm-peak significant wave height, or wind speed) which varies systematically with one or more covariates  $\mathbf{X} = (X_1, X_2, \dots, X_D)$ , (e.g. storm direction, season, ...), such that the tail of the conditional distribution  $Y|\mathbf{X} = \mathbf{x}$  follows a known parametric form. In a peaks over threshold model, we expect  $Y|(\mathbf{X} = \mathbf{x}, Y > u(\mathbf{x}))$  to follow a generalised Pareto (GP) distribution, provided that the extreme value threshold  $u(\mathbf{x})$  is sufficiently large, with shape, scale and threshold parameters  $\xi(\mathbf{x})$ ,  $\sigma(\mathbf{x}) > 0$  and  $u(\mathbf{x})$  varying systematically with  $\mathbf{x}$ , and density

$$f_{\text{GP}}(y|u(\mathbf{x}), \sigma(\mathbf{x}), \xi(\mathbf{x})) = \begin{cases} \frac{1}{\sigma(\mathbf{x})} \left[ 1 + \xi(\mathbf{x}) \left( \frac{y - u(\mathbf{x})}{\sigma(\mathbf{x})} \right) \right]^{-1-1/\xi(\mathbf{x})}_+, & \xi(\mathbf{x}) \neq 0, \\ \frac{1}{\sigma(\mathbf{x})} \exp \left( - \left( \frac{y - u(\mathbf{x})}{\sigma(\mathbf{x})} \right) \right), & \xi(\mathbf{x}) = 0, \end{cases} \quad (1)$$

where  $[\cdot]_+ = \max\{\cdot, 0\}$ . Using a similar approach, we can also model the rate of occurrence of events, dependent on covariates,  $\mathbf{X}$ , assuming a Poisson model for counts of exceedances.

In more applied work, the effects of covariates have been either ignored historically, or accommodated by performing independent extreme value inferences for subsets of the data corresponding to different parts of the covariate domain.

---

\*Corresponding author e.mackay@exeter.ac.uk

Partition of the sample into smaller samples on sub-domains can result in increased and unquantified bias and uncertainties if the sample on each sub-domain is treated independently, without taking into account the systematic variation of the response with the covariate (e.g. Mackay et al. 2010, Mackay and Jonathan 2020a). A better approach is to allow the parameter of the extreme value model to vary with covariate. Numerous such covariate models have been developed for extreme value analysis, including Carter and Challenor (1981), Davison and Smith (1990), Coles and Walshaw (1994), Robinson and Tawn (1997), Scotto and Guedes-Soares (2000), Anderson et al. (2001), Chavez-Demoulin and Davison (2005), Fawcett and Walshaw (2007), Mendez et al. (2008), Northrop and Jonathan (2011), Randell et al. (2015), Sigauke and Bere (2017) and Eastoe (2019). Zanini et al. (2020) provides a general description of semi-parametric covariate representations for extreme value analysis of peaks over threshold, including penalised B-splines, Bayesian adaptive regression splines and Voronoi partitions. These all take the form of a linear combination of basis functions on the covariate domain, but differ (a) in the way that basis functions are constructed and modified, and (b) by additional penalisation of the variability (e.g. variance or roughness) of basis coefficients, for a given sample, to improve inference. Inference using these representations is generally computationally complex, requiring Bayesian inference for all but the simplest representations.

In marked contrast, Ross et al. (2018) use a penalised piecewise-constant (PPC) model. The motivation for this approach is the provision of a simple, pragmatic but useful representation for covariate effects, both in terms of parameter estimation and uncertainty quantification. In the PPC model, the multi-dimensional covariate domain  $\mathcal{D}$  is partitioned into sub-domains or ‘‘bins’’  $B_1, B_2, \dots, B_M$ . In each bin  $B_k$ , the distribution of  $Y|(\mathbf{X} = \mathbf{x}, Y > u_k)$ ,  $\mathbf{x} \in B_k$  is assumed to be stationary and follow a GP distribution with density given in Equation 1, with pre-specified extreme value threshold  $u(\mathbf{x}) = u_k$  and scale  $\sigma(\mathbf{x}) = \sigma_k$ . The shape parameter  $\xi$  is assumed common to all bins. The objective of the PPC inference is simultaneous estimation of the scale parameters and common shape parameter for all  $M$  bins. This is achieved by maximisation of predictive likelihood for a hold-out sample of data, further penalised by the roughness of the set  $\sigma_1, \sigma_2, \dots, \sigma_M$ . The roughness penalty is selected to maximise the predictive likelihood, evaluated using a cross-validation scheme. Bootstrap resampling is used to quantify uncertainties of estimated parameters and other predictions. The PPC model has been used in a number of applications (e.g. Ross et al. (2020), Mackay and Jonathan (2020a), Guerrero et al. 2021).

In Mackay and Jonathan (2020a), it was observed that the piecewise-constant assumption did not provide the most natural, parsimonious representation for parameter variation on the covariate domain for some applications of interest, and that a piecewise-linear covariate parameterisation may be more appropriate. These observations motivated initial research into a penalised piecewise-linear (PPL) model outlined in Mackay and Jonathan (2020b) (using common shape parameter and one-dimensional covariates) and subsequently the full development for arbitrary covariate domains reported here. There is a long history of piecewise-linear or segmentation regression, motivated by the assumption of local smooth (and in particular linear) variation of model parameters with covariates (e.g. Cleveland 1979, Yang et al. 2016). Like PPC, the PPL model is intended to provide a useful compromise between simplicity and flexibility. Like PPC, PPL is simpler to implement and compute than extreme value models incorporating general basis function representations for covariates discussed by Zanini et al. (2020). At the same time, relative to PPC, PPL provides a more flexible and physically-realistic covariate representation at the cost of increased computational complexity.

## 1.2. Motivating application

We apply the non-stationary PPL extreme value model to data from the NORA10 hindcast of Reistad et al. (2011), for a location in the northern North Sea. The hindcast provides time-series of significant wave height, (dominant) wave direction and season (defined as day of the year, for a standardised year consisting of 360 days) for three hour sea-states for the period 1957-2010. Throughout, ‘‘direction’’ refers to the direction from which a storm travels expressed in degrees clockwise with respect to north. Since significant wave height is serially-correlated between sea states, storm-peak significant wave height characteristics are isolated from the hindcast time-series using the procedure described in Ewans and Jonathan (2008). Contiguous intervals of significant wave height above a low peak-picking threshold are identified, each interval now assumed to correspond to a storm event. The peak-picking threshold corresponds to a directional-seasonal quantile of significant wave height with specified non-exceedance probability, estimated using quantile regression. The maximum of significant wave height during the storm interval is taken as the storm-peak significant wave height (henceforth  $H_s$ ). The values of directional and seasonal covariates at the time of storm-peak significant wave height are referred to as storm-peak values of those variables. The resulting storm-peak sample consists of 5388 values of  $H_s$ . The seasonal and directional characteristics of the sample are obvious from inspection of the figures in Sections 4 and 5. The directional influence of the land shadow of Norway is clear on the interval ( $45^\circ, 210^\circ$ ). A more gradual seasonal variation is also present. We expect these sources of non-stationarity to be adequately characterised in the estimated PPL model. Data from the same location have been considered in Randell et al. (2016) and Konzen et al. (2021).

### 1.3. Objectives and layout

The objectives of the current work are to extend the PPL model of Mackay and Jonathan (2020b) to incorporate non-stationary shape parameter and multi-dimensional covariates, and to provide software for applications in one and two dimensions<sup>1</sup>. Further, we seek to demonstrate the usefulness of the PPL model in application to non-stationary extreme value analysis of storm-peak significant wave height with directional-seasonal covariates for a location in the northern North Sea. The layout of the paper is as follows. Section 2 describes the PPL model formulation in detail, and Section 3 the approach to inference. Sections 4 and 5 discuss the North Sea application for covariates in one and two dimensions. Finally, Section 6 provides discussion and conclusions.

## 2. Model and preliminaries

This section introduces the extreme value model with piecewise-linear covariate representation. It also provides a summary of the approaches used to estimate the density  $f_{\mathbf{x}}$  of covariates, and the non-stationary extreme value threshold  $u$ . A key motivation for the modelling approach is that covariate effects are more easily identified, and are more influential, in some aspects of the analysis than others (e.g. Anderson et al. 2001). In particular, covariate dependence of  $f_{\mathbf{x}}$  and  $u$  is relatively-easily identified. It is therefore reasonable to use more flexible non-stationary non-parametric methods to capture these effects. Covariate dependence of GP scale  $\sigma$  is also identifiable in general, but typically not to the same extent as that of  $f_{\mathbf{x}}$  and  $u$ , suggesting a simpler covariate representation for  $\sigma$  may be appropriate. Quantifying the covariate dependence of GP shape  $\xi$  is most problematic; we anticipate that the piecewise-linear representation, or even a stationary estimate, will be suitable in general.

### 2.1. Formulation

The joint density of covariates  $\mathbf{X}$  and response  $Y$  can be written as  $f_{\mathbf{X},Y}(\mathbf{x},y) = f_{\mathbf{X}}(\mathbf{x})f_{Y|\mathbf{X}}(y|\mathbf{x})$ , where  $f_{\mathbf{X}}(\mathbf{x})$  is the density of covariates. We assume that the tail of  $f_{Y|\mathbf{X}}(y|\mathbf{x})$  converges to a GP distribution. Our model for exceedances of high non-stationary threshold  $u(\mathbf{x})$ , can therefore be written  $f_{Y|\mathbf{X}}(y|\mathbf{x}) = \zeta(\mathbf{x})f_{\text{GP}}(y|u(\mathbf{x}), \sigma(\mathbf{x}), \xi(\mathbf{x}))$ ,  $y > u(\mathbf{x})$ , where  $\zeta(\mathbf{x}) = \Pr(Y > u(\mathbf{x}) | \mathbf{X} = \mathbf{x})$  is the threshold exceedance probability, and  $f_{\text{GP}}$  is the density of the GP distribution with shape, scale and threshold parameters  $\xi(\mathbf{x})$ ,  $\sigma(\mathbf{x}) > 0$  and  $u(\mathbf{x})$ , defined in Equation 1. Hence, for  $y > u(\mathbf{x})$ , our model for the joint density is

$$f_{\mathbf{X},Y}(\mathbf{x},y) = f_{\mathbf{X}}(\mathbf{x})\zeta(\mathbf{x})f_{\text{GP}}(y|u(\mathbf{x}), \sigma(\mathbf{x}), \xi(\mathbf{x})), \quad y > u(\mathbf{x}). \quad (2)$$

Inference requires estimation of the covariate density  $f_{\mathbf{X}}(\mathbf{x})$ , a non-stationary threshold  $u(\mathbf{x})$  (typically corresponding to a constant values of  $\zeta$ ), and non-stationary GP model with parameters  $\sigma(\mathbf{x})$  and  $\xi(\mathbf{x})$ . For  $y \leq u(\mathbf{x})$ ,  $f_{\mathbf{X},Y}(\mathbf{x},y)$  can be estimated empirically, e.g. using kernel density estimation. Moreover, although this is not a requirement of the model in general, in the current work, we assume that covariates  $\mathbf{X}$  are periodic and therefore never extreme, so that kernel density estimation is also appropriate to estimate  $f_{\mathbf{X}}(\mathbf{x})$ . Threshold  $u(\mathbf{x})$  will be estimated as a local quantile of  $Y$  on the covariate domain. It then remains to estimate the GP model, with parameters  $\sigma(\mathbf{x})$  and  $\xi(\mathbf{x})$  which vary systematically as a function of covariate, taking (penalised) piecewise-linear forms.

### 2.2. Piecewise-linear covariate representation

The piecewise-linear representation for functions  $\sigma(\mathbf{x})$  and  $\xi(\mathbf{x})$  of covariate  $\mathbf{x} \in \mathcal{D}$  is defined as follows. We first locate  $K$  nodes  $\mathbf{n}_k$ ,  $k = 1, 2, \dots, K$  on the covariate domain  $\mathcal{D}$ , with coordinates  $\mathbf{n}_k = (n_{k,1}, \dots, n_{k,D})$ . Next we define a triangulation on  $\mathcal{D}$ , with nodes as vertices, which partitions  $\mathcal{D}$  into  $M$  covariate bins  $B_m$ ,  $m = 1, 2, \dots, M$ , each of which is a  $D$ -simplex. For problems with a single covariate ( $D = 1$ ), each  $B_m$  is a line segment; for  $D = 2, 3$ ,  $B_m$  is a triangle and tetrahedron respectively. GP model parameter values are specified at the vertices of the triangulation only, and are assumed to vary linearly within each bin  $B_m$ . The relationship between  $M$  and  $K$  depends on  $D$ , and the choice of (non-unique) triangulation of nodes. The assumed periodicity of covariates means that some care is needed when interpolating within bins near the boundary of  $\mathcal{D}$ .

The vertices of bin  $B_m$  are indexed using the indices of the nodes. For bin  $B_m$ , define the index vector  $\mathbf{t}_m = (t_{m,1}, \dots, t_{m,D+1})$ , where  $t_{m,j} \in \{1, \dots, K\}$  and  $t_{m,i} \neq t_{m,j}$  for  $i \neq j$ , so that  $\{\mathbf{n}_{t_{m,1}}, \dots, \mathbf{n}_{t_{m,D+1}}\}$  is the set of nodes defining the vertices of  $B_m$ . For an arbitrary function  $\psi : \mathbb{R}^D \rightarrow \mathbb{R}$ , whose values are specified at the nodes, the linear

---

<sup>1</sup>MATLAB software for the PPL model is available at <https://github.com/edmackay/PPL-model>, along with a user-guide and the data underpinning the current study.

interpolant of  $\psi$  for a point  $\mathbf{x} = (x_1, x_2, \dots, x_D) \in B_m$  is given by  $\psi(\mathbf{x}) = \tilde{\mathbf{x}}\boldsymbol{\beta}_m$ , where  $\tilde{\mathbf{x}} = (1, x_1, x_2, \dots, x_D)$ , and  $\boldsymbol{\beta}_m = (\beta_{m,0}, \dots, \beta_{m,D})^T$  is the solution of  $A_m \boldsymbol{\beta}_m = \mathbf{c}_m$ , where

$$A_m = \begin{pmatrix} 1 & \mathbf{n}_{t_m,1} \\ \vdots & \vdots \\ 1 & \mathbf{n}_{t_m,D+1} \end{pmatrix} = \begin{pmatrix} 1 & n_{t_m,1,1} & \dots & n_{t_m,1,D} \\ \vdots & \vdots & & \\ 1 & n_{t_m,D+1,1} & \dots & n_{t_m,D+1,D} \end{pmatrix} \quad (3)$$

and vector  $\mathbf{c}_m = (\psi(\mathbf{n}_{t_m,1}), \dots, \psi(\mathbf{n}_{t_m,D+1}))^T$  stores the current values of  $\psi$  at the nodes  $\mathbf{t}_m$ . Provided that  $D$  is not too large, it is straightforward to calculate  $A_m^{-1}$  for each bin  $B_m$ , and save it in memory for repeated use, making iterative estimation of a piecewise-linear model computationally efficient.

### 2.3. Estimation of covariate density and extreme value threshold

Consider a sample  $\mathcal{S} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$  of  $N$  conditionally-independent observations, to be used to estimate the model in Equation 2. This section describes inference for covariate density  $f_{\mathbf{x}}(\mathbf{x})$  and threshold  $u(\mathbf{x})$ , with penalised piecewise-linear inference for the GP model described in Section 3.

#### Covariate density

The covariate density  $f_{\mathbf{x}}(\mathbf{x})$  is estimated using kernel density (KD) estimation, for points on a regular grid in  $\mathcal{D}$ . At location  $\mathbf{x}$ , the KD estimate is

$$\hat{f}_{\mathbf{x}}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \prod_{d=1}^D \phi\left(\frac{x_d - x_{i,d}}{w}\right) \quad (4)$$

where  $\phi$  is the standard normal density, and common bandwidth  $w$  is user-specified to give a reasonable compromise between smoothness and resolution. We note that many more sophisticated models for non-parametric density estimation are available, but this approach appeared adequate for the application discussed in Sections 4 and 5.

#### Threshold estimation

In the current work, the threshold exceedance probability  $\zeta$  is set prior to analysis, hence threshold estimation reduces to estimation of conditional quantiles  $\{u(\mathbf{x}) : \Pr(Y > u(\mathbf{x}) | \mathbf{X} = \mathbf{x}) = \zeta\}$ . We choose to use a two-step process comprised of (1) local quantile estimation and (2) subsequent quantile smoothing for this purpose, but again note that many alternative approaches would be suitable. The local quantile is again estimated on a regular grid in  $\mathcal{D}$ . For each location  $\mathbf{x}$ , we find the  $C$  nearest observations in  $\mathcal{S}$ , calculate the empirical quantile corresponding to exceedance probability  $\zeta$ , and smooth using a Gaussian kernel. The value of  $C$  and the bandwidth,  $w$ , of the smoothing kernel are modelling choices.

## 3. Estimation of penalised piecewise-linear generalised Pareto model

This section describes maximum roughness-penalised likelihood estimation of GP shape  $\xi(\mathbf{x})$  and scale  $\sigma(\mathbf{x})$ , taking piecewise-linear forms for  $\mathbf{x} \in \mathcal{D}$ . A critical precursor to successful inference is reasonable node placement on  $\mathcal{D}$ , discussed in Section 3.1. Given nodes, the maximum likelihood inference is discussed in Section 3.2. The use of cross-validation for optimal choice of roughness penalties is discussed in Section 3.3.

### 3.1. Node placement and triangulation

We seek to locate nodes on  $\mathcal{D}$  such that the piecewise-linear representation is able to describe variation of GP parameters well. Intuitively, this suggests that nodes be located where the gradient of parameters changes sharply with respect to covariate. This is particularly important when there are only a few nodes. To guide selection of node positions, we calculate initial estimates of  $\sigma(\mathbf{x})$  and  $\xi(\mathbf{x})$  on a regular grid on  $\mathcal{D}$ , in a similar manner to threshold estimation outlined in Section 2.3. For each  $\mathbf{x}$ , we find the  $C$  nearest threshold exceedances of  $u(\mathbf{x})$ , and use these to calculate (locally-stationary) moment estimators  $\hat{\xi}(\mathbf{x}) = (1 - (m(\mathbf{x}))^2/v(\mathbf{x}))/2$  and  $\hat{\sigma}(\mathbf{x}) = m(\mathbf{x})(1 - \hat{\xi}(\mathbf{x}))$ , where  $m(\mathbf{x})$  and  $v(\mathbf{x})$  are the mean and variance of the  $C$  nearest threshold exceedances,  $z_{j(\mathbf{x})} = y_{j(\mathbf{x})} - u(\mathbf{x}_{j(\mathbf{x})})$ ,  $j(\mathbf{x}) = 1, \dots, C$ , where  $j(\mathbf{x})$  indicates the index of the  $j^{\text{th}}$  closest observation to  $\mathbf{x}$ . If  $\hat{\xi}(\mathbf{x}) < 0$  and  $z_{max}(\mathbf{x}) > -\hat{\sigma}(\mathbf{x})/\hat{\xi}(\mathbf{x})$ , where  $z_{max}(\mathbf{x}) = \max\{z_{j(\mathbf{x})} : j(\mathbf{x}) = 1, \dots, C\}$ , then we set  $\hat{\xi}(\mathbf{x}) = -\hat{\sigma}(\mathbf{x})/z_{max}(\mathbf{x})$ , since for a negative shape parameter the

upper end point of the GP distribution is  $-\sigma/\xi$ . If  $C$  is large, then the true values of the parameters may have a large variation over the range of  $\mathbf{x}$  from which the sample is drawn, leading to bias in the estimates. Alternatively, if  $C$  is small, then the estimates will be subject to higher variance. Nevertheless, it is assumed that it is possible to choose  $C$  such that these initial estimates are adequate to identify gross features of  $\sigma(\mathbf{x})$  and  $\xi(\mathbf{x})$  on  $\mathcal{D}$ . Once node locations have been specified, aided by the local estimates of  $\sigma$  and  $\xi$ , we triangulate the covariate domain. For  $D = 1$ , the partition achieved is unique, with bins corresponding to intervals between adjacent nodes. For  $D > 1$ , the triangulation is not unique, and different options are possible as discussed in Section 5; clearly, the choice of triangulation also affects model performance.

### 3.2. Parameter estimation for given roughness coefficients

Estimates of the GP scale and shape parameters at the nodes are found by maximising the sample likelihood, penalised for parameter roughness on  $\mathcal{D}$ . Optimal roughness penalties are estimated by cross-validation, described in Section 3.3.

With parameters  $\theta = (\sigma(\mathbf{n}_1), \dots, \sigma(\mathbf{n}_K), \xi(\mathbf{n}_1), \dots, \xi(\mathbf{n}_K))$  to be estimated, the sample likelihood for  $\mathcal{S}$  under the PPL model is  $\mathcal{L}(\theta | \mathcal{S}) = \prod_{i \in \mathcal{I}} f_{\text{GP}}(y_i | u(\mathbf{x}_i), \sigma(\mathbf{x}_i), \xi(\mathbf{x}_i))$ , where  $\mathcal{I}$  is the index set of exceedances  $\{i : y_i > u(\mathbf{x}_i)\}$ , and for  $\mathbf{x} \in B_m$ , parameters  $\sigma(\mathbf{x})$  and  $\xi(\mathbf{x})$  take piecewise-linear representations parameterised in terms of the corresponding bin nodes. The sample negative log-likelihood is denoted  $\ell(\theta | \mathcal{S}) = -\log(\mathcal{L}(\theta | \mathcal{S}))$ .

Roughness penalisation is used to smooth the variation of parameters on  $\mathcal{D}$  for optimal predictive performance. This is quantified using the gradient of the parameter function in each covariate dimension, in each bin. The partial derivatives of function  $\psi(\mathbf{x})$  for a point  $\mathbf{x} \in B_m$ , are given by the interpolation coefficients,  $\beta_m$ , defined in Section 2.2. That is,  $[\partial\psi(\mathbf{x})/\partial x_d]_{\mathbf{x} \in B_m} = \beta_{m,d}$ . The penalised negative log-likelihood becomes

$$\ell^*(\theta | \boldsymbol{\lambda}, \mathcal{S}) = \ell(\theta | \mathcal{S}) + \sum_{d=1}^D \sum_{m=1}^M \left( \lambda_{\sigma,d} \left| \frac{\partial \sigma(\mathbf{x})}{\partial x_d} \right|_{\mathbf{x} \in B_m} + \lambda_{\xi,d} \left| \frac{\partial \xi(\mathbf{x})}{\partial x_d} \right|_{\mathbf{x} \in B_m} \right) \quad (5)$$

where  $\boldsymbol{\lambda} = (\lambda_{\sigma,1}, \dots, \lambda_{\sigma,D}, \lambda_{\xi,1}, \dots, \lambda_{\xi,D})$  is the vector of roughness penalties. For given  $\boldsymbol{\lambda}$ , constrained non-linear optimisation is used to minimise  $\ell^*(\theta | \boldsymbol{\lambda}, \mathcal{S})$ . In the current work, we use the MATLAB function *fmincon*, under the constraint  $-0.5 \leq \xi(\mathbf{x}) < 0$ , reasonable for environmental variables (with  $\xi = -0.5$  corresponding to linear decrease in GP density with  $y$ , and  $\xi < 0$  ensuring a finite upper bound for the distribution of  $Y$ ). A large value of  $\lambda$  for some parameter-covariate pair imposes a large penalty on the corresponding parameter gradient, and hence drives smoother solutions for that parameter-covariate pair. The starting solution for the optimisation is found using a Voronoi partition of  $\mathcal{D}$ , with bins corresponding to the set of points closest to each node. Independent GP fits, constrained such that  $-0.5 \leq \xi(\mathbf{x}) < 0$ , are then made per Voronoi bin, and the parameter estimates are used as a starting solution for the PPL estimation. In some applications, we may wish to assume that the GP shape parameter does not vary with covariate. In this case the model form can be simplified in the obvious way.

### 3.3. Estimation of optimal roughness coefficients

Section 3.2 describes how the parameter set,  $\theta$ , is estimated, given roughness coefficients  $\boldsymbol{\lambda}$ . The value of  $\boldsymbol{\lambda}$  is selected to maximise the predictive performance of the model using a  $G$ -fold cross-validation scheme. Cross-validation groups  $\mathcal{S}_g$ ,  $g = 1, 2, \dots, G$  are formed by random partition of the sample  $\mathcal{S}$  into  $G$  groups of approximately equal size.

For each of a large number  $L$  of plausible choices for  $\boldsymbol{\lambda}$ , and cross-validation group  $\mathcal{S}_g$ , we estimate the GP model using  $\mathcal{S} \setminus \mathcal{S}_g$ , finding parameter vector  $\theta_g(\boldsymbol{\lambda})$  that minimises the penalised negative log-likelihood  $\ell^*(\theta | \boldsymbol{\lambda}, \mathcal{S} \setminus \mathcal{S}_g)$ . We then calculate the corresponding unpenalised negative log predictive likelihood  $\ell(\theta_g(\boldsymbol{\lambda}) | \mathcal{S}_g)$  for the excluded group  $\mathcal{S}_g$ , using optimal parameter vector  $\theta_g(\boldsymbol{\lambda})$ . The predictive performance of the model for penalty vector  $\boldsymbol{\lambda}$  is accumulated as  $\mathcal{P}(\boldsymbol{\lambda}) = \sum_{g=1}^G \ell(\theta_g(\boldsymbol{\lambda}) | \mathcal{S}_g)$ . The optimal value of  $\boldsymbol{\lambda}$  is that which minimises  $\mathcal{P}(\boldsymbol{\lambda})$ .

Since random partitioning into groups leads to uncertainty in the predictive performance  $\mathcal{P}(\boldsymbol{\lambda})$ , the procedure is repeated  $R$  times, with different replicates of random partitions of  $\mathcal{S}$ . The overall predictive performance of the model is quantified using  $\bar{\mathcal{P}}(\boldsymbol{\lambda}) = \frac{1}{R} \sum_{r=1}^R \mathcal{P}_r(\boldsymbol{\lambda})$ , where  $\mathcal{P}_r(\boldsymbol{\lambda})$  is the predictive performance estimated for replicate  $r$ . The uncertainty in  $\bar{\mathcal{P}}(\boldsymbol{\lambda})$  is calculated using a jackknife procedure. In outline,  $R$  different jackknife estimates of  $\bar{\mathcal{P}}(\boldsymbol{\lambda})$  are calculated, using all  $R$  estimates  $\mathcal{P}_{r'}(\boldsymbol{\lambda})$ ,  $r' = 1, 2, \dots, R$  except for the  $r^{\text{th}}$  estimate,  $r = 1, 2, \dots, R$ . The uncertainty  $U(\bar{\mathcal{P}}(\boldsymbol{\lambda}))$  in the mean predictive performance  $\bar{\mathcal{P}}(\boldsymbol{\lambda})$  is quantified as the range of the  $R$  jackknife estimates. The optimal choice of  $\boldsymbol{\lambda}$  can be adapted to accommodate uncertainty due to the cross-validation strategy. In practice, we might select  $\boldsymbol{\lambda}^*$  such that  $\bar{\mathcal{P}}(\boldsymbol{\lambda}^*) \leq \bar{\mathcal{P}}(\boldsymbol{\lambda}^\circ) + U(\bar{\mathcal{P}}(\boldsymbol{\lambda}^\circ))$ , where  $\boldsymbol{\lambda}^\circ$  is the value that minimises  $\bar{\mathcal{P}}$ , and the values of the components of  $\boldsymbol{\lambda}^*$  are larger than those of  $\boldsymbol{\lambda}^\circ$ . Hence, the estimated model using  $\boldsymbol{\lambda}^*$  will be stiffer and therefore more parsimonious than that using  $\boldsymbol{\lambda}^\circ$ , whilst constraining the degradation in predictive performance to an acceptable level.

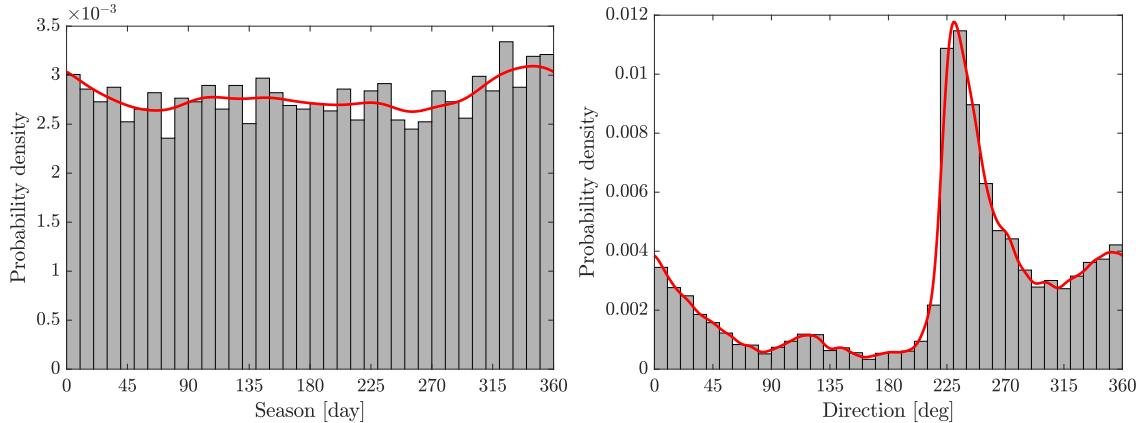


Figure 1: Kernel density estimates (red) of covariate density for 1-D seasonal and directional analyses, with empirical histograms (grey) for comparison. Kernel bandwidth is 15 days for seasonal analysis and  $5^\circ$  for directional analysis.

### 3.4. Computational complexity of predictive inference

The inference procedure involves iterative accumulation of predictive performance measure  $\bar{P}$  over  $L$  roughness coefficients,  $G$  cross-validation groups, and  $R$  repetitions of the cross-validation partition. For a  $D$ -dimensional covariate domain  $\mathcal{D}$ , exploring a reasonable number of values for each of the  $2 \times D$  components of the vector  $\lambda$  is necessary for reliable performance; searching over  $S$  values for  $2D$  roughness coefficients requires  $L = S^{2D}$  iterations. Computational complexity for the full analysis therefore scales as  $G \times R \times S^{2D}$ . Assuming that GP shape is stationary on  $\mathcal{D}$  reduces computational complexity to  $G \times R \times S^D$ .

In light of this, to restrict computational time to reasonable values for physically-reasonable model complexity, we restrict our interest to the following archetypes. Case A is the computationally-simplest analysis undertaken, for which the GP shape parameter is assumed stationary on  $\mathcal{D}$ , and a common roughness coefficient is assumed for GP scale with respect to each covariate dimension (i.e.  $\lambda_{\sigma,1} = \lambda_{\sigma,2} := \lambda_\sigma$  for a  $D = 2$  directional-seasonal problem). Case B assumes that both GP shape and scale are non-stationary on  $\mathcal{D}$ , with common roughness coefficients for each with respect to each covariate dimension (i.e.  $\lambda_{\sigma,1} = \lambda_{\sigma,2} := \lambda_\sigma$ ,  $\lambda_{\xi,1} = \lambda_{\xi,2} := \lambda_\xi$  for a  $D = 2$  directional-seasonal problem). Finally, Case C again assumes that GP shape is stationary on  $\mathcal{D}$ , but estimates separate roughness coefficients for GP scale and each covariate dimension (i.e.  $\lambda_{\sigma,1}$  and  $\lambda_{\sigma,2}$  for a  $D = 2$  directional-seasonal problem). Note that for  $D = 1$ , Cases A and C coincide.

After some experimentation for the North Sea application introduced in Section 1.2, and detailed in Sections 4 and 5 below, we choose to adopt  $G = 5$ ,  $R = 5$  and  $S = 10$  as reasonable values for the numbers of cross-validation groups, partition repetitions and (component) roughness coefficients respectively. The  $S$  values of component roughness coefficients considered are equal logarithmic increments on the interval  $[10^{-1}, 10^5]$ , ensuring reasonable coverage from very flexible to very stiff covariate representations.

## 4. Application with one-dimensional covariate

Here we discuss the estimation of one-dimensional directional and seasonal PPL extreme value models for the data described in Section 1.2, following the procedure outlined in Section 2.1. Two-dimensional directional-seasonal models are presented in Section 5.

Kernel density estimates for the covariate density  $f_{\mathbf{x}}(\mathbf{x})$  are shown in Figure 1 for both seasonal and directional cases. After some experimentation, we chose to set the kernel density bandwidth to  $w = 15$  days, to achieve a reasonable seasonal model, compared to the empirical histogram. For the directional model, a bandwidth of  $w = 5^\circ$  was chosen to better resolve variation around  $225^\circ$ .

To estimate the non-stationary extreme value threshold  $u$ , the threshold exceedance probability  $\zeta$  was set to 0.3 (0.2) for seasonal (or directional) analyses respectively. Inspection of model output confirmed that these choices of  $\zeta$  ensure relatively large samples are retained for GP inference, without incurring too much apparent bias. Sample sizes of  $N = 1584$  (or 1077) threshold exceedances were retained. Estimates of the corresponding seasonal (or directional) quantiles,  $u$ , utilise  $C = 100$  (or 50) nearest observations of threshold exceedance, smoothed using a Gaussian kernel with a bandwidth of  $w = 15$  days (or  $5^\circ$ ). The number of nearest observations,  $C$ , and the smoothing

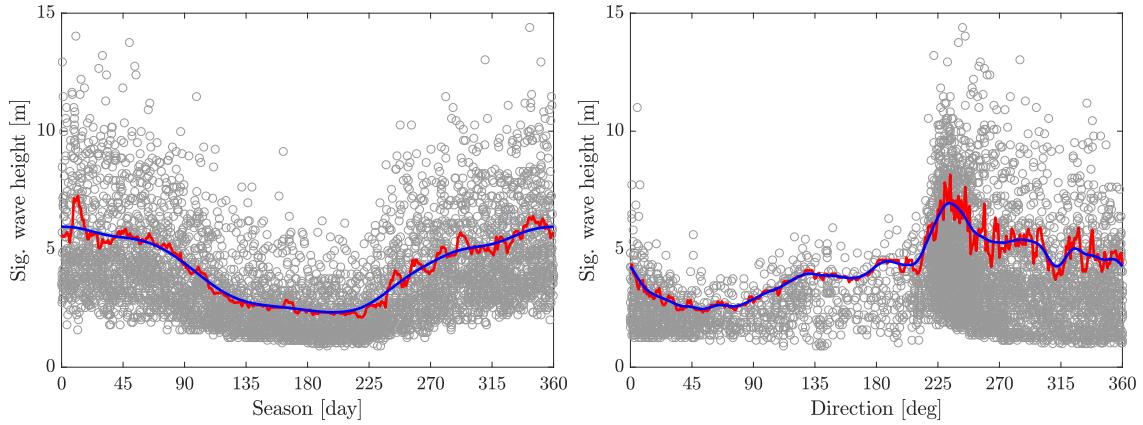


Figure 2: Threshold estimates for 1-D seasonal and directional analyses. Preliminary local quantile estimates in red, smoothed estimates in blue. Observations are shown in circles. Threshold exceedance probability is set at  $\zeta = 0.3$  for seasonal analysis and  $\zeta = 0.2$  for directional analysis.

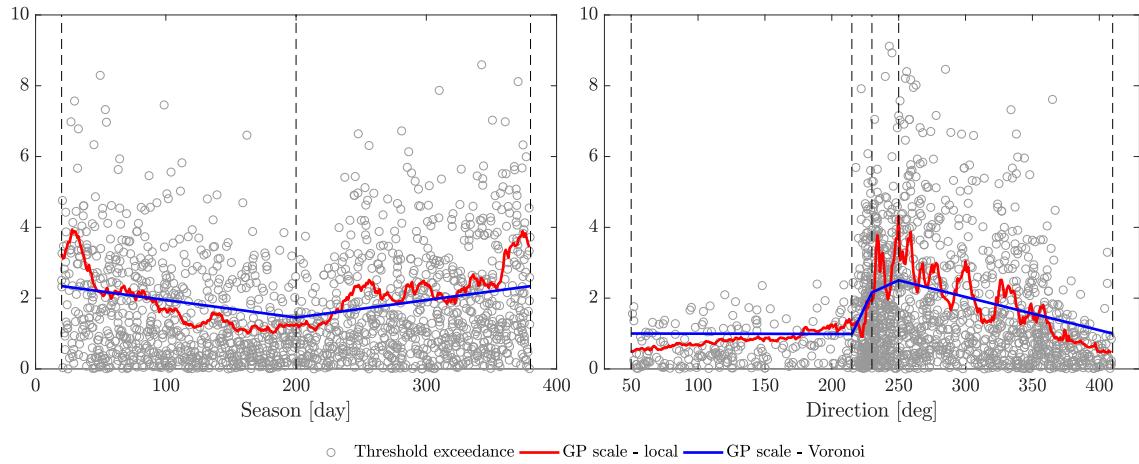


Figure 3: Threshold exceedances together with initial local estimates of GP scale parameter based on nearest observations. Starting values for the PPL estimates of scale based on the Voronoi partition are also shown. The locations of the nodes are shown as vertical dashed lines.

kernel bandwidth,  $w$ , have been selected by experimentation to give a reasonable trade-off between having sufficient observations to obtain an accurate estimate, and over-smoothing of non-stationary behaviour. The larger value of  $C$  and smoothing bandwidth for seasonal analysis reflects smoother variation of  $H_s$  with season compared to direction. Resulting threshold estimates are shown in Figure 2. For the seasonal analysis, the threshold ranges from  $u = 2.3$  m in summer to  $5.9$  m in winter. For the directional analysis, the threshold ranges from  $u = 2.5$  m for storms from the northeast to  $6.9$  m for storms from the southwest.

Seasonal and directional models with  $K = 2, 3, 4$  and  $6$  nodes are estimated, corresponding to each of Cases A, B and C above. The range of values chosen for  $K$  is intended to illustrate the effect of increasing model complexity on predictive performance. It was found that using  $K > 6$  nodes did not give improved predictive performance. For the seasonal analysis, nodes are spaced uniformly on covariate domain  $\mathcal{D}$ , with the first node location at  $n_1 = 20$  days, corresponding to the peak value of the initial local estimate of GP scale  $\sigma$ . For the directional case we place nodes manually, starting with 2 nodes located at the minimum and maximum of the initial local estimates of  $\sigma$ . For increasing  $K$ , additional nodes are added in turn, without changing the locations of existing nodes. Examples of the node placement for the  $K = 2$  seasonal model and  $K = 4$  directional model are shown in Figure 3, together with the local estimates (red) of  $\sigma$ , and corresponding initial values (blue) for the PPL model from a Voronoi partition. The blue curve appears to provide a reasonable representation of the variation in scale apparent in the local red estimates.

The left panel of Figure 4 illustrates the choice of  $\lambda_\sigma^\circ$  for directional Case A with  $K = 4$  nodes. The minimum mean

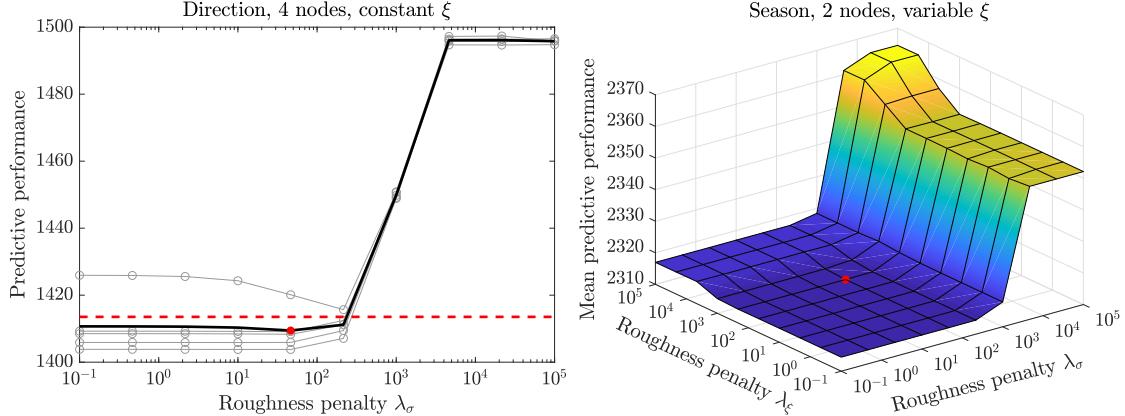


Figure 4: Optimal roughness coefficients for 1-D models. Left: Predictive performance (negative log-likelihood) against roughness coefficient  $\lambda_\sigma$ , for directional Case A with  $K = 4$ . Grey lines indicate results for  $R = 5$  replicate runs, and the solid black line is the mean over replicates. The dashed red line shows the value of the optimal predictive performance  $\bar{\mathcal{P}}(\boldsymbol{\lambda}^*)$  recorded. Right: Mean predictive performance over  $R = 5$  replicates for seasonal Case B with  $K = 2$ , with roughness coefficients  $\lambda_\xi$ ,  $\lambda_\sigma$ . In both panels, the red dot shows the location of penalty vector  $\boldsymbol{\lambda}^\circ$  yielding optimal predictive performance.

Case	Variable	Number of nodes, $K$			
		2	3	4	6
A	Optimal performance $\log_{10}(\lambda_\sigma^*)$	2316 (0.7) 2.3	2321 (1.3) 2.3	2314 (2.1) 2.3	2324 (0.6) 3
	Optimal performance $\log_{10}(\lambda_\xi^*)$	2314 (2.0) 2.3	2316 (3.3) 2.3	2311 (1.2) 1.7	2308 (2.5) 2.3
B	Optimal performance $\log_{10}(\lambda_\sigma^*)$	3	3	3	2.3
	Optimal performance $\log_{10}(\lambda_\xi^*)$	3	3	3	2.3

Table 1: Optimal predictive performance,  $\bar{\mathcal{P}}(\boldsymbol{\lambda}^*)$ , and optimal penalties,  $\boldsymbol{\lambda}^*$ , for 1-D seasonal model with  $K = 2, 3, 4$  and 6 nodes. Numbers in brackets are approximate uncertainties on performance from jackknife analysis.

negative log predictive likelihood  $\bar{\mathcal{P}}(\lambda_\sigma^\circ)$  occurs at  $\lambda_\sigma = 10^{1.67}$ . However, values of  $\lambda_\sigma \leq 10^{2.33}$  are all within the range of acceptable predictive performance, as quantified by the jackknife uncertainty in  $\bar{\mathcal{P}}(\lambda_\sigma^\circ)$ . Hence we set the optimal penalty as  $\lambda_\sigma^* = 10^{2.33}$  for this case. The utility of replicating the cross-validation procedure  $R = 5$  times is evident: for four of five repeats, the predictive performance is relatively constant up to the optimal penalty, whereas on one repeat, the performance deteriorates for  $\lambda_\sigma < \lambda_\sigma^*$ . The sharp increase in  $\bar{\mathcal{P}}(\lambda_\sigma)$  around  $\lambda_\sigma = 10^3$  indicates that the corresponding models are too stiff to capture the variation of the scale parameter. For  $\lambda_\sigma \geq 10^{3.67}$ , predictive performance asymptotes as the estimated GP scale  $\sigma$  is effectively constant on  $\mathcal{D}$ .

The corresponding plot for seasonal Case B with  $K = 2$  nodes is shown in the right panel. Variation of mean predictive performance  $\bar{\mathcal{P}}$  with  $\lambda_\sigma$  is similar to that in the left panel. There is less variation in  $\bar{\mathcal{P}}$  with  $\lambda_\xi$ , indicating that allowing  $\xi$  to vary on  $\mathcal{D}$  provides only small improvement in model performance.

For case A, values of optimal predictive performance  $\bar{\mathcal{P}}(\boldsymbol{\lambda}^*)$ , associated uncertainties and optimal values for roughness coefficients  $\boldsymbol{\lambda}^*$  for the 1-D seasonal models with  $K = 2, 3, 4$  and 6 are given in Tables 1. Values of predictive performance scores are directly comparable between seasonal models, since all are based on the same sample.

Although there is some variability in  $\bar{\mathcal{P}}(\boldsymbol{\lambda}^*)$ , given its uncertainty we can conclude that model performance is generally independent of  $K$ . Performance at  $K = 6$  is somewhat worse, probably because that analysis suggests that a larger value of  $\lambda_\sigma^* = 10^3$  is appropriate. Further investigation revealed that for smaller values of roughness coefficient, optimal parameter estimates from a particular cross-validation on sample  $\mathcal{S} \setminus \mathcal{S}_g$  (i.e. omitting subset  $\mathcal{S}_g$  for some  $g = 1, 2, \dots, G$ ) yielded a predicted likelihood of zero on prediction set  $\mathcal{S}_g$ , or infinite predictive negative log-likelihood; this is a common problem with prediction from responses defined on bounded domains (e.g. Northrop et al. 2017). Thus, flexible models tend to be rejected, and a large roughness coefficient is required for plausible predictive performance; we surmise that the 6-node model with  $\lambda_\sigma^* < 10^3$  induces some over-fitting.

There is a slight improvement in performance, relative to the uncertainty associated with the predictive performance estimates, for Case B over Case A. We note that, generally, a larger roughness coefficient is selected for GP shape  $\xi$  than for scale  $\sigma$ . Conditional quantiles corresponding to different non-exceedance probabilities, estimated under the

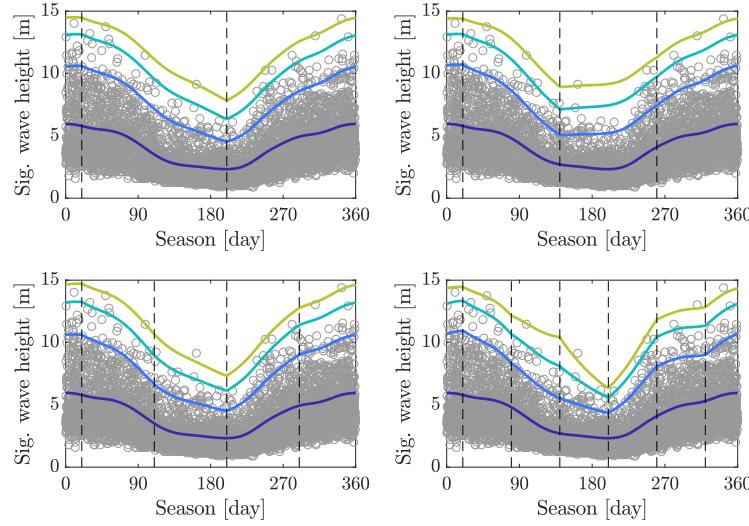


Figure 5: Conditional quantiles for 1-D seasonal Case B model with  $K = 2, 3, 4$  and  $6$ . Coloured lines show quantiles at non-exceedance probabilities of  $0.7$  (corresponding to threshold,  $u$ ),  $0.9$ ,  $0.99$  and  $0.999$ . Circles are original observations. Vertical lines show node locations.

fitted model for different  $K$ , are shown in Figure 5. Unsurprisingly, given the similar predictive performances listed in Table 1, quantile levels are similar for all choices of  $K$ , with  $K = 6$  suggesting somewhat lower extreme quantile levels in the interval  $[140, 200]$  days.

Estimates of GP scale and shape for Case B with  $K = 4$  are shown in Figure 6, for each of 100 bootstrap resamples of the original sample  $\mathcal{S}$ . Parameter estimates for individual bootstrap resamples are shown in grey, and their mean in black. Seasonal variation is clear, with larger scale in winter and somewhat more positive shape in the summer. Uncertainty in the scale parameter estimate is much lower than the magnitude of the seasonal variation. The corresponding uncertainty in the shape parameter estimate is higher relative to the seasonal variability; for some bootstrap resamples, the estimated shape parameter was approximately constant with season (these are highlighted with dashed red lines).

Comparisons of the tail probabilities  $\Pr(Y > y)$  estimated directly from the sample, and from simulation under the fitted 1-D seasonal Case B model with  $K = 4$  are shown in Figure 7. Corresponding “monthly” tails are also compared. The number of the simulated points is  $1000 \times N$ , (where  $N$  is the original sample size). This reduces sampling uncertainty in the empirical distribution function for the simulated data to a reasonable level at the exceedance

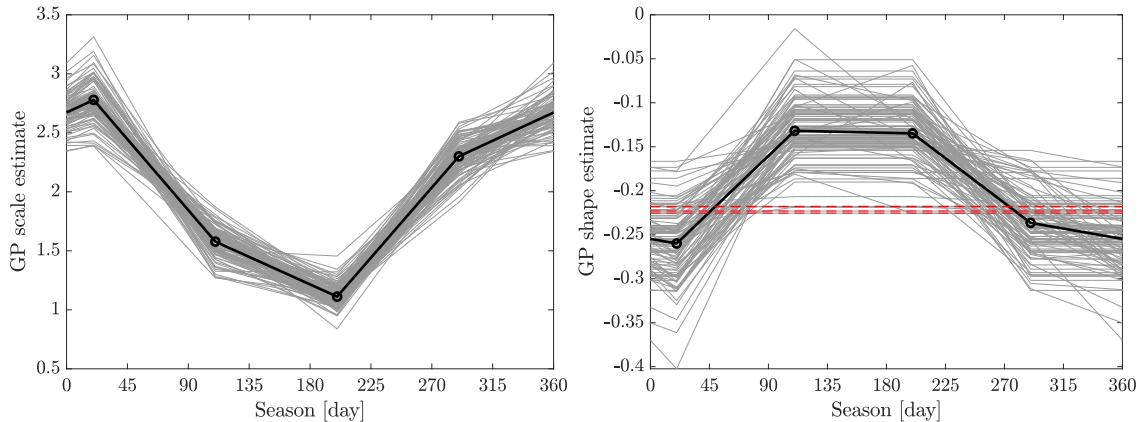


Figure 6: Estimates of GP scale and shape parameters for 1-D seasonal Case B model with  $K = 4$ . Grey lines show estimates from each of 100 bootstrap resamples of the original sample. Bold lines show bootstrap means. Dashed red lines highlight cases where the shape parameter estimate was approximately constant with season.

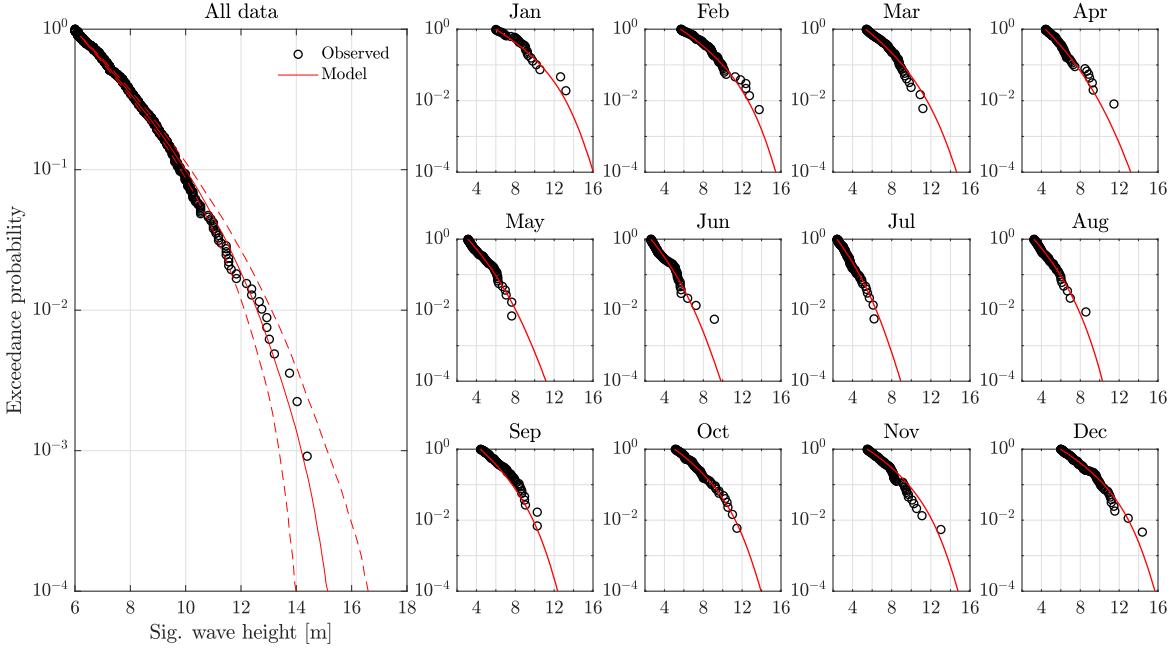


Figure 7: Annual and monthly exceedance probabilities for the 1-D seasonal Case B model with  $K = 4$ . Empirical estimates in black, and model-based estimates in red. Dashed lines on annual exceedance probability plot represent 95% uncertainty bands based on 100 bootstrap trials.

Case	Variable	Number of nodes, $K$			
		2	3	4	6
A	Optimal performance	1438 (0.7)	1410 (2.6)	1409 (4.1)	1409 (2.2)
	$\log_{10}(\lambda_\sigma^*)$	2.3	2.3	2.3	1.7
B	Optimal performance	1432 (0.7)	1403 (1.5)	1406 (2.0)	1413 (5.2)
	$\log_{10}(\lambda_\sigma^*)$	2.3	1.7	1.7	2.3
	$\log_{10}(\lambda_\xi^*)$	2.3	1.7	5	5

Table 2: Optimal predictive performance,  $\bar{\mathcal{P}}(\boldsymbol{\lambda}^*)$ , and optimal penalties,  $\boldsymbol{\lambda}^*$ , for 1-D directional model with  $K = 2, 3, 4$  and 6 nodes. Numbers in brackets are approximate uncertainties on performance from jackknife analysis.

probability associated with the largest observation (see Mackay and Jonathan 2021 for details). For the left-hand all-year plot, a 95% bootstrap uncertainty band (based on 100 bootstrap resamples) is also shown; for clarity, uncertainty bands on monthly plots are suppressed). There is good agreement between observations and the fitted model.

Corresponding results for 1-D directional models for Cases A and B are summarised in Table 2. There is little difference in performance for 1-D directional models for Cases A and B, with  $K > 2$ . However,  $K = 2$  suggests poorer performance, indicating that at least three nodes are required to capture directional variability. Allowing shape to vary gives a small improvement. However, for Case B with  $K = 4, 6$  the optimal roughness coefficients correspond to effectively constant shape on  $\mathcal{D}$ , again resulting from occurrences of infinite negative log predictive likelihoods for smaller choices of roughness coefficient. Conditional quantiles corresponding to different non-exceedance probabilities, estimated under the fitted Case B model for different  $K$ , are shown in Figure 8;  $K = 2$  does not adequately capture behaviour in the land shadow of Norway at angles in the interval  $[50, 200]$ ; for larger values of  $K$ , quantile levels are generally comparable. For comparison with Figure 6, estimates of GP scale and shape for Case A with  $K = 6$  are shown in Figure 9, for each of 100 bootstrap resamples of the original sample,  $\mathcal{S}$ . The scale parameter estimates are generally tightly-banded, with the most variability in the node at  $225^\circ$ , where there are few observations and there is a sharp change in the parameter gradient. The shape parameter estimates are well-behaved and generally fall within a relatively small range of values.

Exceedance probability plots were generated for all 1-D models. The plots for the directional models showed comparable levels of agreement to that illustrated in Figure 7, so are not shown here. Note that the predictive performance

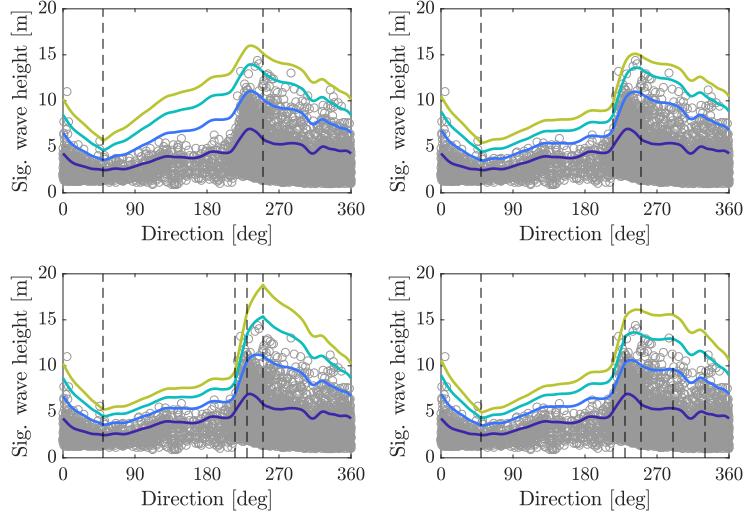


Figure 8: Conditional quantiles for 1-D directional Case B model with  $K = 2, 3, 4$  and  $6$ . Coloured lines show quantiles at non-exceedance probabilities of  $0.8$  (corresponding to threshold,  $u$ ),  $0.9$ ,  $0.99$  and  $0.999$ . Circles are original observations. Vertical lines show node locations. Note that for  $K = 4, 6$ , optimal models correspond to constant GP shape.

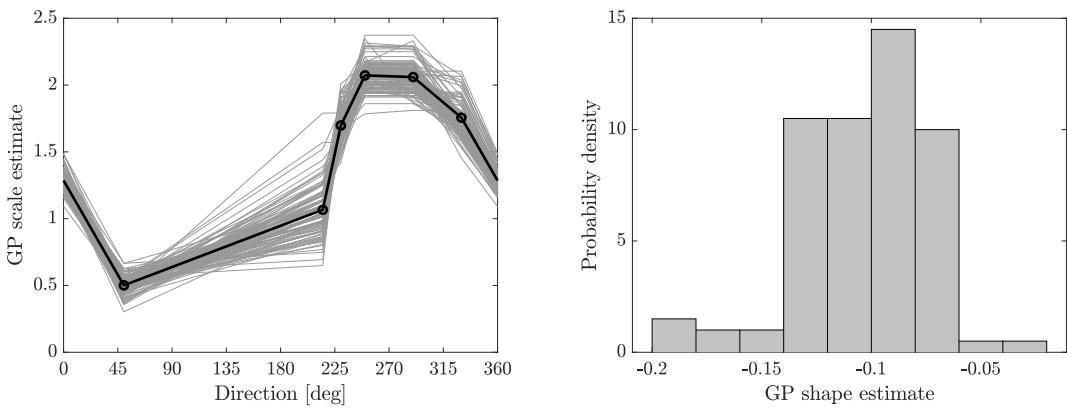


Figure 9: Estimates of GP scale and shape parameters for 1-D directional Case A model with  $k = 6$ . In the left panel, grey lines show estimates from each of 100 bootstrap resamples of the original sample, and bold lines show bootstrap means. The right panel gives a histogram of the corresponding shape estimates.

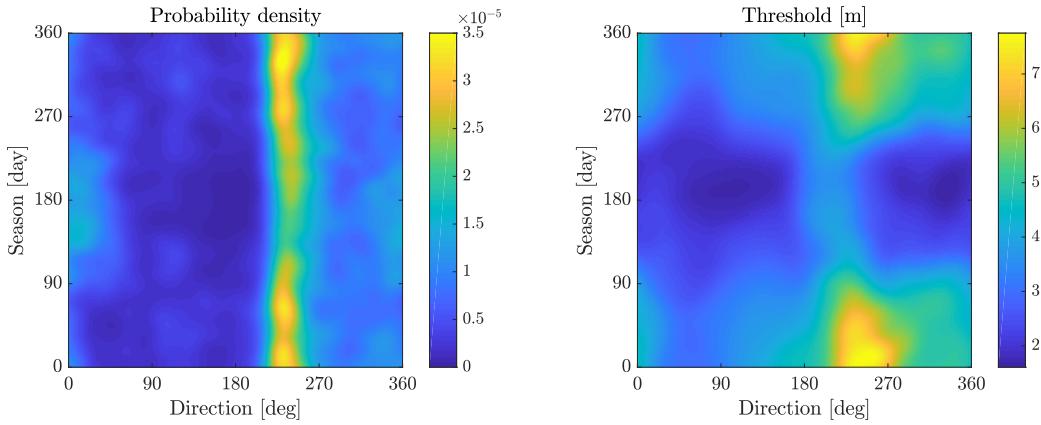


Figure 10: Estimates of covariate density (left) and GP threshold parameter (right) for 2-D directional-seasonal analysis. Kernel bandwidth  $w = 10^\circ$  (directionally) and 10 days (seasonally) for both kernel density estimate (left) and threshold smoothing (right). Initial local threshold estimates based on 50 nearest observations.

of the directional and seasonal models cannot be compared, as the set of threshold exceedances is different in each case, due to the different choices of directional and seasonal non-stationary thresholds, with different exceedance probabilities,  $\zeta$ .

## 5. Application with two-dimensional covariate

Here we discuss the estimation of two-dimensional directional-seasonal models, again following the procedure outlined in Section 2.1. Note that the definition of the Cases A, B, C archetypes is given in Section 3.4.

Kernel density estimates for the covariate density  $f_{\mathbf{X}}(\mathbf{x})$  are shown in the left panel of Figure 10, using a common kernel density bandwidth to  $w = 10^\circ$  directionally and 10 days seasonally, ensuring that directional details can be resolved. The estimated threshold corresponding to a local exceedance probability  $\zeta = 0.3$  is given in the right panel of the Figure 10, using the  $C = 50$  nearest neighbours, smoothed with a Gaussian kernel with bandwidth 10 degrees (directionally) and 10 days (seasonally).

Node placement in 2-D presents a greater challenge than in 1-D. We adopt two approaches to node placement. The first approach (henceforth “regular grid”) is defined in terms of  $D$  ( $= 2$  here) sets of marginal node locations (on  $[0, 360]$  here), used together to produce a regular lattice on  $\mathcal{D} = [0, 360] \times [0, 360]$ . The first node location in each set is also used to locate an additional “wrapping node”, located at the location of the first node plus 360 degrees or days, to accommodate periodicity on  $\mathcal{D}$ . The sets of marginal nodes are then used to create a regular rectangular grid of nodes on  $\mathcal{D}$ . Extra nodes are then added at the centres of each rectangle created, producing a (periodic) triangulation of  $\mathcal{D}$ . For a 2-D representation with  $K_1$  nodes directionally, and  $K_2$  seasonally, the resulting total number  $K$  of nodes is  $K = 2K_1K_2$ , and the corresponding number of triangular bins is  $M = 4K_1K_2$ . An example of a  $3 \times 2$  regular grid (with three marginal nodes in direction and two in season, and a total of 12 nodes and 24 bins) is shown in the left panel Figure 11. We refer to such a representation as a “ $K_1 \times K_2$ ” grid. The intuitive approach explained in Section 3.1 is again used to specify each set of marginal nodes. The regular grid is attractive because it is specified in terms of sets of marginal nodes which are relatively straightforward to locate.

The second approach to node specification (henceforth “irregular grid”) permits “freehand” node location on  $\mathcal{D}$ . This approach is obviously more flexible, and potentially more parsimonious than the regular grid, but requires judicious choice of node locations in  $D$  dimensions; this specification may be relatively straightforward in 2-D, but in general may prove challenging. Once the irregular node locations are specified, we create “wrapping nodes” with co-ordinates shifted by  $\pm 360$  (degrees or days) relative to those of the specified nodes in each covariate dimension. We then triangulate  $\mathcal{D}$  using the specified and wrapping nodes. The right panel of Figure 11 illustrates an irregular grid with 6 specified nodes and 12 unique triangular bins.

As for 1-D covariates, irregular nodes in  $D$  dimensions should be placed where the local gradient of a model parameter with covariate is expected to change, i.e. at local turning points. We select node locations by inspection, again using initial local estimates of GP scale to guide node placement. An example is shown in Figure 12 (corresponding to a 2-D extension of Figure 3). The surface shows initial local GP scale estimates from a local stationary GP fit to the  $C = 50$

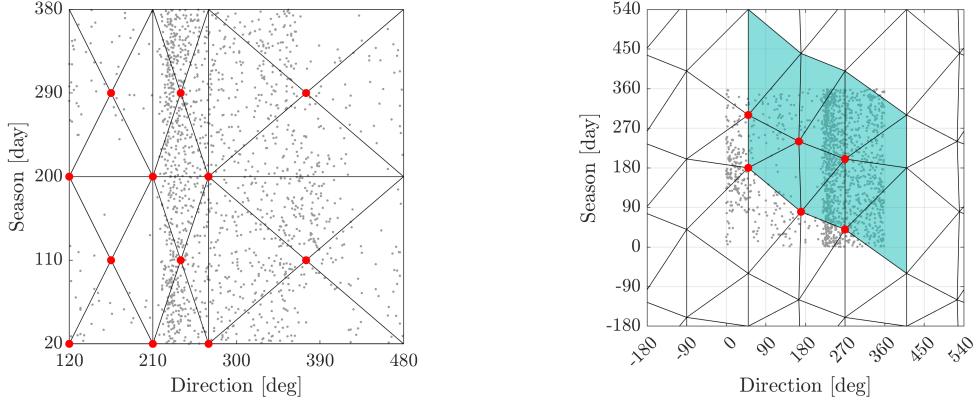


Figure 11: Examples of regular and irregular 2-D grids (left and right respectively) on extended covariate domain. Unique nodes are shown as red dots; all other nodes are repeats, included for ease of comprehension. Threshold exceedances are shown in grey. For ease of illustration on the regular grid, locations of observations with  $x$ -coordinate  $\leq 120^\circ$  and/or  $y$ -coordinate  $\leq 20$  days, where  $(120^\circ, 20 \text{ days})$  corresponds to the node location nearest the origin) have been shifted by  $+360$ . For the irregular grid, unique bins are highlighted in green. All other bins are repeats of highlighted bins, with vertices shifted by  $\pm 360$  in one or both covariate dimensions.

nearest neighbours any location on  $\mathcal{D}$ , Gaussian-kernel-smoothed with common bandwidth  $w = 10$  degrees or days. Black grid lines emanating from nodes represent the irregular grid triangulation, and a piecewise-linear representation for GP scale. Node placement can be adjusted manually until reasonable agreement is obtained between actual value of scale, and that suggested by the piecewise-linear triangulation.

We assess the predictive performance of non-stationary PPL extreme value models on 2-D covariate domains using different choices of regular and irregular grids. For regular grids, we consider  $K_1 = 2, 3$ , or  $4$  marginal directional nodes and  $K_2 = 2$  or  $3$  marginal seasonal nodes. We consider irregular grids with  $K = 4, 6, 8, 10, 12$  and  $14$  nodes. In the present examples, node sets for irregular grids are nested, in the sense that nodes for a  $K$ -node grid are also nodes for the  $K'$ -node grid when  $K \leq K'$ .

The cross-validation procedure used to estimate optimal values of roughness coefficients is the equivalent to that discussed in Section 4 for 1-D covariates, adapted to accommodate the appropriate number of roughness coefficients. Figure 13 shows an example for a (regular)  $3 \times 3$  Case A model, involving the estimation of a single roughness parameter  $\lambda_\sigma$  for GP scale in both direction and season. Grey lines show negative log predictive likelihood  $\bar{\mathcal{P}}$  on roughness penalty  $\lambda_\sigma$ , for each of  $R = 5$  replicate analyses, and the solid black line represents the mean  $\bar{\mathcal{P}}$  over the  $R$  replicates. As for the 1-D case, the dashed red line shows the ‘‘accepted’’ value of predictive performance. The inset panel shows a magnified view in the vicinity of the minimum of  $\bar{\mathcal{P}}$ . The optimal roughness coefficient is found to be  $\lambda_\sigma^* = 10^{1.67}$  in this case.

Predictive performance for different regular and irregular grids is summarised in Tables 3 and 4. As for 1-D covariates, we find that more complex models sometimes show inferior performance, corresponding to overly-heavy roughness penalisation (and hence over-fitting when not sufficiently penalised) to avoid problems of infinite negative log predictive likelihood during cross-validatory assessment. For the regular grids, allowing separate penalties for  $\sigma$  in each covariate dimension (Case C) did not significantly improve the predictive performance compared to having a common penalty for  $\sigma$  in both covariate dimensions (Case A). Therefore, for the irregular grids, only Cases A and B were considered. Allowing shape to vary with covariates (Case B) led to improved performance for some grids, but worse performance for others. In general the difference in performance was less than the uncertainty range in the estimated performance. Moreover, for most grids, the optimal penalties for the shape parameter force the shape to be effectively constant with covariate.

Irregular grid models with six or more nodes give better performance than regular grid models. Even the simplest 4-node irregular grid model gives comparable performance to the regular grids; this feature is attributed to careful irregular grid node placement, allowing more parsimonious description of parameter variation. Although there is some improvement in the predictive performance for irregular grids with additional nodes, these improvements are small compared with jackknife performance uncertainties.

Figure 14 shows directional-seasonal quantiles under the regular  $3 \times 3$  and irregular 6-node Case A models, corresponding to non-exceedance probability 0.99. The effects of both the piecewise-linear grid for  $\sigma(\mathbf{x})$ , and the nonlinear

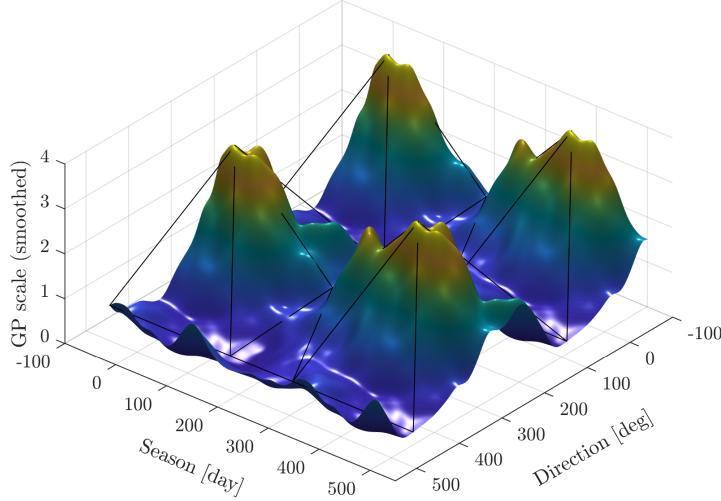


Figure 12: Gaussian-kernel-smoothed initial local GP scale estimates. Grid lines between nodes (black lines) visualise a piecewise-linear scale representation. Reasonable agreement between actual value of scale, and that suggested by the piecewise linear triangulation, suggests good node placement. An extended covariate domain is used to aid interpretation.

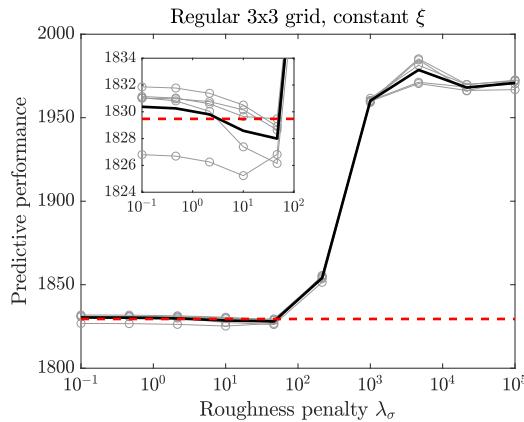


Figure 13: Illustrative results for estimation of optimal roughness coefficient for a  $3 \times 3$  Case A model. Grey lines show negative log predictive likelihood  $\mathcal{P}$  on roughness coefficient  $\lambda_\sigma$ , for each of  $R = 5$  replicate analyses. The solid black line is the mean  $\bar{\mathcal{P}}$  over the  $R = 5$  replicates. The dashed red line indicates the ‘accepted’ value of predictive performance. Inset panel gives a magnified view near the minimum of  $\bar{\mathcal{P}}$ .

Case	Variable	Numbers of marginal nodes, $(K_1 \times K_2)$					
		$(2 \times 2)$	$(3 \times 2)$	$(4 \times 2)$	$(2 \times 3)$	$(3 \times 3)$	$(4 \times 3)$
A	Optimal performance $\log_{10}(\lambda_\sigma^*)$	1837 (2.8) 1.67	1828 (2.9) 1.67	1839 (0.9) 2.33	1830 (3.1) 1.67	1828 (0.8) 1.67	1863 (2.5) 2.33
	$\log_{10}(\lambda_\xi^*)$	1833 (1.5) 0.33	1832 (3.2) 1.67	1824 (7.6) 1.67	1836 (3.6) 1.67	1824 (2.3) 1.67	1964 (3.7) 3
B	Optimal performance $\log_{10}(\lambda_\sigma^*)$	1833 (1.5) 0.33	1832 (3.2) 1.67	1824 (7.6) 1.67	1836 (3.6) 1.67	1824 (2.3) 1.67	1964 (3.7) 3
	$\log_{10}(\lambda_\xi^*)$	2.33	5	3	3	2.33	3
	Optimal performance $\log_{10}(\lambda_{\sigma_1}^*)$	1838 (1.3) 2.33	1834 (2.3) 1.67	1833 (2.8) 1.67	1829 (1.7) 1.67	1830 (3.0) 1.67	1840 (2.7) 1.67
C	$\log_{10}(\lambda_{\sigma_2}^*)$	2.33	1	2.33	1.67	1.67	2.33

Table 3: Optimal predictive performance,  $\bar{\mathcal{P}}(\lambda^*)$ , and optimal penalties,  $\lambda^*$ , for 2-D regular grid models. Numbers in brackets are approximate uncertainties from jackknife analysis.

Case	Variable	Number of nodes, $K$					
		4	6	8	10	12	14
A	Optimal performance	1836 (1.5)	1816 (2.3)	1815 (2.1)	1814 (4.6)	1814 (3.6)	1813 (4.8)
	$\log_{10}(\lambda_\sigma^*)$	2.33	2.33	1.67	2.33	1.67	1.67
B	Optimal performance	1834 (1.4)	1816 (2.6)	1815 (3.1)	1809 (3.1)	1809 (2.8)	1815 (2.2)
	$\log_{10}(\lambda_\sigma^*)$	1.67	2.33	-0.33	1.67	1.67	1.0
	$\log_{10}(\lambda_\xi^*)$	2.33	2.33	5.0	5.0	4.33	5.0

Table 4: Optimal predictive performance,  $\bar{P}(\boldsymbol{\lambda}^*)$ , and optimal penalties,  $\boldsymbol{\lambda}^*$ , for 2-D irregular grid models. Numbers in brackets are approximate uncertainties from jackknife analysis.

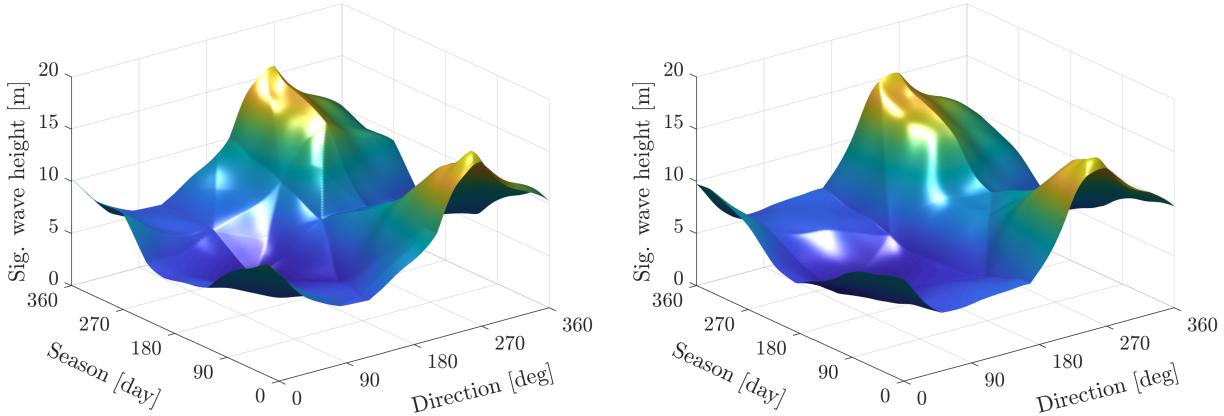


Figure 14: Directional-seasonal conditional quantile of  $H_s$  for non-exceedance probability 0.99, for regular  $3 \times 3$  and irregular 6-node Case A models (left and right respectively).

non-parametric threshold estimate are visible. The irregular grid provides a more adequate description of the sharp transition in  $H_s$  around  $210^\circ$ , due to judicious node placement. However, quantile estimates are generally similar, indicating that inferences are not strongly dependent on grid choice.

Parameter estimates for the regular  $3 \times 3$  Case A model are shown in Figure 15. The left panel shows the mean scale estimate of inferences using each of 100 bootstrap resamples. The right panel shows the corresponding distribution of stationary shape estimate showing a narrow spread of estimates. The general features of the left panel reflect those of the corresponding covariate density and threshold in Figure 10. Values of parameter estimates are further similar to those for the corresponding 1-D directional model illustrated in Figure 9. An illustrative directional tail plot for the irregular 6-node directional-seasonal Case A model is given in Figure 16. Agreement between observations and the fitted model is good.

## 6. Discussion and conclusions

Capturing covariate effects in ocean environmental extreme value models can be important for optimising the design of an asymmetric structure with respect to direction, or assessing risk associated with an operation in a given season. It can also potentially improve inference for extreme quantiles. The level of sophistication required to describe covariate effects depends on the characteristics of the sample. We expect that a relatively simple penalised piecewise-linear (PPL) representation for the functional forms of extreme value model parameters will often be sufficient. In this paper, we demonstrate that the PPL generalised Pareto (GP) model for peaks over threshold of significant wave height provides good characterisation of directional and seasonal effects at low computational cost. Roughness penalisation, regulated using a cross-validation scheme, ensures that parsimonious models are estimated. Uncertainties in parameter and return value estimates are quantified using bootstrap re-sampling.

In some circumstances a decision will be required between 1-D or 2-D models. Some design situations require an understanding of seasonal extremes (e.g. for planning operations with a limited duration), and/or directional extremes (e.g. if the loading on a structure is dependent on the wave, wind or current direction), dictating the type of model

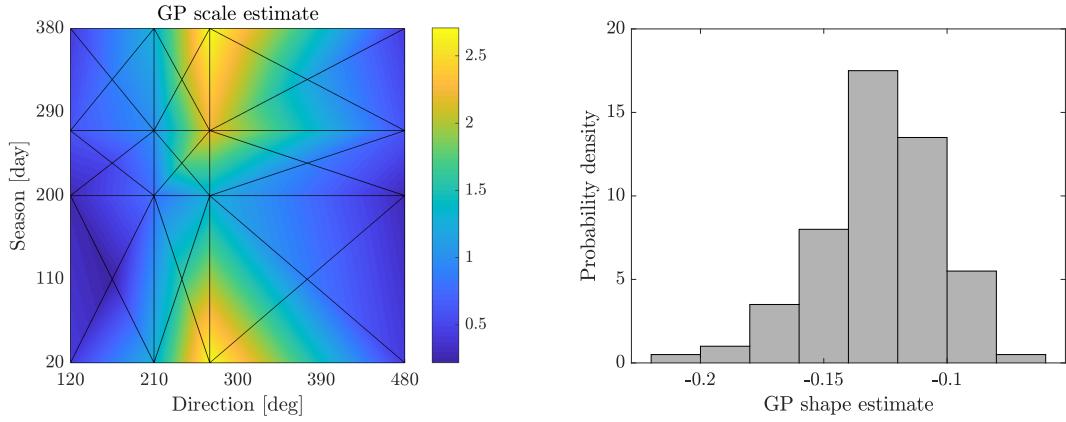


Figure 15: Estimates of GP scale and shape for the regular  $3 \times 3$  Case A model. Left: mean directional-seasonal scale estimate. Right: distribution of (stationary) shape estimates.

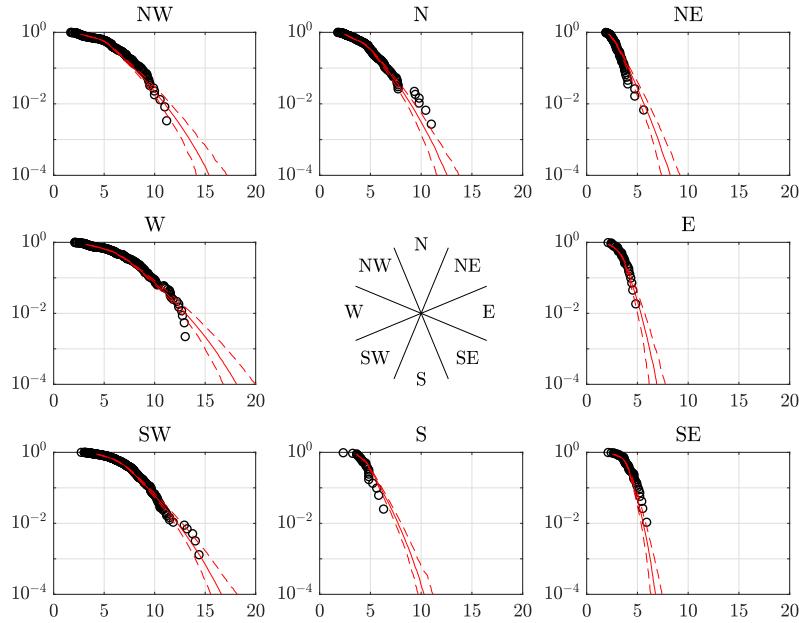


Figure 16: Tail plots per directional octant from the irregular 6-node directional-seasonal Case A model. Abscissa is significant wave height in metres. Ordinate is exceedance probability. Empirical estimates in black, and model-based estimates in red. Dashed lines represent 95% uncertainty bands based on 100 bootstrap trials.

required. However, even if seasonal or directional return values are not required, there can still be value in using a non-stationary model (Mackay and Jonathan, 2020a). Typically, formal statistical model selection methods assume a common data sample. In our case, the sample to which the threshold-exceedance model is fitted, differs between model types. This is because the thresholds differ between 1-D seasonal, 1-D directional and 2-D models. However, a subset of observations that is common to all models could be defined, for which the (predictive) likelihood could be compared across models. Comparison of diagnostic plots, such as those shown in Figures 7 and 16, can also be useful in selecting a model. In some cases, where there is limited data available, or covariate effects are relatively weak, using a stationary model may be more appropriate.

Relative to more sophisticated covariate representations (e.g. Wood 2011, Zanini et al. 2020), the PPL model is conceptually simple to understand, and facilitates computationally-simple inference, both of which we consider to be attractive features for the metocean practitioner. Relative to the simpler penalised piecewise-constant (PPC) approach of Ross et al. (2020), the additional flexibility of the PPL model allows a physically more plausible representation of continuous parameter variation on the covariate domain. Better use of the sample is made, with parameter values at each PPL node informed by observations in all bins adjacent to that node. In contrast, the PPC model is only penalised for the total variance of parameter values over the bins, ignoring bin proximity.

In the current work, as outlined in Section 2.1 and Equation 2, we are interested in estimating the joint distribution  $f_{\mathbf{X},Y}(\mathbf{x}, y)$  for  $y > u(\mathbf{x})$ . This requires models for the extreme value threshold ( $u(\mathbf{x})$ ), in order to extract threshold exceedances, and estimate their density  $f_{GP}(y|u(\mathbf{x}), \sigma(\mathbf{x}), \xi(\mathbf{x}))$  and the density  $f_{\mathbf{X}}(\mathbf{x})$  of covariates. Unlike the GP model for threshold exceedances, estimation for both  $u(\mathbf{x})$  and  $f_{\mathbf{X}}(\mathbf{x})$  utilises the full sample. It is therefore reasonable to seek more detailed descriptions, using flexible local estimators, for covariate density and threshold, than for the density of threshold exceedances. In this sense, the current hierarchical approach imposes increasing regularity on non-stationary effects as the statistical efficiency of the estimation decreases, incorporating kernel density estimation for covariate density and local quantile estimation for threshold, but a more rigid piecewise-linear model for GP scale, and either a constant or piecewise-linear model for GP shape. However, as implemented here, uncertainties from the estimation of covariate density and threshold are not propagated into the extreme value analysis; estimates of extreme value models here are therefore conditional on the estimated extreme value threshold. Extreme quantile estimates are further conditional on the estimated covariate density. The effect of threshold uncertainty on return values can be large; in the current work, an appropriate threshold non-exceedance probability is selected to ensure that inferences are stable with respect to small variation in this choice. More complex approaches (e.g. Randell et al. 2016) seek to estimate covariate density, threshold and extreme value parameters in a single inference, at the cost of increased model complexity and computational burden. Due to the relative sparsity of information for the GP shape parameter, its roughness coefficient is often relatively large, implying a smooth variation of the estimate on the covariate domain. Nevertheless, models admitting non-constant shape perform slightly better in general compared to those with constant shape: allowing a little variation in shape improves performance.

Choice of the number and locations of nodes is key to the success of the PPL methodology. Diagnostic tools can aid these choices. In 2-D, it appears that judiciously-chosen “irregular” nodes provide the best predictive performance. Nevertheless, models exploiting nodes located on a “regular” rectangular grid, specified in terms of nodes for marginal components, also perform relatively well. In the current work, we have assumed that the same node locations are appropriate for GP scale and shape estimation; clearly this may not always be the case, and we might expect that fewer nodes might be sufficient to describe shape parameter variation. Nevertheless, the optimal choice of roughness coefficients goes some way to providing balanced parameterisations across covariate dimensions and GP parameters. In general, extreme value modelling using Bayesian inference, exploiting reversible jump or similar methodologies in 1-D (e.g. Zanini et al. 2020) and 2-D (e.g. Jonathan 2021) provide approaches to estimate the number and location of nodes automatically as part of the inference, at increased computational cost for a more complex model formulation. Experience also suggests that the Bayesian paradigm which permits, via stochastic sampling schemes (e.g. the Metropolis-Hastings or mMALA algorithms), an exploration of the full posterior distribution of the parameters, is more reliable than the frequentist approach. The latter relies on accurate identification of the global mode of a high-dimensional and often highly multi-modal surface; in many cases it is difficult to either identify the mode or to verify that the mode is indeed global.

In the current analysis, the number of nodes is kept to a reasonably small value, since we are motivated by the idea of providing the conceptually simplest but useful extreme value model for the sample. However, the statistical literature on penalised B-splines (e.g. Eilers and Marx 2010) recommends specification of a large number of spline nodes (or knots), with roughness penalisation subsequently imposing the optimal smoothness on the solution, reducing the effective number of degrees of freedom in the covariate representation. In our case, with small numbers of nodes, penalisation plays a less important role in general, but is nevertheless important when the number of nodes is over-

specified. Moreover, as a relatively simple optimisation method has been used in the present work, using a small number of nodes makes finding an optimal set of parameter estimates feasible, without excessive computational effort. Further work might include a systematic comparison of PPL performance relative to its peers, and perhaps also relative to methods involving transformation of data to standard scales prior to extreme value analysis (e.g. Eastoe and Tawn 2012). The current work has addressed periodic covariates only; inference for non-periodic covariates is in some senses simpler. However, in other applications (e.g. Mackay and Jonathan 2020b) a non-periodic covariate may itself become extreme, requiring joint extreme value modelling of “covariate” and response.

## Acknowledgement

This work was funded by the UK EPSRC Supergen Offshore Renewable Energy Hub, project EP/S000747/1 on “Improved Models for Multivariate Metocean Extremes (IMEX)”.

## References

- Anderson, C., Carter, D., and Cotton, P. (2001). *Wave climate variability and impact on offshore design extremes*. Report commissioned from the University of Sheffield and Satellite Observing Systems for Shell International.
- Carter, D. J. T. and Challenor, P. G. (1981). Estimating return values of environmental parameters. *Quarterly Journal of the Royal Meteorological Society*, 107:259.
- Chavez-Demoulin, V. and Davison, A. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society C: Applied Statistics*, 54:207–222.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.
- Coles, S. and Walshaw, D. (1994). Directional modelling of extreme wind speeds. *Applied Statistics*, 43:139–157.
- Davison, A. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society B*, 52:393.
- Eastoe, E. and Tawn, J. (2012). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society C: Applied Statistics*, 58:25–45.
- Eastoe, E. F. (2019). Nonstationarity in peaks-over-threshold river flows: A regional random effects model. *Environmetrics*, 30(5):1–18.
- Eilers, P. H. C. and Marx, B. D. (2010). Splines, knots and penalties. *Wiley Interscience Reviews: Computational Statistics*, 2:637–653.
- Ewans, K. C. and Jonathan, P. (2008). The effect of directionality on northern North Sea extreme wave design criteria. *Journal of Offshore Mechanics and Arctic Engineering*, 130:041604:1–041604:8.
- Fawcett, L. and Walshaw, D. (2007). Improved estimation for temporally clustered extremes. *Environmetrics*, 18:173–188.
- Guerrero, M. B., Huser, R., and Ombao, H. (2021). Conex-Connect: Learning Patterns in Extremal Brain Connectivity From Multi-Channel EEG Data. *Pre-print*.
- Haver, S. (1987). On the joint distribution of heights and periods of sea waves. *Ocean Engineering*, 14:359–376.
- Jonathan, P. (2021). A non-parametric representation for multidimensional covariates in an extreme value model. *Gwerddon*, 33:68–84.
- Jonathan, P., Flynn, J., and Ewans, K. C. (2010). Joint modelling of wave spectral parameters for extreme sea states. *Ocean Engineering*, 37:1070–1080.
- Konzen, E., Neves, C., and Jonathan, P. (2021). Modeling nonstationary extremes of storm severity: Comparing parametric and semiparametric inference. *Environmetrics*, 32:e2667.

- Mackay, E., Challenor, P. G., and Bahaj, A. B. S. (2010). On the use of discrete seasonal and directional models for the estimation of extreme wave conditions. *Ocean Engineering*, 37:425–442.
- Mackay, E. and Jonathan, P. (2020a). Assessment of return value estimates from stationary and non-stationary extreme value models. *Ocean Engineering*, 207.
- Mackay, E. and Jonathan, P. (2020b). Estimation of environmental contours using a block resampling method. *Proceedings of the International Conference on Offshore Mechanics and Arctic Engineering - OMAE2020*, 2A-2020.
- Mackay, E. and Jonathan, P. (2021). Sampling properties and empirical estimates of extreme events. *Ocean Engineering*, 239:109791.
- Mendez, F. J., Menendez, M., Luceno, A., Medina, R., and Graham, N. E. (2008). Seasonality and duration in extreme value distributions of significant wave height. *Ocean Engineering*, 35:131–138.
- Northrop, P., Attalides, N., and Jonathan, P. (2017). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society C: Applied Statistics*, 66:93–120.
- Northrop, P. and Jonathan, P. (2011). Threshold modelling of spatially-dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics*, 22:799–809.
- Randell, D., Feld, G., Ewans, K., and 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.
- Randell, D., Turnbull, K., Ewans, K., and Jonathan, P. (2016). Bayesian inference for non-stationary marginal extremes. *Environmetrics*, 27:439–450.
- Reistad, M., Breivik, O., Haakenstad, H., Aarnes, O. J., Furevik, B. R., and 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.
- Robinson, M. E. and Tawn, J. A. (1997). Statistics for extreme sea currents. *Applied Statistics*, 46:183–205.
- Ross, E., Astrup, O. C., Bitner-Gregersen, E., Bunn, N., Feld, G., Gouldby, B., Huseby, A., Liu, Y., Randell, D., Vanem, E., and Jonathan, P. (2020). On environmental contours for marine and coastal design. *Ocean Engineering*, 195:106194.
- Ross, E., Sam, S., Randell, D., Feld, G., and Jonathan, P. (2018). Estimating surge in extreme North Sea storms. *Ocean Engineering*, 154(January):430–444.
- Scotto, M. and Guedes-Soares, C. (2000). Modelling the long-term time series of significant wave height with non-linear threshold models. *Coastal Engineering*, 40:313–327.
- Sigauke, C. and Bere, A. (2017). Modelling non-stationary time series using a peaks over threshold distribution with time varying covariates and threshold: An application to peak electricity demand. *Energy*, 119:152–166.
- Towe, R., Eastoe, E., Tawn, J., and Jonathan, P. (2017). Statistical downscaling for future extreme wave heights in the North Sea. *Annals of Applied Statistics*, 11:2375–2403.
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(1):3–36.
- Yang, L., Liu, S., Tsoka, S., and Papageorgiou, L. G. (2016). Mathematical programming for piecewise linear regression analysis. *Expert Systems with Applications*, 44:156–167.
- Zanini, E., Eastoe, E., Jones, M., Randell, D., and Jonathan, P. (2020). Covariate representations for non-stationary extremes. *Environmetrics*, 31:e2624.