# Non-stationary environmental extremes

Elena Zanini, B.Sc.(Hons.), M.Res

# Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Unless stated otherwise, I developed my own computing code for the analysis in this thesis and used R and Matlab as coding languages.

The word count for this thesis is 68,015 words.

Elena Zanini

# Acknowledgements

Finally, a sincere thank you to Jethro, whose patience I stretched more times than I can count, yet he stood beside me during the most tiring times and always tried to cheer me up.

# Abstract

Our physical environment gives rise to extreme events (river floods, heatwaves, atmospheric pollution, hurricanes, earthquakes) that have the potential to endanger lives and cause large economic loss. Understanding the characteristics of extreme ocean environments in particular, and their interactions with marine and coastal structures, is critical to the safety of all who inhabit coastal regions, or depend on the ocean for their livelihoods. Extreme value analysis presents a useful framework to quantifying the extreme ocean environment from samples of observations.

Reliable extreme value models for the ocean environment must accommodate known sources of systematic variation in ocean storm severity due to covariates such as wave direction and season. This motivates the two main areas of research addressed in this thesis. The first area considers the introduction of covariates in the generalised Pareto and non-homogeneous Poisson point process models for extremes. The relative performance of non-stationary forms of these models is investigated in terms of ease of implementation, parameter estimation and predictive performance on both simulated and hindcast samples. Both approaches have their merits. The key finding is the importance of employing a model formulation that captures the covariate-response relationship appropriately.

The second area of research is the development and evaluation of approaches to estimate the covariate-response relationship. Covariate effects in extreme ocean storms are often intricate, requiring a flexible framework to estimate the variation of extreme value model parameters as a function of covariates. Parameterisations need to be sufficiently complex to be physically realistic, but sufficiently parsimonious to be practically useful. Semi- and non-parametric models are ideal candidates. A covariate parameterisation often needs to accommodate rapid change or specific covariate interactions in one part of the covariates domain, but smoother variation in other parts. The covariate parameterisation needs to be sufficiently flexible and reliable to facilitate the study of extreme ocean environments from different ocean basins with fundamentally different physical characteristics. Model sophistication usually comes at the expense of computational stability and efficiency. Modelling procedures are required which are capable of incorporating multiple covariates (suche as direction, season, location, water depth), ensuring appropriate smoothness of model parameter variation with covariates, to maximise predictive performance whilst avoiding overfitting.

Extreme value models within which model parameter variation with respect to covariates is described using penalised splines, Bayesian adaptive regression splines, radial basis functions and covariate domain partitioning are considered for 1-D and 2-D covariates. The performance of each is evaluated using simulated and hindcast samples. Spline-based models perform relatively well, in terms of predictive performance and computational stability, across a range of applications. Radial basis functions and covariate domain partitioning formulations appear to provide a promising parsimonious alternative when covariates interact strongly.

# Contents

## 4    One-dimensional covariate modelling      111

## Appendices      176

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Extreme value theory is defined as the study of the extremal properties of random processes. This field of statistics has been applied in many different areas, including environmental processes, finance, drug safety and insurance. This thesis focuses in particular on modelling environmental phenomena. Hence, while the theory and models developed would be valid for different applications, the examples and datasets presented are of environmental origin.

## 1.1 Environmental extremes

Environmental processes which may be of interest include rainfall, river flow, pollution levels, wind speed and wave height. Although these processes are not extremes, we may be interested in the extreme events that arise from them, as these often yield disastrous consequences for the environment, with associated heavy costs and risk to human lives. For some processes, such as wave height and wind

speed, extreme value theory can be used to estimate the probabilities of extreme levels of the processes, which in turn can be used to help, for example, in the design of structures such as oil platforms or water defences. They can also be used to assess the danger associated with extreme events, including the risk of flooding, structural damage, environmental contamination and potential loss of life. Both univariate and multivariate processes have been studied extensively, although in this work we focus on the former.

As one of the six oil and gas "supermajors", Shell is interested in the design, construction and maintenance of offshore structures. It then becomes essential to understand the extreme conditions that marine structures are likely to experience in their lifetime in order to manage the risk of structural damage, which can lead to losses in revenue, environmental pollution and staff endangerment. In particular, a variety of institutions, such as the American Petroleum Institute (API), the Norwegian Petroleum Safety Authority (PSA) or the International Organization for Standardization (ISO), stipulate offshore regulation guidelines. These dictate that structures must be designed to withstand a 1 in $N$ year storm i.e. a storm so bad that it is only observed on average once in $N$ years, where $N$ is usually chosen to be between 100 and 10000 years.

Extreme value analysis provides the ideal framework for modelling and predicting phenomena with very low probabilities. Nevertheless, a variety of issues arise in extreme value modelling, especially since only the most extreme observations are considered for modelling purposes. The focus is then on a relatively small amount of data in comparison with the long term predictions desired, and the resulting uncertainty needs to be accounted for in a consistent and reliable manner. Moreover,

the phenomena underlying the processes of interest are often very complex, and observations tend to be non-homogeneous in space, time or both. Hence, considering covariates becomes essential to proposing a realistic model. For example, one cause of structural damage is storm waves, with the most severe ones being wind generated. A variable commonly used to summarise the impact of a storm and its severity in the metocean engineering is the significant wave height, often denoted by Hs, which is defined as the mean wave height of the highest third of waves. We may suspect that the height of waves is affected by physical considerations such as seasonal cycles of storm severity, variation of water depth or fetch variability with direction; then, the inclusion of the corresponding covariates is likely to be important for good model fit. Unsurprisingly, it is then common for design guidelines for offshore sites to require directional and omni-directional predictions, while installation, maintenance and manning schedules expect seasonal (e.g. monthly) ones. Jonathan et al. (2008) and Jonathan and Ewans (2011) demonstrate that, for this type of data, a directional or a seasonal extreme value model generally explains the observed variation significantly better than a model which ignores directionality and seasonality, and that omni-directional criteria developed from a covariate-dependent model are different from a constant model which ignores them. Hence, constructing a model that adequately captures these effects is necessary, and statistical tools to analyse and extrapolate from such a model become essential.

This thesis concentrates on *peaks over threshold* (POT) methods, where we focus on the tail of the distribution by only considering observations that are above an arbitrary large value. In the stationary case, two models are available from

the literature, namely the generalised Pareto distribution (GPD) and the non-homogeneous Poisson point (NHPP) process formulation, which are introduced in Section 2.1. Although these models are theoretically equivalent, they both require adaptations to account for covariate effects, with each method having advantages and limitations.

Once a suitable extreme value model has been chosen, further important considerations are necessary. When analysing extreme characteristics of the ocean environment, often multiple sources (e.g. direction, season and two-dimensional location) of covariate-dependence are at work, and a flexible and scalable framework for covariate-dependent modelling is then needed. There are different methods to incorporate covariate effects into an extreme value model, and in this thesis we focus on representing the parameters of a chosen statistical model as functions of covariates. First of all, an ideal approach would be capable of being applied to, for example, samples from neighbouring or distant locations without requiring extensive ad hoc changes. Furthermore, computational efficiency and stability for single- or multi-dimensional covariates is essential. Different parameterisations are available from the literature. In oceanographic applications, it is often necessary to consider one that is complex enough to be practically useful, such as semi-parametric and non-parametric models. One then needs to consider a series of key aspects. First of all, while semi- and non-parametric formulations yield flexible models, they can lead to a large number of correlated parameters. It is also important to implement suitable strategies to avoid overfitting while fully capturing the covariate information. Furthermore, computational efficiency and stability are paramount, especially when high-dimensional covariates affect the process.

This thesis considers different existing models and formulations in light of the issues mentioned above, and focuses on proposing methods with the desired characteristics that are capable of yielding accurate predictions for non-stationary extreme processes.

## 1.2 Hindcast datasets

The approaches and methodologies considered in this thesis are applied, in different chapters, to two hindcast wave height datasets from different ocean basins.
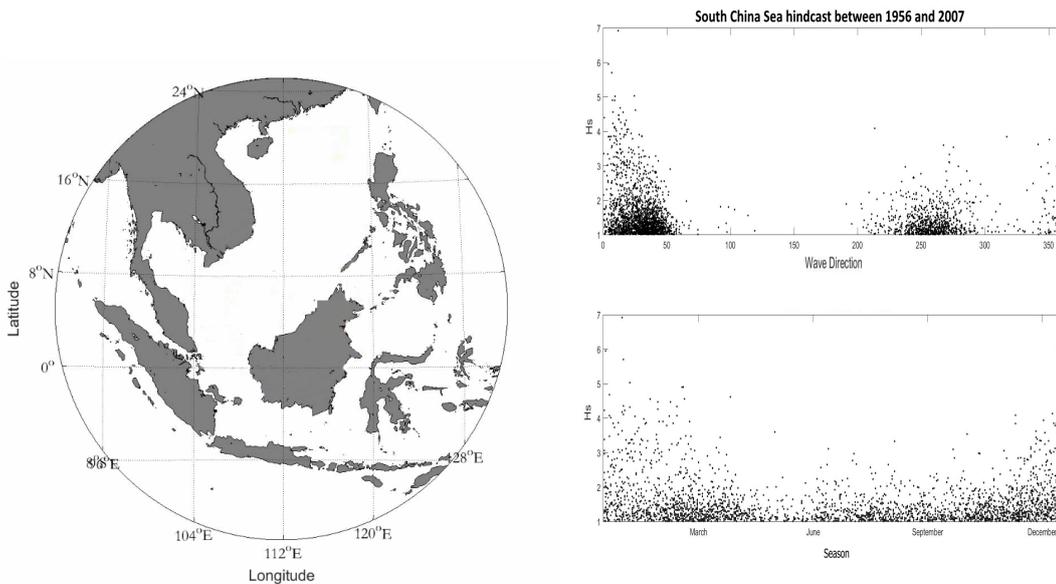


Figure 1.2.1: Map of the hindcast locations in the South China Sea (left panel), with peak significant wave height Hs (m) on wave direction (right upper panel) and season (in day of the year, right lower panel) for one of the datasets.

Two of the datasets come from locations in the South China Sea, and consist of a hindcast time-series for sea state significant wave height Hs. They include time-series for the dominant wave direction and season, the latter defined as day of the year, for a notional year consisting of 360 days. They are part of the SEAFINE

hindcast database by Oceanweather, which produced wind and wave hindcast data on a fine grid of the southern part of the South China Sea for a continuous period between 1956 and 2015 (SEAFINE, 2015). For the purpose of the work in this thesis, we only consider the storm peak significant wave height, such as the one shown in Figure 1.2.1 for one of the South China Sea locations. This is obtained from the hindcast by setting a low threshold for peak-picking, in order to identify separate storm events, and then selecting the maximum of significant wave height during the storm interval.

The North Sea hindcast was previously reviewed by Breivik et al. (2007). It was produced using calibrated meteorological computer models based on historical data from existing offshore sites, and covers the Norwegian Sea, Barents Sea and the North Sea. The model makes use of the ERA40 global hindcast project (see Uppala et al., 2005 for a comprehensive overview), the HIRLAM atmospheric model (see Undén et al. (2002)) and a modified version of the WAM Cycle 4 model for the wave components Gunther et al. (1992). The storm peak significant wave heights in Figure 1.2.2 come from a specific location off the coast of Norway, from September 1957 to December 2008.
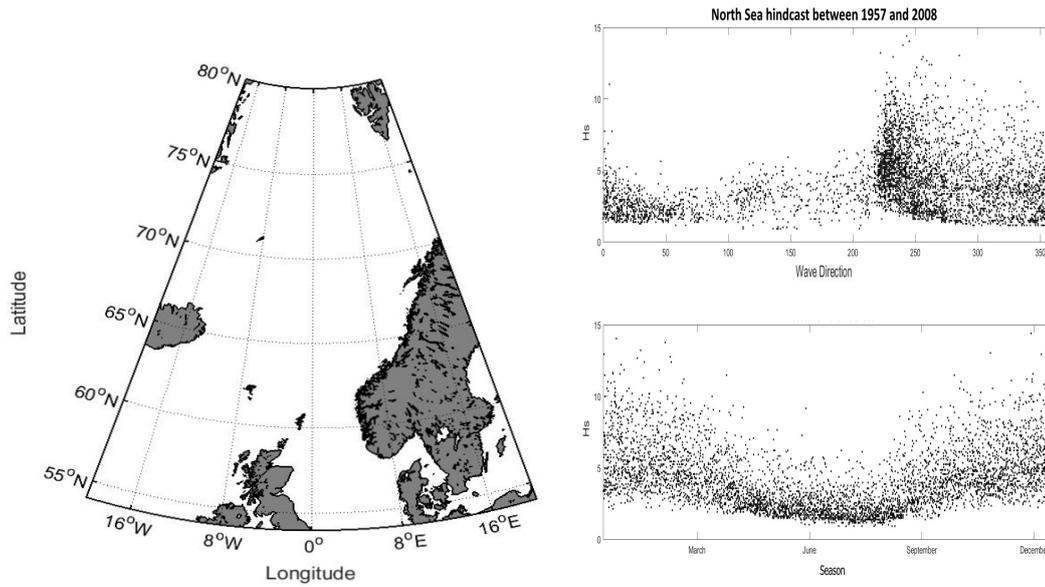
Figure 1.2.2: Map of the hindcast location in the North Sea (left panel), peak significant wave height Hs (m) on wave direction (right upper panel) and season (in day of the year, right lower panel).

## 1.3 Motivating questions

- In what ways do covariates affect extreme value peaks over threshold models? What advantages and disadvantages do the different models show in handling non-stationarity?

- Can extreme value models incorporate covariates in a flexible way that does not require ad hoc formulations for each dataset?

- How do different model formulations perform with respect to balancing flexibility, accuracy and ability to avoid overfitting?

- How do models performances differ when one or multiple covariates are considered?

## 1.4 Thesis overview

The aim of the thesis is to assess and improve inference for existing extreme value methodology in order to properly model covariate effects. To do this, first we consider different extreme value models and assess their performance on non-stationary data. Secondly, we focus on formulations for model parameters and aim to develop some approaches that capture this non-stationarity appropriately, in terms of flexibility, efficiency and accuracy. Notation is defined and consistent within each chapter, although not necessarily across different chapters.

Chapter 2 provides an overview to the existing theory and techniques this thesis builds upon. In particular the material is divided into three main sections. First, we review models from univariate extreme value theory which are used as the basis of statistical methodology in this thesis, as well as standard approaches to capture covariate effects. Next, we introduce the main notation and concepts of Bayesian inference approaches used in later chapters. Finally, we introduce the notation for basis formulations, which can be used to represent the model parameters as functions of covariates. We then provide an overview of a few specific known formulations, some of which provide the basis for the ones developed and presented in later chapters.

Chapter 3 focuses on the two known peaks over threshold methods, namely the generalised Pareto model and the non-homogeneous Poisson point process approach. The research within this chapter aims to compare the necessary adaptations and performances of these approaches in the presence of covariates, and maximum likelihood methods are used for inference. Simulation studies, as well as an application to the hindcast data from the South China Sea introduced in Section 1.2, show

how the models perform and highlight the relative advantages and disadvantages of both methods.

In Chapter 4, we first explore model parameter formulations aimed at capturing covariate effects, and in particular we focus on the case where only one covariate is considered. We review penalised spline bases and their previous application to extreme value data. We use the P-spline set-up as a benchmark model, and we then propose an alternative basis formulation consisting of a linear combination of kernels, in a way which is related to radial basis functions. Inference is performed using Markov chain Monte Carlo methods, which are applied to a fixed-dimensional basis as well as additional models where the number of basis components is allowed to change. The models are applied to simulated data, and performance is assessed based on parameter and return level estimates, as well as test statistics comparing the return level distributions to the pre-determined underlying ones of the simulated data in order to assess the accuracy of return level estimates. Further analysis is presented by applying the models to the hindcast data from the North Sea introduced in Section 1.2.

In Chapter 5, we consider model parameter formulations for two-dimensional covariates. The standard penalised spline basis approach is illustrated for higher dimensional covariates and again used as a benchmark. We then propose a tessellation formulation which yields a piecewise constant representation of the model parameters on the covariate space. We also extend the linear combination of kernels approach from Chapter 4 by considering suitable adaptations of two-dimensional kernels. Only models with a fixed number of components are considered, and Markov chain Monte Carlo methods are again used for inference. The models are

applied to a new simulation study, and performance is assessed using the same techniques and methods as in Chapter 4. Further analysis is presented by applying the models to both the South China Sea and the North Sea hindcast datasets. The chapter concludes with some observations about extending the approaches considered to higher dimensional covariates and opportunities for further work.

# Chapter 2

# Background material and

# literature review

The novel approaches and methodologies introduced in this thesis build upon theoretical results and techniques well established in the literature from different statistical fields. The reader can find here a broad survey of the most relevant concepts and recent advances, which provide a solid background for the work presented in the rest of this thesis. The review is divided into three main areas of interest, namely extreme value theory (EVT), covariate parameterisations and inference methods, where all of the above focus on univariate random variables.

## 2.1 Univariate extreme value theory

Extreme value theory is a relatively recent branch of statistics which differentiates itself from other statistical fields by focusing on the tail behaviour of probability

distributions. Starting in the middle part of last century, it was developed to better describe rare events which could not be fully and accurately characterised by existing statistical methods. In general, most data are concentrated in the body of distributions, which means that estimates such as the mean and standard deviation are typically driven by these central values. Usually, a fit to the centre of the distribution also allows many different extrapolation models for the tail regions. This uncertainty reduces the reliability of estimates of high quantiles, which are more relevant in the study of the distribution tails, and produces results that are biased by the most common values observed. It is also typical for applications focused on extreme values to require estimates for levels that are beyond the range of the current data. Extreme value theory relies on asymptotic theory to provide probabilistic results that apply specifically to the tail of distributions and disregard their body, yielding a more accurate representation of the tail behaviour and allowing this type of extrapolation.

Two main types of modelling approaches are available in the literature, depending on how univariate extreme observations are identified. In Section 2.1.1 we introduce the block maxima model, while Section 2.1.2 focuses on methods for threshold exceedances. Finally, we review extensions for non-stationary series in Section 2.1.3. We refer the reader to Leadbetter et al. (1983) and Embrechts et al. (1997) for thorough overviews of the mathematical background and probabilistic results, while Coles (2001) and Beirlant et al. (2004) provide excellent reviews of EVT applications and data analysis approaches.

### 2.1.1   Block maxima model and the generalised extreme value distribution

Let us consider a sequence $Y_1, \ldots, Y_n$ of independent and identically distributed (i.i.d.) random variables over some given temporal interval (block) with same unknown distribution function $F$ such that $Y \sim F$. If we define $M_n = \max(Y_1, \ldots, Y_n)$ to be the maximum value recorded, then it is possible to infer about the upper tail of $F$ by focusing on the statistical behaviour of $M_n$. By simple manipulation and use of the i.i.d. characteristic, we find that the distribution of these sample maxima can be expressed as

$$
\begin{aligned}
P(M_n \leqslant y) &= P(Y_1 \leqslant y, \ldots, Y_n \leqslant y) \\
&= P(Y_1 \leqslant y) \cdots P(Y_n \leqslant y) \\
&= \{F(y)\}^n.
\end{aligned}
\tag{2.1.1}
$$

Note, before we proceed any further, that the work presented in this thesis focuses on right tails of distributions, so we will use maxima notation for the remaining of this section. Nevertheless, the same approach can be applied to define minima, so that all the theory and methodology on extremes data can be applied for both tails of a distribution.

The distribution $F$ as defined in Eq. 2.1.1 is not usually known. One can instead consider families of distributions that ensure that $F^n$ converges to the tails of the distribution $F$. Nevertheless, the distribution of $M_n$ is degenerate, that is, in the

limit $n \to \infty$, it collapses to a point mass at the upper end of $F$ such that

$$M_n \to y^F, \qquad \text{where} \qquad y^F = \sup\{y : F(y) < 1\}.$$

In order to overcome this issue, we can use a linear normalisation of $M_n$ which converges to a non-degenerate limit distribution. In particular, let $a_n > 0$ and $b_n$ be some sequences of constants. Then the Unified Extremal Types Theorem (Fisher and Tippett, 1928) ensures that, for appropriate choice of normalising constants $a_n > 0$ and $b_n$ and for $n \to \infty$,

$$P\left(\frac{M_n - b_n}{a_n} \leqslant y\right) \to G(y), \tag{2.1.2}$$

where $G$ is non-degenerate and can take one of the following forms:

**Negative-Weibull**: $G(y) = \begin{cases} \exp\{-(-y)^\alpha\} & y < 0, \alpha > 0 \\ 1 & y \geqslant 0; \end{cases}$

**Gumbel**: $G(y) = \exp\{-\exp(-y)\}, \qquad -\infty < y < \infty$

**Fréchet**: $G(y) = \begin{cases} 0 & y \leqslant 0 \\ \exp\{-y^{-\alpha}\} & y > 0, \alpha > 0. \end{cases}$

Although these distributions have been used directly in the literature, it can be restrictive to choose one *a priori* to capture the tail behaviour. The Unified Extremal Types Theorem (UETT) provides a common parameterisation for these

forms, known as the Generalised Extreme Value (GEV) distribution, defined as

$$G(y) = \exp \left\{ - \left[ 1 + \xi \left( \frac{y - \mu}{\psi} \right) \right]_+^{-1/\xi} \right\}, \qquad (2.1.3)$$

where $[\cdot]_+ = \max\{\cdot, 0\}$, and is defined through three parameters, namely a location $\mu$, scale $\psi$ and shape $\xi$, such that $G \equiv GEV(\mu, \psi, \xi)$, where $\psi > 0$. The GEV distribution comprises all the possible limit distribution of $M_n$, and in particular we can link it to the known three formulations from before using the shape as a distinguishing feature. More precisely, if $\xi < 0$, then $G$ will be a **Negative-Weibull** distribution with a finite upper end point, whereas $\xi > 0$ indicates a **Fréchet** distribution with a heavy upper tail, and $\xi = 0$ denotes a **Gumbel** distribution with exponential upper tail.

A detailed proof of the theorem is omitted here, although the reader may find an intuitive justification in Coles (2001), while Leadbetter et al. (1983) provides a more formal overview.

In practice, in order to use the UETT to derive a statistical model for extreme data, we need to first separate this data into $k$ smaller preselected blocks, such as by years, with $n$ observations. It is important to notice that this is an asymptotic result. In other words, it relies on the assumption that the limit in Eq. 2.1.2 holds for some finite value of $n$ as long as $k$ and $n$ are chosen carefully. In particular, $n$ should set as high as possible to ensure sufficiently many observations are grouped before their maximum is computed. Nevertheless, sufficiently many independent maxima are required in order to perform inference, such that $k$ needs to be set as high as possible too, leading to a trade-off. One should notice that the block length

$n$ is given by the context in many practical applications, such as daily, monthly or yearly maxima for environmental applications. Then, one can use some inference methods, such as moments based approaches, maximum likelihood or Bayesian inference, to estimate the three model parameters $(\mu, \psi, \xi)$.

**Prediction**

Once a distribution $G$ has been fitted to the data, the information obtained needs to be used for prediction. As mentioned in Chapter 1, the concept of return levels is the main tool to perform this task. More specifically, the $N$-year return level $z_N$ is the $(1 - 1/N)$ quantile of $F$, i.e. $1 - F(z_N) = 1/N$, and indicates that an event exceeding such a level is expected to occur on average once every $N$ years. In the case where $F$ can be modelled by a $\text{GEV}(\mu, \psi, \xi)$ distribution, the return levels are given by:

$$
z_N = \begin{cases} \mu - \dfrac{\psi}{\xi} \left[ 1 - \left\{ -\log\left(1 - \dfrac{1}{N}\right) \right\}^{-\xi} \right] & \text{for } \xi \neq 0, \\[3mm] \mu - \psi \log\left\{ -\log\left(1 - \dfrac{1}{N}\right) \right\} & \text{for } \xi = 0. \end{cases} \tag{2.1.4}
$$

Different inference methods can be used to obtain estimates of the return values. A common approach is to use maximum likelihood to obtain estimates $(\widehat{\mu}, \widehat{\psi}, \widehat{\xi})$ of the GEV parameters. These can then be substituted in Eq. 2.1.4 to obtain the corresponding return level estimates $\widehat{z}_N$. This is the approach used, for peaks over threshold methods in Chapter 3, while in Chapter 4 we review the estimation of return levels in a Bayesian framework.

While this chapter provides an introduction to the GEV model, given its essential

role in the development of extreme value theory, it is not used in the work presented in the next chapters. For this reason, we omit any further details on this model.

## 2.1.2 Threshold exceedances and the generalised Pareto distribution

As mentioned in Chapter 1, extreme datasets are, by definition, limited in size. As a consequence, statistical modelling and analysis can present additional challenges given the limited information available. Hence, preserving as much data as possible can be desirable, although care should be taken to ensure only observations from the tail are considered, as further discussed further below. The block maxima approach can sometimes be wasteful as it disregards all but the most extreme observations in each block, even if more than one large value is present. Hence, after the selection of block maxima, there are often unselected observations that lie between the smallest and largest block maxima recorded in the data. These tail values, despite being more extreme than some of the block maxima considered, are ignored by the block maxima approach. A more efficient alternative is instead to define as extreme those observations that exceed a sufficiently high threshold $u$. It is then possible to obtain a similar result for the threshold exceedances to the Extremal Types Theorem for block maxima, and develop what is known as a *peaks over threshold* (POT) model.

Assume, as before, that the observations $Y_1, \ldots, Y_n$ are i.i.d. random variables from the same unknown distribution function $F$. Let us also assume that $F$ is in the domain of attraction of the GEV distribution $G \sim \text{GEV}(0, 1, \xi)$, with zero

location parameter, unit scale and normalising constants $a_n > 0$ and $b_n$, such that the asymptotic results in Eq. 2.1.2 and 2.1.3 hold. Then we can think of the (normalised) observations $Y_1, \ldots, Y_n$, $Y_i > u \; \forall i$, as realisations of a point process, where the limiting process leads to an asymptotic model for the upper tail of $F$. More specifically, we can then construct the non-homogeneous point processes $P_1, P_2, \ldots$ on $[0, 1] \times \mathbb{R}$, where

$$P_n = \left\{ \left( \frac{i}{n+1}, \frac{Y_i - b_n}{a_n} \right); i = 1, \ldots, n \right\} \longrightarrow P, \text{ as } n \to \infty$$

can be shown to converge to a non-homogeneous Poisson point process $P$ (Pickands, 1975) with intensity

$$\lambda(t, y) = (1 + \xi y)_+^{-1 - 1/\xi}. \tag{2.1.5}$$

Note that this limiting process is non-degenerate, as it retains the large observations, while smaller points are all normalised to the same value

$$b_l = \lim_{n \to \infty} (y_F - b_n)/a_n,$$

where $b_l \in \mathbb{R}$. To use this result as a basis for a statistical model for data with a finite sample size, a suitable threshold must first be chosen, such that the validity of the asymptotic argument is preserved. One can then use the NHPP result to obtain a conditional model of the distribution of threshold exceedances. In particular, given an appropriate threshold $v > b_l, v \in \mathbb{R}$, we can study the Poisson

point process on the set $B_v = [0,1] \times [v, \infty)$, where the integrated intensity

$$\Lambda(B_v) = \int_0^1 \lambda(v,t)\, dt = (1 + \xi v)_+^{-1/\xi} \tag{2.1.6}$$

represents the expected number of points in $B_v$.

Now let $u_n(v) = a_n v + b_n$ be the threshold on the original scale. Given some $y > 0$,

$$
\begin{aligned}
\Pr(Y_i > a_n y + u_n(v) | Y_i > u_n(v)) \quad &= \quad \Pr\left(\frac{Y_i - b_n}{a_n} > y + v \,\Big|\, \frac{Y_i - b_n}{a_n} > v\right) \tag{2.1.7}\\[2mm]
&= \quad \Pr(\text{a given point in } P_n > y + v \,|\, \text{a given point in } P_n > v)\\[2mm]
&\xrightarrow{n\to\infty} \quad \Pr(\text{a given point in } P > y + v \,|\, \text{a given point in } P > v)\\[2mm]
&= \quad \frac{[1 + \xi(y+v)]_+^{-1/\xi}}{[1 + \xi v]_+^{-1/\xi}}\\[2mm]
&= \quad \left[1 + \xi\left(\frac{y}{1 + \xi v}\right)\right]_+^{-1/\xi}\\[2mm]
&= \quad \left[1 + \xi\frac{y}{\psi_v}\right]_+^{-1/\xi},
\end{aligned}
$$

which is the survival function of the generalised Pareto distribution (GPD), with $\psi_v = 1 + \xi v$. It follows that

$$a_n^{-1}[Y_i - u_n(v)]_+ \,|\, Y_i > u_n(v) \to \text{GPD}(\psi_v, \xi).$$

The result is only exact in the limit, as we need to consider $n \to \infty$ for the third step of Eq. 2.1.7 to hold. In reality, though, we only have finite samples of size $n < \infty$. Hence, we need to shift our focus from the limiting case to a finite sample with unknown normalising constants $a_n$ and $b_n$. Letting $Y_u$ be the excess variable over some sufficiently high threshold $u$, i.e. $Y_u = (Y - u)_+$, we can then

approximate the asymptotic result, for $y > 0$, by

$$Pr(Y_u > y | Y_u > 0) = \left[1 + \xi \frac{y}{\psi_u}\right]_+^{-1/\xi}, \tag{2.1.8}$$

where the scale parameter accounts for the normalising constants $a_n$ and $b_n$. The variable $Y_u | Y_u > 0$ is then said to follow a $GPD(\psi_u, \xi)$. Note that the value of the scale parameter $\psi_u$ is threshold dependent, but can be easily converted to the corresponding scale for a higher threshold $v > u$, by using

$$\psi_v = \psi_u + \xi(v - u). \tag{2.1.9}$$

**Threshold estimation and uncertainty**

The choice of a suitable threshold $u$ is essential for threshold exceedance models. If $u$ is too low, one might include too many observations that are not "extreme", and the asymptotic limiting distribution does not adequately describe the data. Nevertheless, it is advisable to choose as low a threshold as possible to maximise the amount of observations used, obtain a more accurate inference and reduce the uncertainty in the parameter estimates. This trade-off introduces some subjectivity, although some diagnostics are available to help in the choice. We briefly overview the two most common diagnostics, which will be used in Chapters 3 and 4.

The mean residual life (MRL) plot tests the stability of the expected value of the distribution for a range of increasing thresholds. In particular, if we assume a set of threshold exceedances $Y_i | Y_i > u, i = 1, \ldots, m$ follow a generalised Pareto

$(\psi_u, \xi)$ distribution, then the expected value of the threshold excesses $Y - u$ can be written as

$$E[Y - u | Y > u] = \frac{\psi_u}{1 - \xi},$$

for $\xi < 1$. If one then considers a higher threshold $v$, such that $v > u$, then the expectation of this subset of excesses would be

$$E[Y - v | Y > v] = \frac{\psi_u + \xi(v - u)}{1 - \xi},$$

for $\xi < 1$. In other words, if the original threshold $u$ is large enough, then the mean excesses for higher thresholds should be linear in $v$ for all $v > u$. One can then use graphical methods such as the mean residual life plot to assess this stability. From this type of plot, as shown in Chapter 3, one can visually choose a threshold above which the mean residuals and confidence bounds stabilise - small threshold values usually show narrow bounds and an average behaviour that is still changing, while very high thresholds exhibit an erratic behaviour due to the lack of data.

A different diagnostic applies a similar argument to the parameter estimates, which should be stable for a range of increasing threshold values. Let us assume again that, for a given threshold $u$, $Y$ follows a generalised Pareto $(\psi_u, \xi)$ distribution. Then, for any higher threshold $v > u$, we know from Eq. 2.1.9 that

$$Y - v | Y > v \sim GPD(\psi_u + \xi(v - u), \xi).$$

While the scale of the distribution changes for increasing thresholds, we know from the theory that the shape parameter is threshold invariant and hence the parameter estimates should be constant. With the use of reparameterisation, we

can obtain a scale formulation that is also stable over changing thresholds, where this modified scale is given by $\psi^* = \psi_u - \xi v$. A suitable threshold would then retain as much data as possible while preserving parameter stability. That is, it would be the lowest possible value of $u$ above which the estimates of the modified scale and shape parameter remain constant.

While traditionally the threshold is selected before analysis and treated as fixed and known, some work has been done to include the estimation of the threshold and its uncertainty in the inference mechanism. Wadsworth and Tawn (2012) propose a way to account for the uncertainty due to the intrinsic subjectivity for a choice of a threshold. We have just reviewed how a classic approach to pick a threshold uses the stability of the model parameters. Following the same logic, Wadsworth and Tawn (2012) focus on a sub-asymptotic form of extreme value theory (Smith, 1987). The idea is to incorporate the uncertainty due to the choice of threshold in the model, and the authors then provide a likelihood ratio testing procedure for the threshold selection. In particular, second order asymptotic arguments are used to increase the flexibility in the fit of the GPD.

The models above all focus solely on data from the tail of the distribution and are based on the asymptotic theory behind extreme value models. Recently, a different approach has emerged, with authors advocating that all the data can be informative for the choice of threshold. These models are known as "whole-body" approaches and often consist of two parametric components, namely a model for the bulk distribution below the threshold and a POT model above it. These two different components are then joined together at the threshold, which is itself treated as an unknown parameter. In this way, the estimation of the threshold

exploits all of the data, so as to provide a more accurate result, and all uncertainties are accounted for by the Bayesian inference. A noticeable example of this two-component parametric approach can be found in the work by Behrens et al. (2004). More recent work in the area is presented in MacDonald et al. (2011), where a flexible extreme value mixture model is proposed. This model aims to use as much of the information available as possible by combining a non-parametric kernel density estimator for the bulk of the distribution with a GPD model for the tail.

**Prediction and return levels**

In Section 2.1.1 we introduced the mathematical definition of a return level and an approach to prediction for the block maxima model. Using the GPD approach, the result obtained is a conditional distribution for the upper tail, and the conditioning needs to be undone to obtain a marginal model for the observations $Y$. To this end, the proportion of observations which exceed the threshold $u$ can be used as a rate of exceedance parameter $\phi_u = P(Y > u)$. When a GPD is used to conditionally model $F$, then the $N$-year return level $z_N$, given $n_y$ observations per year, is

$$
z_N = \begin{cases} u + \psi_u \log(n_y N \phi_u) & \text{if} \quad \xi = 0, \\[2mm] u + \frac{\psi_u}{\xi} \left[ (n_y N \phi_u)^\xi - 1 \right] & \text{otherwise.} \end{cases} \tag{2.1.10}
$$

As mentioned before in Section 2.1.1, one can use different inference methods to obtain estimates for $(\psi_u, \xi)$. Then, for stationary processes, we can approximate

the probability of exceedance as the proportion

$$\phi_u = \frac{\mathbb{E}\left[N_{\text{exc}}\right]}{N_{\text{obs}}} \approx \frac{N_{\text{exc}}}{N_{\text{obs}}}, \qquad (2.1.11)$$

where $N_{\text{obs}}$ is the total number of observations.

For the NHPP approach, there is no closed form formulation for computing return levels. Instead, one needs to move to either a GEV (following the methodology in Section 3.1.1) or a GPD framework. In view of work presented in later sections, we prefer using the latter. Hence, to obtain return levels $z_n$ for a NHPP, we first need to compute the GPD scale as $\psi_u = \sigma + \xi(u - \mu)$. Then, we can exploit the fact that the expected number of exceedances can be approximated by the integrated intensity $\Lambda(B_u)$ over $B_u = [0, 1] \times [u, \infty)$, and get

$$\phi_u = \frac{\mathbb{E}\left[N_{\text{exc}}\right]}{N_{\text{obs}}} \approx \frac{\Lambda(B_u)}{N_{\text{obs}}}, \qquad (2.1.12)$$

where $\Lambda(B_u)$ is given by Eq. 2.1.6.

### 2.1.3   Accounting for non-stationarity

The theoretical results and statistical models discussed so far rely on the assumption of independent and identically distributed observations. In reality this is rarely the case and both the independence and the identically distributed assumptions may fail. In particular, extreme observations often display signs of a variety of trends. This is especially common in environmental datasets which are affected by a variety of climate processes, with different fluctuating behaviours, as is the case

for the data we introduced in Chapter 1. Due to the physical process by which waves are generated, these generally have a strong seasonal pattern, with an annual periodicity; decadal or semi-decadal climate *phenomena* can also account for large scale variations. For a specific location, we can also detect variability with respect to wave direction; for example, wind is typically stronger from some directions than others, and, together with fetch and water depth, can strongly influence the resulting magnitude of the waves. Further, location itself can be an important factor. For example, a more exposed location will be associated with longer fetches, resulting in a more extreme wave climate.

In a similar way to generalised linear modelling, the standard approach to account for non-stationarity is to build an extreme value model where one or more of the parameters of the statistical distribution considered are described as a function of the covariate (Smith, 1989). These functions can take different forms; amongst the parametric approaches, Davison and Smith (1990) introduce the simplest formulation, where the generalised Pareto distribution parameters are linear functions of covariates, with an exponential link function to ensure positivity of the scale parameter. Note that, while the methodology is presented for threshold exceedances, since these are the type of data analysed in later chapters, the same approach can be applied to block maxima.

Let us consider a process $\{Y_t\}$, with $t = 1, \ldots, n$ and $Y_t \in \mathbb{R}$, and let $\{\boldsymbol{X}_t\}$, $\boldsymbol{X}_t \in \mathbb{R}^m$, be the corresponding $m$ covariates. For example, the response variable may be significant wave height with wave direction as the covariate.

Smith (1989) and Davison and Smith (1990) provide the first full extension of the GPD to the non-stationary case. First, they deal with short-term dependence

using the extremal index to select a mean cluster size and identify cluster maxima (Leadbetter et al., 1983; Leadbetter and Rootzen, 1988). Note that the data is then assumed to have no temporal dependence, and the reader should refer to the literature for more details on methods to check and, if necessary, account for this (see Coles 2001 Section 5.3.2. and Ferro and Segers 2003). Then they apply the peaks over threshold model to these maxima, and introduce non-stationarity in the parameters, which are now modelled as functions of covariates. It is then possible to model the excesses as $\mathrm{GPD}(\psi_u(\boldsymbol{x}), \xi(\boldsymbol{x}))$ with probability of exceedance $\phi_u(\boldsymbol{x}) = \Pr(Y > u | \boldsymbol{X} = \boldsymbol{x})$.

Then, for $y > 0$, we obtain the non-stationary version of Eq. 2.1.8 as

$$\Pr(Y > y + u \mid Y > u, \boldsymbol{X} = \boldsymbol{x}) = \left[1 + \frac{\xi(\boldsymbol{x})y}{\psi_u(\boldsymbol{x})}\right]_+^{-1/\xi(\boldsymbol{x})}.$$

Now let us represent the covariate dependence using some known function $f$ such that, for example, $\phi_u(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{\phi_u})$, where $\boldsymbol{\phi_u}$ is a vector of coefficients for the covariate contribution to the parameter, and we denote the vector of all the unknown coefficients as $\boldsymbol{\theta} = (\boldsymbol{\phi_u}, \boldsymbol{\psi_u}, \boldsymbol{\xi})$. Then, the likelihood function for $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n} (1 - \phi_u(\boldsymbol{x}_t))^{1 - I[y_t > u]} \left[\phi_u(\boldsymbol{x}_t) \frac{1}{\psi_u(\boldsymbol{x}_t)} \left(1 + \xi(\boldsymbol{x}_t) \frac{y_t - u}{\psi_u(\boldsymbol{x}_t)}\right)_+^{-1 - 1/\xi(\boldsymbol{x}_t)}\right]^{I[y_t > u]}.$$

Both Smith (1989) and Davison and Smith (1990) propose a linear model for the parameters, with

$$\log \psi_u(\boldsymbol{x}) = \boldsymbol{\psi_u}' \boldsymbol{x}, \ \xi(\boldsymbol{x}) = \boldsymbol{\xi}' \boldsymbol{x}, \ \mathrm{logit}\, \phi_u(\boldsymbol{x}) = \boldsymbol{\phi_u}' \boldsymbol{x}.$$

Other parametric models, such as Fourier parameterisation, have also been used, especially in the presence of one or more periodic covariates, see for example Coles and Walshaw (1994) who model extreme wind speeds. An alternative approach to the use of link functions would be to constrain the parameter space for the likelihood optimisation algorithms, to match the support of the distribution. There are some definite advantages to parametric models: they are easy to fit, given the low number of unknown model parameters, and it is relatively straightforward to use them for return level prediction. Nonetheless, there are also considerable limitations. First of all, the choice of an appropriate functional form is often not simple and open for debate. Furthermore, these models might not manage to fully capture more "local" behaviour and over-simplify the relationship between the covariate and the response: a parametric formulation is often smoother than what would be considered reasonable from a visual inspection of the data, and its lack of local flexibility might provide poorly-fitting models even when large numbers of parameters are used. One alternative is to use a local likelihood technique, as presented by both Davison and Ramesh (2000) and Hall and Tajvidi (2000), while Chavez-Demoulin (1999) and Pauli and Coles (2001) propose penalised semi-parametric formulations for the peaks over threshold and the block maxima approach respectively. Semi-parametric and non-parametric techniques model the parameters as smooth functions of the covariates, potentially capturing local features better than parametric models. The seminal work by Chavez-Demoulin and Davison (2005) focuses, in particular, on the use of spline smoothers to parametrise the relationship between the covariate and model parameters, while Yee and Stephenson (2007) propose vector generalized additive models (VGAM) models. The reader

should refer to Chavez-Demoulin and Davison (2012) for a general review on non-stationary methods for extremes, while more details on different parameterisations are provided in Section 2.3.

## 2.2 Inference procedures

Different estimation methods for extreme value models are available, including moment based methods, such as probability-weighted moment estimation and L-moments (Hosking, 1990), maximum likelihood (ML) and Bayesian inference. Hosking et al. (1985) show that the L-moments approach can yield better performance than maximum likelihood methods when the sample size is small. Moments-based methods are in general conceptually and practically simple, and these factors, together with their good performance for small samples, have contributed to their popularity in geoscience fields such as hydrology. Nevertheless, moment-based methods are difficult to apply beyond the i.i.d. case, that is when a covariate trend is present such as for the type of data considered in this thesis, while maximum likelihood methods can be easily extended to this case (Katz et al., 2002). Coles (1999) also criticize the L-moments method since it requires an *a priori* assumption that the shape $\xi < 1$, and they suggest that the issue of the ML estimates with small samples can be improved by using a penalized likelihood method. While maximum likelihood methods have been widely used in the literature, Bayesian inference is an increasingly popular alternative, especially with the development of Markov chain Monte Carlo (MCMC) algorithms. Bayesian methods can be easily applied to non-stationary series and show some additional advantages when com-

pared to ML techniques. Coles and Simiu (2003) argue the importance of properly accounting for the uncertainty in an extreme value model to produce accurate prediction, and they suggest it can be better accomplished using a fully Bayesian framework. A possible advantage of Bayesian inference is the option of using prior information to exploit experts' knowledge; while this can aid model fitting in situations where data is sparse, as is the case for extremes observations, in this work we opt to use uninformative priors which allow the algorithms to explore the estimated posterior density freely. There are various noticeable contributions to the literature of fitting extremes with Bayesian inference, from earlier publications of Coles and Powell (1996) and Coles and Tawn (1996) to more recent work by de Zea Bermudez and Turkman (2003), Beirlant et al. (2004) and Cabras et al. (2011). For a more general introduction to Bayesian techniques, we refer the reader to the popular monograph by Casella and Berger (2002). Brooks (1998) and Gilks et al. (1996) provide excellent introductions to Markov chain Monte Carlo methods in particular, while more advanced details on performance and convergence of these algorithms can be found, amongst others, in the works by Roberts et al. (1997), Roberts and Rosenthal (1998) and Roberts et al. (2001).

## 2.2.1 Bayesian model and Markov chain Monte Carlo Markov techniques

Suppose we are interested in some variable $Y$, and we have a model that can be parametrised by a vector of unknown parameters $\boldsymbol{\theta} \in \Theta$. Bayesian inference allows the inclusion of prior beliefs about $\boldsymbol{\theta}$ for a given parametric distribution $f(Y; \boldsymbol{\theta})$.

These beliefs can be expressed in what is known as a prior distribution $\pi(\boldsymbol{\theta})$. Then Bayes theorem states that

$$p(\boldsymbol{\theta}|Y) = \frac{\pi(\boldsymbol{\theta})f(Y|\boldsymbol{\theta})}{\int_{\Theta} \pi(\boldsymbol{\theta})f(Y|\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{2.2.1}$$

where and $p(\boldsymbol{\theta}|Y)$ is known as the posterior. Then we can use this result to perform inference on the parameters via the posterior distribution. It is worth noting that the normalising integral in Eq. 2.2.1 makes the computation of the posterior harder. In some cases, "conjugate" prior distributions can be chosen so that probabilistic properties can be exploited to avoid the need to calculate the integral, but most often a closed form solution for the posterior is not available. A major revolution in Bayesian inference was marked by the development of Markov chain Monte Carlo (MCMC) techniques, which have enabled the use of Bayesian inference where it could not be used before. As an example, consider cases with a high number of unknown parameters, where a closed form solution to the integral is not available and numerical approximation would be too computationally intensive. In particular, MCMC techniques provide a framework that allows the estimation of the posterior when other approaches, such as conjugacy or other sampling techniques e.g. importance sampling, are not suitable. In all Bayesian inference, one can obtain more exhaustive results than the point estimates, their asymptotic distribution and marginal confidence intervals yielded by maximum likelihood methods. Similar summaries to those computed via maximum likelihood, such as posterior mean and credible intervals, can still be obtained. Bayesian inference can provide a more complete picture, as it incorporates both the uncertainty in the parameter and that stemming from sampling, while only the latter is accounted for in maxi-

mum likelihood estimates.

In Sections 2.2.2 and 2.2.3 we present the main MCMC procedures we use in the upcoming chapters, while Section 2.2.4 introduces an extension of MCMC techniques where the size of the unknown parameter vector $\boldsymbol{\theta}$ changes between iterations.

## 2.2.2 Metropolis-Hastings algorithm

The Metropolis sampler (Metropolis et al., 1953) is one of the most general MCMC technique and allows one to work on distributions that are difficult to sample from, such as complex posteriors, where other sampling schemes such as importance or Gibbs sampling might not be suitable. In fact, this method works on the assumption that the target distribution $p(\boldsymbol{\theta}|y)$ is only known up to a constant of proportionality. The main advantage of this method is that it does not require knowledge of such constants of proportionality in order for sampling to be performed, so that the normalising integral need not be computed. Starting with an initial value $\boldsymbol{\theta}^0$, the sampler requires a candidate point to be sampled from a proposal distribution $q(\cdot|\boldsymbol{\theta}^{(i-1)})$, which is dependent on only the previous point $\boldsymbol{\theta}^{(i-1)}$. While the original algorithm called for the proposal density $q$ to be symmetric, Hastings (1970) proposed a generalisation of the algorithm, known as the Metropolis-Hastings sampler (MH), which could use non-symmetric proposals. If we denote the target distribution as $p(\cdot)$ and the proposal as $q(\cdot)$, the resulting approach is shown in Algorithm 1.

In practice, the proposal distribution is frequently chosen to be symmetric. This

---

**Algorithm 1** Metropolis-Hastings algorithm

---

1: **Initialize:**
   $\boldsymbol{\theta}$ as some value $\boldsymbol{\theta}^0$ in the support of the target $p(\boldsymbol{\theta}|y)$
2: **for** i=1, ..., M **do**
3:     Generate a proposal $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})$;
4:     Calculate the acceptance probability

$$a(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*) = \min\left\{ \frac{p(\boldsymbol{\theta}^*|y)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})}{p(\boldsymbol{\theta}^{(i-1)}|y)q(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^*)}, 1 \right\}. \qquad (2.2.2)$$

5:     Draw $u$ from a uniform distribution of a range $[0, 1]$
6:     **if** $a > u$ **then**
7:         Accept: $\boldsymbol{\theta}^{(i+1)} \leftarrow \boldsymbol{\theta}^*$
8:     **else**
9:         Reject: $\boldsymbol{\theta}^{(i+1)} \leftarrow \boldsymbol{\theta}^{(i-1)}$
10:    **end if**
11: **end for**

---

popular version of the algorithm is called a random walk Metropolis (RWM). This is a special case of the MH algorithm above, where at each iteration the update is centred around the current value and from there a symmetric proposal is made, such that $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)}) = q(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^*)$. Then, the resulting acceptance probability simplifies to

$$a(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*) = \min\left\{ \frac{p(\boldsymbol{\theta}^*|y)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})}{p(\boldsymbol{\theta}^{(i-1)}|y)q(\boldsymbol{\theta}^{(i-1)}|\boldsymbol{\theta}^*)}, 1 \right\} = \min\left\{ \frac{p(\boldsymbol{\theta}^*|y)}{p(\boldsymbol{\theta}^{(i-1)}|y)}, 1 \right\}.$$

A typical choice for $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})$ is such that

$$\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)} \sim N(\boldsymbol{\theta}^{(i-1)}, \lambda^2\Sigma), \qquad (2.2.3)$$

where the matrix $\Sigma$ is often chosen in an attempt to match the correlation structure of the target distribution, for example by approximating such correlation from some initial parameter estimates or *a priori* knowledge, or it can simply be taken as the identity matrix. The tuning parameter $\lambda$, commonly referred to as the stepsize,

is the only other user-specific input required. Choosing such a stepsize is a balance between proposing large enough jumps to ensure fast convergence, and ensuring that a reasonable proportion of the proposed parameter updates are accepted, as values that are too different from the current are more likely to lead to a worse model and be rejected. The random walk Metropolis (RWM) is often referred to as a blind algorithm since the proposal density, in principle, does not depend on the target distribution. The algorithm also requires manual tuning of the proposal covariance and can show high autocorrelation in the parameter estimates, with resulting mixing issues in the chain, which can yield slow convergence to the target as well as less reliable estimates of the uncertainty. On the other hand, the RWM algorithm is relatively simple to implement, as it does not require any in depth knowledge of correlation and surface information of the target distribution. Furthermore, it is easily extendible to any dimension of the unknown parameter vector $\boldsymbol{\theta}$, as one can either propose a joint update or iteratively consider each single parameter separately. The latter option can, in particular, aid mixing in high-dimensional cases: randomly proposing multiple parameter updates can lead to very low acceptance rates, or require very small stepsizes and hence suffer from slow convergence. By considering each single unknown parameter in turn, both issues can be avoided, albeit requiring more computations. A different improvement can be obtained by first running a short chain using the RWM algorithm, using the parameter chains to estimate the correlation matrix $\Sigma$ for the proposal distribution, and then running a longer chain with this informed MH algorithm. A large number of extensions have been developed and in the next Section we focus on those used in the rest of this thesis.

## 2.2.3 Derivative-based MCMC algorithms

There are different ways of improving the proposal distribution, and in this section we focus on what are known as derivative-driven MCMC algorithms, which use a similar approach to gradient ascent methods to optimise the parameter estimates in maximum likelihood. For these techniques, information in the form of the gradient of the log-posterior is used in the proposal mechanism. This allows us to explore the parameter space in a more informed manner, yielding higher acceptance rate and faster convergence. The simplest of these approaches use the first derivative of the target distribution to produce proposals towards the centre of its domain. In this thesis, we use in particular the Metropolis-adjusted Langevin algorithm (MALA). This technique is based on approximation methods to discretise the stochastic differential equation in a Langevin diffusion process. Roberts and Stramer (2002) construct a proposal $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})$ through an Euler–Maruyama discretisation of the Langevin diffusion equation and use it as a candidate kernel in a Metropolis–Hastings algorithm. The resulting proposal $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})$ is formulated such that

$$\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)} \sim N\left(\boldsymbol{\theta}^{(i-1)} + \frac{\lambda^2}{2}\nabla\log p(\boldsymbol{\theta}^{(i-1)}|y), \lambda^2 I\right) \qquad (2.2.4)$$

where $\lambda$ is some tuning parameter. Comparing Eq. 2.2.4 to the random walk proposal in Eq. 2.2.3, the Langevin proposal has the additional deterministic term $\frac{\lambda}{2}\nabla\log p(\boldsymbol{\theta}^{(i-1)}|y)$, which is also known as the drift term. Essentially, the drift term moves the current state $\boldsymbol{\theta}^*$ to a point with higher posterior density, provided that the stepsize $\lambda$ is sufficiently small. It is worth noticing that typically, if $\lambda >> \lambda^2$ (e.g. $\lambda << 1$), then the random part of the proposal will dominate and vice versa,

though this also depends on the form of derivative of the log-posterior. Although the algorithm construction often implies markedly faster mixing and convergence than RWM (Roberts et al., 2001), the performance of MALA can sometimes be frustrating for the practitioner and it is still critical to find a good proposal covariance when using this method (Roberts et al., 2001). Girolami and Calderhead (2011) developed a modification of the Metropolis proposal mechanism in which the steps are made using a Riemannian metric instead of the standard Euclidean distance. This procedure is referred to as Riemann manifold MALA (mMALA). In this technique, both the mean vector and covariance matrix in the proposal distribution depend on the current state of the Markov chain, although we refer the reader to Girolami and Calderhead (2011) for further details. While the proposal distribution assumes a complex form in the original mMALA, one can assume constant curvature of the target distribution to obtain a simplified mechanism (Girolami and Calderhead, 2011; Xifara et al., 2014). Then the proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})$ is formulated such that

$$\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)} \sim N\left(\boldsymbol{\theta}^{(i-1)} + \frac{\lambda^2}{2}G^{-1}(\boldsymbol{\theta}^{(i-1)})\nabla \log p(\boldsymbol{\theta}|y), \lambda^2 G^{-1}(\boldsymbol{\theta}^{(i-1)})\right) \qquad (2.2.5)$$

where $G(\boldsymbol{\theta})$ is some metric tensor of choice, for example based on e metric on the Fisher Information. In this simplified version of the mMALA, the state-dependent covariance matrix in the proposal mechanism still allows adaptation to the local curvature of the target distribution depending on the current state (Xifara et al., 2014). There are different options for the tensor metric: in Chapters 4 and 5 we adopt this algorithm as it produced the most accurate estimates and we choose to use the expectation of the second-order derivative of the log target, which again

equates to maximum likelihood techniques using second-order derivative information. Note also that, if a constant metric tensor such that $G(\boldsymbol{\theta}) = cI$, for some scalar $c$, then the mMALA reduces to a MALA algorithm with scaled matrix $G$. One of the main potential issues with this techniques lies in the choice of metric tensor, and in obtaining analytical expressions and stable estimates for both the tensor and its derivatives. Secondly, as shown in Eq. 2.2.5, it is necessary to invert the metric tensor matrix, which can be computationally intensive and potentially unstable for high dimensional problems. Despite these issues, both versions of the mMALA algorithm have proposal mechanism with automated tuning that depends on the current state of the chain. In fact, by exploiting the natural Riemann structure of the parameter space of statistical models, they have the potential to overcome tuning issues of simpler MCMC techniques, which is particularly useful when the unknown parameter vector is of high dimension and exhibits strong correlations. The mixing and convergence of these algorithms can then be considerably better for target densities with complex geometries and behaviours.

### 2.2.4   MCMC methods with dimension-changing state space

Green (1995) introduced one of the most significant recent development of MCMC methods to enable these methods to be applicable to statistical problems where the number of unknowns is itself one of the unknowns. These methods were defined as reversible jump Markov chain Monte Carlo (RJ MCMC) and can be viewed as an extension of the Metropolis-Hastings algorithm onto a more general state space. Nevertheless, these models are often regarded as difficult to implement. In

the original paper, Green (1995) use measure theory to demonstrate the method's validity, although knowledge of this field of probability is not required to apply this technique to real problems. While we refer the reader to the original paper for details on the derivation of the method and its foundations, we present here a more implementation-based review that tries to link the model back to the MH algorithm. More practically-based introductions to the topic can be found in Richardson and Green (1997), Tierney (1998) and Green (2003).

As with the standard Metropolis-Hastings, reversible jump Markov chains use the detailed balance conditions to implement time-reversible Markov chains (Hastings, 1970). Although we omit in depth detail, this condition essentially guarantees that moves from state $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$ are made as often as moves from $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}$, which ensures that the chain will converge to the desired target distribution. As mentioned before, extending the Metropolis-Hastings algorithm to cases where the dimension of the parameter vector changes between iterations presents theoretical challenges and resorts to measure theory. Nevertheless, the resulting algorithm is relatively straightforward. Suppose, as before, that we have some observed data $y$, and that we have a countable collection of candidate models $M = M_1, M_2, \ldots$, where the models are indexed by some parameter $k \in \mathcal{K}$. We can consider the index $k$ as an auxiliary model indicator variable, such that $M_{k*}$ denotes the model with $k = k^*$. As before, we have vectors of unknown parameters, although now each model $M_k$ has an $n_k$-dimensional vector of unknowns $\boldsymbol{\theta}_k \in \Theta^{n_k}$ , where $n_k$ can take different values for different models. As with the standard Metropolis-Hastings algorithm, on some iteration $i$, we have some current state, which now includes the information of the state space dimension, such that $\boldsymbol{\theta}^{(i-1)} = (k, \boldsymbol{\theta}_k^{(i-1)})$ in model $M_{k^{(i-1)}}$.

Markov chain transitions are then performed by first proposing $\boldsymbol{\theta}^* = (k^*, \boldsymbol{\theta}^*_{k^*})$ in model $M_{k^*}$ from a proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})$. Then, the standard MCMC detailed balance condition is enforced through the acceptance probability which is calculated using Eq. 2.2.2 as for the general MH algorithm. This probability needs to be properly adapted to account for the change in dimensionality in the state space. First, consider a set of move types (e.g. birth, death, etc.), which present the essential property of being reversible and can be denoted by $m$. One approach to implementing such moves is by drawing a vector of continuous random variables $u$, independent of $\boldsymbol{\theta}$, and setting $\boldsymbol{\theta}^*$ by using an invertible deterministic function $h(\boldsymbol{\theta}, u)$. The reverse of the move (from $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}^{(i-1)}$) can then be accomplished by using the inverse transformation, so that the proposal is deterministic. Green and Silverman (1994) show that, if we consider the trans-dimensional scenario with some move type $m$ from a lower to a higher dimensional state space, such that $n_{k^*} > n_{k^{(i-1)}}$, Eq. 2.2.2 can be seen as

$$a_m(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*) = \min\{1, A\}, \qquad \text{where} \qquad A = \frac{p(\boldsymbol{\theta}^*|y)r_m(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(i-1)}|y)r_m(\boldsymbol{\theta}^{(i-1)})q(u)}J,$$

where $r_m(\boldsymbol{\theta}^{(i-1)})$ is the probability of choosing move $m$ when in state $\boldsymbol{\theta}^{(i-1)}$, $q(u)$ is the full density function of $u$, and the final term is the Jacobian of the change of variables matrix

$$J = \left|\det\left(\frac{\partial\boldsymbol{\theta}^*}{\partial(\boldsymbol{\theta}^{(i-1)}, u)}\right)\right|$$

Note that, in case of a move to a lower dimensional state, $a_m(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\theta}^*) = \min\{1, 1/A\}$. As standard MCMC practice, if rejected, the chain remains at the current state $\boldsymbol{\theta}^{(i-1)}$.

In practice, a reversible jump sampler often alternates, between moves "within-model", where the state space is unchanged, and "between-models", with a change in dimensionality. Within-model moves follow the standard MCMC practices and methods reviewed in the previous sections. More care needs to be taken for dimension-changing moves, which need to show full reversibility. In Chapter 4 we provide an overview of a standard approach to "between-models" moves, which is based on the work presented Green (1995) and is commonly adopted in the literature.

## 2.3   Covariate parameterisations

Classical regression models and approaches, such as weighted least squares, rely on the assumption that the distribution of interest belongs to the normal distribution or exponential family. Extreme value distributions do not fall into either category, and hence require a different approach. This section reviews different covariate modelling techniques, all of which can be interpreted as extensions of generalised linear models (GLMs). We assume here that the reader is familiar with the basic theory of GLMs, although we recommend the monographs by McCullagh and Nelder (1989) and Dobson and Barnett (2008) for further details.

In Section 2.3.1 we introduce some notation and terminology common to all models. In Sections 2.3.2 and 2.3.3 we provide an overview of constant, linear and Fourier series descriptions, while Section 2.3.4 focuses on spline formulations. Finally, Section 2.3.5 considers the use of kernels to describe covariate-dependent variations.

## 2.3.1 Basis description

Assume, as before, that the observations $Y_1, \ldots, Y_n$ are independent random variables from an unknown distribution function $F(\theta)$, for some distribution parameter $\theta$. Now we consider the case where some non-stationarity is present, in which case the assumption of identically distributed data is no longer fulfilled. Then, this non-stationarity can be captured by assuming that the distribution parameter $\theta$ is covariate-dependent. Let $\{\boldsymbol{X}_t\}$, $\boldsymbol{X}_t \in \mathbb{R}^m$, be some corresponding $m$ covariates. A common approach is then to model the variation of $\theta$ by defining it as a function of the covariates of interest, such that

$$\theta(\mathbf{x}) = f(\mathbf{x}),$$

for some function $f$.

There are many choices of functional forms possible. A particularly useful way of representing such relationships is through a combination of linear basis functions. This set-up provides a common frame of reference across different modelling approaches and a common framework to represent linear and nonlinear relationships alike. In particular, we can assume that the aim is to obtain a linear combination of a set of $p$ basis functions $b_1(\mathbf{x}), \ldots, b_p(\mathbf{x})$, such that

$$\theta(\mathbf{x}) = \beta_1 b_1(\mathbf{x}) + \beta_2 b_2(\mathbf{x}) + \ldots + \beta_p b_p(\mathbf{x}) \tag{2.3.1}$$

for some coefficients $\beta_1, \ldots, \beta_p$. We can see this approach as a generalization of linear regression, where essentially the input is replaced with a function of inputs.

A common way to present the basis formulation is via the use of what is known as "basis matrix". We then rewrite Eq. 2.3.1 as

$$\theta(\mathbf{x}) = B\left(\mathbf{x}\right)\boldsymbol{\beta}, \tag{2.3.2}$$

where $B(\mathbf{x})$ is a matrix of $p$ columns, and its $j^{\text{th}}$ column corresponds to the $j^{\text{th}}$ basis function $b_j$, denoted as in Eq. 2.3.1 and evaluated at $\mathbf{x}$, for $j = 1, 2, ..., p$. Then, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is a $p$-vector of basis coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ to be estimated during statistical modelling.

### 2.3.2   Constant parameterisation

The simplest statistical models assume stationary series, that is, the model parameter is constant. In the constant parameterisation, the value of $\theta$ does not vary with respect to the covariates $\mathbf{x}$. We therefore adopt a scalar basis function, where $p = 1$ and the function is constant across all values of covariate, so that $B(\mathbf{x}) = \mathbf{1}$, where $\mathbf{1} = (1, \ldots, 1)$ is a $1 \times n$ vector of ones.

### 2.3.3   Polynomial and Fourier series formulation

For simplicity in this case, we hereafter assume $m = 1$ (so that $\mathbf{X} = X$ for ease of notation) and $X \in \mathcal{D} = [0, 2\pi)$.

The simplest covariate-dependent formulation is through a polynomial regression

model. Such a model for a predictors $X$ is

$$\theta(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_p x^p,$$

where $p$ is the degree of the polynomial. Note that the case $p = 1$ corresponds to a linear relation.

The Fourier series formulation provides a suitable extension of the polynomial model in the case where the covariates of interest show periodic behaviour. Then we write $\theta(x)$ in terms of a $p$-order Fourier series

$$\theta(x) = c_0 + \sum_{k=1}^{p} \left\{ c_k \cos(kx) + d_k \sin(kx) \right\},$$

where $c_k \in \mathbb{R}$ and $d_k \in \mathbb{R}$ for $k = 1, 2, ..., p$. This can be written in basis notation by defining a basis matrix $B$ with basis column vectors

$$\mathbf{b}(x) = (1,\ \cos(x),\ \cos(2x),\ \ldots,\ \cos(px),\ \sin(x),\ \sin(2x),\ \ldots,\ \sin(px))',$$

and basis coefficients

$$\boldsymbol{\beta} = (c_0,\ c_1,\ c_2,\ \ldots,\ c_p,\ d_1,\ d_2,\ \ldots,\ d_p)'.$$

We refer the reader to the work of Ruppert et al. (2003) for a more in depth presentation of these models as well as other semi-parametric regression approaches.

## 2.3.4 Splines

Spline formulations have developed rapidly in the statistics literature during the past decades. They are of particular interest because they are not restricted by distribution functions as they require no a priori knowledge of it. In these section we introduce the principles of splines, based on the material from Hastie et al. (2001), Green and Silverman (1994), Ruppert et al. (2003), and Brezger and Lang (2006).

Let us again, for simplicity of notation, assume only one covariate is present such that $m = 1$, so that $\mathbf{X} = X$ and $X \in \mathcal{D}$ for some non-periodic domain $\mathcal{D} = [a, b]$. We can then obtain the piecewise polynomial function $f(x)$ by dividing the domain of $X$ into contiguous intervals defining $f$ as a separate polynomial in each subinterval. Then an order-$M$ spline with interior knots $\tau_r$, $r = 1, \ldots, n_\tau$ is piecewise-polynomial of degree $d_S$ such that $M = (d_S + 1)$. The general form for the function $f$ in terms of the the truncated power basis set is then

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \cdots + \beta_{d_S+1} x^{d_S} + \sum_{r=1}^{n_\tau} \beta_{d_S+r+1}(x - \tau_r)_+^{d_S}, \qquad (2.3.3)$$

where $(x - \tau_r)_+$ equals $x - \tau_r$ if $x > \tau$ and 0 otherwise, and $a < \tau_1 \leqslant \tau_2 \leqslant \cdots \leqslant \tau_{n_\tau-1} \leqslant \tau_{n_\tau} < b$. The resulting model consists of $d_S + 1$ polynomials and $n_\tau$ truncated polynomials, such that $d_S + n_\tau + 1$ basis functions exist.

A $B$-spline, or basis spline, is a mathematically equivalent but computationally more stable formulation for the polynomial spline. Under this framework, we can define

$$f(x) = \sum_{j=1}^{p} \beta_j b_j(x) = B(x)\boldsymbol{\beta},$$

with design matrix $B = \begin{pmatrix} b_1'(x) & b_2'(x) & ... & b_p'(x) \end{pmatrix}$ and where $\beta_j$, $j = 1,,\ldots,p$ are the basis coefficients. In order to construct each $B$-spline basis functions $b_j$, we first need to define a sequence of augmented knots $\{\zeta\}$. Let $a = \tau_0 < \tau_1$ and $\tau_{n_\tau} < \tau_{n_\tau+1} = b$ be two boundary knots. These can be used to define the domain over which we wish to evaluate our spline. The augmented set of knots $\zeta$ can then be defined as follows:

- $\zeta_1 \leqslant \zeta_2 \leqslant \cdots \leqslant \zeta_M \leqslant \tau_0$;

- $\zeta_{j+M} = \tau_j, j = 1, \ldots, n_\tau$;

- $\tau_{n_\tau+1} \leqslant \zeta_{n_\tau+M+1} \leqslant \zeta_{n_\tau+M+2} \leqslant \cdots \leqslant \zeta_{n_\tau+2M}$.

Note that the actual values of these additional knots beyond the boundary are arbitrary, and it is common practice to have them equal to $\tau_0$ and $\tau_{n_\tau+1}$ respectively. Using $n_\tau$ interior knots for a spline of order $M$ results in $p = n_\tau + 2M$ $B$-spline basis functions. We can then compute each basis functions $b_j$, $j = 1,,\ldots,p$ recursively Let $b_j^m$ be the $j$th B-spline basis function of order $m$, $m \leqslant M$, for the augmented knot sequence $\{\zeta\}$. The basis function can then be constructed using the Cox-de Boor recursion formula in terms of divided differences of the different $m$-order polynomial pieces. Then

$$
b_j^1(x) := \begin{cases} 1 \text{ if } \zeta_j \leqslant x < \zeta_{j+1} \\ \\ 0 \text{ otherwise,} \end{cases}
$$

for $j = 1, \ldots, n_\tau + 2M - 1$, and

$$b_j^m(x) := \frac{x - \zeta_j}{\zeta_{j+m-1} - \zeta_j} b_j^{m-1}(x) + \frac{\zeta_{j+m} - x}{\zeta_{j+m} - \zeta_{j+1}} b_{j+1}^{m-1}(x),$$

$j = 1, \ldots, n_\tau + 2M - m$ and $m = 1, \ldots, M$.
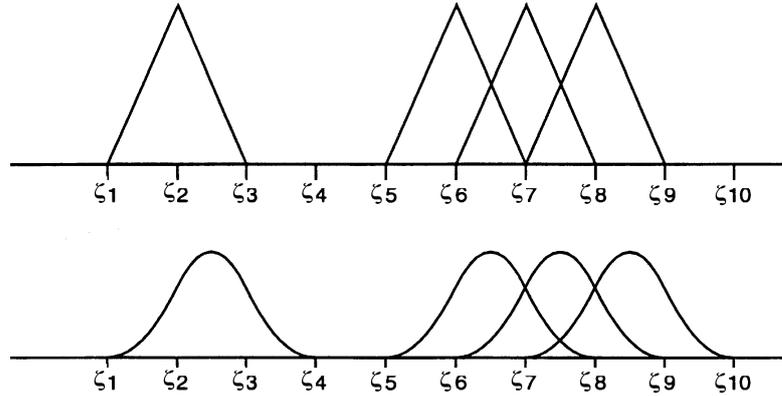


Figure 2.3.1: A representation of first (top) and second (bottom) degree B-spline components (adapted from Eilers and Marx 1996)

B-splines are considered, because they have many desirable attributes (Eilers and Marx, 1996). For example, as shown in the top half of Figure 2.3.1, $b_j^1(x)$ is a triangular function that is zero up to $x = \zeta_j$ included, has a maximum of one at $x = \zeta_{j+1}$ and decreases to zero for $x \geqslant \zeta_{j+2}$ onward. The bottom of the figure illustrates the second-degree case. This shows one of the most advantageous characteristics of B-splines: their support, contrary to truncated polynomials, is bounded. That is, for a B-spline of degree $d_S$, only $d_S + 2$ knots build the support of each single B-spline curve. We can notice this by producing a function $f(x)$ using some random coefficients $\boldsymbol{\beta}$, shown in Figure 2.3.2.
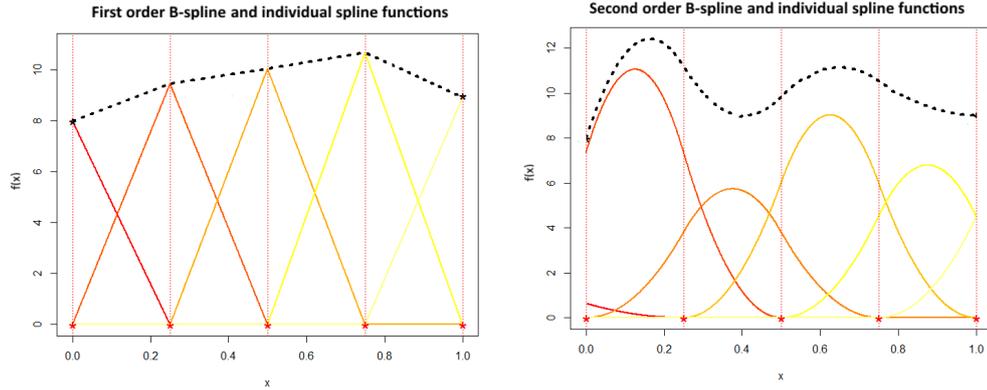
Figure 2.3.2: A representation of first (left) and second (right) degree B-spline components, defined by some interior knots (red *), and resulting function $f(x)$ $(-)$.

From Figure 2.3.2, we can see that each B-spline overlaps only with $d_S + 2$ neighbouring B-splines. Furthermore, one can prove that at the joining points, derivatives up to order $d_S - 1$ are continuous (De Boor et al., 1978). That is, a B-spline basis consists of $d_S + 1$ polynomials of degree $d_S$, which are $d_S - 1$ times continuously differentiable, resulting in different smoothness of regression curves. This effect is shown in Figure 2.3.2, where we obtain a piecewise linear curve for a 1-degree spline and a piecewise quadratic for a 2-degree spline respectively.

Regression splines exploit the formulations above to represent the parameters of a probability distribution as a function of some covariates, that is, the model parameter from the distribution of interest can be modelled using a basis matrix $B$ and a vector of coefficients $\boldsymbol{\beta}$ as in Eq. 2.3.2. The choice of number and location of knots is a crucial issue with regression splines. When too few knots are used, the resulting function might not be flexible enough to capture the variability in the data, while overfitting is a risk when a large number of knots is used. Knot locations can, for similar reasons, have a noticeable impact on the fit obtained. Eilers and Marx (1996) propose a roughness penalty approach as a possible solution:

they suggest choosing a high number of equally spaced knots, and then imposing a difference penalty on adjacent spline basis coefficients. This set up allows enough flexibility while also controlling the smoothness of the fitted curve. We can choose to penalise the roughness, denoted by $R_\theta$, with respect to $x$ during inference, in order to impose smoothness of $\theta$ with respect to $x$. Then, one can incorporate this penalty in the inference mechanism. A quadratic form in $\boldsymbol{\beta}$ is typically assumed for the roughness, e.g.

$$R_\theta = \frac{1}{2}\rho_\theta\boldsymbol{\beta}'Q_\theta\boldsymbol{\beta}, \tag{2.3.4}$$

where $\rho_\theta \geqslant 0$ is a roughness coefficient, and $Q_\theta$ is the roughness penalty matrix expressing the squared magnitude of differences between the values of adjacent basis coefficients $\boldsymbol{\beta}$. For maximum likelihood methods, the penalising term is added, as a positive term, to the negative log-likelihood to be optimised. In Bayesian inference, the penalty is imposed via the prior on the parameters $\boldsymbol{\beta}$, such that

$$\boldsymbol{\beta}|\rho_\theta \sim \rho_\theta^{1/2} \exp\left(-\frac{\rho_\theta}{2}\boldsymbol{\beta}^{\mathrm{T}}Q_\theta\boldsymbol{\beta}\right).$$

Such a penalty ensures that very sharp changes will be discouraged unless there is enough evidence from the data, hence providing a smoother curve which does not overfit to localised random behaviour. The precision matrix $Q_\theta$ often can be formulated in terms of some some $(p-1) \times p$ difference matrix $D_\theta$, such that

$$Q_\theta = D'_\theta D_\theta. \tag{2.3.5}$$

Higher-order differences imply that a larger number of adjacent knots have an impact on each other. In practice (Eilers and Marx, 2010), it is often enough to

use a first-order difference matrix, such that

$$
D_\theta = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix},
$$

as this often enough to control the smoothness, while also being more computationally efficient than higher-order ones. Note that, under this model, inference consists of the estimation of the basis coefficients $\boldsymbol{\beta}$ and the global roughness coefficient $\rho_\theta$, while the number and location of the knots of the given degree spline basis is fixed.

### 2.3.5  Radial functions

First applied to geodesic approximation, radial basis functions (RBF) have since been widely applied in a number of fields such as neural networks, image processing and interpolation, kinetic modelling and solution of differential and integral equations. There is a vast literature on radial basis functions in different fields of statistics. While more details are beyond the scope of this work we refer the reader to the papers by Powell (1987) and Girosi and Poggio (1990) for some thorough reviews of the theory involved, and the monograph by Buhmann (2003) for an overview of their application and adaptations to different fields.

Radial functions are a special class of function whose response changes monotonically with distance from a central point $c$. This distinguishing feature joins many

specific formulations, which differ by the choice of centre, distance scale and the precise shape of the radial function. The most general formula for any RBF in any dimension is

$$h(\boldsymbol{x}) = f\left((\boldsymbol{x} - \boldsymbol{c})'\boldsymbol{R}^{-1}(\boldsymbol{x} - \boldsymbol{c})\right),$$

where $f(\cdot)$ is the specific function used, $\boldsymbol{c}$ is the centre and $\boldsymbol{R}$ is the metric chosen. This metric is often Euclidean, such that $\boldsymbol{R} = r^2\boldsymbol{I}$ for some scalar radius $r$ and identity matrix $\boldsymbol{I}$. Common example of function choices include Cauchy, multiquadric and Gaussian functions, the latter of which in 1-D is

$$h(x) = \exp\left(-\frac{(x - c)^2}{r^2}\right), \tag{2.3.6}$$

for some scalar input $x \in (-\infty, \infty)$. Gaussian density functions can also be classified as an RBF, as they are just a version of Eq. 2.3.6 scaled to ensure the density integrates to 1. Gaussian RBF's have a localised impact, as they show a significant response only in a neighbourhood near the centre, although theoretically they have global support.

RBF's can be employed in any sort of linear and nonlinear regression model in the same way other basis functions types considered in Sections 2.3.3 and 2.3.4. In particular, assume again that we have a one-dimensional predictor $X$, then we can model the variation of a covariate-dependent parameter $\theta(x)$ using a linear combination of RBF's

$$\theta(x) = \sum_{j=1}^{p} \beta_j f_j(x),$$

where each $f_j$ is characterised by its own parameters $c_j$ and $r_j$ and $p$ is the total number of radial function components.

# Chapter 3

# A comparison of peaks over threshold methods

In Chapter 1, we considered the importance of accurate statistical models for extreme data in various application, such as for modelling the extreme conditions offshore structures are subject to. In Chapter 2, we introduced the statistical theory of extreme value modelling and noticed that often the applications where this is used tend to produce data showing non-homogeneity, so that considering covariates becomes essential to proposing a realistic model. For example, one cause of structural damage to offshore sites are storm waves, with the most severe sea states usually being wind generated. Consequently, we may suspect that the height of the waves will change according to, for example, season, geographic location of the sampling site, or direction of the wind. Constructing a model that includes such factors is necessary, and statistical tools to analyse and extrapolate from such a model become essential. Hence, in Section 2.1.3, we reviewed the standard

approach to account for covariate dependence for extremes data.

In this chapter, we focus our attention on *peaks over threshold* (POT) methods, where we study the tail of the distribution by only considering observations that are above an arbitrary value. In the stationary case, two models are available from the literature, namely the generalised Pareto distribution (GPD) and the non-homogeneous Poisson point (NHPP) process formulation. This chapter aims to compare them and shows that, although theoretically they are equivalent (Smith, 1989; Davison and Smith, 1990), in practice each method has its own advantages and limitations. Moving from one model to the other is straightforward in the case of stationary processes. Nonetheless, we are mainly interested in modelling and analysing non-stationary processes. Model parameters are, in this case, no longer directly transferable using the equations in Section 2.1.2. Furthermore, being able to correctly capture covariate effects is complicated, in the EVT framework, by the reduced amount of data available by construction. This may lead to added instability in the optimisation of likelihoods used by maximum likelihood (ML) methods for model fitting. This is particularly true in the case of the Poisson point process formulation, where additional issues arise compared to the equivalent GPD model, as shown in Section 3.1.2.

In the following sections, we systematically review the performance of model fitting and extrapolation for both the GPD and PP models in the non-stationary case, as well as highlighting areas with potential for further development.

# 3.1 Threshold approaches

The peaks over threshold approach is suited to the types of application and datasets described in Chapter 1, which comprise of more than just annual or monthly maxima. In particular, data may be available daily, hourly or even sub-hourly. Among others, it was used in the seminal work by Smith (1989) applied to air pollution data, and soon after popularised by Davison and Smith (1990). Assume that $Y_1, \ldots, Y_n$ are independent and identically distributed (i.i.d.) random variables from the same unknown distribution function $F_Y$ over some domain $\Omega$, where, say, $\Omega = \mathbb{R}$. It seems natural to define as *extreme* those observations that exceed a certain value, that is, the $Y_i$, $i = 1, \ldots, n$, above a chosen threshold $u \in \mathbb{R}$. Two equivalent models exist in the literature to analyse observations within this framework, namely:

1. The generalised Pareto distribution with distribution function, for $y > u$,

$$\mathbb{P}[Y < y] = 1 - \phi_u \left\{ 1 + \xi \left( \frac{y - u}{\psi_u} \right) \right\}_+^{-\frac{1}{\xi}},$$

   with scale $\psi_u > 0$, exceedance probability $\phi_u \in [0, 1]$, both conditional on the threshold $u$, and shape $\xi \in \mathbb{R}$;

2. The non-homogeneous Poisson point process with intensity on $[0, 1] \times [u, \infty)$,

$$\lambda(t, y) = \frac{1}{\sigma} \left\{ 1 + \xi \left( \frac{y - \mu}{\sigma} \right) \right\}_+^{-1-\frac{1}{\xi}},$$

   with location $\mu \in \mathbb{R}$, scale $\sigma \in \mathbb{R}^+$, and shape $\xi \in \mathbb{R}$, and where we use

standard notation

$$[\cdot]_+ := \max\{0, \cdot\}.$$

Although these models have already been introduced in Section 2.1.2, it is worth noticing a few intrinsic characteristics and results for the approaches considered, as both have advantages and drawbacks and neither prevails as an overall "better" model. In particular,

- The Poisson point process parameters are threshold invariant, provided we are far enough in the tail of the distribution for a given threshold $u$. Then, any subset of the extreme observations obtained with a new choice of threshold $v > u$ will follow the same distribution, i.e. the latter will be described by a NHPP with the same parameters $\mu, \sigma, \xi$;

- The Poisson point process parameters are strongly dependent, which makes parameter estimation harder to perform. Wadsworth and Tawn (2012) suggests that introducing an additional factor in the likelihood may help in reducing the correlation between them, as detailed in Section 3.1.1. In a similar manner, Sharkey and Tawn (2017) propose the reparameterisation of NHPP parameters in terms of a tuning parameter for a Bayesian implementation and introduce a method for choosing this additional term so to obtain near-orthogonality of the model parameters for stationary processes, or when a linear trend in one covariate is present in the location parameter only.

- In the GPD model, $\phi_u$ is orthogonal to the parameters $\psi_u$ and $\xi$, which simplifies their estimation, although $\psi_u$ and $\xi$ are still correlated. Threshold

invariance no longer holds, as $\psi_u$ and $\phi_u$ depend on the threshold. The choice of this threshold, as we mentioned earlier, is in itself an issue for both the NHPP and GPD models and may be subjective, which will then have an effect on model fitting and extrapolation.

- There exists a reformulation of the GPD scale parameter $\psi_u$ in terms of a lower threshold $u_0 < u$ for which the assumption of GPD distributed exceedances still holds. This re-parameterisation is often used for threshold selection methods testing for stability of parameters over a range of threshold candidates. It essentially uses Eq. 2.1.9 to re-parametrise the GPD scale parameter as $\psi^* = \psi_u - \xi u$, which is constant with respect to $u$. An alternative re-parameterisation follows the work by Cox and Reid (1987) and Chavez-Demoulin and Davison (2005) to obtain more computationally advantageous formulations of the GPD parameters. It consists of moving from GPD parameters $(\psi_u, \xi)$ to the asymptotically independent pair $(\nu_u, \xi)$, where $\nu_u = \psi_u(1 + \xi)$.

- Although the threshold stability property often leads theoretical statisticians to prefer the NHPP approach, the GPD formulation is usually preferred by applied users, as more immediately interpretable.

## 3.1.1 Non-homogeneous Poisson process interpretation as yearly maxima

One interesting characteristic of the NHPP parameterisation is that it can be easily unified with that of the GEV. In particular, given some observed data over

a number of years $N_{\text{years}}$, using the NHPP likelihood function in Eq. 3.1.2, with corresponding integrated intensity as in Eq. 2.1.6, yields MLE's for NHPP parameters corresponding to the $N_Y$ maxima. It may, instead, be of interest to know the GEV parameters corresponding to the yearly maxima, which is in fact the same time scale that applies to MLE's for GPD parameters. To obtain such estimates, one assumes $N_Y$ independent replications of the Poisson process defined over the set $B_u = [0, 1] \times [u, \infty)$ as before, so that the resulting NHPP model has integrated intensity measure

$$\Lambda(B_u) = N_Y(1 + \xi u)_+^{-1/\xi}. \tag{3.1.1}$$

Let us denote the estimates from fitting a NHPP likelihood with integrated intensity from Eq. 2.1.6 as $(\mu^{(1)}, \sigma^{(1)}, \xi^{(1)})$, where we use the notation $\theta^{(1)}$ to indicate that we are now referring to the case where we take $N_Y = 1$. For the rest of the chapter, we will also be using the notation $(\mu, \sigma, \xi)$ to refer to the yearly-maxima parameter estimates computed by setting $N_Y$ to the actual number of years of data available. This corresponds to using a NHPP likelihood with adjusted integrated intensity as in Eq. 3.1.1. Then, the relationship between the parameters in Eq. 2.1.6, denoted as $\theta^{(1)}$, and those in Eq. 3.1.1 is given by

$$\xi^{(1)} = \xi, \qquad \sigma^{(1)} = \sigma \left(\frac{1}{N_Y}\right)^{-\xi}, \qquad \mu^{(1)} = \mu - \frac{\sigma^{(1)}}{\xi}\left(1 - \left(\frac{1}{N_Y}\right)^\xi\right).$$

Wadsworth and Tawn (2012) found that it may be possible to improve fitting via a careful choice of the value of $N_Y$. In particular, when we have $N_{\text{exc}}$ in the dataset given a chosen threshold, setting $N_Y = N_{\text{exc}}$ in Eq. 3.1.1 improved the mixing properties when Bayesian methods were used to compute parameter estimates. In

a maximum likelihood framework, this corresponds to a reduction in the correlation of parameters. In the stationary case, this change in the value of $N_Y$ presents no major difficulties, so that one could transform estimates from any ad hoc time-scale to, say, estimates for yearly maxima. In particular, the relationship between the $N_Y$-year maxima parameters from Eq. 3.1.1 and estimates resulting from a different choice of $N_Y$, say $N_Y^*$, is given by

$$\xi^{N_Y^*} = \xi, \qquad \sigma^{N_Y^*} = \sigma \left( \frac{N_Y}{N_Y^*} \right)^\xi, \qquad \mu^{N_Y^*} = \mu - \frac{\sigma}{\xi} \left( 1 - \left( \frac{N_Y}{N_Y^*} \right)^\xi \right).$$

It is important to be aware of the meaning of the $N_Y$ term in Eq. 3.1.1, both during the fitting procedure and when analysing results.

For illustration purposes, we produce a single data set of GPD distributed observations. We set $u = 0$ and, given NHPP parameters $\mu = 3, \sigma = 2, \xi = 0.1$, we compute the GPD scale $\psi_u = \sigma + \xi(u - \mu)$. Hence, we simulate 200 years of data, which corresponds, using the integrated intensity from Eq. 3.1.1, to $\Lambda(B_u) = 5079$ expected threshold exceedances. We then compute the joint conditional NHPP negative log-likelihood for the scale and shape parameters, by fixing the value for the location to be the one chosen for the simulation, i.e. $\mu = 3$. Figure 3.1.1 shows the resulting joint conditional negative log-likelihood surfaces, with infeasible regions in grey, and using white to indicate parameter combinations leading to negative log-likelihood values above the $60^{\text{th}}$ percentile of all the values obtained. The plots also show both the original true parameters and the estimates for parameters $\sigma, \xi$ which maximise the conditional negative log-likelihood.
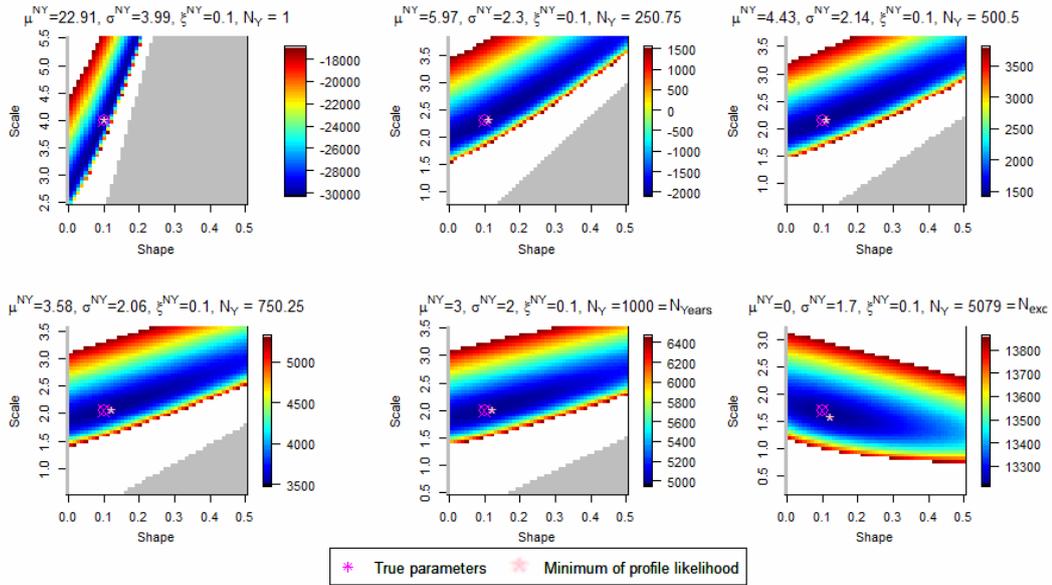
Figure 3.1.1: Scale-shape joint NHPP negative log-likelihood, conditional on $\mu = 3$. The infeasible region is depicted in grey, and parameter combinations leading to negative log-likelihood values above the 60th percentile of all the values computed are shown in white. True parameters and computed estimates leading to the minima of the conditional negative log-likelihood are also given.

Figure 3.1.1 also provides an example of the impact of different choices for the $N_Y$ term. First, we notice how the surface is much narrower for some of the $N_Y$ values, and especially for $N_Y = 1$. The plot clearly shows a strong correlation between the shape and the scale parameters, and the narrower optimal ridge, represented by the darker region, is likely harder to explore for an optimisation routine, which is less likely to step in the most optimal direction. On the other hand, as suggested by Wadsworth and Tawn (2012), a choice of $N_Y = N_{\text{exc}}$ reduces the correlation of the parameters and yields a wider surface, easier to explore. Furthermore, Figure 3.1.1 also highlights the impact of the choice of $N_Y$ over the values of the parameter estimates. Given the visibly large variability of the location (and, partly, scale) estimates for the different $N_Y$ values, it is evident that we need to be clear when reporting results by providing not only the parameter estimates but also the value

of $N_Y$ used, given the strong relationship between them.

## 3.1.2 Peaks over threshold models for non-stationary series

The theoretical results discussed so far rely on the assumption of independent and identically distributed observations. In reality, this is rarely the case, and both the independence and the identically distributed assumptions may fail. In particular, extreme observations often display signs of a variety of trends. In Section 2.1.3, we have introduced the approaches for non-stationary series.

Let us consider a process $\{Y_t\}$, with $t = 1, \ldots, n$ and $Y_t \in \mathbb{R}$, and let $\{\boldsymbol{X}_t\}$, $\boldsymbol{X}_t \in \mathbb{R}^m$, be the corresponding $m$ covariates. For example, the response variable may be significant wave height with $m = 2$ covariates, some dates $x_d$ to account for seasonality and some wave direction $x_w$.

We have already introduced the modelling and prediction methodology for the Generalised Pareto model in Section 2.1.3, although a few additional remarks are necessary. The orthogonality of the parameters is preserved in the non-homogeneous setting, and may indeed simplify their estimation. Furthermore, the estimation of the rate exceedance $\phi_u$ and of the shape and scale parameter occur separately. This simplifies both the fitting and the simulation process, especially for covariate-dependent models. Nevertheless, as previously explained, the GPD characterisation does not allow for threshold stability. Given a different threshold, the scale will in fact change as shown in Eq. 2.1.9. In the non-stationary case,

this becomes $\psi_v(\boldsymbol{x}) = \psi_u(\boldsymbol{x}) + (v - u)\xi(\boldsymbol{x})$. In other words, $\psi_v(\boldsymbol{x})$ no longer has the same form as $\psi_u(\boldsymbol{x})$ and, in fact, it may not even include the same covariates, since $\xi(\boldsymbol{x})$ may include covariates not required in $\psi_u(\boldsymbol{x})$.

Similar problems arise for all the re-parameterisations mentioned for models of stationary series in Section 2.1 for both the GPD and the NHPP approach, so that one may not always be able to easily apply them, and retrace intercept and covariate coefficients between different formulations.

Using the Poisson point process approach may be a possible way around threshold instability, as, although all parameters may be still assumed to be covariate-dependent, they are also threshold invariant, assuming we are far enough in the tail of the distribution. Nevertheless, the non-homogeneous Poisson point process formulation for covariate-dependent cases has added computational complexity. The following section provides an overview of this approach for covariate-dependent models, and aims to highlight the complications arising in this case.

**Poisson point process fit**

Let $\{Y_t\}$, for $t = 1, \ldots, n$, be a set of $n$ independent identically distributed observations $Y_t \in \mathbb{R}$, following a distribution function $F_Y$ defined over the domain $\mathbb{R}$. One can use the threshold approach to fit a Poisson point process with a given intensity $\lambda(t, y)$ as in Eq. 2.1.5 and parameters $(\mu, \sigma, \xi)$ to the dataset.

For inference using the Poisson process approach, the likelihood can be derived for

some high threshold $u$ as

$$L(\mu, \sigma, \xi; y_1, \ldots, y_n) \propto \exp\left\{-N_Y\left[1 + \xi\left(\frac{u - \mu}{\sigma}\right)\right]_+^{-1/\xi}\right\} \prod_{i=1}^{N(B_u)} \sigma^{-1}\left\{1 + \xi\left(\frac{y_i - \mu}{\sigma}\right)\right\}_+^{-1-1/\xi}$$

for threshold exceedances $y_1, \ldots, y_{N(B_u)}$, $B_u = [0,1] \times [u, \infty)$ and factor $N_Y$, often set as the number years of observations $N_{\text{Years}}$. Maximum likelihood methods can hence be used to obtain estimates for the parameters.

Now let $\{\boldsymbol{X}_t\}$, $\boldsymbol{X}_t \in \mathbb{R}^m$, be some covariates that characterise a non-stationary behaviour in the observations, where $m$ is the number of different covariates considered. The parameters and consequently rate of exceedance now depend on the covariates $\boldsymbol{X}$.

This causes a change in the likelihood, as the integrated intensity can no longer be easily computed. In particular, the intensity to be integrated is now

$$\lambda(\boldsymbol{x}, y) = g(\boldsymbol{x})\sigma(\boldsymbol{x})^{-1}\left\{1 + \xi(\boldsymbol{x})\left(\frac{y - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)\right\}^{-1/\xi(\boldsymbol{x})-1}, \tag{3.1.2}$$

where $g(\boldsymbol{x})$ is the density function of the covariates. Note that we omit the $N_Y$ factor in Eq. 3.1.2, as this is just a constant, hence having no repercussions on the integration procedure. Now let $\mathcal{X}$ be the covariate space, where each covariate $x_{i,j}$ is in the space denoted as $\mathcal{X}_j = [\min(X_j), \max(X_j)]$ and $\mathcal{X} = \bigotimes_j \mathcal{X}_j$, the overall cross-product of the single covariate spaces. Define $B_u = \mathcal{X} \times [u, \infty)$, and denote the integrated intensity as $\Lambda(B_u)$. The likelihood function then becomes

$$L(\mu(\boldsymbol{x}), \sigma(\boldsymbol{x}), \xi(\boldsymbol{x}); y_1, \ldots, y_n) \propto \exp\left\{-\Lambda(B_u)\right\} \prod_{i=1}^{N(B_u)} \sigma(\boldsymbol{x})^{-1}\left\{1 + \xi(\boldsymbol{x})\left(\frac{y_i - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)\right\}_+^{-1-1/\xi(\boldsymbol{x})}$$

where the integrated intensity is

$$
\begin{aligned}
\Lambda(B_u) &= \int_{\mathcal{X}} \int_{B_u} \lambda(\boldsymbol{x}, y) dy \, d\boldsymbol{x} \\
&= \int_{\mathcal{X}} \int_{B_u} g(\boldsymbol{x})\sigma(\boldsymbol{x})^{-1} \left\{ 1 + \xi(\boldsymbol{x}) \left( \frac{y - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \right) \right\}_+^{-1/\xi(\boldsymbol{x})-1} dy \, d\boldsymbol{x} \\
&= \int_{\mathcal{X}_m} \cdots \int_{\mathcal{X}_1} g(\boldsymbol{x}) \left\{ 1 + \xi(\boldsymbol{x}) \left( \frac{u - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \right) \right\}_+^{-1/\xi(\boldsymbol{x})} dx_1 \ldots dx_m \quad (3.1.3)
\end{aligned}
$$

**Estimation of the integrated intensity**

Since we do not know the covariate density $g(\boldsymbol{x})$, the integral over the covariate space needs to be approximated. The approach described in this section is based on the theory of Monte Carlo estimation.For a more detailed review on these methods, the reader can refer for instance to Rubinstein (1981), MacKay (1998), Kalos and Whitlock (2009) and Hammersley (2013).

In general, the multidimensional definite integral

$$
I = \int_{\Omega} h(\boldsymbol{y}) d\boldsymbol{y},
$$

where $\Omega$ is the variable space, may be approximated as

$$
I \approx \frac{V}{N} \sum_{i=1}^{N} h(\boldsymbol{y_i}), \quad (3.1.4)
$$

where $V := \int_{\Omega} d\boldsymbol{y} = |\Omega|$ is the volume of the variable space we integrate over, and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are simulated from a uniform distribution on $\Omega$.

Often, the covariates we have in a dataset are not uniformly distributed, so that Eq. 3.1.4 needs to be adapted. Let $p(\boldsymbol{x}) = \left\{ 1 + \xi(\boldsymbol{x}) \left( \frac{u - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \right) \right\}_+^{-1/\xi(\boldsymbol{x})}$, then there

are now two possible ways to proceed.

**Method 1: uniformly spaced covariates.**

Take some covariates $\boldsymbol{x^*}$ uniformly distributed on the covariate space $\mathcal{X}$ according to some distribution $H_X$ with density $h_X$. Then, we can write the integrated intensity in Eq. 3.1.3 as

$$\Lambda(B_u) = \int_{\mathcal{X}} g(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \qquad (3.1.5)$$

$$= \int_{\mathcal{X}} \frac{g(\boldsymbol{x})p(\boldsymbol{x})}{h_X(\boldsymbol{x})} h_X(\boldsymbol{x})d\boldsymbol{x}.$$

We can draw $N^*$ sample from the density $h_X$, so that the integrated intensity becomes

$$\Lambda(B_u) \approx \frac{1}{N^*} \sum_{i=1}^{N^*} \frac{g(\boldsymbol{x}_i^*)p(\boldsymbol{x}_i^*)}{h_X(\boldsymbol{x}_i^*)}.$$

Now, using the fact that $h_X$ was a uniform distribution such that $h_X(\boldsymbol{x*}) = 1/|\mathcal{X}|$, $\forall \boldsymbol{x*} \in \mathcal{X}$, then we obtain

$$\Lambda(B_u) \approx \frac{|\mathcal{X}|}{N^*} \sum_{i=1}^{N^*} g(\boldsymbol{x}_i^*) \left\{ 1 + \xi(\boldsymbol{x}_i^*) \left( \frac{u - \mu(\boldsymbol{x}_i^*)}{\sigma(\boldsymbol{x}_i^*)} \right) \right\}_+^{-1/\xi(\boldsymbol{x}_i^*)}. \qquad (3.1.6)$$

Finally, we can approximate the volume $|\mathcal{X}|$ as

$$|\mathcal{X}| \approx \prod_{j=1}^{m} |\max(X_j) - \min(X_j)|,$$

assuming the covariate space is bounded by construction, for example wave directional covariates would be $x_i \in [0, 360)$, or that reasonable bounds can be imposed.

**Method 2: observed covariates.** Assume we have a dataset with $\tilde{N}$ observed covariates $\tilde{\boldsymbol{x}}$. Then we can compute the integrated intensity in Eq. 3.1.3 by noticing that $\{\tilde{\boldsymbol{x}}\}$ is just a sample from the target distribution $g(\cdot)$, so that Eq. 3.1.5 becomes

$$\begin{aligned}
\Lambda(B_u) &= \int_{\mathcal{X}} g(\boldsymbol{x})p(\boldsymbol{x})dx \\
&\approx \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} p(\tilde{\boldsymbol{x}}_i) \\
&= \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \left\{ 1 + \xi(\tilde{\boldsymbol{x}}_i) \left( \frac{u - \mu(\tilde{\boldsymbol{x}}_i)}{\sigma(\tilde{\boldsymbol{x}}_i)} \right) \right\}_+^{-1/\xi(\tilde{\boldsymbol{x}}_i)}.
\end{aligned}$$

Both methods work well, and have different advantages and drawbacks. **Method 2** is, potentially, less computationally intensive, as there is no need to estimate the covariate density for a given sample. Furthermore, by working on a sample of realisations from the target density, this may provide better estimates of the intensity in areas of higher covariate density. On the other hand, **method 1**, by selecting selecting an arbitrary number of covariates equally spaced over the covariate domain, provides a more uniform coverage of the entire domain, especially where only a small number of observations is available.

**Covariate density approximation** Note that **method 1** requires knowledge of the density $g(\boldsymbol{x})$, which is unknown. Instead, we can use an approximation $\widehat{g}(\boldsymbol{x})$, whose values can be obtained via:

1. **Closed form assumption**, where we propose a parametric distribution for

the covariates. This should try to capture the relationship between covariates as closely as possible, by imposing dependence or independence between the covariates as well as with a careful choice of covariance structures. These assumptions will clearly affect the performance of the model in capturing the tail behaviour. A proper understanding of the process investigated, as well as thorough analysis of the covariates observed, will be fundamental in reducing such impact.

2. **Non-parametric methods**, such as kernel density estimation, can be used. Again, care should be taken to account for the correlation between covariates, which may be needed to choose an appropriate kernel form. This approach may not be possible when a high number of covariates is considered, as kernel density is not only more computationally intensive, but also highly unreliable, with increasing dimensions (Parzen, 1962; Rosenblatt et al., 1956).

Both approaches become increasingly complicated when more than one covariate is considered, i.e. $m > 1$, and will give rise to additional uncertainty and potential computational error, which are likely to affect the model fitting.

**Return values**

For non-stationary processes, there is no straightforward way to make return value predictions, as the parameters of the model are now conditional on some covariates, and consequently the same holds for return values. We need to redefine the concept of return value in the context of non-stationary data. A possible approach is to fix the return period of interest, and obtain covariate-dependent return values. That

is, for the covariate space $\mathcal{X}$ given by the data, take the $n$-year conditional return value to be the value $z_{n,i}$ such that

$$1 - F(z_{n,i}|\boldsymbol{x_i}) = 1/n,$$

for covariate value(s) $\boldsymbol{x_i}$, where $F$ is the distribution of the model considered.

For both the GPD and NHPP models, we use Eq. 2.1.10, where the model parameters are now replaced by the value of the parameters at the covariates we are conditioning on. The main difference stands in the computation of the exceedance probability $\phi_u(\boldsymbol{x})$, which is itself covariate-dependent.

For GPD models, we can use different approaches to estimating this probability. Throughout this chapter, we choose logistic regression methods. Essentially, we define a Bernoulli random variable $z_i$, where

$$z_i = \begin{cases} 1 & \text{if } Y_i > u, \\ 0 & \text{if } Y_i \leqslant u. \end{cases}$$

Then $z_i \sim \text{Bernoulli}(\phi_u(\boldsymbol{x_i}))$, with likelihood

$$L(z;x) = \prod_{i=1}^{n} \phi_u(\boldsymbol{x_i})^{z_i}(1 - \phi_u(\boldsymbol{x_i}))^{1-z_i} \tag{3.1.7}$$

for $z_i \in (0,1)$, and logistic regression can be used to estimate $\phi_u(\boldsymbol{x})$.

For NHPP models, we follow the same idea presented in Eq. 2.1.12. In particular, we approximate the covariate-dependent exceedance probability using the

conditional integrated intensity with respect to $Y$, denoted as $\Lambda(y|\boldsymbol{x})$, given by

$$\Lambda(y|\boldsymbol{x}) = g(\boldsymbol{x}) \left\{ 1 + \xi(\boldsymbol{x}) \left( \frac{y - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \right) \right\}_+^{-1/\xi(\boldsymbol{x})}. \tag{3.1.8}$$

Then, similarly to Eq. 2.1.12, we have

$$\phi_u(\mathbf{x}) \approx \frac{N_Y \Lambda(y|\boldsymbol{x})}{N_{\text{obs}}},$$

where it is essential to include the same $N_Y$ factor that was used during fitting to obtain the right scaling.

An alternative would be to marginalise the return values, so as to obtain estimates that include the entire covariate space. This process requires integrating return values over the covariate space, which often involves numerical methods and is computationally expensive. An empirical approach to the above is performed via simulation. In particular, recall that the $n$-year return level is a value we expect to be exceeded only once in $n$ years. Then, if we repeatedly simulate $N$ years of observations (for simulations methodology, see Section 3.2), we can obtain return values for any $n < N$ as the value $z_n|\boldsymbol{x}$ that is exceeded on average $N/n$ times. For example, if we simulate $N = 1000$ years of data, the 500 and 100 year return levels will be the values exceeded, respectively, twice and 10 times (on average). One can then use bootstrapping or empirical quantiles to obtain confidence bounds for the values computed.

## 3.2 Simulation routine

### 3.2.1 GPD simulation

Simulating data from a generalised Pareto distribution is fairly straight-forward, even when one or more of the parameters is a function of covariates: the following steps summarise the procedure:

1. Given a threshold and scale, shape and rate parameter values, simulate co-variates $\boldsymbol{x_1^*}, \ldots, \boldsymbol{x_n^*}$. This can be done either

    i  *parametrically* by sampling from a pre-specified closed form distribution;

    ii  *non-parametrically* by randomly sampling with replacement from a given dataset, or;

    iii  *non-parametrically with noise*, so as to allow for a certain level of "randomness" by adding noise to the sampled values.

    Note that when more than one covariate is considered, sampling is from the full multi-dimensional covariate space. If parametric sampling is used, the chosen distribution should reproduce the dependence between covariates.

2. Given simulated covariates $\boldsymbol{x^*}$, evaluate the exceedance rate and use this to determine whether or not to simulate an exceedance, via an accept-reject procedure. That is, for each covariate values $\boldsymbol{x_i^*}$ (or $x_i^*$ if only one covariate is modelled),

    • Draw a uniform random variable $a$;

- Compute the exceedance probability $\phi_{u,i}$, by first defining a Bernoulli random variable $z_i$, as detailed in Section 3.1.2. Then, we use logistic regression for the resulting likelihood in Eq. 3.1.7.

- If $\phi_{u,i} > a$, simulate $Y_i$ as an exceedance as described in the following step, otherwise retain the value for $\boldsymbol{x_i^*}$ and repeat the procedure until all covariates have been considered.

3. Given that $Y_i > u$, simulate $Y_i$ from a $\mathrm{GP}(\psi(\boldsymbol{x_i^*}), \xi(\boldsymbol{x_i^*}))$ distribution.

As we saw in the previous section, fitting a NHPP model requires a "complete" dataset, as the distribution of the covariates below the threshold is used in the integrated intensity estimation. As detailed in step (3) above, the procedure already produces observed covariates for non exceedances. It may be of interest to obtain observations below the threshold as well, particularly if we later want to consider using a covariate-dependent threshold for the same dataset. This is more complicated, as assumptions need to be made with respect to the distribution of the body of the observation $Y_i$, and one should be particularly careful with boundary issues at the threshold. It is, instead, often advisable to simulate using a "minimum" threshold, with subsequent constant or covariate dependent threshold above it at all times.

### 3.2.2  Poisson point process simulation

The aim is to produce datasets whose observations come from a non-homogeneous Poisson point process. The additional challenge is to include non-stationary behaviour. In the stationary case, a NHPP is summarised by:

- An integrated intensity $\Lambda_u$, which determines the expected number $N_u$ of exceedances over a high threshold $u$, so that one can obtain a random number of exceedances for a sample as $N_u \sim \text{Po}(\Lambda_u)$;

- A generalised Pareto distribution with threshold $u$ and parameters $\sigma_u = \sigma + \xi(u - \mu)$ and $\xi$, which is used to simulate the magnitude of the corresponding $N_u$ exceedances.

**Intensity estimation**

In the non-stationary case, the expected number of exceedances and the distribution of their magnitudes changes across covariate values. One way to capture this is by writing the parameters as functions of the covariates, resulting in the intensity given in Eq. 3.1.2. The integrated intensity $\Lambda_u$ is now allowed to vary with some covariates and is changing over the space where the points occur. Let $B_u = \mathcal{X} \times [u, \infty)$, where $\mathcal{X}$ is the covariate space, then the calculations required for the integrated intensity are given in Eq. 3.1.3. Although the integral may be evaluated using Monte Carlo integration as described in Section 3.1.2, the process is complicated by the multidimensionality of the covariates.

An alternative approach to obtaining $\Lambda_u$ is to approximate the continuous functions of the parameters in Eq. 3.1.2 by a piecewise constant representation. First, we divide the covariate domain into small subspaces, $\mathcal{S}_1, \ldots, \mathcal{S}_s$. The underlying assumption is that, within each $\mathcal{S}_k$, points occur according to a NHPP$(\mu_k, \sigma_k, \xi_k)$ with corresponding intensity $\lambda_k$, such that we can approximate $\lambda(\boldsymbol{x})$ by $\lambda(\boldsymbol{x}) = \sum_{k=1}^{s} \lambda_k$.

From Eq. 3.1.2, we know that we also need to consider the distribution of the covariates $g(\boldsymbol{x})$. Then, for each of the $k = 1, \ldots, s$ subspaces, we again employ a piecewise constant approximation by using the value of the density at the subspace midpoint $\boldsymbol{x}_k$. Then if we let $\mathcal{X} = \bigcup_k \mathcal{S}_k$, since $g(\boldsymbol{x}_k)$ is just a constant,

$$
\begin{aligned}
\Lambda_{k,u} = \Lambda(\mathcal{S}_k \times (u, \infty)) &= \int_u^\infty g(\boldsymbol{x}_k) \frac{1}{\sigma_k} \left\{ 1 + \xi_k \left( \frac{y - \mu_k}{\sigma_k} \right) \right\}_+^{-1/\xi_k - 1} dy \\
&= g(\boldsymbol{x}_k) \left\{ 1 + \xi_k \left( \frac{u - \mu_k}{\sigma_k} \right) \right\}_+^{-1/\xi_k}
\end{aligned}
\tag{3.2.1}
$$

is the integrated intensity for each of the $s$ subspaces.

Note that the procedure above relies on the assumption of stationarity of the process over each covariate subspace, provided the subspace volume $\Delta_{k,u}$ is small enough. Hence, one should take care in the choice of the number of subspaces $s$, which should be large enough to allow for the assumption of stationarity to hold, whilst still being small enough to limit the effect on the computational cost required. One may want to investigate different choices for $s$, either via graphical methods, by taking into consideration the phenomenon at hand.

Then, in each subregion, the number of exceedances is just $N_{k,u} \sim \text{Po}(\Lambda_{k,u})$. Now recall that the sum of Poisson random variables is, itself, Poisson distributed, and assume independence between subregions. We can hence approximate the distribution of the total number of exceedances as $N_u \sim \text{Po}(\Lambda_u)$, where

$$
\Lambda_u = \sum_k \frac{1}{\Delta_{k,u}} \Lambda_{k,u}
$$

with the integrated intensities in each subspace rescaled by the subspace volume $\Delta_{k,u}$.

**Simulation algorithm**

The Poisson point process is represented as the points $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ are the covariate values with corresponding exceedances $y_i$.

Note that, while similar issues to those in the estimation of the intensity term in the model fit arise, here further complications are present, so that the methodology suggested in Section 3.1.2 is no longer sufficient. First, we cannot determine *a priori* how many exceedances to simulate, as the number of exceedances is a Poisson random variable, whose intensity is a function of the model parameters. Furthermore, simulating the covariates now requires solving for the upper limit of the integral of the intensity function, as detailed below.

The simulation procedure is as follows:

1. Fix a threshold $u$, the total number $n$ of observations required, which includes both exceedances and non-exceedances, and a combination of location, scale and shape parameters.

2. Consider $m$ covariates and define an overall distribution $g(\boldsymbol{x})$ for the covariates, as in step (1) for the GPD simulation in Section 3.2.1. Recall that a multivariate distribution may be needed and issues may arise in estimating the dependence structure of this distribution, unless independence of covariates is assumed. There are then three possible approaches to proceed, where in particular the first two aim to simulate covariate values conditional on the response being an exceedance.

3. Use one of the methods listed below to simulate covariates $\boldsymbol{x}^*$.

4. Compute $\mu(\boldsymbol{x^*}), \sigma(\boldsymbol{x^*}), \xi(\boldsymbol{x^*})$, and hence the GPD parameter $\sigma_u(\boldsymbol{x^*}) = \sigma(\boldsymbol{x^*})+$ $\xi(\boldsymbol{x^*})(u-\mu(\boldsymbol{x^*}))$. Then, simulate excesses $(y_1, \ldots, y_n)$ as $Y \sim GPD(\sigma_u(\boldsymbol{x^*}), \xi(\boldsymbol{x^*}))$.

5. If **method 3** was used, covariate values for non-exceedances are already available. For the others, simulate covariates below the threshold using any of the methods listed for covariate sampling, considering this time only the distribution of the covariates below the threshold.

**Method 1**

- Compute $\Lambda_u$ and hence compute the number of exceedances $N_{exc} \sim \text{Po}(\Lambda_u)$;

- Use numerical integration and root finding as follows:

- Simulate $z_1, \ldots, z_{N_{exc}}$ from an i.i.d. Uniform $(0,1)$ distribution;

- Define the cumulative distribution function (cdf) as

$$F(\boldsymbol{x^*}) := P(\boldsymbol{X} \leqslant \boldsymbol{x^*} | Y > u) = \int_{\boldsymbol{X}_{min}}^{\boldsymbol{x^*}} \frac{g(\boldsymbol{W}) \left\{ 1 + \xi(\boldsymbol{W}) \left( \frac{u - \mu(\boldsymbol{W})}{\sigma(\boldsymbol{W})} \right) \right\}_+^{-1/\xi(\boldsymbol{W})}}{\Lambda_u} d\boldsymbol{W}$$

$$(3.2.2)$$

- In order to obtain covariate estimates $\boldsymbol{x^*}$, use the probability integral transform and solve the integral in Eq. 3.2.2. In particular, use root finding methods to obtain the value of the upper limit $\boldsymbol{x^*}$ such that $F(\boldsymbol{x^*}) = z$ for each of $N_{exc}$ cases.

**Method 2**

- Perform the first step from Method 1;

- Proceed with the following empirical approach: define $s$ covariate subspaces, and take $\{(\boldsymbol{x_k}) : k = 1, \ldots, s\}$ to be the midpoints of the subspaces;

- For each subspace, determine $\mu(\boldsymbol{x_k}), \sigma(\boldsymbol{x_k}), \xi(\boldsymbol{x_k})$ (e.g. $\mu(\boldsymbol{x_k}) = \mu_0 + \mu_1 \times x_{k,1}$), and consequently the integrated intensity in the the subspace $\Lambda_{k,u}$ as in Eq. 3.2.1;

- Compute the empirical cdf $CDF_\Lambda$, then use the probability integral transform to simulate covariate values $\widehat{\boldsymbol{x}}$ for $(\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_s)$, as shown in Figure 3.2.1; note that this method essentially provides a discretised or "binned" version for solving Eq. 3.2.2 in **method 1**.
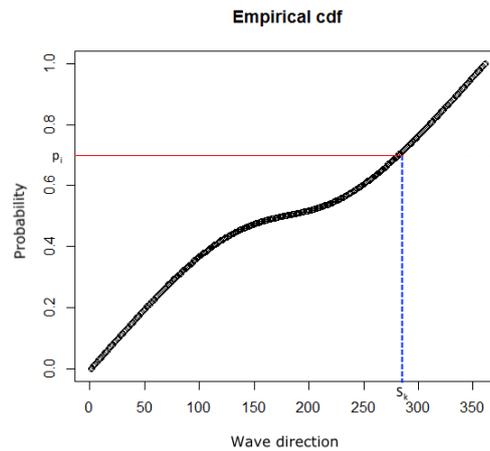


Figure 3.2.1: Empirical cdf and covariate subspace selection for wave direction.

In order to account for the fact that covariates are actually continuous, given that $\widehat{\boldsymbol{x}}_k$ has been simulated, a value for $\boldsymbol{x}^*$ is then simulated at random from $\mathcal{S}_k$.

**Method 3**

- Simulate $n$ covariates $\boldsymbol{x}^*$, following any of the methods mentioned in Section 3.2.1.

- Compute the point-wise integrated intensity

$$\Lambda_{\boldsymbol{x}_i^*, u} = g(\boldsymbol{x}^*_i) \left\{ 1 + \xi(\boldsymbol{x}_i^*) \left( \frac{u - \mu(\boldsymbol{x}_i^*)}{\sigma(\boldsymbol{x}_i^*)} \right) \right\}_+^{-1/\xi(\boldsymbol{x}_i^*)}.$$

- Proceed as in step (3) of the GPD simulation algorithm in Section 3.2.1 to determine for which of the $\boldsymbol{x}_i^*$ to simulate an exceedance, where we assume the conditional integrated intensity can be used to approximate the exceedance probability, following Eq. 3.1.8.

## 3.3   Simulation study

The non-homogeneous Poisson point process and the generalised Pareto formulations are known to be equivalent in the case of stationarity, as we have reviewed in Section 2.1.2. More complications arise where covariates are introduced, and we wish to consider in more details the case where there is non-stationarity in the data. The study presented in this section aims to determine if there is any noticeable difference in performance between the two methods in a simple case.

We consider three models for both the case where a positive shape and a negative shape is used, which is always assumed to be constant. This should provide enough insight for the performance of the methods in the case of finite and infinite upper limits. We then focus on the remaining NHPP parameters, taking $\mu$, $\sigma > 0$, or both to be covariate dependent in the form of a linear trend. We limit the study to the case where a single positive continuous covariate $X \in \mathbb{R}^+$ is considered. To avoid overcomplicating matters, we also do not consider the use of a varying threshold

neither at the simulation nor at the fitting stage, using instead the constant values $u_{\text{sim}} > 0$ and $u_{\text{fit}} \geqslant u_{\text{sim}}$ respectively, where the choice for the latter is motivated and detailed in Section 3.3.1.

We simulate $N = 200$ datasets for each of the parameter combinations shown in Table 3.3.1. We use the notation $\text{NHPP}_\theta^{+/-}$, for a generic model parameter $\theta$, to denote a NHPP process, where the subscripts shows the parameter(s) chosen to be covariate-dependent and the superscript refers to the sign of the shape. We follow the algorithm presented in method 3 in Section 3.2.2 to perform the simulation. As mentioned there, we cannot pre-determine the number of exceedances to be simulated. Instead, we fix the total number of observations $N_{\text{obs}} = N_{\text{years}} \times N_{\text{py}}$, where in our case we account daily data covering $N_{\text{years}} = 200$ years, such that $N_{\text{py}} = 365$ (for simplicity we do not account for leap years). Finally, the covariate $X$ is assumed to be following a Gamma distribution of choice, i.e. $X \sim \text{Gamma}(4, 1)$.

| Model | Npy | $N_{\text{Years}}$ | $N_{\text{Obs}}$ | $\mathbb{E}[\Lambda(B_{u_{\text{sim}}})]$ | Location | Scale | Shape | $u_{\text{sim}}$ | $u_{\text{fit}}$ |
|---|---|---|---|---|---|---|---|---|---|
| $\text{NHPP}_\mu^-$ | | | | 1125 | (30,3) | 20 | | | 23.68 |
| $\text{NHPP}_\sigma^-$ | 365 | 200 | 73000 | 588 | 30 | (20,0.5) | -0.1 | 5 | 8.91 |
| $\text{NHPP}_{\mu,\sigma}^-$ | | | | 951 | (30,3) | (20,0.5) | | | 21.05 |
| $\text{NHPP}_\mu^+$ | | | | 1688 | (30,3) | 20 | | | 25.68 |
| $\text{NHPP}_\sigma^+$ | 671 | 200 | 73000 | 690 | 30 | (20,0.5) | 0.1 | 5 | 10.77 |
| $\text{NHPP}_{\mu,\sigma}^+$ | | | | 1286 | (30,3) | (20,0.5) | | | 23.15 |

Table 3.3.1: Overview of parameter values used during simulation and fitting procedure.

Given the parameter values chosen (see Table 3.3.1), we can calculate the expected number of exceedances that are likely to be simulated by the accept-reject procedure outlined in the algorithm. An approximate value can be obtained by computing the expected integrated intensity following method 1 in Section 3.1.2. In particular, we cover the covariate space by taking $N_{\text{obs}}$ evenly spaced covari-

ates $x^*$ on $(0, 15]$, which provides good coverage for the covariate space, as shown in Figure 3.3.1. Then, we estimate the densities $g(x^*)$ from a Gamma$(4, 1)$, and using Eq. 3.1.6 we obtain the expected integrated intensity $\mathbb{E}[\Lambda(B_{u_{\text{sim}}})]$, where $B_{u_{\text{sim}}} = (0, 15] \times (u_{\text{sim}}, \infty)$. The resulting values are denoted by $\mathbb{E}[\Lambda(B_{u_{\text{sim}}})]$ in Table 3.3.1.

### 3.3.1 Model fitting

We consider the two peaks over threshold approaches presented in Section 2.1.2 for each of the parameter combinations, denoted in the following plots as $\text{GPD}_\theta^{+/-}$ and $\text{PP}_\theta^{+/-}$ respectively, following the notation introduced in Section 3.3. For ease of notation, we choose to use the abbreviation "PP" in the case of the specific models fitted following the NHPP approach and corresponding estimates and results, especially in the case of plotting and diagnostics. In order to allow a fair comparison across different parameter combinations, we select fitting thresholds that differ from the one chosen at the simulation stage. The aim is to ensure that each model is fitted, on average, to 500 exceedances. This is achieved by considering the 200 simulated datasets from each of the six models from Table 3.3.1, and calculating the median of the $500^{\text{th}}$ largest observation, yielding the values $u_{\text{fit}}$ shown in Table 3.3.1. Note that all the results that are presented here illustrate the optimal case for fitting, as we assume the "true" form of the covariate–response relationship has been correctly identified in all cases. That is, the same parameters are taken to be covariate dependent during fitting as selected at the simulation stage. It is worth noticing that, in terms of the GPD models being fitted, all of the

parameters combinations considered result in a covariate-dependent scale. Recall that we can compute the GPD scale as $\psi_u = \sigma + \xi(u - \mu)$, so that $\psi_u$ can be expected to be covariate dependent when any of the original NHPP parameters is. Finally, it is worth noticing that no link-function form is imposed on the scale parameters. Instead, negative values are disregarded by assigning an undesirable value, e.g. $-10^6$. to the resulting log-likelihood being maximised. This prevents the need to impose link functions, thus helping in general to obtain more directly interpretable parameter estimates. We also avoid the use of constrained optimisation by restricting the optimal support for the parameters instead of the feasible set. That is, an undesirable value is assigned to the log-likelihood being maximised when an infeasible parameter choice is encountered, e.g. when maximising, setting the log-likelihood to $-10^6$ if $\sigma(x_i) < 0$ for any observed covariate $x_i$.
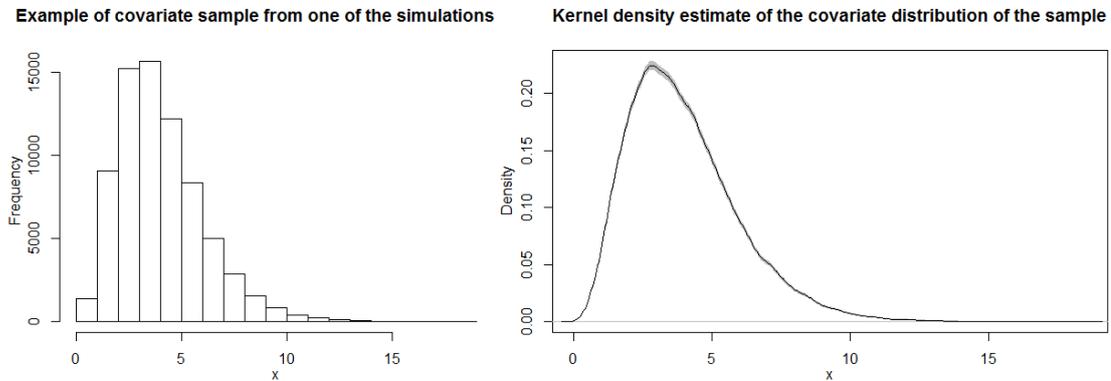


Figure 3.3.1: Histogram of a sample of covariates (left) and corresponding covariate density (right) with 95% confidence bands from kernel density estimation.

The distribution of the covariates plays an important role for the fitting stage, especially when an NHPP model is considered. In fact, standard kernel density methods need to be used in order to estimate the covariate distribution needed to compute the integrated intensity, as we previously discussed in Section 3.1.2. Figure 3.3.1 shows, on the left, an overview of the covariate sample for one of

the simulations, while on the right the estimated kernel density is plotted with 95% bootstrap confidence bands. The plots show the density being appropriately estimated, and hence we can dismiss this potentially affecting in any way the NHPP fitting procedure.

Maximum likelihood methodology was used for model fitting and inference, and the Nelder-Mead algorithm (Nelder and Mead, 1965, as implemented by the `stats:optim()` function in R) was chosen to perform optimisation, because of its robustness and derivative-free formulation, and bootstrapping is used to estimate uncertainty.

**Parameter estimation**

For the fitting procedure, a suggested initial value is provided for the shape parameter only. This is taken to be $\xi^{\text{init}} = +/-0.01$, with the correct sign being given in each case, although preliminary tests showed no differences when a generic small non-zero shape (e.g. $\xi^{\text{init}} = +0.01$) was used. For the remaining parameters, the intercept coefficients are computed using the L-moments methods (Hosking, 1998). In order to not provide any pre-determined guess in terms of the parameter dependence on the covariate, with, for instance, a positive, negative, strong or weak trend, any coefficient terms for the covariate were set equal to 0.

Since the probability of exceedance in the GPD model is computed separately, we first consider the resulting efficiency across the six different models, by considering a single simulation for each of them.
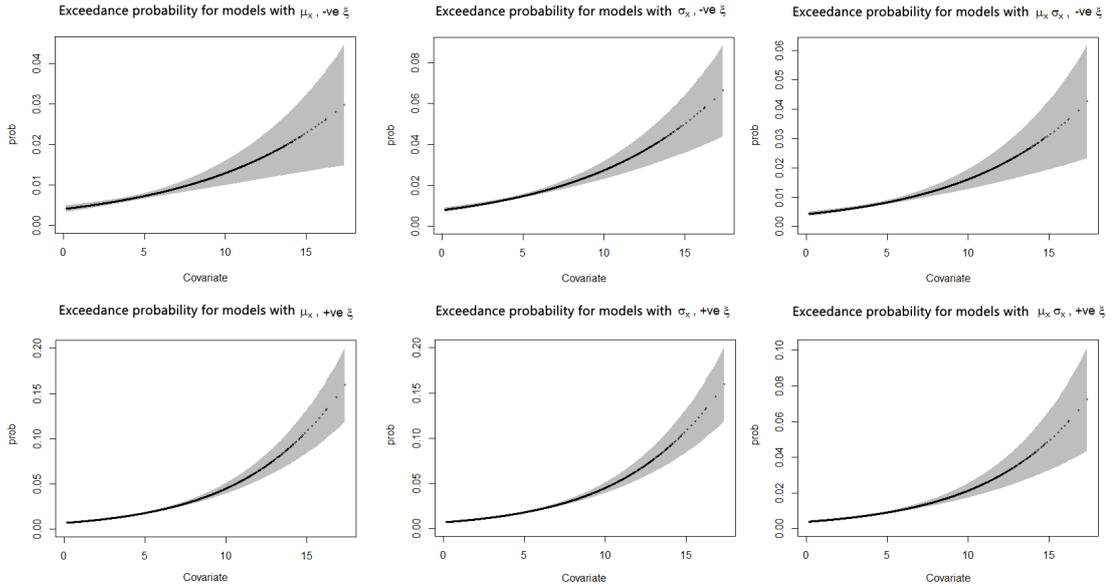
Figure 3.3.2: Probability of exceedance for the GPD model with 95% bootstrap confidence intervals, given simulations from all of the parameter combinations shown in Table 3.3.1.

The estimates are consistent, in size, across both the finite and infinite upper limit case. Higher exceedance probabilities are estimated for samples simulated from the latter case, in agreement with the higher estimates of the expected integrated intensity in Table 3.3.1. In general, wider confidence bounds are found for larger covariate values. This is expected, since a Gamma distribution was chosen, so that fewer large values are observed, providing less information for estimating the exceedance probability with confidence.

For each of the six models, we then compute mean squared error (MSE) and bias for the results of the 200 NHPP fits compared to the original simulation parameters.
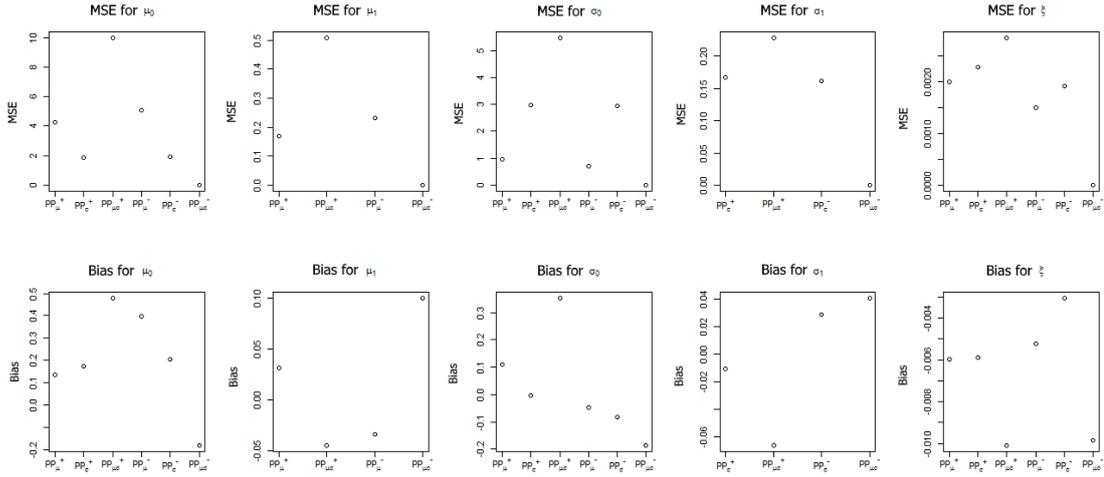
Figure 3.3.3: MSE and bias results for maximum likelihood estimates for each of the models from Table 3.3.1, as compared to the true model parameters.

The results, shown in Figure 3.3.3, show consistently good fits across the different parameters and models, with marginally larger biases, for all coefficients, detected when the data considered has been simulated with both location and scale parameters depending on covariates, given the added complexity of the model.
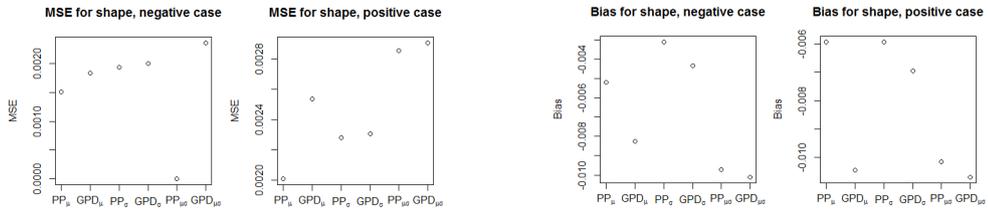


Figure 3.3.4: MSE and bias results for maximum likelihood estimates for the shape parameter for each of the models from Table 3.3.1, as compared to the true value using during simulation.

We consider the shape again separately in Figure 3.3.4, as this is the only parameter for which we can directly compare MLE from GPD fits with the original parameters. MSE and bias results show satisfying results across both models, with marginally better PP estimates, especially in the cases where only the NHPP location parameter is covariate-dependent.

In order to test whether fitting results from either POT approaches could be improved, we also repeated the investigation providing more informative starting values for the optimisation. For the GPD fits, we set the scale intercept to the mean value of the scale vector used at the simulation stage, adjusted for the new threshold using Eq. 2.1.9. That is

$$\psi_{u_{\mathrm{fit}}}^{\mathrm{init}} = \psi_{u_{\mathrm{sim}}}^{\mathrm{init}} + \xi(u_{\mathrm{fit}} - u_{\mathrm{sim}}), \qquad \text{where} \qquad \psi_{u_{\mathrm{sim}}}^{\mathrm{init}} = \frac{1}{n}\sum_{i=1}^{n}\{\sigma(x_i) + \xi(u_{\mathrm{sim}} - \mu(x_i))\}$$

where the $x_i, i = 1, \ldots, n$ are a sample of $n = 1000$ covariate realisations following the covariate distribution chosen for the simulations. The remaining starting values for the covariate term in the GPD scale and the shape parameter are set, respectively, to 0 and $+/-0.01$, the latter respecting the sign of the shape used for the simulations. To fit a NHPP, we instead use, for each simulated set, the best parameter combination emerging from a profile likelihood search for the shape of each specific simulation. That is, for each dataset simulated, we perform 10 optimisations for the shape profile likelihood, where the location and scale initial values are taken from the corresponding fit from Section 3.3.1. The 10 shape values considered are taken to be equally spaced, with magnitudes ranging between 0.01 and 0.3, and sign agreeing with the simulation set-up. We again compute mean squared error and bias for all of the sets of simulations and fits. The results did not show any consistent improvement for any of the model parameters, which suggests that both models succeed in computing, for most cases, parameter estimates leading to the optimum of the respective log-likelihood for the sample considered.

**Model prediction**

We follow the the methodology summarised in Section 3.1.2 to obtain conditional and marginal return values.

In this case, we use conditional return levels to investigate whether the fitting methods have been able to properly capture the covariate-response relationship imposed. To do so, we use the equations from Section 3.1.2 to compute the point-wise return value for a specific return period $N_{\mathrm{RP}}$, in this case 200 years, conditionally on the covariate value. These return values are taken equally spaced every 0.1 interval over the space $(0, 15]$, which we showed in Figure 3.3.1 provides a good coverage of the covariates space. This approach is used to compute both "true" return values, where each of the six parameter combinations from Table 3.3.1 is used, and estimated return values, using MLE's obtained during fitting. For the latter, given the consistency of results across the 200 datasets and the computational requirements for simulations, only the MLE's for 100 of the datasets were used.

Figure 3.3.5 then shows the median and upper and lower 5% empirical quantiles for the estimated return levels. Overall, the results obtained from using the NHPP procedure appear to be superior on two counts. First of all, the median return values generally agree with the "true" values, and properly capture the trend in the data. Secondly, narrower confidence intervals are obtained as compared to the GPD model fits. Both models achieve the best results for the cases where only the NHPP location parameter was, during simulation, covariate dependent. Wider confidence intervals are present for the four cases where the scale parameter was

covariate dependent, especially when this was in addition to a trend in location, due to the higher complexity in both the underlying process and the fitted model. In all cases, nevertheless, the "true" return levels always lie within the confidence bounds.
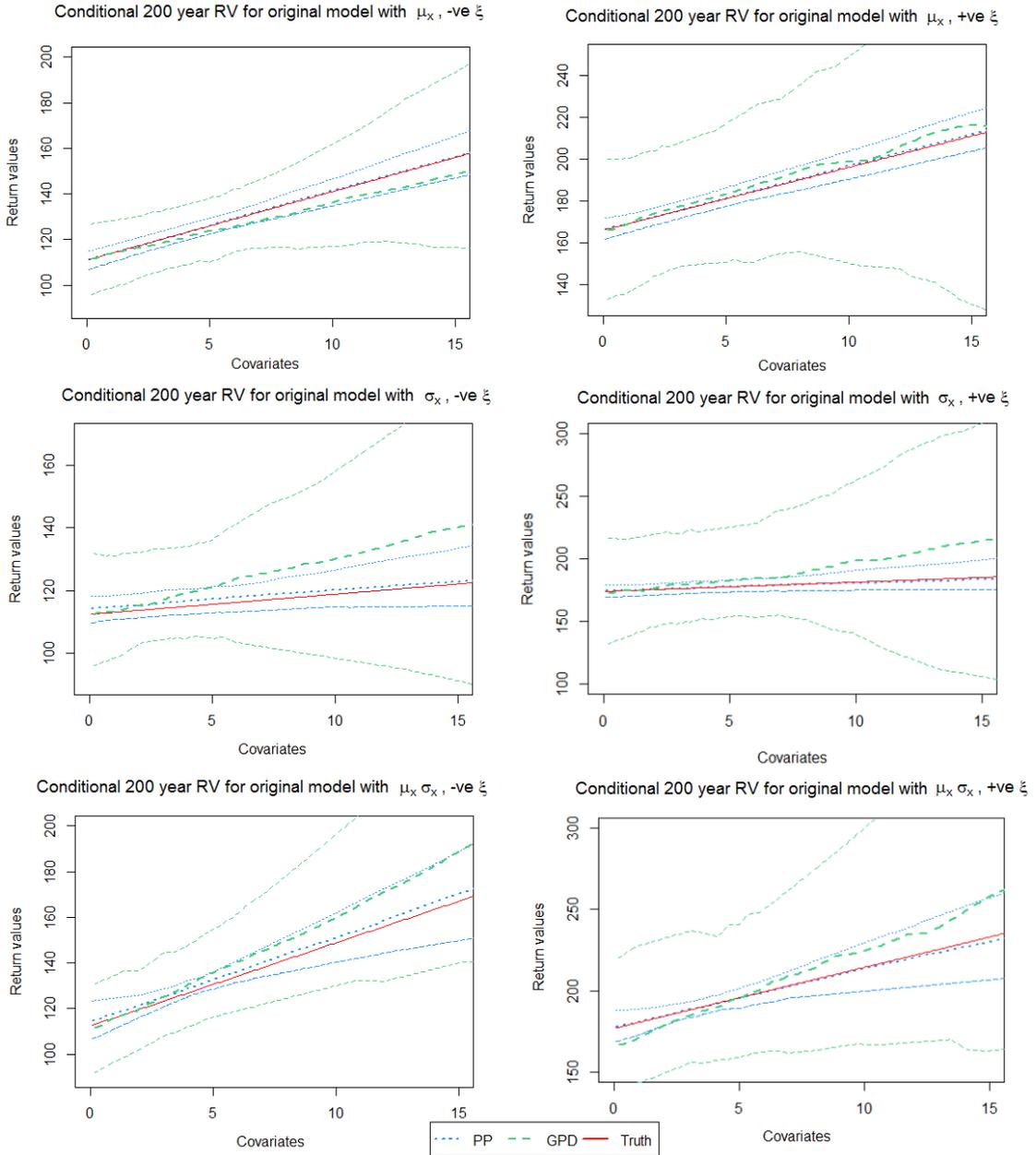


Figure 3.3.5:  Conditional return values for covariates on $(0, 15]$, and using a covariate-dependent exceedance probability for the GPD model.

It is possible that, as mentioned before in Section 3.3.1, the GPD model struggles to

identify with certainty the trend because of the simulation set up. As this followed the NHPP approach, separate location and scale parameter effects were introduced, yielding an implicit advantage to models fitted under the same structure.

To obtain marginal return levels, we produce 100 sets of $N_{\mathrm{Years}} = 10000$ years worth of data for each of the 200 fitted models, using the computed MLE's for the parameters. Then, recall the $N_{\mathrm{RP}}$-year return value is the value exceeded on average every $N_{\mathrm{RP}}$ years, and let

$$ R = \left\lfloor \frac{N_{\mathrm{Years}}}{N_{\mathrm{RP}}} \right\rfloor $$

Then, the $R$- largest observation in each dataset can be seen as a realisation of a sample of $N_{\mathrm{RP}}$ return values. Note that we take the floor of the division $\frac{N_{\mathrm{Years}}}{N_{\mathrm{RP}}}$ if the result is not an integer. In that case, we can only have bounds for what the $N_{\mathrm{RP}}$-year return value, and we take the smallest value for $R$, by using a floor function denoted as $\lfloor R \rfloor$, since this corresponds to the largest of two return values known. This is the more conservative choice, in line with usual policy in many of standard applications of extreme value analysis, where overestimation, if necessary, is preferred to underestimation for safety purposes. Simulating for a larger number of years partly reduces the need for this approximation. After observations for the desired return periods are collected, we can take quantiles to achieve a better overview. In this case, we calculate 5%, 50% and 95% quantiles, denoted as $q_{0.05}, q_{0.5}, q_{0.95}$, to obtain an estimated for the $N_{\mathrm{RP}}$-year return value we would expect on average ($q_{0.5}$) with 95% empirical confidence intervals, as shown in Figure 3.3.6.

We compare these results to what we take as "true" return values. We compute these using the same procedure, where this time we produce 1000 repetitions over the same time period $N_{\text{Years}} = 10000$, for each of the six original parameter combinations from Table 3.3.1. We then again compute the median for the resulting return values, which then shows the "true" average behaviour.
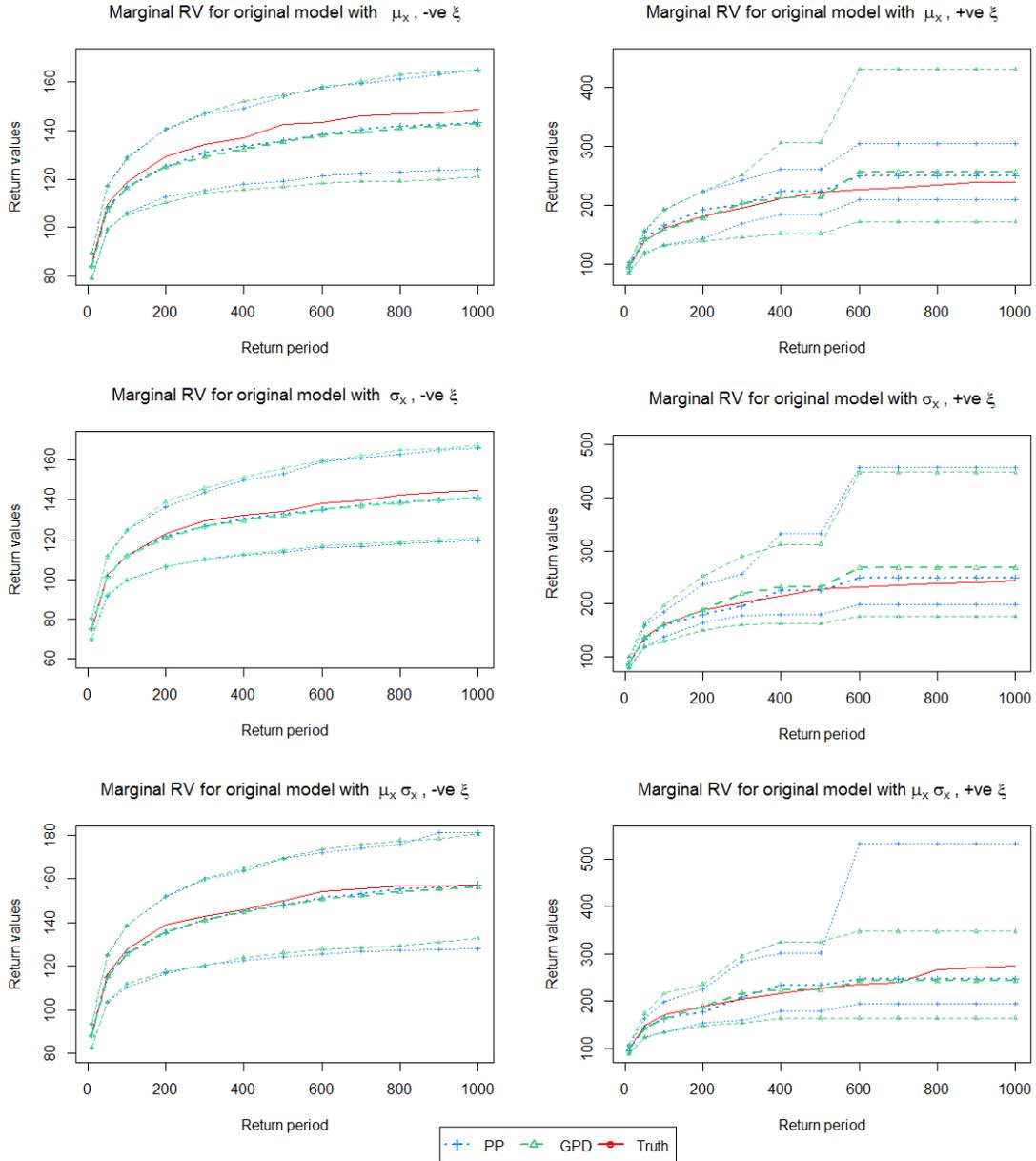


Figure 3.3.6: Marginal return values, where the exceedance probability for the GPD model is taken to be covariate-dependent.

Results are shown in Figure 3.3.6. Both models fitted seem to behave similarly, with the median values proving to be in general very close to the "true" average behaviour. When a negative shape is considered, 95% confidence intervals from both models are similar. The issues found for GPD conditional return levels seem to be less evident in this case, suggesting that both methods achieve similar return values (on average and for confidence bounds) when we extrapolate beyond the length of the original dataset. On the other hand, we notice a some differences when a positive shape is present, especially for the two cases where the location is covariate-dependent. In fact, it seems that, while the NHPP approach provides tighter bounds when the location only is taken to depend on the covariates, the GPD proves to be superior when the scale parameter depends on the covariates as well.

### 3.3.2 Discussion

The two models seem to behave similarly, especially within the context of the extrapolation we are most interested in. For conditional return levels, the fits obtained using a NHPP approach yield better estimates and tighter confidence intervals, as expected from the MSE and bias detected in Section 3.3.1. For marginal return levels, while at times one or the other provides marginally better estimates, neither of them shows to be overall superior given the simulation study provided.

It is worth reflecting on the ability of the study to represent data available in real-life situations. While the choice of a Gamma distribution is common in many applications, one may also want to consider different cases, both within and outside

this family of distributions. Furthermore, although a choice of shape parameter $\sigma = 20$ was made to ensure a discernible trend in the data, we would like to investigate more realistic cases, to address whether the performance of the NHPP and GPD formulation changes in the presence of a more subtle trend. Finally, we chose to simulate 200 years of data to ensure we had enough exceedances for fitting across all of the six different models considered. Nevertheless, environmental data is unlikely to cover such a long span of time.

A few possible extensions for the current study, although outside the scope of this chapter, may provide further ground for investigation and for comparison of the two models. So far, we have only introduced one covariate with a linear formulation. It would be interesting to study the case where more than one covariate is present and potentially correlated. More (e.g. Fourier series) formulations should also be considered, as well as possible ways to address different behaviours in the data, such as censoring. This would, for example, allow us to simulate sets of data which can replicate the land-shadow effect in oceanography applications, such as for the dataset analysed in Section 3.4. Both extensions will likely complicate the log-likelihood surface, hence providing a chance to further test the efficiency of the models and the sensitivity to different starting values. Furthermore, we have so far preferred restricting the optimal space over introducing link functions. It would be useful to compare the two approaches, especially in a more complicated setting, in order to asses whether space restrictions may cause some issues as compared to a re-formulation via link functions which yields an optimal space without internal discontinuity regions.

Finally, as we mentioned before, all of the simulations are constructed using a

NHPP structure. Although the two POT approaches are equivalent for stationary processes, complications arise when covariates are introduced. Model parameters and results are not as easily transferable across different approaches, so that constructing the simulated datasets with a NHPP procedure may provide some implicit advantage to the NHPP fitting approach, both in terms of parameter and return values estimation. Although this may only be marginal, for completeness one may want to repeat this study using a GPD procedure for the simulation stage.

## 3.4 Hindcast data study

In order to further investigate the performance of the NHPP and GPD models, we compare them for both fitting and inference given a real-life application. In particular, we start from a dataset of South China sea wave height observations (Hs), introduced in the following section. We then proceed to fit a covariate-dependent model under both approaches. In the case of stationary datasets, diagnostic models are considered to assess the fit obtained, especially quantile-quantile (QQ) plots, as they focus on the fit at the tail of the distribution.

Standard diagnostic plots are not as informative in the case of covariate-dependent models. In fact, it is important to assess whether the model has been able to capture accurately the relationship between covariate and response variables. We attempt to investigate this point by considering simulations and return values, noting that the former can be the basis for the latter, as explained in Section 3.4.2. In particular, using simulated datasets aids in considering model fitting

under controlled conditions.

### 3.4.1   South China Sea

The data consists of hourly hindcasts of significant wave hight (Hs) between the $1^{st}$ January 1957 and the $31^{st}$ December 2006 in the South China Sea. As we outlined in Section 2.1, we have reasons to believe the hindcast data would not be well summarised by a stationary model. In this case, it is important to understand the phenomena that cause high waves and, hence, the physical characteristics of the area of interest. These, in fact, may shed light on some key factors and be helpful in building a more accurate model.

The wave climate in the South China Sea is determined by a contribution of both waves from local winds and swell, that is waves caused by distant weather systems. Noticeably, the most severe sea states are wind generated, so that it is worth considering the overall climate of the area, which is of monsoonal type.
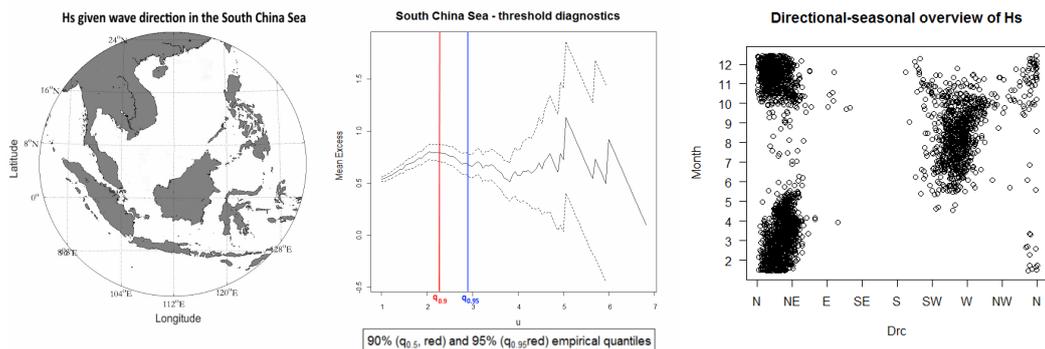


Figure 3.4.1: Overview of the South China Sea location(left), mean residual life plot for threshold selection (middle), with $q_{0.9} := 90\%$ $q_{0.95} := 95\%$ empirical quantiles of the data shown, and an overview of the directional-seasonal relationship (right).

In summer, monsoonal winds blow mainly from the South-West, while in winter,

monsoons occur in the North-East, with the consequent monsoonal surges causing increased sea states. The South China Sea is also affected by typhoons, which cause the most extreme conditions in the area.

For the South China sea data, using both Figure 3.4.1 and knowledge of the local climate, we have identified both seasonal and directional behaviour, so that in summer, monsoonal winds blow mainly from the South-West, while in winter, winds blow mainly from the North-East. Borneo also affects wind flows and wave height, causing what is known as "land-shadow" effect from the South-East. Hence, we decide to consider wave direction as a covariate for the model. We notice three distinct behaviour seem to arise, from different wave directional sectors. In particular, we select the following sector division, shown on the map in Figure 3.4.1, which agree with the climate behaviour described:

- **Sector 1**: $315^o - 60^o$: capturing north-east winter winds;

- **Sector 2**: $60^o - 200^o$: south-east quadrant, affected by land-shadow;

- **Sector 3**: $200^o - 315^o$: capturing south-west summer winds.

This selection is merely for investigation purposes, as the models we are considering are continuous and should be able to capture this sectoral behaviour through the covariate-response relationship. It is, nevertheless, interesting to perform model fitting for the first and third sector only, using the same threshold value as the one used further on for covariate-dependent models, i.e. taking the 90% quantile of the entire $Hs$ dataset as threshold. Note that we ignore the second sector as, due to the land-shadow, there are no exceedances above the threshold chosen. Then, if indeed there is a difference in behaviour in the two sectors, we will obtain

different parameter estimates. This is, in fact the case, giving the following NHPP (henceforth denoted, for ease of notation, as PP when used to indicate a model fit rather than the general modelling approach) and GPD maximum likelihood estimates:

**Sector 1** : PP with $\mu = 1.873\,(0.097), \quad \sigma = 0.966\,(0.046), \quad \xi = -0.122\,(0.046),$

and

GPD with $\psi^{(u)} = 0.9\,(0.071), \quad \xi = -0.122\,(0.046), \quad \phi^{(u)} = 0.13;$

**Sector 3** : PP with $\mu = -2.654\,(0.138), \quad \sigma = 1.852\,(0.135), \quad \xi = -0.241\,(0.203),$

and

GPD with $\psi^{(u)} = 0.663\,(0.188), \quad \xi = -0.241\,(0.203), \quad \phi^{(u)} = 0.03,$

where the rates $\phi^{(u)}$ are approximated following Eq.2.1.11. The results do differ noticeably, with the confidence intervals not overlapping for either the location or scale PP estimates, which is suggestive of non-stationarity in the series. Wider confidence intervals for the third sectors are due to much fewer exceedances (only 24) for fitting, so that there is not enough reliable evidence to believe in a change in the shape behaviour. Nevertheless, the results support the hypothesis of an underlying directional behaviour, and hence the use of a covariate dependent model. Note that this, being continuous in nature, will allow us to avoid fitting a specific model to a reduced number of observations for one of the sectors.

We perform likelihood ratio tests to assess which mathematical form to use to describe the covariate-dependent behaviour, as well as to determine which parameters should be covariate-dependent. We consider Fourier series formulations for

the parameters, as it is a natural approach given the periodicity of the covariate. We also test different orders of Fourier series, with the second order description being the most relevant according to test results. This process leads to the following two models for observations $\boldsymbol{Y}_{SCS}$:

(1) $Y_{SCS} \sim GPD(\psi^{(u)}(x), \xi)$, where

$$\psi^{(u)}(x) = \psi_0^{(u)} + \psi_1^{(u)} \cos(x_{WavDrc}) + \psi_2^{(u)} \sin(x_{WavDrc})$$

$$+ \psi_3^{(u)} \cos(2x_{WavDrc}) + \psi_4^{(u)} \sin(2x_{WavDrc})$$

$$\log\left(\frac{\phi^{(u)}(x)}{1-\phi^{(u)}(x)}\right) = \phi_0^{(u)} + \phi_1^{(u)} \cos(x_{WavDrc}) + \phi_2^{(u)} \sin(x_{WavDrc})$$

$$+ \phi_3^{(u)} \cos(2x_{WavDrc}) + \phi_4^{(u)} \sin(2x_{WavDrc})$$

(2) $Y_{SCS} \sim PP(\mu(x), \sigma, \xi)$, where

$$\mu(x) = \mu_0 + \mu_1 \cos(x_{WavDrc}) + \mu_2 \sin(x_{WavDrc})$$

$$+ \mu_3 \cos(2x_{WavDrc}) + \mu_4 \sin(2x_{WavDrc}),$$

where the exceedance probability $\phi^{(u)}(x)$ is also dependent on the covariate. This is obtained by performing logistic regression on the observations (as exceedances and non-exceedances) given the covariate, and is formulated using a link function as common in the literature. Note that we do not impose link functions on any of GPD parameters. Instead, we insure the parameter estimated are feasible by artificially restricting the optimal space, as previously explained in Section 3.3.1. Due to the nature of logistic regression, $\phi^{(u)}(x) \in [0,1]$ by construction.

A theme that emerges throughout this work is the sensitivity of the Poisson point process to initial values chosen, especially when a non-stationary model is considered. This also emerged when we considered introducing covariates in the scale parameter of the PP model, that is the case $PP(\mu(x), \sigma(x), \xi)$. Here, the strong

dependence between the location and shape parameter led to terrible fits when standard starting values were used in numerical optimisation (L-moments). It was necessary to fine-tune the choice by using the MLE's from the model $PP(\mu(x), \sigma, \xi)$ to obtain a suitable fit, although the improvement was still not enough to justify considering a covariate-dependent scale.

We then proceed to use models (1) and (2) above, using the 90% quantile of the observations as threshold. Both seem to fit the data adequately, as shown in Figure 3.4.2.
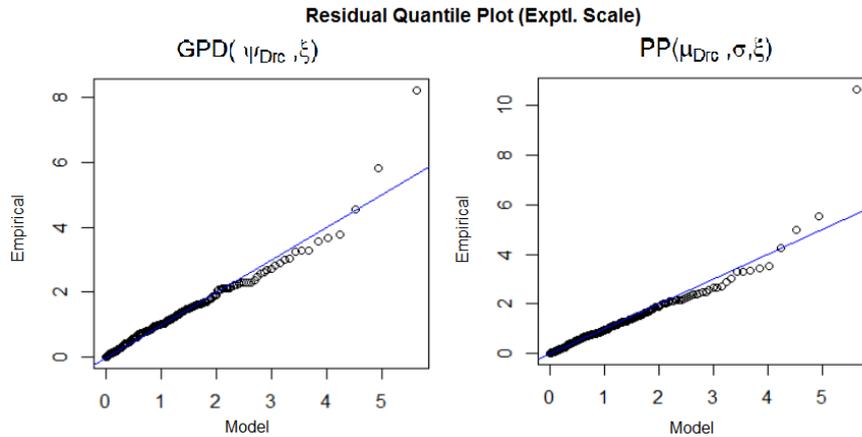


Figure 3.4.2: QQ plot diagnostics for fitting GPD and PP directional models.

As mentioned earlier, quantile-quantile plots are not the only useful tool used to assess the fit of covariate-dependent models. For the next step, we proceed to use simulation to see if the model fit captures the underlying behaviour in the data.

## 3.4.2  Simulation study

The next few sections implement or refer to some of the methods available to compare the models considered and their ability to properly capture the non-

homogeneous characteristics of the data.

For the following section, we use simulation. While the original dataset contained 2797 observations, we round up the total number of observations to 2800 data points for all of the simulations produced, where we would expect the number of exceedances to be approximately 10% given the choice of threshold from the previous section. Furthermore, no assumption is made in terms of distributions of the covariate, which are approximated, both during the fitting and the simulation stages, via kernel density estimation. The covariate distribution is needed to compute the integrated intensity for the NHPP approach, as we previously discussed in Section 3.1.2. Figure 3.4.3 shows, on the left, an overview of the covariate sample for the dataset, while on the right the estimated kernel density is plotted with 95% bootstrap confidence bands. The plots show the density being appropriately estimated in this case, and hence we can dismiss this potentially affecting in any way the NHPP fitting procedure.
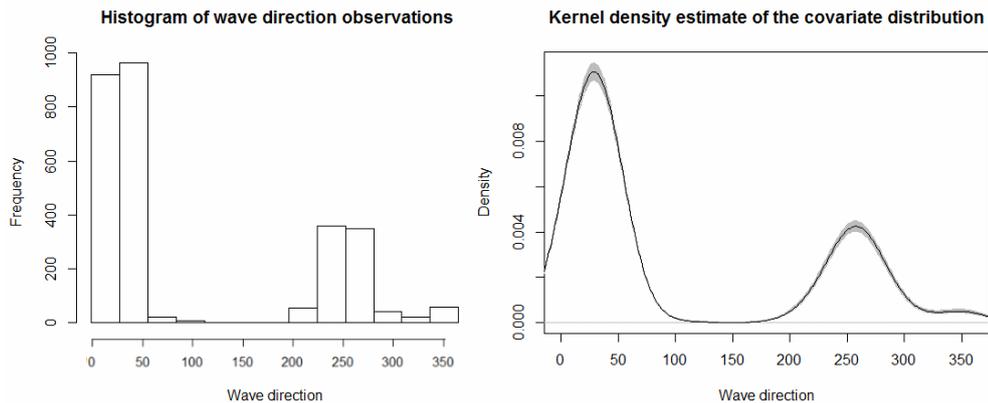


Figure 3.4.3: Histogram of wave direction covariate (left) and covariate density with 95% confidence bands from kernel density estimation.

Similarly, it is important to understand how well we can estimate the exceedance probability $\phi^{(u)}(x)$ in the GPD model. In fact, although this will not affect the

estimate of the remaining GPD parameters, which are orthogonal to it, it will

impact the quality of inference and prediction.
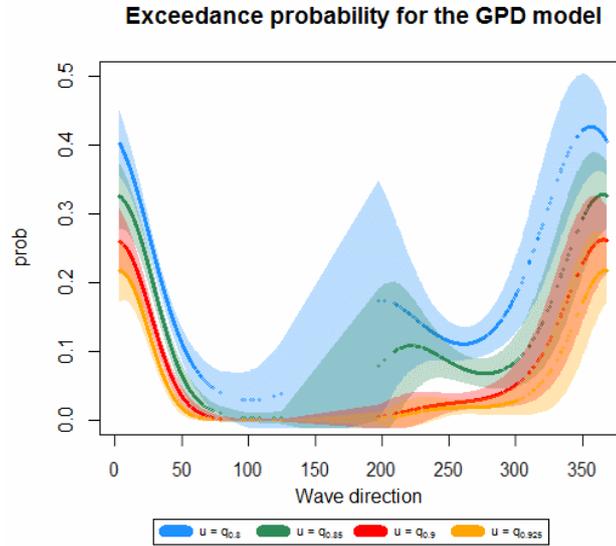


Figure 3.4.4: Probability of exceedance estimated from logistic regression given the wave direction observations, with 95% confidence intervals. Results are shown for different choices of threshold, all taken as empirical quantiles of the original dataset, which will be used further in Section 3.4.2

.

The exceedance probability in Figure 3.4.4 seems to be able to capture the covariate-response relationship well. As expected, wider confidence intervals arise in the land-shadow sector, due to the sparsity of information available. Lower threshold values also lead to higher exceedance probabilities, as is especially visible in sector 3, where we see more clearly the impact of summer climate for the 80% and 85% quantile thresholds.

We then produce two sets of simulations:

Sets 1 We fit models (1) and (2) from Section 3.4.1 to the observed data, and use parameter estimates to simulate N=1000 datasets, which we will refer to as

GPD1 and PP1.

**Sets 2** We fit again models (1) and (2) to each simulated dataset, and use the new parameter estimates to produce a new set of simulations, referred to as GPD2 and PP2 respectively.

All of the sets should highlight similarities and difference with the observed data. In particular, inconsistencies may be due to:

1. Issues with the models fitted and the functions chosen to represent the dependence of the parameters on the covariate;

2. Computational error from model fitting;

3. Computational error from the simulation stage.

We used two rounds of simulations in order to try and disentangle the effects of these three main causes of inconsistency. For example, issues with the choice of functional form for the distribution parameters will mainly affect results in Set 1. The second round of simulation is then an attempt to reduce the potential model misspecification issues from point 1 and focus on the latter two. In fact, by using the same structure and assumptions during the simulation and inference stages, model misspecification is no longer an issue and equal footing is provided to both the Generalised Pareto and the NHPP approaches. We can then compare the results from Set 1 and Set 2: any inconsistency in results from the two sets within each approach are then most likely linked to points 2 and 3 above, and could highlight a potential superiority of one modelling approach over the other.

**Frequency and size of exceedances**

A preliminary way to assess the methods at hand is to investigate the average ability of reproducing, via simulation, an "observed" set of data. In particular, both the number and the size of exceedances should be reproduced. Furthermore, given the interest in non-stationary processes, we want to analyse these traits in relation to the covariates considered.

We have reviewed both seasonal and directional behaviour for the South China Sea in Section 3.4.1, where we identified three different directional sectors of interest. For each sector, we compare exceedance counts and sizes for a series of simulations, against those from the original dataset.

In particular, we focus on the exceedances above the 90% quantile (also used as a threshold in the model fitting), as well as the 95% quantile, where the latter is considered to investigate how well the most extreme behaviour is captured. These are also shown in Figure 3.4.1, in red and blue respectively
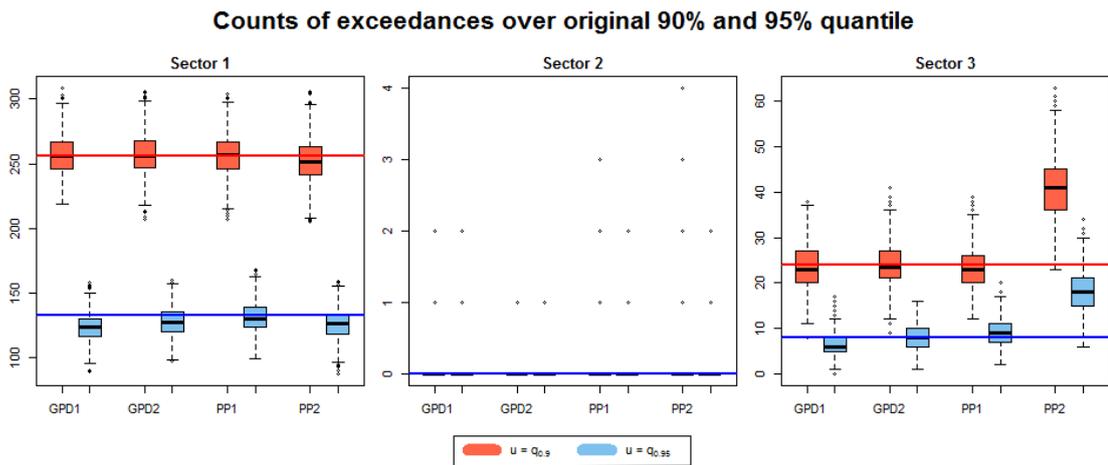


Figure 3.4.5: Overview of the spread of simulated exceedances per sector, with **sector 1** $= [315^o, 60^o)$, **sector 2** $= [60^o, 200^o)$ and **sector 3** $= [200^o, 315^o)$. The corresponding number of exceedances of the original dataset over the 90% and 95% empirical quantiles are shown by the red and the blue line respectively.

Boxplots of exceedance counts show consistent results for both models, although the second set of simulations using a NHPP model, denoted as PP2, seems to lead to an overestimate of the number of exceedances in **sector 3**. The land-shadow effect in **sector 2** seems to be well captured across all sets of simulations. Hence, we focus on the remaining sectors to analyse sizes of exceedances. In particular, we want to investigate how well we can reproduce the tail behaviour of the exceedances distribution in both sectors. Hence, we compare 50%, 90% and 95% empirical quantiles of the simulated exceedances with the observed ones.
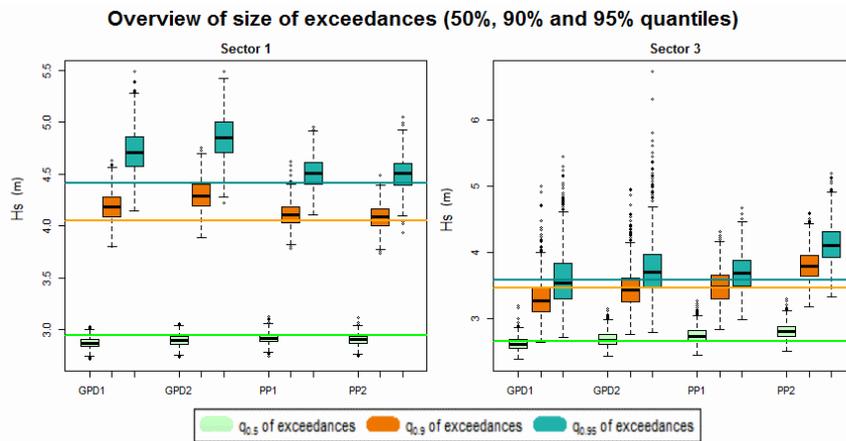


Figure 3.4.6: Overview of the size of simulated exceedances per sector, with **sector 1** $= [315^o, 60^o)$, **sector 2** $= [60^o, 200^o)$ and **sector 3** $= [200^o, 315^o)$. The lines show the 50%, 90% and 95% empirical quantiles of the original exceedances.

Exceedance size in **sector 1** is consistent with the original observations, although slightly overestimated by the GPD fits. Nevertheless, more variability and overestimation arise in **sector 3**, as later confirmed by conditional return values in the next Section (see Figure 3.4.7). While the GPD fits seem to be capturing the size of exceedances in this sector better on average, they also yield a much wider range of values. On the other hand, PP fits show a smaller range of results, although they overestimate on average the Hs values for the second set of fits. As expected,

simulated datasets are more consistent under the same model for the GPD, which suggest that the issue may lie in the model assumed for the covariate-response relationship. We also need to note that a non-parametric approach has been used for the intensity estimation in the PP model fits. This is, in general, computationally intensive, especially if simulation is used to estimate return values, and in most cases we may have to specify a parametric model for the covariates. The stronger connection to the original dataset, nevertheless, may be the reason for the ability of the NHPP approach to capture with less uncertainty the behaviour in the third sector. In fact, a non-parametric model for the covariate distribution could reduce the impact of a too-rigid model for the response-covariate relationship, which may be the cause of the overestimate of the size of exceedances for the GPD approach in **sector 1**.

To explore these issues further, we proceed to consider return values, as they summarise the estimates of all parameters. In the next section we also review some issues that arise for return value estimation for non-stationary processes.

**Return values**

We follow the the methodology summarised in Section 3.1.2 to obtain conditional and omni-directional return values. For the latter, we produce 1000 sets of $N = 100$ years worth of data, in order to also compute empirical confidence intervals.

Both plots show a tendency of the model to underestimate return values. This can be noticed by also comparing the 50 year return levels to the maximum observation in the original 50-year data (6.915m), which lies outside both the confidence

intervals, with the PP fits yielding a more noticeable underestimate.

The conditional return levels plot shows an interesting feature: the model does not seem to capture the covariate-response relationship in a satisfactory manner. In fact, it leads to underestimates in what we defined as the first sector, possibly due to the occurrences in the third. Vice versa, values for the third sector seem to be overestimated, especially for the PP fit results, while here the GPD results seem to impose a behaviour in disagreement with the climate and morphology of the location.
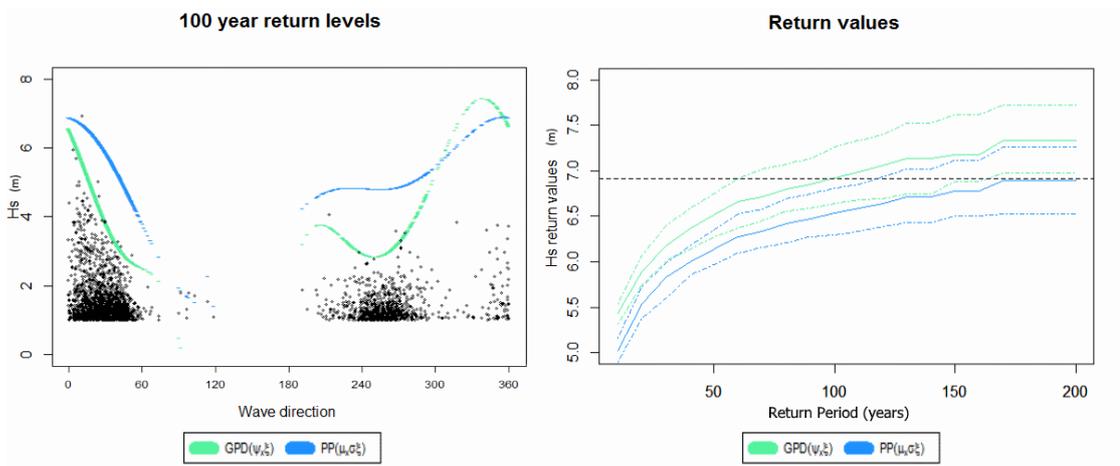


Figure 3.4.7: 100 year conditional return levels (left) and omnidirectional (right) return levels with 95% empirical confidence intervals, computed from models (1) and (2) fitted in Section 3.4.1. The dashed line identifies the maximum observation recorded in the 50 years of data, namely 6.915 m.

This "balancing out" ultimately leads to lower overall return levels, as the highest exceedances, occurring in the first sector, are underestimated. As mentioned before, this may be due to the rigidity of the covariate function for the model parameters, with the Fourier series model possibly failing to capture its complexity.

**Threshold choice and model sensitivity**

The choice of threshold is a potential source of controversy, with no clean-cut rule to select a value so that we are far enough in the tail to ensure the limiting theory is a reasonable approximation. It is then interesting to consider the effect that such a choice has on model estimates and, hence, extrapolation. The next sections focus on comparing results from a few selected constant thresholds illustrated in Figure 3.4.8, as well as provide an preliminary attempt to consider a covariate-dependent threshold.

**Multiple thresholds compared** A theoretical advantage of the Poisson point process model is threshold stability. That is, provided that we are considering extreme enough observations, any subset of those obtained from a higher choice of threshold should be described by the same distribution, i.e. the location, scale and shape parameters will remain consistent.
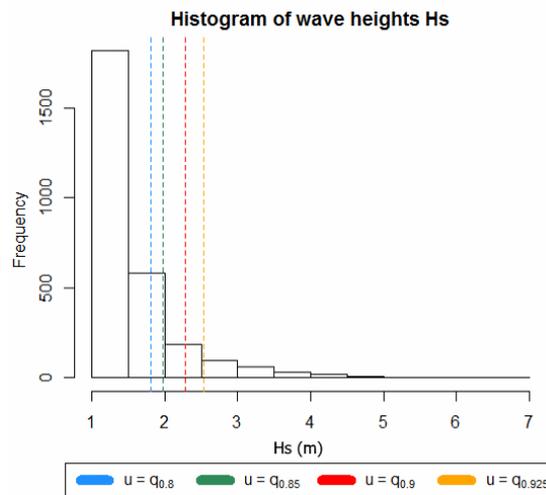


Figure 3.4.8: Overview of empirical quantiles of the original dataset, to be used as threshold values.

It is then interesting to see whether, indeed, that is the case, and similarly consider

what effect the same choice of thresholds has when fitting the GPD model. We fit

| | | PP MODEL | | GPD MODEL |
|---|---|---|---|---|
| | No.Exc | Scale $\sigma$ | Shape $\xi$ | Shape $\xi$ |
| $q_{0.8}$ | 559 | 0.587 (0.54,0.65) | -0.146 (-0.19,-0.1) | -0.154 (-0.21,-0.06) |
| $q_{0.85}$ | 418 | 0.607 (0.57,0.66) | -0.19 (-0.23,-0.13) | -0.207 (-0.29,-0.1) |
| $q_{0.9}$ | 280 | 0.629 (0.58,0.69) | -0.207 (-0.26,-0.14) | -0.225 (-0.29,-0.14) |
| $q_{0.925}$ | 210 | 0.631 (0.58,0.71) | -0.197 (-0.25.-0.12) | -0.212 (-0.25,-0.19) |

Table 3.4.1: Estimates for constant parameters for both GPD model (1) and PP model (2), together with profile likelihood confidence intervals.

models (1) and (2) for a series of threshold values corresponding to the empirical quantiles shown in Figure 3.4.8. As shown in Table 3.4.1, constant parameters have estimates similarly consistent across threshold values for both models. The same stands for the covariate-dependent parameters shown in Figure 3.4.9, although we notice more variation for the estimates of the PP location parameter. Both models show similar uncertainty in the land-shadow sector, given the lack of information available. It is necessary to notice, nevertheless, that part of the covariate behaviour is captured, for the GPD model, by the exceedance probability (see Figure 3.4.4). This shows consistency across different thresholds, while an increased ability to detect the local behaviour in **sector 3** appears for lower threshold values.
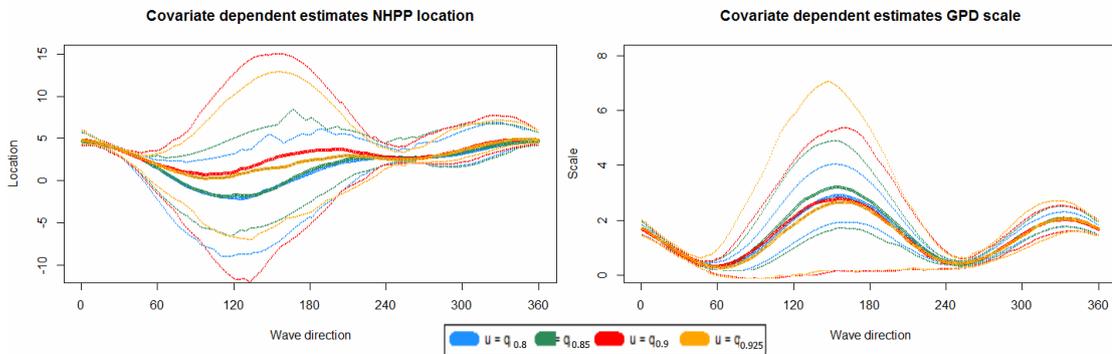


Figure 3.4.9: Covariate-dependent parameter estimates from a repeated simulation-fitting routine, using empirical quantiles of the original dataset as thresholds, for both GPD model (1) and PP model (2)

Then, we want to focus on shape estimates as a source of comparison, since location and scale parameters are not easily comparable for covariate-dependent GPD and PP model fits, but both approaches should lead to consistent shape estimates, if the shape is taken to be constant, as in our case. In order to do so, for both models we obtain profile likelihood confidence intervals for the parameter estimates, as shown in Table 3.4.1. Furthermore, we consider the overall result of parameter estimates for the GPD model by estimating return values given a range of thresholds. In this case, we produce 100 sets of $N = 100$ years worth of data, in order to also compute bootstrap confidence intervals.
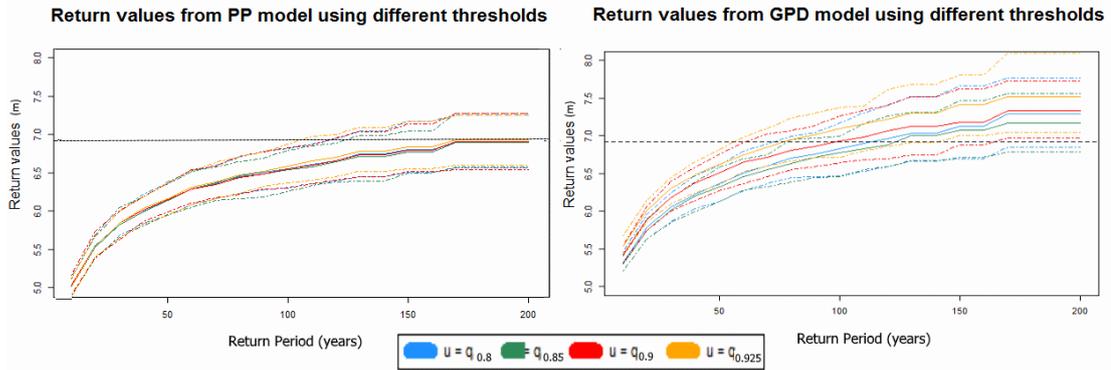


Figure 3.4.10: Marginal GPD (right) and PP (left) return values from fitting model (1) and (2) respectively with empirical quantiles of the original dataset as thresholds, with 95% empirical confidence intervals.

Both return values in Figure 3.4.10 and parameter estimates in Table 3.4.1 seem to be consistent amongst different thresholds. Shape estimates, in particular, seem to follow a similar pattern for changing thresholds under both models. In other words, the NHPP model does not stand out as superior in terms of threshold stability. It may nevertheless be worth noticing that the GPD estimates seem to have wider confidence intervals, especially for higher quantile choices with fewer observations selected.

**Non-stationary threshold**   We have noticed the potential consequences of choosing a fixed threshold during the previous stages of our simulation study. In particular, this may be one of the reasons for a levelling of the predicted behaviour over the covariate space, with overestimation of the behaviour in the third directional sector, and underestimation of the wave heights in the first one.

It may also be the case that a constant threshold fails in selecting appropriately extreme observations, in other words, the threshold should be large enough that, for any given covariate value, we are far enough in the tail to model the exceedance magnitudes with a GPD. However, a threshold that is sufficiently large for one covariate value may be too low for another one. Furthermore, from a statistical point of view, the estimation of a covariate effect would be more precise if the exceedances considered are spread as far across the observed values of the covariate as possible. A constant threshold, instead, will likely reduce the range of covariates that lead to exceedances.

Hence, we consider using a covariate-dependent threshold in an attempt to improve the model. In this case, we decide to follow the approach introduced by Northrop and Jonathan (2011), who promote the use of quantile regression to set a threshold for which the probability $p$ of threshold exceedance, chosen as $p = 0.9$ in our case, is approximately constant across different values of the covariate(s), namely wave direction.

Some care should be taken in the computation of threshold values. In particular, we used spline models to perform the quantile regression. These require a choice of the number of knots, which essentially determines how many curves are used to

perform the fit and how smooth the final result will be. Figure 3.4.11 shows, on the left, the 90% quantile threshold selected for a few choices of number of knots. Also shown in Figure 3.4.11 are the different data points selected by (a) a constant threshold, taken as the 90% empirical quantile, and (b) a covariate-dependent threshold, corresponding to 90% quantile regression values obtained using spline fitting with $k$=36 knots (see Section 2.3.4 for further details on spline formulation).
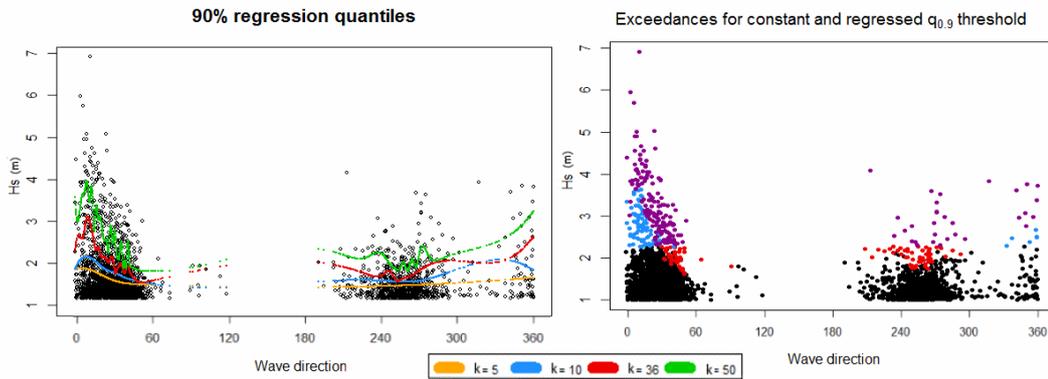


Figure 3.4.11: Values of 90% quantile regression over wave direction using different number of spline knots (left). Exceedances selected only using a constant 90% quantile threshold (blue), only from a non-homogenous 90% quantile regressed threshold with $k = 36$ knots (red), and from both thresholds (purple).

We then proceed to fit both stationary and covariate-dependent models, as summarised in Section 3.4.1.
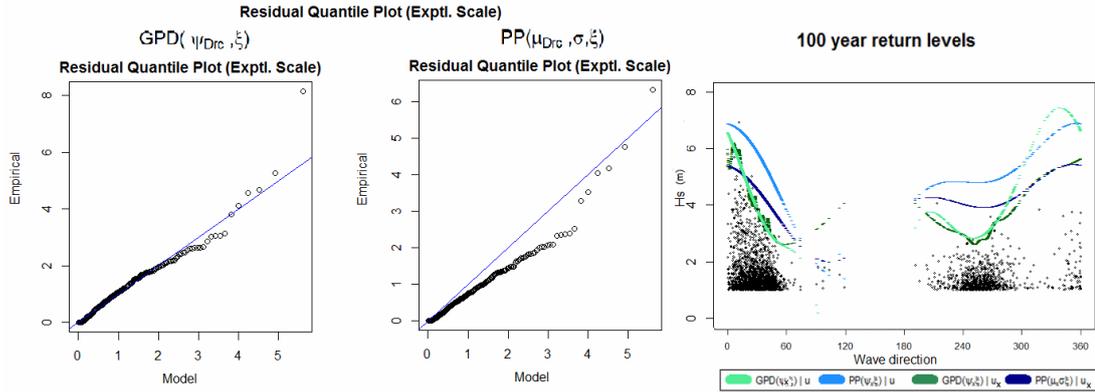
Figure 3.4.12: QQ plot diagnostics for fitting GPD model (1) and PP model (2) with a non-homogenous 90% quantile threshold regressed over wave direction observations (left). 100 year conditional return values compared for GPD and PP models for constant (90% quantile) and non-homogeneous thresholds (right).

The PP model does not provide an adequate fit. It is likely that the use of a covariate-dependent threshold has further complicated the likelihood surface, as well as the selection of starting values, which rely on the L-moments method. The return value plot shows a slight improvement compared to the constant threshold case.

The GPD model seems to lead to a more desirable fit. If we consider conditional return values in Figure 3.4.12, we see that the model seems to be doing slightly better than for a constant threshold for GPD fits, although the covariate-dependent behaviour is still not fully captured. As for constant threshold, we notice there is a risk for underestimation of the return values in the **sector 1**, which is especially evident for in the PP fit.

It is worth considering the parameter estimate and confidence intervals. Table 3.4.2 provides an overview for the constant parameters estimates, which seem fairly stable across different threshold values. Note that the positive estimate for the GPD shape shown in Table 3.4.2, and obtained using a covariate-dependent

| | No.Exc | PP MODEL | | GPD MODEL |
| | | Scale $\sigma$ | Shape $\xi$ | Shape $\xi$ |
|---|---|---|---|---|
| $q_{0.8}$ | 578 | 0.589 (0.46,0.62) | -0.011 (-0.13,0.1) | -0.062(-0.15,-0.04) |
| $q_{0.85}$ | 420 | 0.542 (0.5,0.64) | -0.085 (-0.15,0.01) | -0.113 (-0.2,-0.01) |
| $q_{0.9}$ | 273 | 0.547 (0.49,0.63) | -0.087 (-0.16,0.02) | -0.084 (-0.2,0.09) |
| $q_{0.925}$ | 221 | 0.532 (0.46,0.62) | -0.031 (-0.13,0.1) | 0.232 (0.12,0.4) |

Table 3.4.2: Estimates of constant parameters for both GPD model (1) and PP model (2), when fitting is performed using a range of covariate-dependent threshold obtained via quantile regression. Profile likelihood confidence intervals are shown in brackets.

threshold, should not be considered a contradiction to the previous estimates. In fact, by using a covariate-dependent threshold, we have selected observations which will have a slightly different distribution than before. As Figure 3.4.13 shows, the distribution of threshold excesses exhibits, given a covariate-dependent threshold, a shorter and heavier tail.
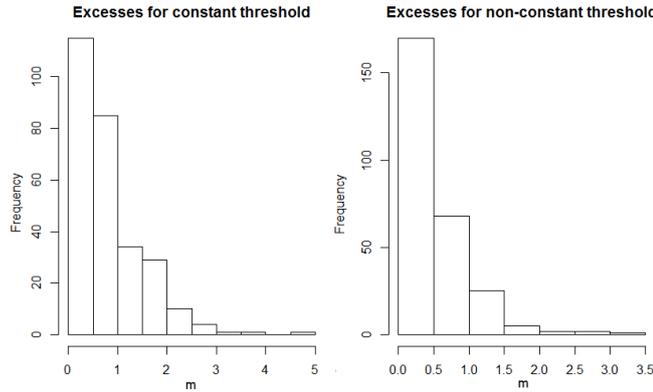


Figure 3.4.13: Overview of threshold excesses for the original South China Sea wave heights, using a constant 90% quantile threshold (left) and non-homogenous 90% quantile threshold regressed over wave direction observations (right).

We can also consider covariate-dependent parameters in Figure 3.4.14. Again, they seem to be consistent across thresholds, although it is interesting to notice the spread of values has reduced in both cases. This shows the choice of a covariate-dependent threshold has succeeded in capturing part of the covariate-dependence.
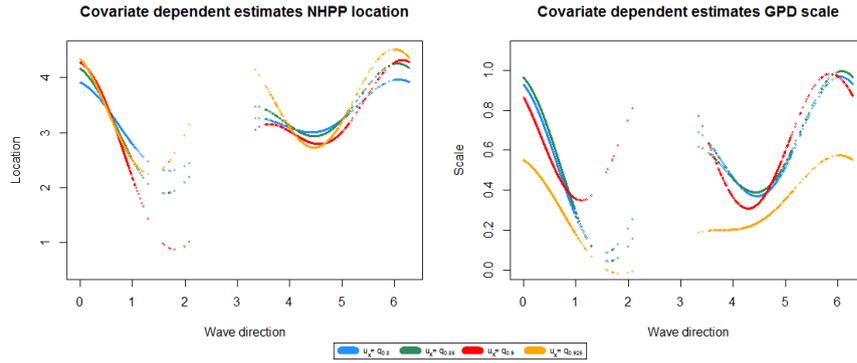
Figure 3.4.14: Covariate-dependent parameter estimates from a repeated simulation-fitting routine, using quantile regression on the original dataset to obtain thresholds, for both GPD model (1) and PP model (2).

As a result, the covariate-response dependence has been reduced for the resulting sample of exceedances. In this case, we could not provide empirical confidence intervals, as a repeated simulation-fitting procedure is complicated by the use of non-stationary thresholds. In fact, especially in order to use 80% quantile regressed threshold values, we would need to first use a constant, lower threshold, which may not lead to informative results given the mean residual life plot in Figure 3.4.1.

## 3.5 Discussion and further work

We have considered two different approaches to modelling threshold exceedances, the generalised Pareto and the non-homogeneous Poisson point process model, and compared them on simulated and hindcast data in Section 3.3 and 3.4 respectively. We found that neither of the models proves to be consistently superior to the other. Each has its own advantages and limitations, with both models presenting a more complicated likelihood space when covariates are introduced, and hence a higher risk of convergence to a local optima for maximum likelihood parameter estima-

tion. While results from both simulations studies show wider confidence intervals for GPD parameter estimates, the extrapolation results from both models seem to be mostly in agreement, in terms of average behaviour as well as uncertainty. Conditional return levels from both studies show the importance of providing a model formulation that can capture the covariate-response relationship appropriately.

The NHPP process is often praised because the parameters are *threshold invariant.* Nevertheless, although the final results from both models are mostly in agreement, the PP approach incurs a series of additional issues. It has proved to be more unstable and sensitive to starting values in the numerical optimisation of the likelihood. This is likely due to the dependence between model parameters, while both the scale and shape are orthogonal to the exceedance rate for the GPD. Furthermore, as discussed in Section 3.1.2, we need to know the distribution of covariates in order to use the NHPP model. We saw that this can be estimated using a closed form or non-parametric methods, with both approaches adding additional uncertainty to the model. Further computational error is also introduced by the use of numerical methods, which need to be employed to estimate the process integrated intensity, for which no close form exists in the non-homogeneous framework. All of these additional issues for the NHPP approach have also an effect on computational cost, both during the model fitting and the inference stage.

The orthogonality of model parameters and relative simplicity in implementation of the GPD approach appear to be strong advantages, despite the dependence of the model on the threshold selected. It is especially worth noticing that, in this case, no assumption is needed, at the fitting stage, about the distribution of covariates considered. On the other hand, estimates for NHPP model parameters

showed to have smaller confidence intervals in both Section 3.3 and 3.4, although the approach seemed to eventually yield potential underestimates of the return levels in Section 3.4.

For both approaches, potential improvement could be achieved using a more flexible formulation to model the link between parameters and covariates. In particular, the use of higher order Fourier series or semi-parametric models, such as splines, may better represent this relationship, although both would require the use of penalised likelihood methods. We consider these and more extensions in Chapters 4 and 5. It would be worth considering the effect of such a parameterisation on model fitting and inference, as a higher number of parameters would need to be estimated. For both models, this would likely complicate the likelihood space, so that there would be a higher probability to converge to local optima. This may be particularly detrimental for the PP model, the use of which may be discouraged by the sensitivity to initial parameter solutions and the additional computational cost.

# Chapter 4

# One-dimensional covariate modelling

## 4.1 Introduction

Metocean and coastal engineers are often interested in the estimation of design conditions for offshore facilities. It is a primary concern to understand the extreme conditions marine structures are likely to experience in their lifetime, in order to prevent structural damage, and hence losses in revenue, environmental pollution and staff endangerment. This generally involves estimating the extreme behaviour of meteorological and oceanographic variables that capture the various environmental loading quantities, that is topographic or weather conditions, such as waves, winds and currents, that contribute to the maximum load that a structure can withstand. Of primary interest for these variables is the estimation of return values corresponding to long return periods. Since data are only available

for several decades, this requires extrapolation. Extreme value analysis provides an ideal framework for this type of inference.

Statistical methods for modelling extremes of stationary univariate sequences have been thoroughly investigated in the literature. As we have seen in Chapter 2, there are two main modelling strategies: the block maxima and the peaks over threshold (POT) approaches (Davison and Smith, 1990; Coles, 2001; Beirlant et al., 2004). In this work, we focus on the latter: observations exceeding a high threshold are modelled by an asymptotically motivated statistical model for the tail of the underlying probability distribution. This model is based on the generalised Pareto distribution, as described in Section 4.2.1, and, in the case of stationary data, it has two main components, with a total of three parameters: the first component is a Poisson model that captures the rate at which the threshold is exceeded, and then a generalised Pareto distribution (GPD) with two further parameters describes the distribution of the sizes of the threshold exceedances.

In many cases, however, an analysis of the extremes of a series is required where there is clear non-stationarity in the series. This is especially common for environmental datasets where the response variable is affected by a variety of climate processes, with different fluctuating behaviours. The focus of this chapter will be on extreme wave heights. Due to the physical process by which they are generated, wave heights generally have a strong seasonal pattern, with an annual periodicity, as well as, decadal or semi-decadal cycles generated by climate phenomena, such as the climate variations captured by the ISO and ENSO indices. For a specific location, we can also detect variability with respect to wave direction; for example, wind is typically stronger from some directions than others, and,

together with fetch and water depth, can strongly influence the resulting magnitude of the waves. Furthermore, a more exposed location will be associated with longer fetches, resulting in a more extreme wave climate. These factors are directly relevant from a design perspective, and it is common for mandated criteria to account for these physical considerations by stipulating directional requirement. Metocean engineers are then often required to specify return values for wave or wind directional sectors, where these directional return value estimates must be consistent with the estimated omnidirectional return values. Numerous authors have demonstrated the importance of incorporating these physical features and the resulting underperformance of models which ignore directional effects. Amongst others, Jonathan and Ewans (2011) examine this issue through the use of a series of case studies, while Coles and Walshaw (1994) and Robinson and Tawn (1997) focus, respectively, on directional modelling of extreme wind speeds and extreme sea currents.

In this chapter, we investigate different modelling approaches for wave heights exhibiting non-stationarity, such as the one shown in Figure 4.1.1, which is obtained from hindcast datasets introduced in Chapter 1.
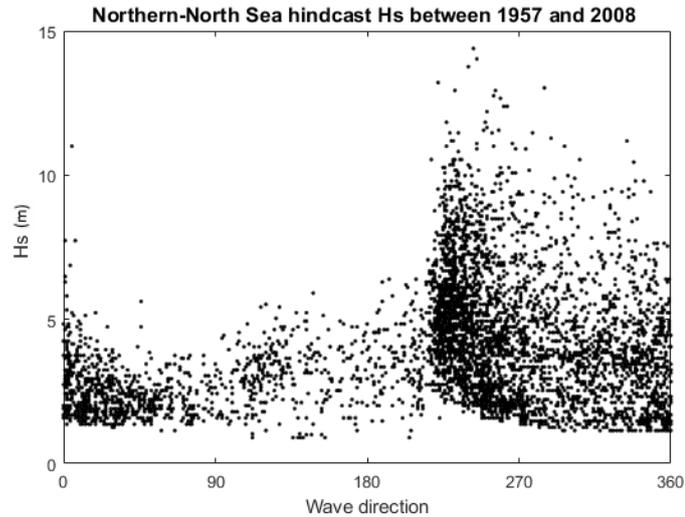
Figure 4.1.1: Storm peak significant wave heigh Hs hindcast (in meters) for a location in the North Sea, with corresponding wave direction (in degrees, with 0 corresponding to waves approaching from the north, and degrees increasing clockwise).

In cases with clear variation in wave height depending on covariates, it is essential to have a statistical model that can appropriately capture the non-stationarity in the series. Models which incorporate an understanding of underlying physical process are more likely to yield reliable return level and uncertainty estimates. This is particularly important in extreme value analysis due to the long-term extrapolation needed and the limited size of datasets. It might also be necessary to estimate return levels for specific covariates, such as directional sector, which can only be reliably estimated from a non-stationary model.

Different regression-based models exist to capture non-stationarity by considering some observed covariates available with the data. In a similar way to generalised linear models, the standard approach is to build a model where one or more of the parameters of the statistical distribution considered are described as a function of the covariates (Smith, 1989). These functions can take different forms; amongst the parametric approaches, Davison and Smith (1990) introduce the simplest for-

mulation, where a linear model describes the generalised Pareto (GP) distribution parameters as functions of covariates, with a log-link function to ensure positivity of the scale parameter. Other parametric models, such as the Fourier parameterisation, have also been used in the presence of one or more periodic covariates, see for example Coles and Walshaw (1994) who model extreme wind speeds.

There are some definite advantages to parametric models: they are easy to fit, given the low number of unknown model parameters, and it is relatively straightforward to use them for return level prediction. Nonetheless, there are also considerable limitations. First of all, the choice of an appropriate functional form to capture the covariate-dependence of a parameter is often not simple and open for debate. Furthermore, these models might not manage to fully capture more "local" behaviour and over-simplify the relationship between the covariate and the response: a parametric formulation is often smoother than what would be considered reasonable from a visual inspection of the data, and its lack of local flexibility might provide poorly-fitting models even when large numbers of parameters are used. One alternative is to use a local likelihood technique, as presented by both Davison and Ramesh (2000) and Hall and Tajvidi (2000), while Chavez-Demoulin (1999) and Pauli and Coles (2001) propose penalised semi-parametric formulations for the POT and the block maxima approach respectively. Another alternative is to use semi-parametric and non-parametric techniques to model the parameters as smooth functions of the covariates, potentially capturing local features better than parametric models. The seminal work by Chavez-Demoulin and Davison (2005) focuses on the use of spline smoothers to parametrise the relationship between the covariate and model parameters, while Yee and Stephenson (2007) propose

vector generalized additive models (VGAM) models. Bottolo et al. (2003) propose a different non-parametric formulation for NHPP model parameters, capturing the effect of a categorical covariate through mixture model representation. The reader is referred to Chavez-Demoulin and Davison (2012) for a general review of non-stationary methods for extremes.

Penalised splines (P-splines) (Chavez-Demoulin and Davison, 2005) are particularly interesting as they can allow a more flexible representation of the distribution parameter with respect to the covariate, when compared to fully parametric models. Jones et al. (2016) compare different parametrizations for the functional form of a model parameter for extreme value analysis of non-stationary series, following the approach we described in Sections 2.1.3 and 2.3. They find the P-spline description to be at least as good as the Fourier series and Gaussian process formulations (Rasmussen and Williams, 2006). Each of these models has their own drawbacks, mainly in relation to computational efficiency and ability to obtain an appropriate level of smoothness. The objective of our work is to develop a flexible and parsimonious model capable of extending efficiently to instances with multiple covariates. In particular, the model should be capable of fully incorporating the covariate information with similar local properties and flexibility to the spline formulations, while also avoiding overfitting. On the other hand, it should have limited computational costs comparable to simpler and parametric approaches. A major consideration in the development of our methodology is ease-of-use and computational efficiency for full-scale oceanographic applications, which may include data from multiple locations on a grid as well as the inclusion of more than just one covariate.

In this chapter, we introduce some novel methodology and new developments of existing methods with respect to covariate parameterisation and inference. In Section 4.2.2 we introduce Bayesian adaptive regression splines (BARS) and propose a new approach for obtaining updated coefficients. Furthermore, while BARS were introduced two decades ago, we are not aware of any other work in extreme value analysis which makes use of this approach. Similarly, in Section 4.2.2, we propose using radial basis functions to represent model parameters. Inference is then performed using reversible jump MCMC (RJ-MCMC): while this approach is based on the work of Green (1995) and Richardson and Green (1997), a series of innovations were required for the model and application considered, as detailed in Section 4.3.3. Finally, to our knowledge, the work by Boldi and Davison (2007), El Adlouni and Ouarda (2009) and Ouarda and El-Adlouni (2011) are the only implementations of reversible jump methodology in the extreme value analysis literature, as further detailed in Section 4.3.3. Our approach differs from all of these papers, as we use RJ-MCMC for the first time to estimate generalised Pareto distribution parameters as functions of a continuous covariate.

The outline of the chapter is as follows. Section 4.2 outlines the different model parameterisations, while in Section 2.2 we present the inference schemes used. Section 4.4 introduces the case studies used, which are selected from the paper by Jones et al. (2016). We briefly describe underlying model forms used to generate samples for inference, outline the procedure for estimation of return value distributions and their comparison, and present results of those comparisons. In Section 4.5, we apply the methodology tested to the hindcast data from the northern North Sea shown in Figure 4.1.1. Finally, Section 4.6 provides a discussion of

results, and conclusions of the study. The appendix describes key elements of the inference schemes in more detail.

## 4.2   Estimating non-stationary extremes

### 4.2.1   Extreme value model

Suppose we are interested in a process $\{Y_t\}$ with corresponding covariate $\{X_t\}$. For example, the response variable may be significant wave height, with wave direction as the covariate. In Section 2.1.3 we reviewed the approach proposed by Smith (1989) and Davison and Smith (1990) to account for non-stationarity in the generalised Pareto distribution. This corresponds to formulating the GPD parameters as functions of the covariate of interest. We also follow the approach by Chavez-Demoulin and Davison (2005), previously introduced in Chapter 4, and re-parametrise the GPD parameters $(\psi, \xi)$ as the asymptotically independent pair $(\nu, \xi)$, where $\nu = \psi(1 + \xi)$. Note that we will be using this re-parameterisation for estimation purposes only. The results and analysis are presented in terms of $\xi$ and $\psi$, following standard practice in the literature.

As for the stationary case, we need to also model the rate of exceedance. It is necessary to carefully reconsider how to capture this, since the process now has a covariate that changes over time. A way to model this dependence on the covariate is by dividing the covariate domain into $k_b$ small subsets, for example through the process known as "binning". Then, we can define $N_i^u$ as the number of threshold exceedances $Y_t | Y_t > u, X_t \in \mathcal{B}_i$, where $\mathcal{B}_i$ is the $i^{th}$ bin, and we can describe it

as $N_i^u \sim \text{Poiss}\left(\phi\left(x_t\right)\right)$ for $i = 1, \ldots, k_b$. Note that it is advisable to consider the model parameters to be modelled as covariate-dependent in order of decreasing ease of computation. That is, the shape parameter is known to be harder to estimate than the scale and rate, as we discussed in Chapter 2, and in the literature this is often modelled as constant. In this chapter, we allow the Poisson parameter and both the GPD parameters to vary with respect to the covariate.

As mentioned in Section 2.1, threshold selection is also essential to reliable estimates, and Northrop and Jonathan (2011) argue that one should start to account for non-stationarity by including covariates in the threshold formulation itself. This concept has been previously explored, at least to some extent, by setting up different thresholds for specific subsets of the data (see Smith, 1989 and Coles, 2001), or modelling the threshold with the use of trigonometric functions (Coles et al., 1994). Northrop and Jonathan (2011) propose a quantile regression model to estimate a covariate-dependent threshold, and show how to parameterise the full GPD model in order for it to be compatible with this formulation. While for the case studies in Section 4.4 the threshold is known and constant, we will be following the covariate-dependent threshold approach for Section 4.5.

### 4.2.2 Covariate parameterisation

Assume that we want to parameterise the GPD model parameters depending on a covariate, as presented in Sections 2.1.3 and 2.3. For generality, we use $\theta\left(x\right)$ to denote any one of the model parameters as a function of the covariate $x$, in order to represent different models in later sections. We can then use a basis

representation to model $\theta\left(x\right)$ as shown in Eq. 2.3.2, such that $\theta\left(x\right) = B_\theta\left(x\right)\boldsymbol{\beta_\theta}$, for some generic row vector of basis functions $B_\theta\left(x\right)$ with $p$ components and a vector of coefficients $\boldsymbol{\beta_\theta} \in \mathbb{R}^p$.

In our case, we parametrise $\phi\left(x\right)$, $\xi\left(x\right)$ and $\nu\left(x\right)$ as linear combinations of unknown parameters $\boldsymbol{\beta_\phi}$, $\boldsymbol{\beta_\xi}$ and $\boldsymbol{\beta_\nu}$ respectively and some row vectors of basis functions $B_\phi\left(x\right)$, $B_\nu\left(x\right)$ and $B_\xi\left(x\right)$ are evaluated at $x$. Note that we can use this notation even if we want to model the shape parameter as constant, in which case the basis matrix is reduced to a scalar basis function which is constant across all of the covariate values, e.g. $B_\xi(x) = \mathbf{1}$. We consider three different forms of basis function in this chapter, corresponding to an evenly-spaced P-spline (Chavez-Demoulin and Davison, 2005; Brezger and Lang, 2006; Eilers and Marx, 2010), an unevenly-spaced P-spline (Denison et al., 1998b; DiMatteo et al., 2001; Eilers and Marx, 2010; Sriram et al., 2015), and a novel radial basis function (RBF) parameterisation. Further details and an overview of each formulation are presented below.

**Splines**

In Section 2.3.4, we introduced regression splines. Under this model, the vector of basis functions for each model parameter is made up of $p$ B-spline functions. For this work, we use cubic splines, and the first-order smoothness penalty expressed in Eq. 2.3.4. We use de Boor's algorithm to construct splines with both evenly and unevenly spaced knots, and we consider both the case where the number of spline components $p$ is fixed and where it changes, in which case we follow a method called Bayesian Adaptive Regression Splines (BARS), introduced by DiMatteo

et al. (2001).

Recall that the design matrix $B$ is composed by basis spline functions $b_{\theta,j}$, $j = 1, 2, ..., p$, such that $B = (\boldsymbol{b}'(x_1) \quad \boldsymbol{b}'(x_2) \quad ... \quad \boldsymbol{b}'(x_{n_I}))$ for some indexed set of $n_I$ covariate values $x_1, x_2, \ldots x_{n_I} \in \mathbb{R}$. A fundamental property of spline functions is that they have compact support, that is, they are zero outside of a small specific interval. This provides computationally efficient and stable inference, since for a given design matrix $B$, the corresponding precision matrix $B'B$ is band diagonal, which reduces the computational costs of inverting this matrix during a MALA or mMALA MH step, as illustrated in Section 2.2.3. The computational efficiency and stability of spline inference is in marked contrast with fully parametric models that do not have compact support.

**Bayesian adaptive radial basis functions**

We now consider an adaptation of the radial basis function formulation introduced in Section 2.3.5, where we propose using fewer components and allow the location and scale of each RBF to change as part of the inference procedure. In fact, a linear combination of radial basis functions provides an alternative to the spline description for the model parameters, with the potential of requiring fewer components to achieve a similar inference performance. In particular, we can model the bases for $\phi(x)$, $\xi(x)$ and $\nu(x)$ as vectors of $p$ kernels. Then we can represent the parameter $\theta(x)$ as the linear combination of these kernel, such that

$$B_j(x) = f\left(\frac{x - \tilde{m}_j}{\tilde{s}_j}\right), \ j = 1, \ldots, p,$$

where, in our case, $f$ denotes the density function for the Normal distribution with location $\tilde{m}_j \in \mathbb{R}$ and standard deviation $\tilde{s}_j > 0$ for all $j$. Under this model, we need to estimate the coefficient $\beta_j$, location $\tilde{m}_j$ and standard deviation $\tilde{s}_j$ for each of the $p$ density components. Note that we also introduce a baseline factor, corresponding to a free, standing coefficient $\beta_0$, which aims to capture the average behaviour of interest, with the different RBF components detecting the local variation.

The number of these components may differ considerably between the P-spline and the RBF model. For the former, the standard approach consists of using a high number of components, and controlling the parameter smoothness with the use of a penalty, as presented in Section 2.3.4. For the proposed RBF model, it is the number of components that helps control the smoothness of the fit, so we expect a much lower value for $p$. This is indeed one of the potential advantages of this model, as it has the potential to improve computational efficiency, especially when considering possible extensions to multidimensional covariates, as discussed in Section 4.6.

In the same manner as for splines, we use this formulation to represent the covariate dependence in the Poisson-GPD model, and we use Bayesian inference to perform model fitting. We call this model Bayesian adaptive radial basis functions (BARBaR), and we assume that the number of RBF components, as well as the RBF model parameters, such as their location and variance, are unknown. In particular, we do not fix the number of basis components in advance, and instead we estimate this as part of the inference procedure. By allowing movement between models with different numbers of basis functions, we hope to better control the

smoothness of the overall fit and produce a more parsimonious model. Inference is carried out using the reversible jump (RJ) algorithm, as detailed in Section 4.4. We also consider the simpler case where the number of RBF components is fixed, which we denote as BARBaR-f.

## 4.3 Inference procedures

As mentioned in Section 2.2, Bayesian methods are used in the inference process for the estimation of the unknown parameters. For the proposed extreme value model, it is not possible to express the posteriors of interest in a simple closed form, hence MCMC methods are used to sample the desired posterior distribution. In Section 4.3.1 we illustrate the structure of our Bayesian model and we discuss the prior and posterior structures for different parameterisation approaches. In Section 4.3.2 we provide an overview of the MCMC techniques used for inference in the case of problems with a fixed number of components. Then, in Section 4.3.3, we focus on the case where the number of components is also an unknown, and in particular we introduce a sampling approach based on reversible jump (RJ) MCMC.

### 4.3.1 Bayesian model

Following the approaches and notation from Section 4.2.2, model fitting corresponds to estimating, for each of the distribution parameters, the unknown parameters $\Omega_\theta = \{\boldsymbol{\beta_\theta}, \rho_\theta\}$ and $\Omega_\theta = \{\boldsymbol{\beta_\theta}, \boldsymbol{m_\theta}, \boldsymbol{s_\theta}, \gamma_\theta\}$ for the P-spline and the BARBaR

models respectively. Here $\gamma_\theta$ is an additional hyperparameter for the variances of

the kernels $\boldsymbol{s}_\theta$, following a recommendation by Richardson and Green (1997) to

obtain a higher degree of non-informativeness.

In order to fit the models described in Section 4.2.2, we need the joint posterior

distribution for the Poisson-GPD model parameters. We can express this in terms

of the relevant conditional posterior distributions. For all the above models, we

let $N_i^u$ be the number of threshold exceedances $Y_t | Y_t > u, X_t \in \mathcal{B}_i$, where $\mathcal{B}_i$ is the

$i^{th}$ bin and $N_i^u \sim \text{Poiss}\left(\phi\left(x_t\right)\right)$ for $i = 1, \ldots, k_b$. The threshold exceedances are

$$Y_i | Y_i > u, X = x \sim \text{GP}\left(\psi\left(x\right), \xi\left(x\right)\right),$$

where the GPD parameters $\psi\left(x\right), \xi\left(x\right)$ vary continuously as functions of the co-

variate $x$. We use the notation $\theta\left(x\right)$ to refer, more generally, to the Poisson rate,

GPD scale and shape, as each will have the same basis model structure from Eq.

2.3.2. Then for the P-spline, with model parameter defined as in Section 4.2.2, we

have

$$\boldsymbol{\beta_\theta} | \rho_\theta \approx \rho_\theta^{1/2} \exp\left(-\frac{\rho_\theta}{2}\boldsymbol{\beta_\theta}^{\text{T}} Q_\theta \boldsymbol{\beta_\theta}\right), \qquad (4.3.1)$$

$$\rho_\theta \sim \text{Gamma}(10^{-3}, 10^{-3}),$$

This is the Bayesian approach to penalised splines, where the coefficients have a

prior density proportional to the improper Gaussian density in Eq. 4.3.1, where

the density is improper as $Q_\theta$ is rank deficient. Here, the term $-\frac{\rho_\theta}{2}\boldsymbol{\beta_\theta}^{\text{T}} Q_\theta \boldsymbol{\beta_\theta}$

can be seen as the roughness penalty in Eq. 2.3.4 with a precision parameter $\rho_\theta$

controlling the smoothness and a precision matrix $Q_\theta$ defined as in Eq. 2.3.5. The

precision parameter $\rho_\theta$ is given an uninformative Gamma prior distribution, which

is conjugate with the prior Gaussian distribution for $\beta_\theta$.

For the BARBaR formulation defined in Section 4.2.2, for each component $j = 1, \ldots, k$, we suggest the following prior distributions:

$$\beta_{\theta,j} \sim \mathrm{N}(0, R)$$

$$m_{\theta,j} \sim \mathrm{Uniform}(d_l, d_r),$$

$$s^2_{\theta,j} | \gamma_\theta \sim \mathrm{Gamma}(a, \gamma_\theta)$$

$$\gamma_\theta \sim \mathrm{Gamma}(g, h). \tag{4.3.2}$$

Here, $[d_l, d_r]$ denotes the range of the domain of the covariate, $R$ denotes the range of the observations $Y_t$, the coefficients $\beta_{\theta,j}$ are mutually independent of each other and of the other parameters, and $\gamma_\theta$ is common across the different kernel components. It is worth noticing from Eq. 4.3.2 that, while it seems natural that the range of the data would be informative about the location of the kernels, this is not the case for their widths, so we introduce an additional hierarchical level, which controls the degree of similarity between the kernel widths, without being informative about their absolute size. This yields better mixing of the chain (Richardson and Green, 1997). Note that, in the case of the kernel widths estimation, we can exploit the gamma-gamma conjugacy.

**Full Bayesian inference approach**

We can then illustrate the full inference framework as follows. Let us consider some sequence of i.i.d. random variables $Y_t$, $Y_t \sim f(\theta_1(x_t), \ldots, \theta_{n_p}(x_t))$, which are dependent on some covariate $X_t$ and are modelled by the density $f(\cdot)$ described

by $n_p$ density model parameters (e.g. Poisson rate, and GPD scale and shape). The full Bayesian inference approach is presented in Algorithm 2 for all the models considered, and Sections 4.3.2 and 4.3.3 provide more details on the specific inference methods used.

---

**Algorithm 2** Full Bayesian inference approach

---

**Initialize:**

    The vector of unknowns $\boldsymbol{\Theta}$ as some suitable value $\boldsymbol{\Theta}^0$

**for** iter$= 1, \ldots,$ max_iterations **do**

    **for** $l = 1, \ldots, n_p$ **do**              ▷ *Loop over model parameters of $\theta_l(x_t)$*

---

        Standard MCMC steps for all models          ▷ *See Section 4.3.2*

---

▷ *Update the basis parameters*

        **if** Spline basis model **then**

            Generate a joint mMALA proposal for all the coefficients $\boldsymbol{\beta_{\theta_l}}$;

        **else if** BARBaR model **then**

            Generate a joint mMALA proposal for all the basis parameters, e.g. the

            coefficients $\boldsymbol{\beta_{\theta_l}}$, locations $\boldsymbol{m_{\theta_l}}$ and variances $\boldsymbol{s_{\theta_l}}$ (see Appendix 4.A);

        **end if**

        Calculate the acceptance probability

        Accept or reject the proposal;

▷ *Update the basis hyper-parameters*

        **if** Spline basis models **then**

            Gibbs sample the hyperparameter $\rho_\theta$ for the coefficients;

        **else if** BARBaR models with Gaussian kernels  **then**

            Gibbs sample the hyperparameter $\gamma_\theta$ for the kernel variances;

        **end if**

---

        Reversible jump moves for dimension-changing models ▷ *See Section 4.3.3*

---

        Given the current number of basis components $k$

        **if** BARS model **then**

            Randomly select a between-model move (1)-(4) or a within-model step;

            Propose a new basis

        **else if** BARBaR model **then**

            Randomly select a between-model move (1)-(4);

        **end if**

        Propose a new basis

        Calculate the acceptance probability

        Accept or reject the proposal;

    **end for**

**end for**

---

## 4.3.2 Standard MCMC inference algorithms

Estimates for all the unknown parameters are obtained using a Metropolis-within-Gibbs MCMC algorithm (e.g. Gamerman and Lopes, 2006). At each iteration of the MCMC, each of the parameters is sampled in turn conditionally on the values of others. The full conditional posterior distribution of the precision parameter $\rho_\theta$ in the spline model, and of the shape $\gamma_\theta$ for the kernel widths in the BARBaR models, is Gamma by conjugacy, and is sampled exactly in a Gibbs step. A general Metropolis-Hastings (MH) scheme is, instead, used to estimate the coefficients $\boldsymbol{\beta_\theta}$ for all models, since full conditional distributions are not available in closed form for the posterior of these parameters. The sampling is performed according to the approach presented in Section 2.2.2 and following Algorithm 1. Note that we focus on the other basis parameters of the BARBaR models later in Section 4.3.2.

A simple random walk MH update yields limited results in this case for all models: we found it to be inefficient, as mixing was much slower than the more advanced algorithms presented in Section 2.2.3. This is not surprising for a problematic function such as the GPD likelihood, for which the negative log likelihood is likely to be pitted with local minima. As reviewed in Section 2.2.3, gradient based methods such as the Riemann manifold Metropolis-adjusted Langevin algorithm (mMALA), have been demonstrated to improve mixing and yield faster convergence. These exploit gradient and curvature information of the log-posterior at the current state to improve the quality of proposals.

Hence, we decide to use the simplified mMALA algorithm of Girolami and Calderhead (2011) and Xifara et al. (2014) to estimate the coefficients. This accounts

for the local curvature of the likelihood surface with proposals that are partly deterministic and partly stochastic. For sampling, we follow the approach in Section 2.2.3. Furthermore, for stability and computational reasons, we use the expected rather than the observed information from the log-posterior (see Appendix 4.C for further details). It is worth recalling that, while the mMALA scheme is expected to explore the posterior with considerably higher efficiency than a RWM algorithm, it requires knowledge of likelihood derivatives, leading to an additional computational cost. We refer the reader to Appendix 4.C for further details on the necessary derivatives for the models presented here.

**Further adaptations required for the BARBaR model**

In the most general BARBaR model, we do not just need to estimate the coefficient for each component density. Instead, we also need to estimate some additional model parameters (e.g. location and variance in the case of Gaussian kernels). While we initially used RWM, and tried updating jointly (i) all kernel parameters for each single component (i.e. jointly $\beta_{\theta,j}, m_{\theta,j}, s_\theta$ for $j = 1, \ldots, k$), (ii) the same kernel parameter type across different components (e.g. first $\boldsymbol{\beta}_\theta$, then $\boldsymbol{m}_\theta$, then $\boldsymbol{s}_\theta$) and (iii) each element of $\boldsymbol{\beta}_\theta$, $\boldsymbol{m}_\theta$ and $\boldsymbol{s}_\theta$ individually. Overall, while it might be possible to use a random walk Metropolis algorithm, we found all the above options showed slow convergence and required careful fine-tuning of the stepsize for the different model parameters, with fitting performance also very sensitive to the initial values. For a fixed number of kernel components, adaptive MCMC methods (Roberts et al., 2001) can be used with some success, as evident from some limited testing we performed. Nevertheless, once reversible jump methods

are involved, we found that mixing noticeably deteriorated.

Hence, as discussed earlier for the spline models, we investigated the option of exploiting gradient information to improve the convergence and mixing of the chains. Recall from Section 2.2.3 that the mMALA algorithm can use the second order derivative of the posterior to automatically adapt the stepsize for the update. This property holds even in the dimension-changing case, as this stepsize adaptation is valid regardless of the number of component densities in the formulation. This should remove the need for further pre-determined adaptation when jumping to a domain of different size, and yield better convergence. Hence, we decide to implement an mMALA approach to estimate the density components in both the BARBaR and BARBaR-f models. This sampling method improves the observed slow convergence of RWM for the fixed-dimensional BARBaR-f, and considerably aids parameter estimation for the more general BARBaR model. This approach is possible because of the closed form formulation of each kernel component, and we refer the reader to Appendix 4.A for more details and a specific example using Gaussian kernels.

### 4.3.3   Moving between dimensions

In some situations, it might be the case that the number of unknown parameters in a statistical problem is, itself, unknown. "Trans-dimensional" problems can be formulated as the joint inference of a parameter vector $\theta_k$ and the dimension $k$ of this parameter, which is allowed to change and produce a corresponding set of models $M_k$. Reversible jump Markov chain Monte Carlo (RJ MCMC, Green 1995)

is a method for simulating from a Markov chain to obtain a posterior distribution in cases like these, where the state and its space do not have a fixed dimension. In Section 2.2.4 we introduced the main notation and theory for this method, while here we review specifically the adaptations used to perform fitting in Sections 4.4 and 4.5.

**Reversible jump MCMC inference**

As mentioned in Section 2.2.4, the reversible jump algorithm (Green, 1995) is a further extension of the Metropolis-algorithm, designed to allow movement between different dimensional spaces. Of particular interest, among many others, are the papers by Richardson and Green (1997), Dellaportas and Forster (1999) and Denison et al. (1998b), who applied the reversible jump sampler to mixture models, variable selection, and curve fitting respectively. We also explore the extensions of the RJ MCMC algorithm to spline models specifically. Seminal literature for these methods, known as Bayesian adaptive regression splines (BARS), can be found in Denison et al. (1998a), DiMatteo et al. (2001), and Wallstrom et al. (2008), while Brooks (1998) review some essential issues such as output analysis and convergence assessment.

It is worth considering some previous application of RJ MCMC methods to extreme value data. We mentioned before the representation by Bottolo et al. (2003) of NHPP model parameters as a mixture depending on some categorical covariate. In this paper, the authors use RJ MCMC to fit the model, in order to find if any of the categories have similar parameter estimates. Boldi and Davison

(2007) propose a mixture of Dirichlet distributions to represent multivariate extremes spectral functions, and use a reversible jump scheme for inference. More recently, El Adlouni and Ouarda (2009) and Ouarda and El-Adlouni (2011) apply a RJ MCMC approach for covariate selection in a GEV regression model, fitted to hydrological data.

Our works focuses on capturing covariate-dependent behaviour by representing the parameters in an extreme value model as a linear combination of kernel functions. We mentioned in Section 4.1 some commonalities with the work by Bottolo et al. (2003), although in our work we focus on a Poisson-GPD model rather than a NHPP one. While they also proceed to use RJ methodology, there are some key differences both in the implementation and in the aim of their approach. In our case, we deviate from the set-up in Richardson and Green (1997), where the mixture model definition imposes that the weights of the components must sum to one, and these are assigned a Dirichlet prior. On the other hand, as mentioned in Section 4.2.2, we use a combination of kernel components in a semi-parametric regression model, and therefore the coefficients are not constrained in size, sign or sum. The approach in Richardson and Green (1997) also has an additional latent allocation variable, which is introduced to facilitate the implementation of the RJ steps by exploiting conjugacy in a Gibbs update. These key differences require some adaptation of the approach proposed by Richardson and Green (1997). Furthermore, Bottolo et al. (2003) apply the model to a categorical covariate, namely insurance claim types, in order to detect any commonality in the NHPP location, scale or shape parameters across the different categories. In our work we are interested in continuous covariates, and propose the use of RJ MCMC with the

aim of obtaining a parsimonious representation of the covariate-dependent model parameters.

While Section 2.2.4 showed how the RJ MH algorithm works in general, we will now focus on its implementation for the models introduced in Section 4.2.2. Recall that, for each iteration $i$, we have some current state, which includes information on the state space dimension $k$, such that $\boldsymbol{\theta}^{(i-1)} = (k, \boldsymbol{\theta}_k^{(i-1)})$ in model $M_{k^{(i-1)}}$. Markov chain transitions are then performed by first proposing $\boldsymbol{\theta}^* = (k^*, \boldsymbol{\theta}_{k^*}^*)$ in model $M_{k^*}$ from a proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(i-1)})$. Eq. 2.2.2 shows how to calculate the corresponding acceptance probability, which ensures the standard MCMC detailed balance conditions are met. Within this framework, in addition to the structure shown in Section 4.3.2, we have an additional prior for models with varying number of basis components for both formulations. In this case, the number of components is represented by an additional unknown $k_\theta$, which is assumed to follow a Poisson distribution with some chosen rate $\eta_\theta > 0$, such that

$$k_\theta = k|\eta_\theta \sim \text{Poisson}(\eta_\theta). \qquad (4.3.3)$$

In this work, we consider two different pairs of jumps, also known as "between-models" moves. Each pair consists of one move which increases the dimension of the state space and one which reduces it. Furthermore, each move in a pair is the exact reverse of the other one. In our work, we consider two different pairs, with a visual example shown in Figure 4.3.1 namely:

- **birth-death**, which introduce and remove, respectively, a new basis component at random;

- **split-combine**, which respectively divide the effect of a basis component into two different ones or combine two components into one capturing the behaviour of both.
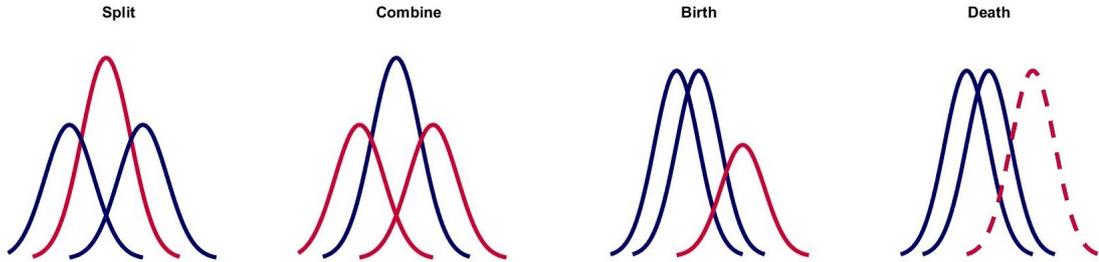


Figure 4.3.1: Intuitive representation of RJ moves, where the components shown in red are the ones modified (split or combined), birthed or killed.

In practice, one can either consider all moves systematically (see Richardson and Green 1997), or chose independent move types randomly with probabilities $r_m(\theta)$. In our case, we proceed with the latter option, using transition probabilities

1. $r_m^b(\theta)$ for the addition of a component (a **birth step**),

2. $r_m^d(\theta)$ for the deletion of a component (a **death step**),

3. $r_m^s(\theta)$ for the separation of a component into two (a **split step**), and

4. $r_m^c(\theta)$ for the merger of a two components into one (a **combine step**),

where $r_m^b(\theta) + r_m^d(\theta) + r_m^s(\theta) + r_m^c(\theta) = 1$. Sections 4.3.3 and 4.3.3 provide details on the implementation of these moves for the BARS and BARBaR model respectively, as well as the computation of the probability of each of the moves considered in order to estimate the acceptance probability as shown in Eq. 2.2.6.

**Spline basis**

For the spline basis model, we follow the methodology illustrated by Denison et al. (1998a), DiMatteo et al. (2001) and Wallstrom et al. (2008), and implement one of the jumps at each of the MCMC sweeps. In addition to the between-models moves above, we construct the algorithm to incorporate "within-model" moves. That is, for some iterations, a standard MH update for the spline knots locations is chosen rather than a reversible-jump on the basis components. This additional option allows us to consider the relocation of one of the spline basis knots, which are fixed in standard P-spline models.

For the purpose of illustration, let us denote the current state $\theta^{(i-1)}$ as $\theta$, with corresponding $k$ basis component functions $\boldsymbol{b}'_i(x), i = 1, \ldots, k$. Then, let $M_K$ denote the model parameterised by $(\boldsymbol{b}'_1(x) \quad \boldsymbol{b}'_2(x) \quad \ldots \quad \boldsymbol{b}'_k(x), k)$ and let $r^*_m(k)$ be the probability of move type $*$ when the model has $k$ components. Following the suggestion from DiMatteo et al. (2001), we set a maximum and a minimum number of basis components, such that $k \in [k_{\min}, k_{\max}]$. Then, for moves to a higher dimension we have

$$
\begin{aligned}
r^{\text{birth}}_m(k) &= r^{\text{B}}_m(k) = \frac{1}{2}\omega \min\left\{1, \frac{p(k+1)}{p(k)}\right\}, \\
r^{\text{split}}_m(k) &= r^{\text{S}}_m(k) = \frac{1}{2}\omega \min\left\{1, \frac{p(k+1)}{p(k)}\right\},
\end{aligned}
\tag{4.3.4}
$$

and for moves to a lower dimensional state, we have

$$r_m^{\text{death}}(k) \;=\; r_m^{\text{D}}(k) = \frac{1}{2}\omega \min\left\{1, \frac{p(k-1)}{p(k)}\right\}, \qquad (4.3.5)$$

$$r_m^{\text{combine}}(k) \;=\; r_m^{\text{C}}(k) = \frac{1}{2}\omega \min\left\{1, \frac{p(k-1)}{p(k)}\right\},$$

where the probability $p(k)$ for any possible number of components $k$ is given by a prior distribution, such as that shown in Eq. 4.3.3. Since the probabilities of all the moves need to sum to 1, then the probability of a (within model) relocation move is just $r_m^{\text{relocation}}(k) = r_m^{\text{R}}(k) = 1 - \{r_m^{\text{B}}(k) + r_m^{\text{D}}(k) + r_m^{\text{S}}(k) + r_m^{\text{C}}(k)\}$. The constant $\omega \in [0, \frac{1}{2}]$ controls the rate at which dimension changing moves are proposed, where the upper end of the domain ensures that the sum of the probabilities of the moves never exceeds 1. Note that the choice for this constant $\omega$, as well as for the candidate densities (and parameters) for the jump moves are arbitrary. In this case, after limited testing of different options, we choose $\omega = 0.4$ (Denison et al., 1998a; DiMatteo et al., 2001; Wallstrom et al., 2008). In comparison with the proposal by DiMatteo et al. (2001), a factor of $\frac{1}{2}$ is added in Eq. 4.3.4 and 4.3.5 to account for the fact that two different pairs of moves are proposed, which is equivalent to imposing $\omega \in [0, \frac{1}{4}]$.

In terms of candidate densities for proposed knots, Beta densities are often a convenient choice for the proposal moves, as they can encompass different behaviours and have a limited domain by definition, as compared to densities such as the normal, which might need to be truncated to be applied to a finite covariate domain. Nevertheless, since the support of Beta distributions is the interval $[0, 1]$, we normalise the covariates to this range during the fitting stage, in agreement

with the set-up by Wallstrom et al. (2008), avoiding the need for rescaling of the domain at each proposal. Below, we illustrate which specific candidate density is used for each of the jumps considered.

**Birth step** Let the current model $M_k$ be described by $k$ spline knots $\tau_1, \ldots, \tau_k$. We can then birth a new knot at random anywhere on the domain of the covariate. That is, we draw a new knot $\tau_{k+1}^*$ uniformly on the normalised covariate domain. For consistency with the split step, we use a $\text{Beta}(1, 1)$ distribution, such that $\tau_{k+1}^* \sim h_B(\tau_{k+1}^*)$, where $h_B$ is the density of a $\text{Beta}(1, 1)$ distribution. The resulting probability for the jump is given by

$$q(M_{k+1}|M_k) = r_m^{\mathrm{B}}(k)\frac{1}{k}h_B(\tau_{k+1}^*).$$

**Death step** One of the existing knots is chosen at random and eliminated, with resulting jump probability

$$q(M_{k+1}|M_k) = r_m^{\mathrm{D}}(k)\frac{1}{k}.$$

**Split step** The split step proceeds in a similar way to the birth one. The main difference is that in this case, a knot can only be birthed in close proximity to an existing one. That is, we select a knot $\tilde{\tau}$ from the $k$ existing ones, and we generate a new knot $\tau_{k+1}^*$ from a narrow symmetric distribution centred around $\tilde{\tau}$. In particular, we let $\tau_{k+1}^* \sim h_S(\tau_{k+1}^*|\tilde{\tau}, d)$, where $h_S$ is the density of a $\text{Beta}(d\tilde{\tau}, d(1 - \tilde{\tau}))$, with spread parameter $d$, such that $\mathrm{E}\left(\tau_{k+1}^*\right) = \frac{d\tilde{\tau}}{d\tilde{\tau}+d(1-\tilde{\tau})} = \tilde{\tau}$. After some

testing, we opted for $d = 100$. The resulting probability for the jump is given by

$$q(M_{k+1}|M_k) = r_m^{\text{S}}(k)\frac{1}{k}\sum_{j:\tau_j \neq \tilde{\tau}} h_S(\tau_{k+1}^*|\tau_j, d),$$

where the summation takes into account than any one of the $k$ existing knots could have been chosen. Hence, one needs to account for the probability of the candidate knot $\tau_{k+1}^*$ being created by each of the $k$ distributions $h_S(\cdot)$.

**Combine step** The combine step proceeds in a similar way to the death one, but it only allows the elimination of a knot that could have been the product of a split step. First, we select a knot $\tilde{\tau}$ from the $k$ existing ones. Then, we consider the probability that it comes from a narrow density centred around one of the remaining knots, such that as $\tilde{\tau}^* \sim h_C(\tilde{\tau}|\tau_j, d)$ for all values of $j$ such that $\tau_j \neq \tilde{\tau}$, where $h_C$ is the density of a $\text{Beta}(d\tau_j, d(1 - \tau_j))$, where the spread parameter $d$ has the same value as in the split step. The resulting probability for the jump is given by

$$q(M_{k+1}|M_k) = r_m^{\text{C}}(k)\frac{1}{k}\sum_{j:\tau_j \neq \tilde{\tau}} h_S(\tau_{k+1}^*|\tau_j, d),$$

where again the summation accounts for the probability of the candidate knot $\tilde{\tau}*$ merging with any of the remaining knots.

**Relocation step** This is a standard MCMC within-model update, although it is not implemented at each step but, instead, selected with probability $r_m^{\text{R}}(k)$ which depends on the current number of spline components as defined earlier in this section. One of the existing knots, $\tilde{\tau}$, is selected at random, and a new location

$\tilde{\tau}^*$ is proposed from a symmetric distribution centred at the current value. We propose using again a Beta distribution, such that $\tilde{\tau}^* \sim h_R(\tilde{\tau}^* | \tilde{\tau}, d(1 - \tilde{\tau}))$, where $h_R$ is the density of a Beta$(d\tilde{\tau}, d(1 - \tilde{\tau}))$, and again choose $d{=}100$. Then we need to compute the probability of the move

$$q(M_k^{\tilde{\tau}^*} | M_k^{\boldsymbol{\tau}}) = r_m^{\mathrm{R}}(k) \frac{1}{k} h_R(\tilde{\tau}^* | \tilde{\tau}, d),$$

which is required for Eq. 2.2.2 to accept or reject the MH step.

**Coefficients update**   It is worth noticing that small changes in the spline basis can have a considerable effect on the curve it generates. In fact, while a knot can be inserted into a B-spline basis without changing the geometry of the curve (de Boor, 2001; Ruppert et al., 2003), the reverse is not true as the resulting system of equations is overdetermined. These changes can affect the acceptance of jump moves, if they cause the parameter function, and hence the proposed posterior, to be too different from that of the current state. In order to increase the acceptance rate and the convergence rate of the algorithm, one can adapt the coefficients so that the resulting curve is more likely to be accepted. Denison et al. (1998a) propose using an adjusted least squares approach to improve the resulting fit, although this might affect the reversibility of the move. DiMatteo et al. (2001) and Wallstrom et al. (2008) follow a different approach and separate the inference into two stages: first they focus on obtaining a suitable basis using reversible jumps, and later they introduce an additional Metropolis update to optimise the coefficients and fully capture the corresponding uncertainty. Biller (2000) and Sharef et al. (2010) opt for a different strategy, by exploiting the insight

from de Boor (2001) to propose curves which are similar, albeit not identical, to the original one. We decide to follow the same intuition, but propose steps that slightly differ from both papers above for the following reasons:

- the algorithm proposed by Biller (2000) always edits the same number of coefficient (three) independently of the original degree of the spline basis. Although this can produce suitable solutions for quadratic and possibly cubic splines, which are most common in the literature, ideally one would want the number of coefficients to be updated to depend on the degree of the basis. In fact, as we have seen in Section 2.3.4, the degree of the spline basis affects how many components overlap and, consequently, how many neighbouring ones are affected by any local change;

- in their work, Sharef et al. (2010) focus on Poisson-distributed observations and introduce an exponential link inside the spline basis, and hence of any basis coefficient updates, to ensure the positivity of the resulting parameter estimates. We are, instead, interested in a general approach which is able to handle both positive and negative coefficients.

It is also worth noticing that if we were to use purely deterministic rules based on de Boor (2001), the required symmetry between the birth and the death move would be destroyed. Instead, we use these as a basis and introduce a uniform random variable $u \in (0, 1)$ to preserve the reversibility of the move as well as the dimension-matching requirement between different states. One then needs to be careful in updating the coefficients correctly. To do so, we need to first consider how a B-spline basis is defined, and need to make a distinction between interior knots

and end points. In fact, while we are normally interested in the interior knots, which cover the support domain, some additional points are required to construct the full basis. Let us consider a B-spline of degree $d_S$ and order $Q_S = d_S + 1$, and recall that it extends over $Q_S + 1$ knots by definition. It is then necessary to extend the internal knots with $Q_S - 1 = d_S$ end points on each side, to give full support to the initial and final B-splines in the internal knot intervals. Note that the values of these endpoints does not matter, and can be assigned the same values as the initial and final end point, or evenly spaced beyond the two limits of the support domain.

The coefficient adjustments are then performed as follows.

**Knot addition** : Given the $k$ knots $\boldsymbol{\tau} = (\tau_l, ..., \tau_k)$, in the birth and split moves we add a new knot $\tilde{\tau}^*$ lying within some interval $(\tau_j, \tau_{j+l})$, with $j \in 1, \ldots, K$. The resulting model is now defined by the new model indicator $k + 1$, the new knots $\tilde{\boldsymbol{\tau}}^* = (\tilde{\tau}_l^*, \ldots, \tilde{\tau}_{K+1}^*)$, where $\tilde{\tau} = \tau_i$ for $i \leqslant j$, $\tilde{\tau}_{j+1} = \tilde{\tau}^*$ and $\tilde{\tau}_i = \tau_{i-1}$ for $i \geqslant j + 2$. The candidate spline coefficients vector $\tilde{\boldsymbol{\beta}}$ can be computed as follows:

$$
\tilde{\beta}_i = \begin{cases}
\beta_i & \text{if } i \leqslant j + 1 \\
\beta_{i-1} & \text{if } i \geqslant j + Q_S + 1 \\
r_i \beta_i + (1 - r_i)\beta_{i-1} & \text{if } j + 1 < i < j + Q_S \\
u\beta_i + (1 - u)\beta_{i-1} & \text{if } i = j + Q_S
\end{cases}
\tag{4.3.6}
$$

where

$$r_i = \frac{\tilde{\tau}^* - \tau_i}{\tau_{i+Q_S-1} - \tau_i},$$

and $u \sim \text{Unif}(0,1)$. The latter is used, in place of the knot ratio $r_{j+Q_S}$, in updating the corresponding coefficient, to introduce randomness and meet the dimension-matching requirements. The Jacobian of the corresponding transformation reduces to

$$J = \left| (\beta_{j+Q_S} - \beta_{j+Q_S-1}) \prod_{i=j+1}^{j+Q_S-1} r_i \right|. \qquad (4.3.7)$$

**Knot deletion** In the death and combine moves, a knot $\tau_j$ is selected to be removed from the vector of knots $\boldsymbol{\tau}$. The spline coefficients are the adjusted by the inverse of the transformation in Eq. 4.3.6, by first deleting $\beta_{j+Q_S-1}$ and adjusting the remaining coefficients to be

$$\tilde{\beta}_i = \begin{cases} \beta_i & \text{if } i \leqslant j+1 \\ \beta_{i+1} & \text{if } i \geqslant j+Q_S+1 \\ \frac{1}{r_i}\beta_i + \frac{1-r_i}{r_j}\tilde{\beta}_{i-1} & \text{if } j+1 < i < j+Q_S \end{cases} \qquad (4.3.8)$$

Given the symmetricity of reversible jump moves, the Jacobian determinant for knot deletion is the inverse of Eq. 4.3.7.

From exploratory testing, we noticed these coefficient adjustments increased the speed of convergence of the algorithm. It is also worth remembering that we are mainly interested in the case where the covariate is periodic. In order to apply the above approach to a periodic domain, one needs to implement some specific

adaptations. In the BARS model, as for the spline model with fixed-dimension, the spline basis itself needs to be periodic. This is fairly common in the literature, and we refer the reader to Eilers and Marx (1996) and Ruppert et al. (2003) for further details on the implementation.

Candidate densities for the split, combine and relocation moves also need to be periodic, although we prefer having a common approach independent of the nature of the covariate. Hence, we decide to use a Beta distribution as a candidate for the periodic covariate case as well. In this case, we introduce a periodic version of the Beta distribution then compute the density for some transformed values of the knot locations. In particular, we propose a transformation such that the knot of interest $\tilde{\tau}$ is shifted to 0.5, the middle of the support of a Beta density. Then, for the chosen spread parameter $d$, we simulate a new location on the periodic domain as follows:

$$t \sim \text{Beta}\left(\frac{1}{2}d, \frac{1}{2}d\right),$$
$$t^* = \tilde{\tau} + \text{range}(\mathcal{D})\left(t - \frac{1}{2}\right),$$
$$\tilde{\tau}^* = d_l + \quad \text{mod}\left(t^* - d_l, \text{range}(\mathcal{D})\right).$$

Finally, the coefficient update algorithm would work as shown in Eq. 4.3.6 and 4.3.8, with some care taken into considering the correct indices for both the new and old vectors of coefficients.

**BARBaR basis**

For the BARBaR model, we follow the methodology proposed by Richardson and Green (1997) and Nobile and Green (2000), with some key adaptations. Firstly, in our models we do not use a categorical variable to allocate observations to any of the components. The update of such an allocation is, in Richardson and Green (1997), a separate step, which can be omitted without affecting the way the remaining parameters are computed. In a similar way to Nobile and Green (2000) and Bottolo et al. (2003), we use weighted Gaussian kernels for the purpose of regression, although, as mentioned before, there is no imposition that these weights have to sum to 1.

For the BARBaR model, we only need to add between-model moves to our inference approach. In fact, the basis component parameters, such as the location and variance for Gaussian kernel components, are all updated as part of the standard MCMC inference in both the BARBaR-f and the BARBaR models. Then, we can follow the same approach as for the BARS model, and find the probability of between-model jumps from Eq. 4.3.5 and 4.3.4, where in this case the constant $\omega = 1$ to ensure that only dimension-jumping moves are considered. For the split-combine moves, we follow a similar approach to Richardson and Green (1997) to ensure sensible suggestions, and propose changes that still preserve the moments of the kernels considered. In particular, for a combine move, the density component parameters are assigned by matching the zeroth, first and second moments of the new component to those of the combination of the two that it replaces, and vice versa for a split move. While we do allow both positive and negative

kernel coefficients, we assume that a positively weighted kernel must be split into two components with positive coefficients, and similarly for a negatively weighted kernel. This assumption preserves the validity of the original moment matching proposals.

**Combine step**   First we choose a component at random, with corresponding parameters $\{\beta_{j_1}, m_{j_1}, s_{j_1}\}$. We can then select an adjacent component with $\{\beta_{j_2}, m_{j_2}, s_{j_2}\}$, where $m_{j_1} < m_{j_2}$ and no other components are located in the interval $[m_{j_1}, m_{j_2}]$. In order for the two kernels to be merged using the moment matching approach, we need to also ensure that the coefficients $w_{j_1}, w_{j_1}$ have the same sign: if this is not the case, the jump is rejected without proceeding any further. If the condition is satisfied, we can then combine them into a new kernel with parameters $\{w_{j^*}, m_{j^*}, s_{j^*}\}$, which can be obtained from the following

$$
\begin{aligned}
w_{j^*} &= w_{j_1} + w_{j_2}, \\
w_{j^*} m_{j^*} &= w_{j_1} m_{j_1} + w_{j_2} m_{j_2}, \\
w_{j^*}(m_{j^*}^2 + s_{j^*}^2) &= w_{j_1}(m_{j_1}^2 + s_{j_1}^2) + w_{j_2}(m_{j_2}^2 + s_{j_2}^2).
\end{aligned}
\tag{4.3.9}
$$

Note that, once suitable adjacent components are selected, this combine proposal follows the approach by Richardson and Green (1997) and is deterministic.

**Split step**   The split step needs to be able to reverse the combine approach above, so it can be deduced from it. In particular, we first need to select a component $j^*$ at random, which can be split into adjacent components $j_1, j_2$ satisfying Eq. 4.3.9. Since three new parameters need to be chosen, we need a vector $\boldsymbol{u}$ of three

random variables, as mentioned in Section 2.2.4. We draw each component of $\boldsymbol{u}$ from a Beta distribution, so that $u_i \in [0, 1]$, $i = 1, 2, 3$. In particular, we let

$$u_1 \sim \text{Beta}(a_1, b_1) \qquad u_2 \sim \text{Beta}(a_2, b_2) \qquad u_3 \sim \text{Beta}(a_3, b_3), \qquad (4.3.10)$$

where, after some testing, we let $a_1 = a_3 = b_3 = \frac{60}{k}$, $b_1 = b_2 = \frac{120}{k}$ and $a_2 = 2$, where $k$ is the current number of components. These choice allow the proposals to change in width and shape depending on the number of kernels already present. Furthermore, note that these densities will need to be evaluated for the values of $\{w_{j^*}, m_{j^*}, s_{j^*}\}$ to obtain the acceptance probability for the combine step. Hence, we opt for a choice of advantageous parameters for the Beta densities. In particular, the choices for $(a_1, b_1)$ yield a slightly negatively skewed density, which favours smaller coefficients. On the other hand, $(a_2, b_2)$ yield a more strongly skewed and narrow density, favouring proposal of components which become closer as the number of existing components increases. Finally, the choices for $(a_3, b_3)$ yield a symmetric density centred at 0.5.

After we draw values for $u_1, u_2, u_3$, we can set

$$
\begin{aligned}
w_{j_1} &= w_{j^*} u_1, \\[4pt]
w_{j_2} &= w_{j^*}(1 - u_1), \\[4pt]
m_{j_1} &= m_{j^*} + u_2 s_{j^*} \sqrt{\left(\frac{w_{j_2}}{w_{j_1}}\right)}, \\[4pt]
m_{j_2} &= m_{j^*} - u_2 s_{j^*} \sqrt{\left(\frac{w_{j_1}}{w_{j_2}}\right)} \\[4pt]
s_{j_1}^2 &= u_3(1 - u_2^2) s_{j^*}^2 \frac{w_{j^*}}{w_{j_1}}, \\[4pt]
s_{j_2}^2 &= (1 - u_3)(1 - u_2^2) s_{j^*}^2 \frac{w_{j^*}}{w_{j_2}},
\end{aligned}
$$

where again the resulting components $j_1, j_2$ will have coefficients with the same sign as the initial kernel $j^*$. We also need to check the adjacency condition is satisfied, otherwise the jump move is rejected before proceeding any further. In order to compute the acceptance probability from Eq.(2.2.6), we need to obtain the Jacobian for the moves above, which is

$$J = \frac{|w_{j^*}(m_{j_1} - m_{j_2})|s_{j_1}^2 s_{j_2}^2}{u_2(1 - u_2^2)u_3(1 - u_3^2)s_{j^*}^2}.$$

The acceptance probability, denoted as $a$, for the split move then follows Eq.(2.2.6), with $q(u)$ being the product of each density function in Eq.(4.3.10) for the respective entry of $u$. The reciprocal of acceptance probability $a$ is used for the combine move.

**Birth step**  Let the current model $M_k$ be described by $k$ kernel components with corresponding kernel locations and variances $m_j, s_j$, $j = 1, \ldots, k$. We can then birth a new kernel at random anywhere on the domain of the covariate. That is, we draw a new kernel location uniformly from the covariate domain. We then need to assign a weight and width to the new kernel, by drawing at random from a modified prior density of each kernel parameter. Similarly to the BARS case, we want to propose a move to a state yielding a similar enough curve to the existing one, as this will increase the chances of acceptance. In the case of additive kernels, introducing a new component with zero weight produces no changes in the resulting curve. Nevertheless, we do not want to assign a zero weight since, in order to ensure the reversibility of moves, only kernels with weight of exactly zero could then be removed. Instead, we draw from a normal distribution that is

narrower than the uninformative coefficient prior, so as to reduce the chance of selecting a large new coefficient which would result in undesirably large change in the parameter estimate. Then, the three new parameters are

$$u_1 = w^* \sim N(0, t^2), \ u_2 = m^* \sim \text{Unif}(d_l, d_r) \text{ and } u_3 = s^* \sim \text{Gamma}(a, b),$$

$$(4.3.11)$$

for $x \in [d_l, d_r]$ and where, after some testing, we set $t = 0.1$, $a = 3.3$ and $b = 0.3$.

**Death step** For a death, we choose one of the existing components with location $\tilde{m}$ at random and remove it, while the remaining components remain unchanged. Given the complete randomness of both the birth and death choices, the Jacobian is simply $J = 1$. The acceptance probability for the birth move follows Eq. 2.2.6, with $q(u)$ being the product of density functions in Eq.(4.3.11) for the respective entry of $u$. The reciprocal of this acceptance probability $a$ is used for the death move.

It is worth noticing again that we are mainly interested in modelling periodic covariates. For the BARBaR model, the adaptation for a periodic covariate is very similar to the one necessary for the BARS model, which we summarised at the end of Section 4.3.3. First, the basis components themselves need to be periodic, which is achieved in this case by using approximated wrapped normal kernels. In particular, we write the wrapped normal distribution of $x \in [0, 2\pi)$ as $WN(m, s^2)$ with the probability density function,

$$\mathring{F}\left(x, m, s^2\right) = \frac{1}{s\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left\{-\frac{(x - m - 2\pi k)^2}{2s^2}\right\}.$$

Because the terms of the series converge to zero, it is possible to approximate the pdf with a truncated series. The number of terms of the series that should be included depends on the size of $s$, as narrower densities will have a more limited support than wider ones (Mardia and Jupp, 2009) and hence require fewer terms. In our case, we note that wide densities should not be expected, as the baseline coefficient $\beta_0$ introduced in Section 4.2.2 should effectively capture the average behaviour of the parameter being modelled. After some inspection, we opted to enforce $s < \pi$, as that allowed the ability to capture a sufficient range of density shapes, and the resulting pdf could be approximated by considering 5 terms without loss of information. We denote this pdf as $\mathring{F}(\cdot)$, such that

$$\mathring{F}\left(x, m, s^2\right) = \frac{1}{s\sqrt{2\pi}} \sum_{k=-4\pi}^{6\pi} \exp\left\{-\frac{(x - m - 2\pi k)^2}{s^2}\right\}. \qquad (4.3.12)$$

Note that, in order to model parameters that are dependent on periodic covariates, further adjustments are required. Similarly to the case where a spline formulation is used, prior or proposal densities related to the component locations need to also be periodic. While this is not an issue when a uniform distribution is used, for Beta densities we use the same shifting approach as proposed for the BARS models.

## 4.4 Evaluation of methods

This section introduces the case studies used to evaluate the relative performance of the models introduced in Section 4.2.2. As previously mentioned in Section

4.1, the cases studies are selected from the paper by Jones et al. (2016) to allow comparison with this previous study. To assess performance, we first focus on estimation of the statistical model parameters. These estimates are then used to calculate distributions of return values corresponding to long return periods, which can be directly compared with the corresponding values obtained from the original parameters used to simulate the datasets. Finally, summary distributions of return values are compared using appropriate statistical tools, as described in Section 4.4.1.

## 4.4.1 Return value distributions and comparison methods

From an engineering perspective, the main inferences from the current analysis are estimates of marginal and conditional return values corresponding to some long return period. In their paper, Jones et al. (2016) use Monte Carlo simulation to obtain empirical return level distributions for each case considered. When the return period is small, say 100 years or less, their simulation approach is computationally feasible. On the other hand, when longer return periods are required, Monte Carlo simulation becomes computationally demanding. In these circumstances, numerical integration schemes yield dramatic reductions in the computational complexity of return value estimation (Ross et al., 2017). We implement the latter approach.

**Return value integration scheme**

The inference procedures introduced in Section 4.3 produce samples from the posterior distributions of the model parameters for each dataset. In order to obtain

estimated return values from these results, one needs to find a way to summarise

the contribution from each iteration of this chain and across the different simulated

datasets for each of the cases introduced in Section 4.4.2. As mentioned above,

while Monte Carlo simulation provides a natural framework, it can also be compu-

tationally intensive. An alternative approach consists of using numerical methods

to obtain a summary distribution of the return levels for each of the cases and

models considered, by extracting the information obtained across MCMC runs for

each data sample and then across all simulated datasets for the same case (Ross

et al., 2017).

Let us first focus on a single MCMC draw from the posterior of the parameters.

First we need to subdivide the covariate domain into small subsets (bins), which

provide a set of binned covariates common to all the samples of each case consid-

ered. Assume we want to divide the covariate domain in $n_b$ total bins, and denote

a covariate bin $S_j$, for $j = 1, 2, ..., n_b$. It is essential to choose a large enough

number of bins such that $S_j$ is sufficiently small and we can assume each model

parameter $\phi(x_i), \psi(x_i), \xi(x_i)$ is constant within the bin, that is for all observed

covariates $x_i \in S_j$. Let us denote the set of Poisson-GPD parameters to be esti-

mated by $\Theta(x) = \{\phi(x), \psi(x), \xi(x)\}$. For each specific bin $S_j$, let $\Theta_j$ indicate

the estimates of these parameters in the bin. For each bin, we can then compute

the cumulative distribution function (cdf) of any storm peak event with such a

covariate value by using the estimates of the distribution parameters $\Theta_j$ in the

bin. We denote this by $F(y_p|\Theta_j)$, which corresponds to the GPD cdf

$$
F(y_p|\Theta_j) = \begin{cases} 1 - \left(1 + \xi_j \frac{y_p - u_j}{\psi_j}\right)^{-1/\xi_j} & \text{if } \xi_j \neq 0, \\[2ex] 1 - \exp\left(-\frac{y_p - u_j}{\psi_j}\right) & \text{if } \xi_j = 0. \end{cases}
$$

Furthermore, the parameter $\phi_j = \phi(x_j)$ can be interpreted as the mean number of storm peak events in the covariate bin $S_j$ in a year. We can then obtain the cumulative distribution function $F_{M_T}(y_p|\Theta_j)$ of the maximum $M_T$ observed in a period of $T$ years in $S_j$ as

$$
\begin{aligned}
F_{M_T}(y_p|\Theta_j) &= P(M_T < y_p) \qquad\qquad\qquad\qquad\qquad\qquad\qquad (4.4.1)\\
&= \sum_{k=0}^{\infty} P(k \text{ events in } S_j \text{ in } T \text{ years}) P^k(\text{size of an event in } S_j < y_p)\\
&= \sum_{k=0}^{\infty} \frac{(T\phi_j)^k}{k!} \exp(-T\phi_j) \times F^k(y_p|\Theta_j)\\
&= \exp\left\{-T\phi_j\left[1 - F(y_p|\Theta_j)\right]\right\}.
\end{aligned}
$$

Eq. 4.4.1 provides estimates of the cdf of the $T-year$ maxima for each of the $n_b$ bins.

As mentioned before, it is often the case that "omni-covariate" return values, as well as results for some specific subset of the covariates, are required. In order to obtain these distributions, we need first to exploit the fact that, given some covariates, storm peak events are considered to be independent. By the law of probability, the joint cdf across all covariate values can then be obtained by taking the product of the individual cdf's for each of the bins. Mathematically, this is

obtained by taking the product over all covariate bins, such that

$$F_{M_T}(y_p|\Theta) = \prod_{j=1}^{n_b} F_{M_T}(y_p|\Theta_j). \tag{4.4.2}$$

Practically, this corresponds to computing an "omni-covariate" (e.g. omni-directional) cumulative distribution function $F_{M_T}(y_p|\Theta)$ of the overall storm peak maximum $M_T$ over all covariate bins.

We can then repeat the above operations given the posterior samples from each MCMC iteration, denoting the resulting cdf's as $F_{M_T}(y_p|\Theta^i)$ for $i = 1, \ldots, n_{\text{iter}}$. Then, we can obtain the posterior expected "omni-covariate" (e.g. omni-directional) cumulative distribution function as

$$\widehat{F}_{M_T}(y_p) = \frac{1}{n_{\text{iter}}} \sum_{i=1}^{n_{\text{iter}}} F_{M_T}(y_p|\Theta_i). \tag{4.4.3}$$

If we have more then one realisation, such as in the case of a simulation study, we then need to proceed to a second stage of the approach. In this case, we consider the posterior expected "omni-covariate" (e.g. omni-directional) cdf indexed by simulated sample $l$ as $\widehat{F}^l_{M_T}(y_p)$. Then, we obtain a summary of the behaviour across the $n_{\text{sim}}$ samples as

$$\widehat{F}^{\text{sim}}_{M_T}(y_p) = \frac{1}{n_{\text{sim}}} \sum_{l=1}^{n_{\text{sim}}} \widehat{F}^l_{M_T}(y_p). \tag{4.4.4}$$

Algorithm 3 summarises the whole approach. Let us apply it to Case 2 as an example. For each of the 100 data sample realisations available, we can implement the first stage of this approach on the 20000 MCMC iterations. We then obtain

---

**Algorithm 3** Posterior expected cdf of return values

---

<u>**Stage 1**</u>:

Consider the l$^{\text{th}}$ observation sample $\{\boldsymbol{y}^l, \boldsymbol{x}^l\}$, for $l = 1, \ldots, n_{\text{sim}}$.

**Input**:

- The posterior estimates for the GPD and Poisson parameter, denoted as $\Theta_j^i$ and evaluated at each bin $j = 1, \ldots, n_b$ and MCMC iteration $i = 1 \ldots, n_{\text{iter}}$;

- A return period $T$ of interest;

- Some future return level (e.g. Hs) of interest $y_p$;

**for** each covariate bin $j = 1, \ldots, n_b$ **do**

    **for** each MCMC run $i = 1 \ldots, n_{\text{iter}}$ **do**

        Compute $F_{M_T}(y_p | \Theta_j^i)$ as in Eq. 4.4.1

    **end for**

    Compute $F_{M_T}(y_p | \Theta^i)$ as in Eq. 4.4.2

**end for**

Obtain the posterior expected "omni-covariate" cdf $\widehat{F}_{M_T}^l(y_p)$ as in Eq. 4.4.3

<u>**Stage 2**</u>:         ▷ *For hindcast datasets, $n_{sim} = 1$ and Stage 2 is not necessary.*

If more the one sample realisation $l$ is available:

1. Repeat **Stage one** for each sample realisation $l = 1, \ldots, n_{\text{sim}}$;

2. Compute the summary (average) mean expected posterior "omni-covariate" cdf $\widehat{F}_{M_T}^{\text{sim}}(y_p)$ as in Eq. 4.4.4

---

the corresponding values for Eq. 4.4.3, which give the expected behaviour across

the MCMC draws and are shown in grey in Figure 4.4.1. If we proceed to stage 2

of the algorithm, we can obtain the average posterior expected omnidirectional cdf

$\widehat{F}_{M_T}^{\text{sim}}(y_p)$, which is shown for the results of the case study in Figure 4.4.4 in Section

4.4.2. This can then be compared to the true underlying return value distribution

shown by the black line in both Figure 4.4.1 as well as the omni-directional plots

in Figure 4.4.4.
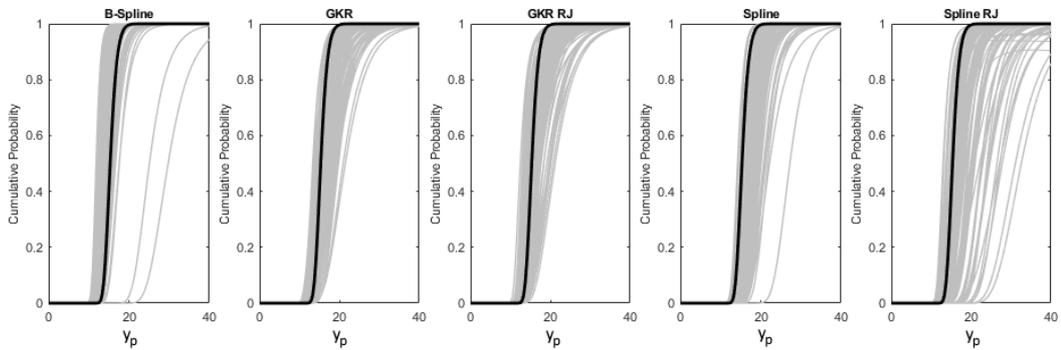


Figure 4.4.1: Posterior expected omnidirectional cdf $\widehat{F}_{M_T}^{l}(y_p)$ (in grey) for each of the 100 sample realisations of Case 2 shown in Figure 4.4.2, corresponding to a return period of ten times the period of the original sample, with the true return value cdf in solid black.

The same two-stage approach can be used to obtain cumulative distribution func-

tions for subsets of the covariates, e.g. for some given directional sector of interest,

by taking the product only across a selection of sectors rather than for all $n_b$ sec-

tors. The sectoral plots (right) in Figures 4.4.4 are the sector-specific equivalent

to the omnidirectional $\widehat{F}_{M_T}^{\text{sim}}(y_p)$ curves (left).

**Distribution divergence tests**

In order to quantify the quality of return value inference, one can use different

statistics that compare a sample with a reference probability distribution. In this

case, for each sample realisation we can compare the empirical cumulative distribution function from the fitted model, obtained following the method described above, with the 'true' one from the known underlying case. The Kolmogorov-Smirnov (KS) and the Kullback-Leibler (KL) criterion have been used in the literature to quantify the discrepancy between distributions. Given two distribution functions, the KS test uses the maximum vertical distance between their cumulative distributions, defined as

$$D_{\mathrm{ks}}\left(F_0, F_1\right) = \sup_x |F_1\left(x\right) - F_0\left(x\right)|,$$

to obtain a measure of their discrepancy. In particular, perfect agreement between $F_1\left(x\right)$ and $F_0\left(x\right)$ yields a minimum KS criterion value of zero, and the larger the value, the more the two distributions differ.

The Kullback-Leibler divergence compares distributions using the average ratio of the logarithms of the density functions

$$D_{\mathrm{kl}}\left(F_0, F_1\right) = \int_{-\infty}^{\infty} \log\left(\frac{f_0\left(x\right)}{f_1\left(x\right)}\right) f_0\left(x\right) dx, \qquad (4.4.5)$$

where there is no upper bound to $D_{\mathrm{kl}}\left(F_0, F_1\right)$ and, similarly to the KS criterion, perfect agreement yields a minimum KL divergence of zero. In order to compute the KL criterion values, we approximate the integral in Eq. 4.4.5 following the approach of Perez-Cruz (2008).

In Section 4.4.3, we illustrate the performance of the different models considered for both cases introduced in Section 4.4.2. For all the 100 samples in each of these cases, we obtain a posterior predictive cdf, corresponding to the grey curves shown

in Figure 4.4.1. Then, we compare it with the 'true' underlying cumulative distribution to obtain the corresponding KS and KL divergence, and we use box-plots to summarise these criterion values across the 100 samples. The same approach is then repeated for the each of the directional sectors of interest.

## 4.4.2 Case studies and inference

The samples were the same as those analysed by Jones et al. (2016) and were obtained directly from the authors. We consider here the first two cases, which we denote as Case 1 and 2 respectively. The data were obtained by simulating 100 sample realisations of size 1000 from each of the underlying models. Given the parameter estimates, the previous authors assumed that all sample realisations correspond to an observational period of $\mathcal{T}$ years.

**Case 1** : For extreme value threshold $u(x) = 0$, Jones et al. (2016) simulate 1000 observations with a uniform Poisson rate $\phi(x) = 1000/360$ per degree covariate, and a low order Fourier parameterisation of GPD shape $\xi(x) = \sin(x) + \cos(2x) + 2$ and scale $\psi(x) = -0.2 + (\sin(x - 30))/10$.

**Case 2** : For extreme value threshold $u(x) = 0$ and the same Fourier parameterisation of GPD shape and scale as in Case 1, a non-uniform Poisson rate $\phi(x) = \max(\sin(x) + 1.1, 0) \times 1000/c_\rho$, where $c_\rho = \int_0^{360} \max(\sin(x) + 1.1, 0)\, dx$ is used to simulate 1000 observations.
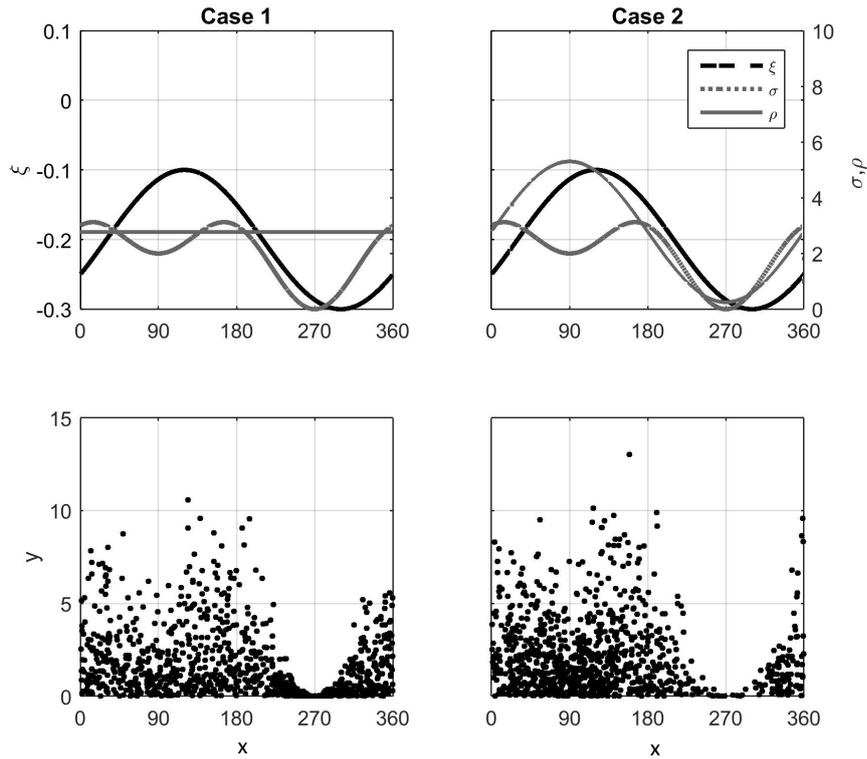
Figure 4.4.2: Illustrations of sample realisations from each of Cases 1 (left) and 2 (right).   Upper panels show parameter variation of GPD shape $\xi(x)$, scale $\psi(x)$ and Poisson rate $\phi(x)$ with direction $x$ for each case. Lower panels show a realisation of the corresponding simulated samples.

Figure 4.4.2 illustrates typical sample realisations of both cases.   Note that the parameter variation of the GPD shape $\xi(x)$ and scale $\psi(x)$ with direction $x$ are identical in Cases 1 and 2. The main and only difference between them lies in the Poisson rate $\phi(x)$, which is constant in Case 1 only.  Cases 2, instead, allows us to focus on a more difficult inference problem, since the very small rate $\phi(x)$ at $x \approx 270°$ leads to a sparsity of corresponding observations.  It is worth noticing that the shape $\xi(x)$ is negative throughout, and reaches its largest negative values at $x \approx 120°$, which yields larger observations.

For each case, the extreme value threshold $u$ was fixed at zero during inference, while both the GPD shape $\xi$ and scale $\psi$ of exceedance size vary as a function of

the covariate $x$. As far as the Poisson rate $\phi$ of threshold exceedance is concerned, we use a constant model in Case 1, while this is covariate dependent in Case 2. We proceed to perform inference using the methodology described in Section 2.2 for the formulations introduced in Section 4.2.2, denoted as follows:

Model 1 **P-spline** : penalised spline with 30 equidistant and fixed knots. Note that, since Model 1 is known in the literature to perform well (Jones et al., 2016), it is here used as a benchmark.

Model 2 **BARBaR-f**: Gaussian kernels RBF's (3 kernels for the exceedance rate parameter and the GPD shape, 5 for the GPD scale), with kernels initiated according to the density of observations.

Model 3 **BARBaR**: Gaussian kernels RBF's, with kernels initially allocated according to the density of observations. During inference, RJ MCMC is used, so both the number and location of kernels are allowed to vary.

Model 4 **P-spline-u**: penalised spline with 8 unevenly spaced fixed knots, with the first half allocated according density of observations, and the remaining positioned halfway between the initial knots.

Model 5 **BARS**: penalised spline, where the knots are initiated as for Model 4, but during inference, RJ MCMC is used, so both the number and location of knots are allowed to vary.
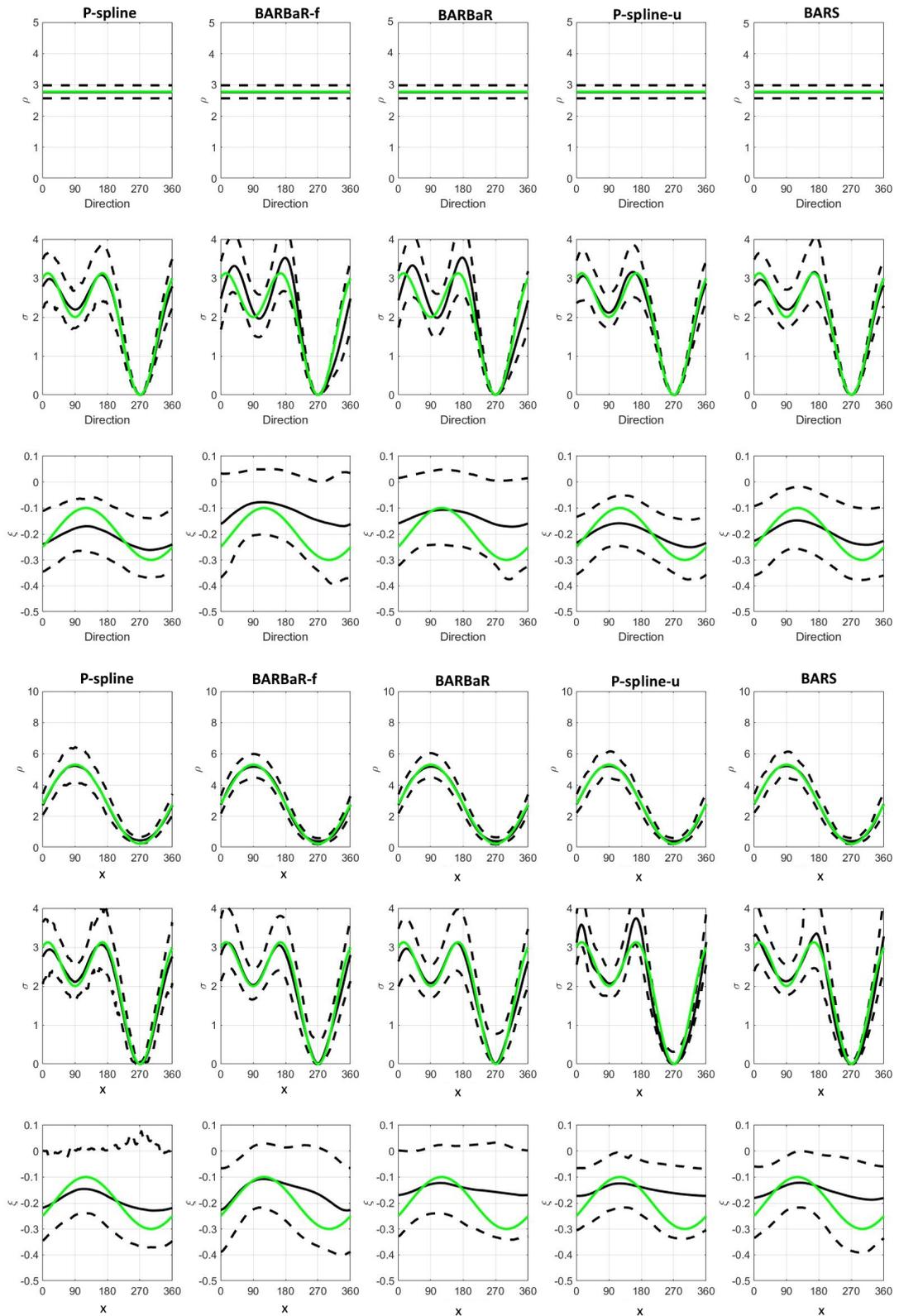
Figure 4.4.3: Parameter estimates for rate of occurrence $\phi(x)$ of the exceedances (upper), the GPD shape $\xi(x)$ (middle) and GPD scale $\psi(x)$ (lower) for the sample realisation of Case 1 (top) and Case 2 (bottom) shown in Figure 4.4.2, for Models 1-5. Each panel illustrates the true parameter (solid green), posterior median estimate (solid black) with 95% credible interval (dashed black).

For both cases we produce plots of parameter estimates based on MCMC run lengths of 40000 in total, with the first 20000 samples removed as burn-in. Note that for Case 1, we fit a constant model for the Poisson parameter, with the same resulting chains being later used to produce posterior expected cdf estimates for all Models 1-5.

We first inspect the median parameter estimates over the sample realisations for the three model parameters in both Case 1 and Case 2 across models 1-5, shown in Figure 4.4.3. Visual inspection suggests that estimates are of similar quality across models, indicating all models are producing good results. As expected, wider intervals between the upper and lower quantile suggest that the identification of $\xi(x)$ is more difficult than $\psi(x)$ and $\phi(x)$. Despite slightly wider interquantile ranges for the BARBaR and BARS formulations, the models seem to produce similar fitting results.

We proceed to consider the posterior cumulative distribution functions of return values, corresponding to a return period of ten times the period of the original sample. These are shown in Figure 4.4.4. Here we notice that omnidirectionally there is a similarity in performance by model type (e.g. between splines formulations and similarly between BARBaR models). In Case 1, the wider interquantile range for the BARBaR models estimates of the GPD shape affect the return values, which consequently are generally higher than both the truth and the spline models estimates. On the other hand, all spline based models underestimate the return value cdf for lower probability values.
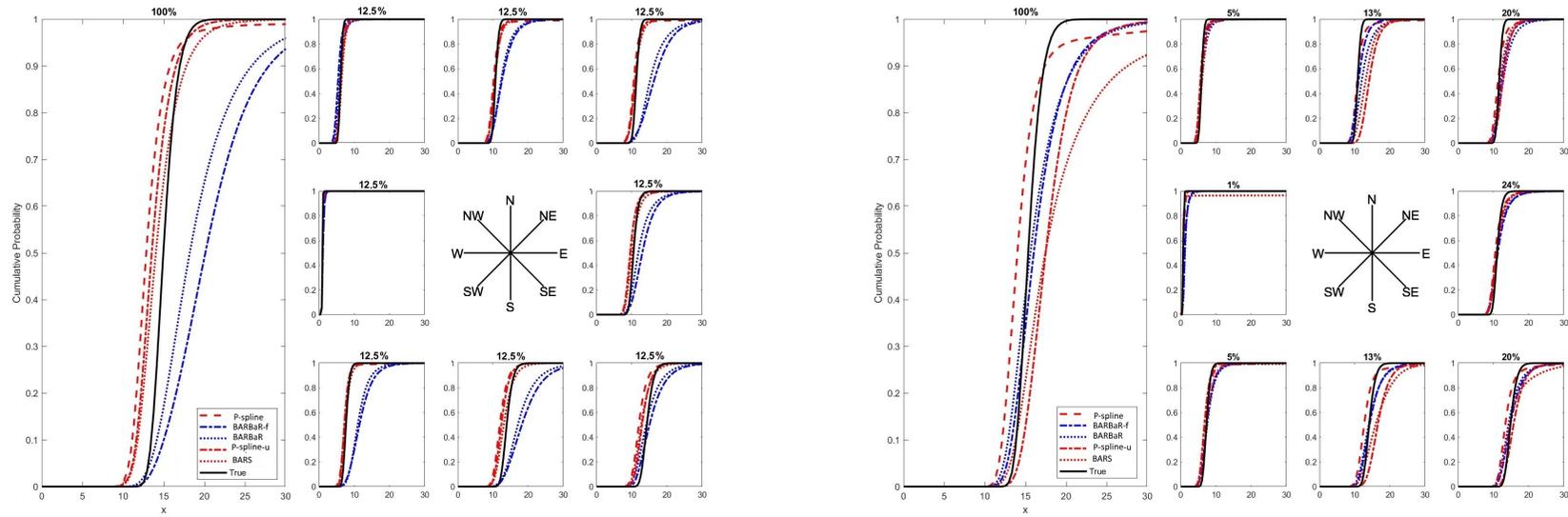
Figure 4.4.4: Average posterior expected return value cdf for the sample realisations of of Case 1 (left) and Case 2 (right) shown in Figure 4.4.2, corresponding to a return period of ten times the period of the original sample. The left hand panel shows the omnidirectional return value distribution, and right hand panels the corresponding directional estimates. The title for each panel gives the expected percentage of individuals in that directional sector. The true return value distribution is given in solid black.

If we inspect the results by directional octant, the spline models show consistently good results across the different directions, while the BARBaR models yield closest agreement with the truth model when both the shape and scale have their lowest values (west and north-west sectors). Nevertheless, none of the models outperform the others across all sectors. One also needs to remember than, in Case 1, the same constant model is used to estimate the Poisson rate, such that the latter does not have any different impact on the return values across different model bases.

Focusing on Case 2, no model seems to consistently outperform the others. The spline basis with a high number of knots, used as a benchmark, seems to again underestimate return values for low probability values, as well as having some issues for higher probabilities. If we look back at Figure 4.4.1, we notice that in a couple of occasions the return levels seem to differ considerably from the truth. We considered the specific sample realisations, and notice this is a consequent of an ill-fitting starting solution, which can be difficult and sometimes unstable for models with a high number of correlated parameters.The BARS model seems to overestimate return levels and similarly struggle with some of simulated data samples, which we found to be the result of the model exploring some locally unlikely value of the GPD parameters for a relatively small number of runs in the respective MCMC chain. Overall, the BARBaR models and the spline model with a low fixed number of knots seem to be in agreement the most, both amongst each other and with respect to the true value. Nevertheless, they still yield marginally higher estimates for the return levels cdf when compared to the density obtained from the true underlying parameters used to simulate the data study.

If we consider the results by directional octant, while the models perform similarly

across most sectors, the highest disagreement is registered in the northern and southern octant, which correspond to the scale parameter highest values. Furthermore, the most noticeable issues for the BARS model are localised to the western sector, which has the smallest number of observations available, and might be a consequence of some instability in the RJ moves proposed in such a situation.

### 4.4.3 Assessing quality of inference

We then employ the criteria illustrated in Section 4.4.1 to compare the distributions of return values statistically and consider how they perform across different realisation samples. In particular, we investigate both the Kolmogorov-Smirnov and the Kullback-Leibler criterion, for both the omnidirectional and directional return values distribution shown in Figure 4.4.4.

Figures 4.4.5 and 4.4.6 summarise the characteristics of the distributions for KS and KL divergence criteria for Cases 1 and 2 respectively. In particular, it is worth noticing that, given the definitions provided in Section 4.4.1, the KL criterion captures the maximum divergence between two different distributions, while the KL test provides a more general summary as it considers all point-wise distances across the different probability levels.
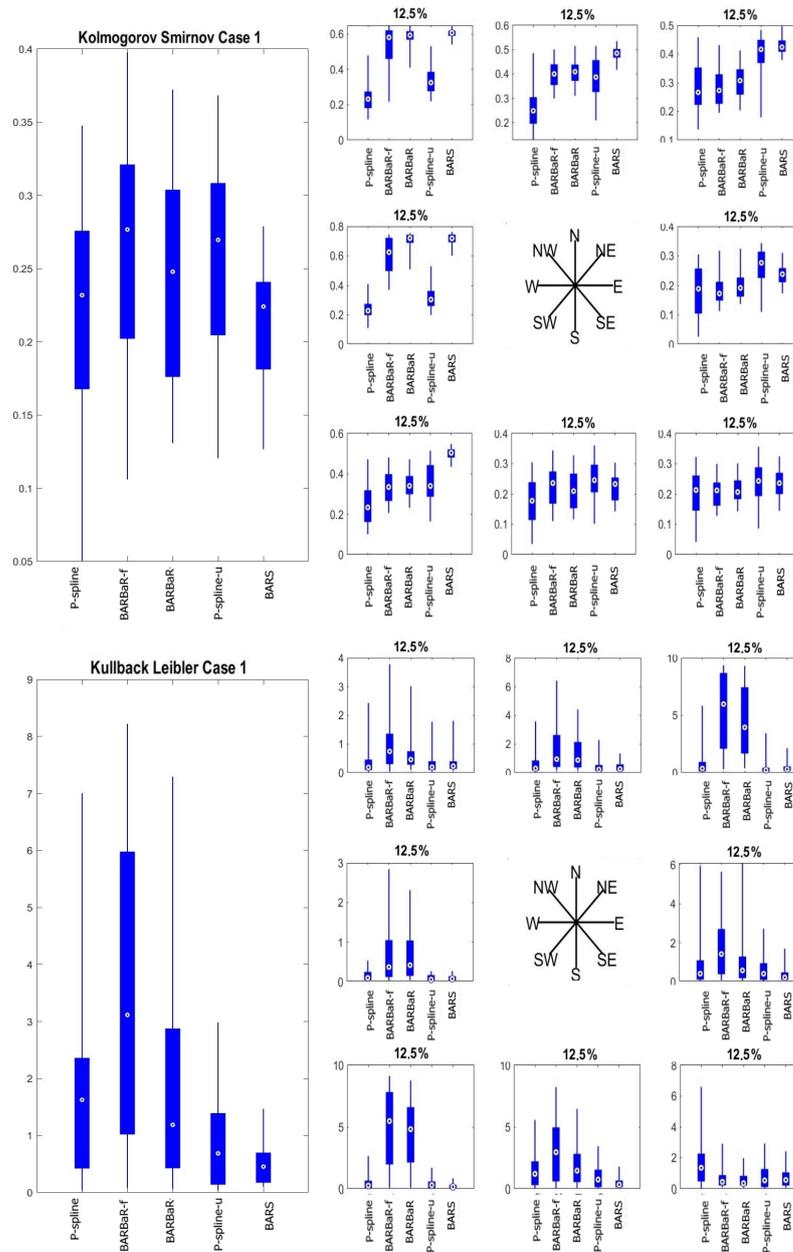
Figure 4.4.5: Box-whisker comparison of samples of the Kolmogorov-Smirnov (KS, top) and Kullback-Leibler (KL, bottom) divergence criteria between omnidirectional posterior expected return value cdf's (corresponding to a return period of ten times that the original sample) and by directional sector, estimated under samples from the true return value distribution and those estimated under models of each of 100 sample realisations for Case 1. The sample of the KS and KL divergence criteria are summarised by the median (white disc with black central dot), the interquartile range (blue rectangular box, with vertical lines showing the 2.5%, 97.5% interval).
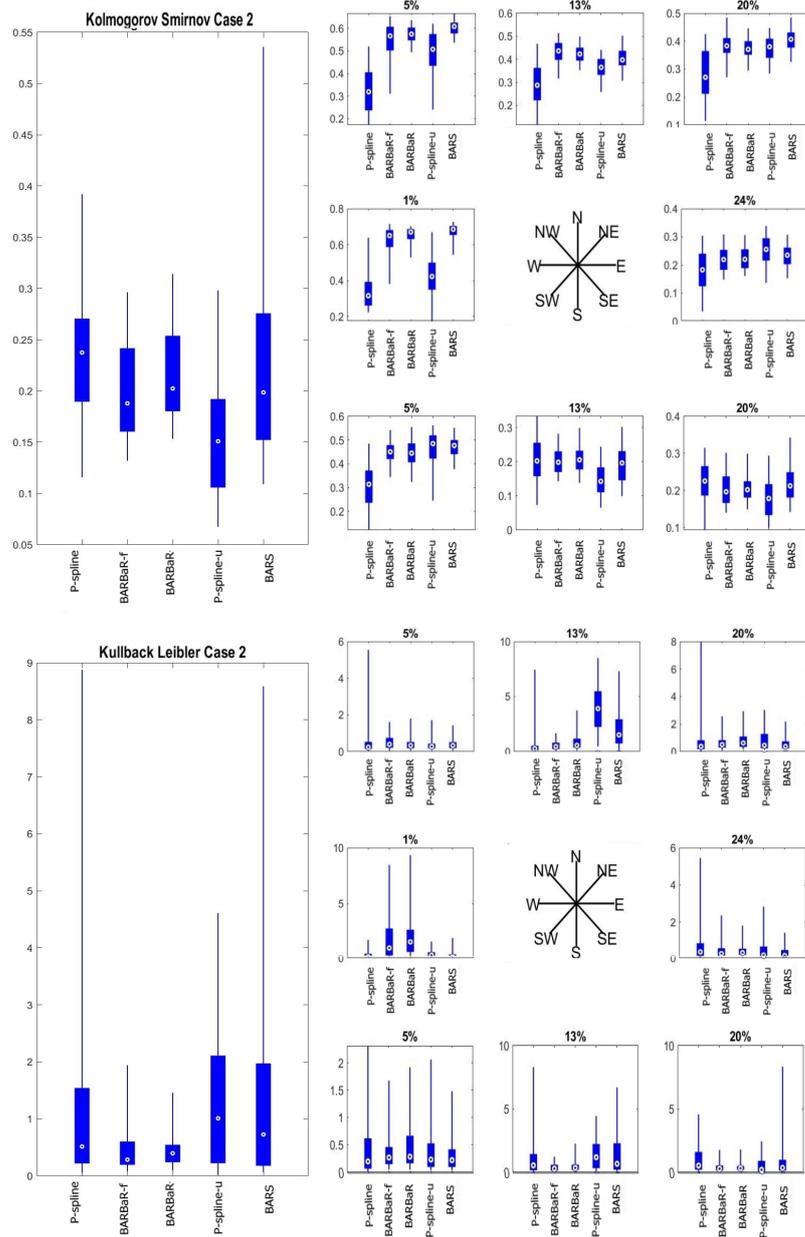
Figure 4.4.6: Same as Figure 4.4.5 but applied to Case 2.

In Case 1, we note that all models perform similarly, in terms of maximum divergence, according to the KS test when looking at the omni-directional distributions. Most differences are noticeable in the western and north-western sectors, where both $\xi(x)$ and $\psi(x)$ are near their minimum values and where the two spline models with fixed number of components achieve the lowest score. If instead we consider the overall divergence between cdf's using the KL criterion, the two spline

models with more parsimonious bases achieve better results omnidirectionally. In general, the BARBaR models seemed to result in more varied results across different realisations, as shown by wider interquantile ranges in the KL criterion, and all but the BARBaR-f model match the performance of the benchmark spline approach for omnidirectional cdf's.

In Case 2, we notice that all non-stationary parameterisations yield lower KS values than the benchmark spline model. We can notice a less stable performance from the BARS model across different sample realisations, as shown by wider interquantile ranges. This is also in agreement with the sample posterior cumulative distributions shown in Figure 4.4.1. This is visible both in terms of maximum (KS test) and average (KL test) divergence. Considering the criteria by octant, the BARBaR models struggle the most for the western sector, where the rate of occurrence of events is relatively low, and both $\xi(x)$ and $\psi(x)$ are near their minimum values. Most differences are noticeable in the western and north-western sectors, where both $\xi(x)$ and $\psi(x)$ are near their minimum values and where the two spline models with fixed number of components achieve the lowest score. This is likely due to the fact that, although contributions from observations that far from the function centre effectively tend to zero, RBF's have a global support, and neighbouring areas will have a higher influence on sectors with a small number of observations, especially when compared to spline based models.

## 4.5 Northern North Sea hindcast

In Section 4.4 we tested and assessed the models on some case studies. We now proceed to apply them to the dataset introduced in Section 4.1, which considers hindcast significant wave heights in the North Sea. Extreme sea states in the North Sea are dominated by winter storms originating in the Atlantic Ocean, which propagate eastwards across the northern part of the North Sea. Different factors contribute to the directionality of the extreme seas, mainly the land shadows of the UK, Scandinavia, and the coast of mainland Europe, and the fetches associated with the Atlantic Ocean, Norwegian Sea, and the North Sea itself.

In particular, in Figure 4.1.1, we can see that the land shadow of Norway (approximately the directional interval $45° - 210°$ has a considerable effect on the rate and size of occurrences with direction. In fact, there is a dramatic increase in both rate and size of occurrences with increasing direction at around $210°$, corresponding to Atlantic storm events from the South-West being able to pass the Norwegian headland. We therefore should expect considerable directional variability in the model parameter estimates for the sample.

### 4.5.1 Inference

For all models discussed here, the extreme value threshold needs to be estimated. Given the non-stationarity present, it is advisable to consider covariate-dependent thresholds. Here we decide to consider thresholds corresponding to a non-exceedance probability varying between 0.6 and 0.8. Note that higher and lower values were also considered initially, although these were excluded following the same crite-

rion illustrated below, and we opt to only show the results for a smaller range
of non-exceedance probability values, reduced to the most interesting, in order to
produce clearer and more informative plots. We aim to choose a threshold that is
low enough to allow as much data to be preserved as possible, while still ensuring
we are only considering extreme data and yielding stable and reliable results.
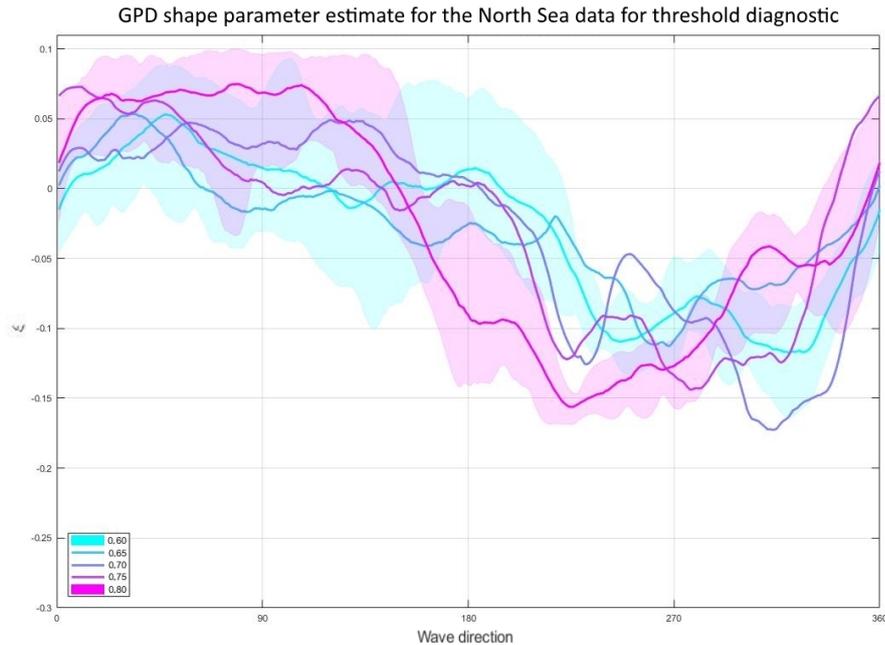


Figure 4.5.1: GPD shape estimates for the North Sea hindcast HS data shown in
Figure 4.1.1. Here, Model 1 from Section 4.4.2 is used to fit the exceedances of
thresholds obtained from non-exceedance probability ranging between 0.6 and 0.8,
with interquantile ranges shaded for the highest and lowest value.

While different methods are available to diagnose constant thresholds for extreme
data (see Section 2.1.2 and Coles, 2001), such diagnostics are harder to interpret in
the presence of non-stationarity. We found it useful to consider stability of fit by
comparing the estimates for the GPD shape parameter across different thresholds.
For consistency these estimates were all achieved using the same model, namely
the spline Model 1 from Section 4.4.2. Figure 4.5.1 shows these estimates, with
shaded areas corresponding to the interquantile ranges for the lowest (0.6) and

highest (0.8) non-exceedance probability. One can visibly notice a more erratic estimate under the highest probability around 180°, which can be expected given the marked sparsity of data in this sector. On the other hand, probabilities of 0.65 or lower seem to miss some interesting behaviour in the sector between $220° - 270°$, which is instead consistently detectable for higher values. Hence, we decide to proceed using a non-exceedance probability of 0.7.

We consider the same models for parameter formulations as in Section 4.4, using the methodology described in Section 2.2. We then inspect the parameter estimates for all model parameters across Models 1-5, which are shown in Figure 4.5.2, based on run lengths of 40000 run lengths, with the first 20000 samples removed as burn-in. Again we use Model 1 as a benchmark. Estimates seem to be mostly consistent across the different models, with higher discrepancy and wider interquantile ranges for the shape parameter, as expected. The most interesting performance perhaps is shown in the exceedance rate estimate (top). Here we notice how both the benchmark spline model and the BARS formulation better capture the sharp change shown by the data in the sector between $180° - 220°$ , with the latter also using a more parsimonious model. The superiority of the spline model in this case can be explained by the nature of the model itself. In fact, a spline formulation can capture a behaviour similar to a step change in the data with the use of two very closely located knots, which is possible if a high number of knots is considered, or by allowing knots to move in the case of a more parsimonious model. For stability reasons, the priors used in the Bayesian inference discourage very small values for the kernel widths, and consequently a very sharp step-change in the resulting parameter function.
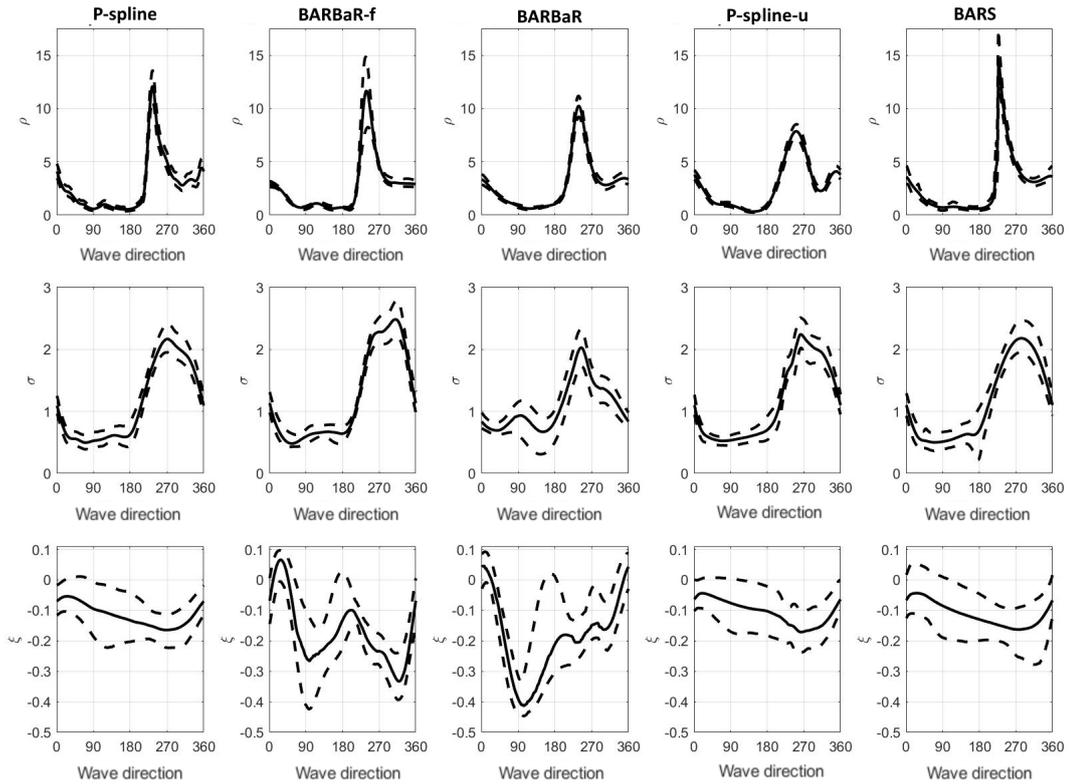
Figure 4.5.2: Parameter estimates for rate of occurrence $\phi(x)$ of the exceedances (upper), the GPD scale $\psi(x)$ (middle) and GPD shape $\xi(x)$ (lower) for the North Sea hindcast dataset shown in Figure 4.1.1, for different model parameterisations. Each panel illustrates posterior median estimate (solid) with 95% credible interval (dashed).

As the scale and shape parameters of the generalise Pareto distribution are negatively correlated, this means that different values of the two parameters can result in the same return value estimates. Hence, we then proceed to consider the posterior cumulative distribution functions of return values, corresponding to a return period of ten times the period of the original sample, which corresponds, in this case, to approximately 500 years. These are shown in Figure 4.5.3.
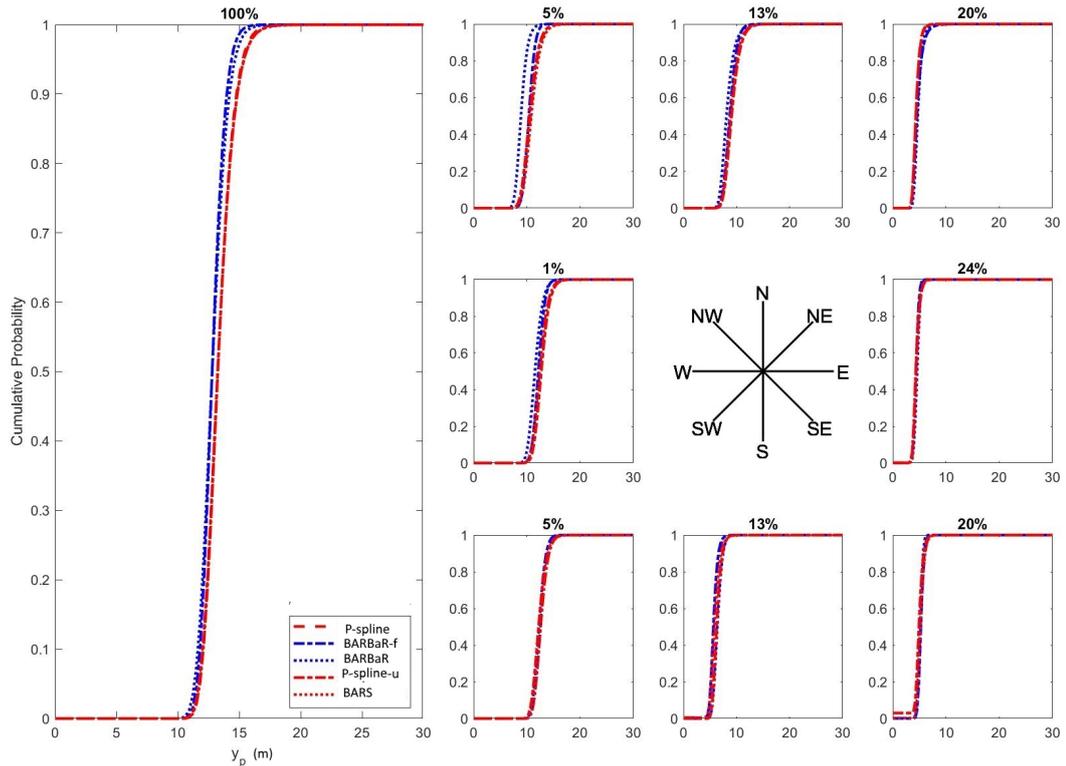
Figure 4.5.3: Posterior expected return value cdf for the NNS hindcast dataset shown in Figure 4.1.1, corresponding to a return period of ten times the period of the original North Sea sample. The left hand panel shows the omnidirectional return value distribution, and right hand panels the corresponding directional estimates. The title for each panel gives the expected percentage of individuals in that directional sector.

Similarly to the results in Section 4.4.3, all models seem to produce comparable estimates both omnidirectionally and when considering specific directional sectors. A systematic variation of return values with direction is visible for all models, with clear agreement in the sectors with highest number of observations, namely corresponding to eastern, south-eastern and southern wave direction. Most differences in the GPD scale and shape parameter estimation seem to unsurprisingly yield similar return values due to the correlation between the two parameters. We can also notice that the spline-based models differ slightly from the BARBaR ones in two of the sectors with lowest occurrence rate (west and north-west). These direc-

tions correspond to the sectors of the domain where the non-exceedance probability undergoes a sharp change, which, as we observed from Figure 4.5.2, is captured differently by models with different basis type. This discrepancy, together with the added uncertainty in sectors with fewer observations, explains the negligible difference in the omnidirectional return level.

## 4.6   Discussion

The aim of this chapter was both to introduce some novel methodology and compare its performance to existing models. In particular, the focus is on non-stationary data, where covariate dependence needs to be appropriately captured by a model to provide realistic environmental extreme value inference: models which ignore or fail to capture the effects of covariates lead to unreliable and unrealistic analysis.

Different parameterisations can be used to represent the generalised Pareto model parameters. As far as existing models are concerned, here we focus on the penalised spline formulation. We use this in its standard form (with a high number of knots and smoothing penalty) as a benchmark and consider a further two variations, both with a lower number of knots. In one of these models, the BARS approach, we allow the number and location of knots to change using reversible jump MCMC. We also introduce a novel approach, which uses a combination of radial bases, in this case Gaussian kernels, to represent the model parameters. We propose two models (BARBaR-f and BARBaR), both with a lower number of components when compared to the standard spline approach. Similarly to the spline cases, we

consider and develop the methodology for both a formulation with a fixed number of components, and one with an unknown number of components fitted using RJ MCMC.

We assessed model performance using simulation studies and later applied all approaches to a hindcast wave height dataset. When applied to hindcast data, all models seem to perform well in terms of quality of inference, yielding realistic and consistent estimates of both the omnidirectional return values distribution and capturing well the directional variation present. Overall, they all match the performance of the benchmark model, and hence generally outperform other standard alternative formulations from the literature (Jones et al., 2016).

Nevertheless, some variation in performance was captured during the simulation study, which needs to be considered together with some observations on ease of implementation. The BARBaR formulations we proposed have then performed well, especially in the most commonly available case where the rate of occurrence is covariate dependent. It is worth noticing that this novel approach can sometimes become computationally intense. In fact, 3 parameters are needed to represent each kernel, with all kernel parameters being proposed jointly by the mMALA algorithm. This yields a model where the total number of parameters to be estimated is soon similar to the spline formulations, especially when considering parsimonious models with fewer knots. This partly contradicts our original aim to produce a more parsimonious model, at least when a one-dimensional covariate is considered. Furthermore, the BARBaR models proved to be more unstable and required careful fine-tuning and more informative priors. A possible extension of these models would introduce a hyperparameter for the coefficients prior, similar

to $\rho_\theta$ in Eq.(4.3.1), which could allow us to impose a varying level of smoothness through, for example, a lasso or, more simply, a ridge penalty term. We have also observed spline inference computations being often faster, potentially outperforming the BARBaR models in computational efficiency. The BARS formulation, having fewer knots and using reversible jump inference, has the potential of combining the advantages of spline models with a more parsimonious representation. In fact, it allows a lower correlation between the basis formulation parameter, as well as faster inference due to the low number of total unknowns to be estimated. The variability of results across different sample realisations shows, nonetheless, that there is scope for refining the BARS methods, especially with respect to approach to update coefficients introduced in Section 4.3.3.

It is important to remember that environmental data is often more complicated than the cases we considered in this work. In particular, it is often the case that multiple covariates are present in the data considered. Carefully and efficiently capturing the non-stationarity is equally important with multiple covariates, while presenting additional challenges. We believe that formulations with a lower number of components would match the less parsimonious models in accuracy, while potentially outperforming them in speed and efficiency. As an example, a standard spline model is normally extended to two dimensions by spacing knots on a grid with $n_1$ and $n_2$ knots to capture the marginal behaviour of each covariate, so that the total number of unknown model parameters is now $n_1 \times n_2$. Most of the time, this will be considerably larger than the number of formulation parameters from a RBF parameterisation. Hence, this approach has the potential to be considerably more efficient in higher dimensions when compared to competing ones.

# Appendix

## 4.A  Gaussian kernels BARBaR model and derivatives with respect to the basis parameters

Let $Y \sim GPD(\nu(x), \xi(x))$. For the sake of illustration, let us focus on the GPD parameter $\nu(x)$, and assume that $\nu(x)$ has a BARBaR form with only one kernel component as a basis, that is:

$$\nu(x) = \beta \, \widehat{F}\left(\frac{x-m}{s}\right)$$

for a Gaussian kernel density $\widehat{F}$ with location $m$ and standard deviation $s$, and coefficient $\beta$. We will refer to each of these three BARBaR basis component parameters as $\theta^c$ to generalise the results.

In essence, this process exploits the chain rule to obtain first and second order derivatives of the kernel density with respect to both remaining basis parameters, location and width. More precisely, by the chain rule, the first order derivative of

the log-likelihood with respect to a formulation parameter will be:

$$\frac{\partial l}{\partial \theta^c} = \frac{\partial l}{\partial \nu} \frac{\partial \nu}{\partial \theta^c},$$

the second order will be:

$$\frac{\partial^2 l}{\partial \theta^{c2}} = \frac{\partial^2 l}{\partial \nu^2} \left( \frac{\partial \nu}{\partial \theta^c} \right)^2 + \frac{\partial l}{\partial \nu} \frac{\partial^2 \nu}{\partial \theta^{c2}}, \qquad (4.A.1)$$

and the second order cross derivative with respect to the parameters $\theta_1^c, \theta_2^c$ will be:

$$\frac{\partial^2 l}{\partial \theta_1^c \partial \theta_2^c} = \frac{\partial^2 l}{\partial \nu^2} \frac{\partial \nu}{\partial \theta_1^c} \frac{\partial \nu}{\partial \theta_2^c} + \frac{\partial l}{\partial \nu} \frac{\partial^2 \nu}{\partial \theta_1^c \partial \theta_2^c}. \qquad (4.A.2)$$

Let us focus on the coefficient $\beta$ as an example, and use the standard notation with the basis form $\nu(x) = \beta B$, then:

$$\frac{\partial \nu}{\partial \beta} = B$$

$$\frac{\partial^2 \nu}{\partial \beta^2} = 0$$

$$\frac{\partial l}{\partial \nu} = B^T \frac{\partial l}{\partial \nu}$$

$$\frac{\partial^2 l}{\partial \nu^2} = B^T \frac{\partial^2 l}{\partial \nu^2} B$$

For the remaining BARBaR basis parameter, we have to compute the different components from eq. 4.A.1.

Then, let $A = \frac{1}{\sqrt{2\pi}s} e^{\frac{-(x-m)^2}{2s^2}}$, and we have:

$$\frac{\partial \nu}{\partial \beta} = A$$

$$\frac{\partial^2 \nu}{\partial \beta^2} = 0$$

$$\frac{\partial \nu}{\partial m} = \beta A \frac{x-m}{s^2}$$

$$\frac{\partial^2 \nu}{\partial m^2} = \beta A \left( \frac{(x-m)^2}{s^4} - \frac{1}{s^2} \right)$$

$$\frac{\partial \nu}{\partial s} = \beta A \left( \frac{(x-m)^2}{s^3} - \frac{1}{s} \right)$$

$$\frac{\partial^2 \nu}{\partial s^2} = \beta A \left( \frac{2}{s^2} - \frac{5(x-m)^2}{s^4} + \frac{(x-m)^4}{s^6} \right)$$

Note that for second order derivatives of the log-likelihood, we are normally interested in the expected derivatives, as by integrating out the data, we can ensure more stability when the Hessian matrix needs to be inverted in the sampling algorithm (see Eq. 4.B.1). If we then take the expectation of Eq. 4.A.1 with respect to the data $y$ and conditioned on the covariate $x$, we get

$$\mathrm{E}\left[ \frac{\partial^2 l}{\partial \theta^{c2}} \,\Big|\, x \right]_y = \mathrm{E}\left[ \frac{\partial^2 l}{\partial \nu^2} \left( \frac{\partial \nu}{\partial \theta^c} \right)^2 \,\Big|\, x \right]_y + \mathrm{E}\left[ \frac{\partial l}{\partial \nu} \frac{\partial^2 \nu}{\partial \theta^{c2}} \,\Big|\, x \right]_y.$$

The derivatives of the distribution parameters $\nu$ with respect to the model parameters $\theta^c$ are not directly dependent on the data $y$, yielding

$$\mathrm{E}\left[ \frac{\partial^2 l}{\partial \theta^{c2}} \,\Big|\, x \right]_y = \mathrm{E}\left[ \frac{\partial^2 l}{\partial \nu^2} \,\Big|\, x \right]_y \left( \frac{\partial \nu}{\partial \theta^c} \right)^2 + \mathrm{E}\left[ \frac{\partial l}{\partial \nu} \,\Big|\, x \right]_y \frac{\partial^2 \nu}{\partial \theta^{c2}}$$

$$= \mathrm{E}\left[ \frac{\partial^2 l}{\partial \nu^2} \,\Big|\, x \right]_y \left( \frac{\partial \nu}{\partial \theta^c} \right)^2,$$

as by definition $\mathrm{E}\left[\frac{\partial l}{\partial \nu} \mid x\right]_y = 0$.

The same principle applies to Eq.4.A.2, where the second term cancels out, hence avoiding the need to compute the second order cross derivatives $\frac{\partial^2 \nu}{\partial \theta_1^c \partial \theta_2^c}$.

Note that despite this simplification, none of the cross-derivatives are zero, meaning that the formulation parameters $\theta^c$ are not orthogonal to each other.

# 4.B Simplified mMALA approach for the Gaussian kernel BARBaR models

Let us look at the simplified mMALA as a type of MH algorithm, and assume the standard implementation and the acceptance probability calculations for non-symmetric proposals hold.

To implement mMALA, proposals $p$ are generated from a Normal distribution. Once we have specified a metric tensor matrix G, given a current state $\theta^c$ we can sample a proposal $\theta^{c*}$ from $N(a(\theta^{c*}), \Sigma(\theta^{c*}))$, where

$$a(\theta^{c*}) = \theta^c - \frac{\epsilon^2}{2} G^{-1}(\theta^c) \frac{\partial}{\partial \theta^c} (L + L_{\mathrm{prior}})$$

$$\Sigma(\theta^{c*}) = \epsilon^2 G^{-1}(\theta^c).$$

There are different options for the tensor matrix $G$, and we choose to use the expectation of the Hessian, denoted as $\bar{H}$, which allows to include stable second order derivative information.

Now, as noted at the end of section 4.A, the formulation parameters are not orthogonal. To fully account for the correlation amongst them, we then should make

a full joint proposal for all the parameters together.

Let us use a case with two different kernels, denoted as 1 and 2, to fully expand the algorithm above. Take a model parameter $\nu(x)$ and assume a BARBaR formulation with two components with unknown parameters be $\beta_1, \beta_2, m_1, m_2, s_1, s_2$. Then, to obtain a joint update, Eq. 4.B.1 becomes

$$a(\theta^{c*}) = \begin{bmatrix} \beta_1 \\ \beta_2 \\ m_1 \\ m_2 \\ s_1 \\ s_2 \end{bmatrix} - \frac{\epsilon^2}{2} G^{-1}(\theta^c) \nabla (L + L_{\text{prior}})$$

$$\Sigma(\theta^{c*}) = \epsilon^2 G^{-1}(\theta^c),$$

where

$$\nabla(\cdot) = \begin{bmatrix} \frac{\partial}{\partial \beta_1}(\cdot) \\ \frac{\partial}{\partial \beta_2}(\cdot) \\ \frac{\partial}{\partial m_1}(\cdot) \\ \frac{\partial}{\partial m_2}(\cdot) \\ \frac{\partial}{\partial s_1}(\cdot) \\ \frac{\partial}{\partial s_2}(\cdot) \end{bmatrix}$$

and

$$
G = \begin{bmatrix} \bar{H}_{\beta^2} & \bar{H}_{\beta,m} & \bar{H}_{\beta s} \\ \hdashline \bar{H}_{\beta,m}^T & \bar{H}_{m^2} & \bar{H}_{ms} \\ \hdashline \bar{H}_{\beta s}^T & \bar{H}_{ms}^T & \bar{H}_{s^2} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 \nu}{\partial \beta_1^2} & \frac{\partial^2 \nu}{\partial \beta_1 \partial \beta_2} & \frac{\partial^2 \nu}{\partial \beta_1 \partial m_1} & \frac{\partial^2 \nu}{\partial \beta_1 \partial m_2} & \frac{\partial^2 \nu}{\partial \beta_1 \partial s_1} & \frac{\partial^2 \nu}{\partial \beta_1 \partial s_2} \\ \frac{\partial^2 \nu}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 \nu}{\partial \beta_2^2} & \frac{\partial^2 \nu}{\partial \beta_2 \partial m_1} & \frac{\partial^2 \nu}{\partial \beta_2 \partial m_2} & \frac{\partial^2 \nu}{\partial \beta_2 \partial s_1} & \frac{\partial^2 \nu}{\partial \beta_2 \partial s_2} \\ \hdashline \frac{\partial^2 \nu}{\partial m_1 \partial \beta_1} & \frac{\partial^2 \nu}{\partial m_1 \partial \beta_2} & \frac{\partial^2 \nu}{\partial m_1^2} & \frac{\partial^2 \nu}{\partial m_1 \partial m_2} & \frac{\partial^2 \nu}{\partial m_1 \partial s_1} & \frac{\partial^2 \nu}{\partial m_1 \partial s_2} \\ \frac{\partial^2 \nu}{\partial m_2 \partial \beta_1} & \frac{\partial^2 \nu}{\partial m_2 \partial \beta_2} & \frac{\partial^2 \nu}{\partial m_2 \partial m_1} & \frac{\partial^2 \nu}{\partial m_2^2} & \frac{\partial^2 \nu}{\partial m_2 \partial s_1} & \frac{\partial^2 \nu}{\partial m_2 \partial s_2} \\ \hdashline \frac{\partial^2 \nu}{\partial s_1 \partial \beta_1} & \frac{\partial^2 \nu}{\partial s_1 \partial \beta_2} & \frac{\partial^2 \nu}{\partial s_1 \partial m_1} & \frac{\partial^2 \nu}{\partial s_1 \partial m_2} & \frac{\partial^2 \nu}{\partial s_1^2} & \frac{\partial^2 \nu}{\partial s_1 \partial s_2} \\ \frac{\partial^2 \nu}{\partial s_2 \partial \beta_1} & \frac{\partial^2 \nu}{\partial s_2 \partial \beta_2} & \frac{\partial^2 \nu}{\partial s_2 \partial m_1} & \frac{\partial^2 \nu}{\partial s_2 \partial m_2} & \frac{\partial^2 \nu}{\partial s_2 \partial s_1} & \frac{\partial^2 \nu}{\partial s_2^2} \end{bmatrix}
$$

# 4.C  Derivatives of the generalised Pareto distribution

The log likelihood of the observed data under the generalised Pareto distribution is

$$
\pi\left(y|\nu,\xi\right) = \begin{cases} \sum_{i=1}^{N} \left[ -\log\left(\frac{\nu_i}{1+\xi_i}\right) - \left(\frac{1}{\xi_i}+1\right) \log\left(1+\frac{\xi_i}{\nu_i}(1+\xi_i)y_i\right) \right] & \text{for } \xi_i \neq 0 \\[2ex] \sum_{i=1}^{N} \left[ -\log\left(\frac{\nu_i}{1+\xi_i}\right) - \frac{(1+\xi_i)y_i}{\nu_i} \right] & \text{for } \xi_i = 0 \ . \end{cases}
$$

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

$$
\nabla_{\beta_\theta}\pi\left(y|\nu,\xi\right) = \left(\nabla_{\beta_\theta}(B_\theta\beta_\theta)\right)^{\mathrm{T}}\left(\nabla_\theta\pi\left(y|\nu,\xi\right)\right)
$$

$$
= B_\theta^{\mathrm{T}}\left(\nabla_\theta\pi\left(y|\nu,\xi\right)\right) \ .
$$

The components of $\nabla_\xi \pi (y|\nu, \xi)$ are computed as

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

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

$$
\frac{\partial}{\partial \nu_i} \pi (y) = \begin{cases} \frac{1}{\nu_i}\left(1 - \left(\frac{1}{\xi_i} + 1\right)\frac{G_i - 1}{G_i}\right) & \text{for } \xi_i \neq 0 \\[2mm] \frac{1}{\nu_i}\left(1 - \frac{G_i - 1}{\xi_i}\right) & \text{for } \xi_i = 0 \ . \end{cases}
$$

Differentiating $\nabla_{\beta_\theta} \pi (\beta_\theta)$ $(\theta = \xi, \nu)$ again gives the Hessian matrix

$$
\nabla_{\beta_\theta} \nabla_{\beta_\theta}^{\mathrm{T}} \pi (\beta_\theta) = \nabla_{\beta_\theta} \nabla_{\beta_\theta}^{\mathrm{T}} \pi (y|\nu, \xi) \ .
$$

Applying the chain rule

$$
\nabla_{\beta_\theta} \nabla_{\beta_\theta}^{\mathrm{T}} \pi (y|\nu, \xi) = B_\theta^{\mathrm{T}} \left( \nabla_\theta \nabla_\theta^{\mathrm{T}} \pi (y|\nu, \xi) \right) B_\theta \ .
$$

Note that the components of $\nabla_\theta \pi (y|\nu, \xi)$ and $(\nabla_\theta \nabla_\theta^{\mathrm{T}} \pi (y|\nu, \xi))$ are computed separately for $\theta = \xi$ and $\theta = \nu$. Further, the expected values of the second derivatives of the likelihood with respect to $\xi$ and $\nu$ are

$$
-\mathbb{E}_Y \left[ \frac{\partial^2}{\partial \xi_i \partial \xi_j} \pi (y|\nu, \xi) \right] = \begin{cases} \frac{1}{(1+\xi_i)^2} & \text{for } i = j \\[2mm] 0 & \text{for } i \neq j \end{cases}
$$

and

$$-\mathbb{E}_Y \left[ \frac{\partial^2}{\partial \nu_i \partial \nu_j} \pi\left(y|\nu, \xi\right) \right] = \begin{cases} \frac{1}{\nu^2(1+2\xi_i)} & \text{for } i = j \\ \\ 0 & \text{for } i \neq j \end{cases}$$

such that Hessian matrices are diagonal. Moreover, the expectations of all of the cross derivatives $\frac{\partial^2}{\partial \xi_i \partial \nu_j} \pi\left(y|\nu, \xi\right)$ are zero, since estimates of $\xi$ and $\nu$ are asymptotically independent by construction (e.g. Chavez-Demoulin and Davison, 2005).

# Chapter 5

# Two-dimensional covariate modelling

## 5.1 Introduction

Metocean and coastal engineers are often tasked with the estimation of the extreme conditions that marine structures are likely to experience in their lifetime, which need to be considered for both design and maintenance purposes. Extreme value analysis provides an ideal framework for this type of inference. However, the characteristics of extreme ocean environments tend to vary systematically with known covariates such as storm direction, location and water depth. Then, capturing the effects of underlying physical phenomena is essential to obtain a reliable and trustworthy analysis. This requires building models which can include such covariates, as well as statistical tools for analysis and extrapolation.

In Chapter 4, we focused on modelling non-stationary series by including a single
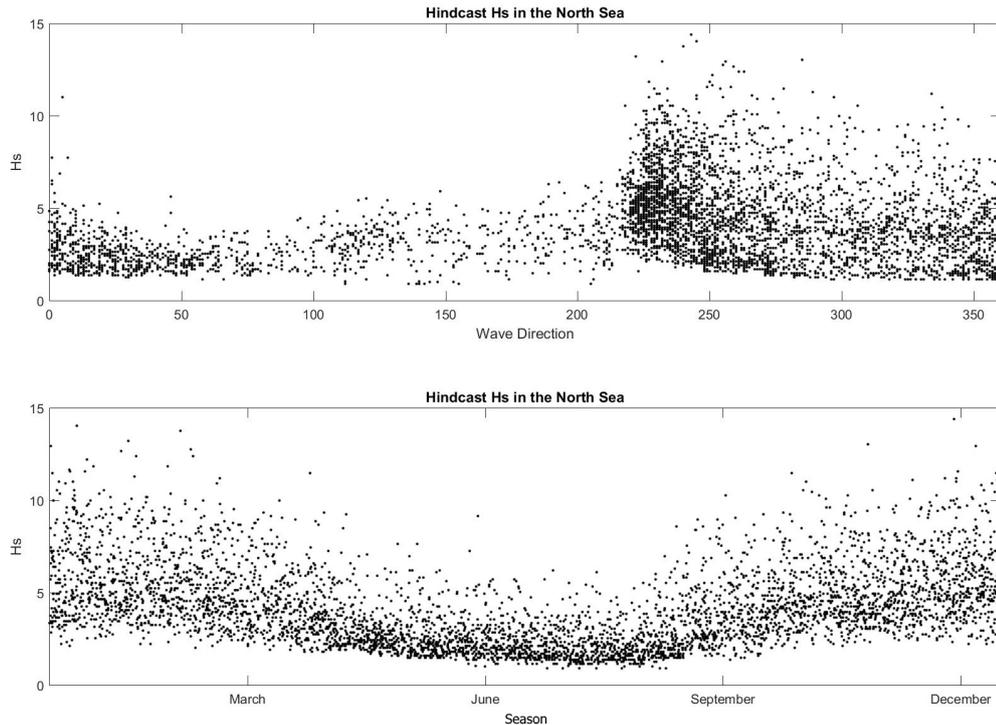
Figure 5.1.1: Storm peak significant wave heigh Hs hindcast (in meters) for a location in the North Sea, with corresponding wave direction (in degrees, with 0 corresponding to waves approaching from the North, and degrees increasing clockwise) and season (in day of the year, for a year consisting of 360 *seasonal days*).

covariate in the analysis. Nevertheless, it is often the case that multiple covariates simultaneously contribute to the characteristics of extreme observations. This is especially common for environmental datasets where a variety of climate processes are present. For example, Figure 5.1.1 shows how the hindcast dataset used in Chapter 4 exhibits both a directional and a seasonal pattern. Given the importance of having a statistical model that can appropriately capture the non-stationarity in a series, especially for extreme observations, clearly multiple covariates should be included in the analysis if more than one covariate has a significant effect on the response variable. For the remainder of this chapter, we focus on extreme significant wave heights, which, as mentioned in Section 4.1, usually exhibit a variety of covariate effects. We concentrate here on the observed variability with

respect to season and wave direction, although some remarks on extensions to higher dimensions can be found in Section 5.6.

The presence of a second additional covariate introduces considerable complications, from a modelling, computational and prediction perspective. If one was to use a simple linear regression model, the introduction of additional covariates would be straightforward, with a single extra term and corresponding unknown regression coefficient. Nevertheless, this approach assumes the covariate effects modelled are independent and of a linear nature, and this model, which was already inadequate for one-dimensional covariates, is incapable of capturing nonlinearities and interaction between multiple ones. As reviewed in Section 4.1, a variety of parametric and non-parametric formulations provide more flexible approaches, each with their own advantages and disadvantages. Generalised additive models would allow the introduction of multiple covariates with nonlinear behaviours, although they would neglect to model the interaction between them. Hence, extending the models from Section 4.1 to a two-dimensional covariate domain is not straightforward.

First of all, one needs to produce a model which is able to properly capture the effect of one or more covariate on the response. Scalability of the model becomes an essential factor, as the computational cost involved increases considerably when a new covariate is included.

Let us, for example, consider the P-spline formulation. As detailed further in Section 5.2.2, a standard approach consists of obtaining the model basis matrix as a tensor product of bases in each covariate dimension. The resulting basis has now $n_1 \times n_2$ knots, with corresponding unknown coefficients, where $n_1$ and $n_2$ refer to

the number of knots for the basis of the first and the second covariate respectively.

The total number of unknown parameters, then, increases rapidly, as each knot

added, say, for the first covariate, involves the addition of $n_2$ knots in the total

joint basis.  This effect can be seen in Figure 5.1.2.  Then, even starting from

two fairly low dimensional bases, say with 10 knots for each covariate space, the

resulting joint basis would have 100 knots with corresponding highly correlated

unknown coefficients to be estimated.
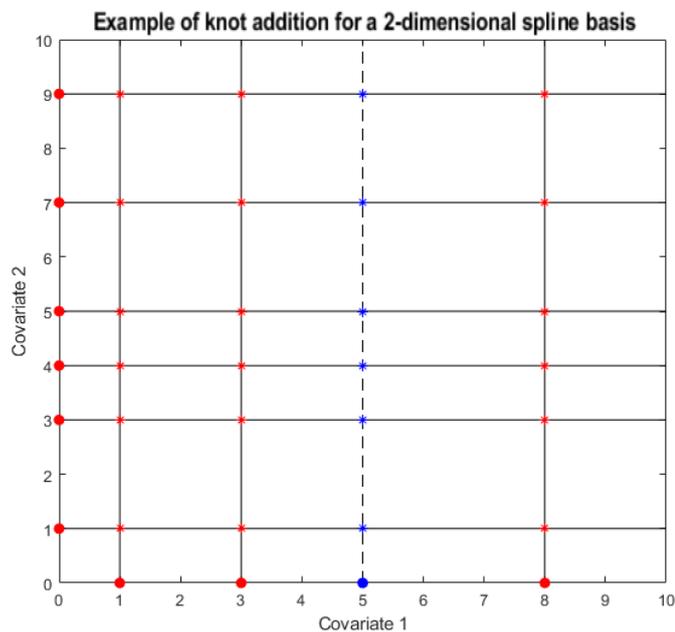


Figure 5.1.2: Example of a knots locations for a 2-D P-spline basis obtained with
the approach from Section 5.2.2. The knot locations for each individual covariate
domain are shown as •, while ∗ indicates the resulting knots on the 2-D domain.
Here we show the effect of adding a knot, denoted as a blue • in the first covariate
domain, requiring the addition of all the knots denoted by blue ∗ in the final 2-D
basis.

This issue affects both smoothing and regression analysis, and arises even when

only one covariate is present. Let us consider a one-dimensional smoothing appli-

cation on the motorcycle crash helmet impact data from Silverman (1985), as this

has often been used in the spline literature for illustration purposes (Eilers and

Marx, 1996). Smoothing consists in finding an estimate of some smooth function $f(x)$ modelled such that

$$Y|X = x \ \sim f(x) + \epsilon \ = B(x)\boldsymbol{\beta} + \epsilon,$$

for some response $Y_1, \ldots, Y_n$ with corresponding covariates $X_1, \ldots, X_n$, some basis matrix $B(x)$ and some random noise $\epsilon \sim N(0, \sigma^2)$. This corresponds, for penalised splines, to finding explicit solutions $\boldsymbol{\beta}$ to the linear system $(B'B + \rho Q)\boldsymbol{\beta} = B'\boldsymbol{y}$, for the spline basis $B = B(x)$, penalty matrix $Q$ and roughness coefficient $\rho$. In practice, the model covariance

$$(B'B + \rho Q)^{-1} \tag{5.1.1}$$

then needs to be computed. Inverting large matrices is complicated and can incur numerical instabilities especially if the matrix for inversion is rank deficient. To assess this stability, one can compute the condition number for a matrix inverse to measure the worst-case sensitivity to small perturbations (Higham, 2002). Larger condition numbers indicate "ill-conditioned" matrices, where relatively small changes in the input matrix can cause large changes in the solution to the system of equations. It is often the case that ill-conditioned matrices stem from overdetermined systems with highly correlated components. Hence, a different way to investigate this issue consists of considering the correlation between terms in the inverse matrix.

We then decide to examine both of these factors for a series of basis with an increasing number of spline components $p$: for each we compute the conditioning

number as well as correlation matrix, visualised using the `imagesc` function in
MATLAB, for the inverse in Eq. 5.1.1. In all cases, a value for the roughness
coefficient term $\rho$ was obtained using cross-validation to maximise the predictive
performance of the model. We can then proceed to consider the inverse matrices,
defined as in Eq.5.1.1, for a range of basis components, where in each case we con-
sider the inverse from the optimum smoothing fit. As we can see in Figure 5.1.3,
bases with more components require higher values of roughness penalty in order
to contain overfitting. As a roughness penalty controls the influence that adjacent
components have on each other, we notice that when a stronger smoothness is
imposed, more components have high correlation values.
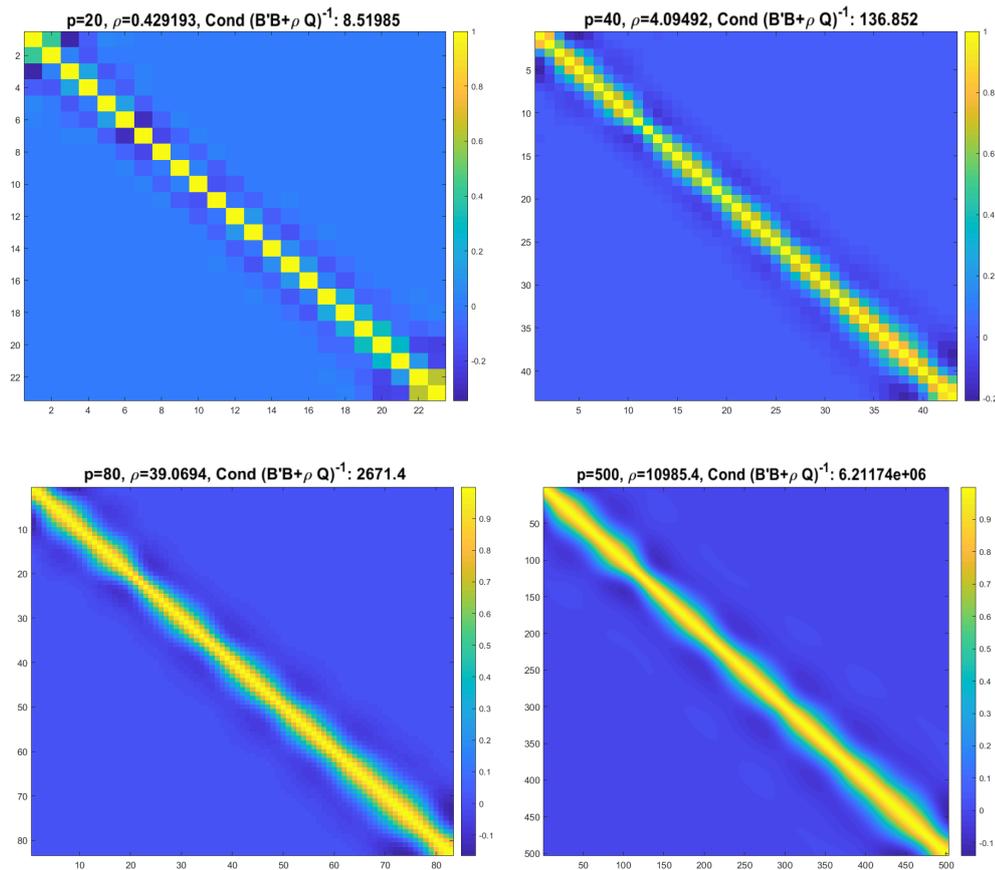


Figure 5.1.3: Smoothing was performed for the motorcycle crash helmet impact
data from Silverman (1985). Correlation values and conditioning numbers are
shown for the inverse matrix, defined as in Eq.5.1.1, for bases with (left to right,
top to bottom) $p = 20, 40, 80, 500$ spline components.

For example, by simple visual inspection, a correlation value larger than 0.8 can be noticed between any component and at most two adjacent ones for the basis with 20 knots. This number increases to at least 4 in a basis with 80 components, and over 30 in the basis with 500 components. Similarly, the conditioning number for the inverse matrix increases noticeable with the number of knots. These issues only become more marked when more than one covariate dimension is considered, in part due to the increased number of knots required. For example, to produce a directional-seasonal analysis, Randell et al. (2015a) use 32 knots in direction (one every 11.25°) and 24 in season (approximately one per fortnight), for a total of 768 components. Furthermore, the correlation and instability issues highlighted above only worsen when we move from smoothing to regression. For example, inference for the generalised Pareto distributions is affected by inherent difficulties, such as the sparsity of the data and its complicated likelihood surface. In fact, since the latter is not convex, it is difficult for methods to converge to the global optimum solution rather than to a local one, with high correlation and instability further hindering the process. In the case of MCMC methods, these factors result in poor mixing and convergence of the chain, and are further complicated by the computation of the roughness parameters as part of the Bayesian approach.

These issues are potentially more marked as we are using a gridded approach to a two-dimensional covariate domain. In fact, while gridded representations are easy to codify, they might be an inefficient way of describing the underlying surface. Other models, which are constructed to work specifically on 2-D spaces, might be more efficient and stable. These include, for example, the use of 2-D radial basis functions to extend the models from Chapter 4 or a partition approach such

as a Voronoi tessellation. Partition models have become increasingly popular for function estimation given their ability to detect spatial discontinuity (Knorr-Held and Raßer, 2000) and their inherent parsimony. This nonlinear approach makes very few assumptions about the underlying covariate structure and allows the data to dictate the nature of the partition, by providing enough flexibility to capture features of the data produced by the observed covariates. This flexibility might provide a further advantage with respect to a different important issue: the ability of any of these models to fully capture more "local" behaviour and avoid over-simplifying the relationship between the covariates and the response. It is clear from an oceanographic perspective that partition models for covariates are likely to provide parsimonious representations of covariate effects in some cases. Such an advantage might become even more marked for higher-dimensional covariates. Although these are beyond the scope of the work developed in this thesis, scalability of models to higher dimensions is one of the issues motivating this research and is worth considering even in the case of 2-D covariates.

Partition models have been used in a variety of applications, including image analysis (Green, 1995), cluster detection (Knorr-Held and Raßer, 2000), disease mapping (Denison and Holmes, 2001), and modelling of spatial NHPP intensities using a Voronoi tessellation (Heikkinen and Arjas, 1998, 1999). In general, these models have found a natural application in the modelling of spatial behaviour, and we refer the reader to the monograph by Okabe et al. (2009) for further details and application examples of Voronoi tessellation. In this work, we propose using a partition based on Voronoi tesselation to capture the covariate-dependent behaviour in the parameters of the statistical distributions fitted.

Each of these models has their own limitations, often in relation to computational efficiency and ability to obtain an appropriate level of smoothness. The objective of our work is to propose a flexible and parsimonious approach for a two-dimensional covariate. With this aim, we propose extensions of the models from Chapter 4, as well as introduce new ones. In particular, we extend the use of radial basis functions to periodic covariates in a 2-D domain. Relevant adaptations required for this application are detailed in Section 5.2. Following the same approach as for RBFs in Chapter 4, we also consider a Voronoi partition model to obtain piecewise constant estimates for the parameters of the Poisson and generalised Pareto distribution (GPD) fitted. These techniques are used to propose some new formulations and relevant implementations, as detailed in Section 5.2. To our knowledge, there have been no previous studies applying Voronoi tessellation to extreme value analysis and the characterisation of model parameters.

The outline of the chapter is as follows. Section 5.2 introduces the different model parameterisations, while in Section 5.3 we present the final models built upon them and summarise the inference approach and settings. Section 5.4 introduces the case studies used, which were created with the aim of reproducing realistic covariate-dependent behaviour. We briefly describe underlying model forms used to generate samples for inference and we analyse model performance by considering parameter estimates, return value distributions and statistical tests that compare the latter to the known "true" values. In Section 5.5, we apply the methodology to the hindcast data from the Northern North Sea analysed in Chapter 4 as well as from a location in the South China Sea neighbouring the site considered in Chapter 3. Finally, Section 5.6 provides a discussion of results, with conclusions

of the study and prospects for further research.

## 5.2 Non-stationary extremes

Suppose we are interested in a process $\{Y_t\}$ with two-dimensional covariate vector $\boldsymbol{X}_t \in \mathbb{R}^2$. For example, the response variable may be significant wave height, with wave direction and day of the year to account for seasonality as the covariates. We can then proceed to incorporate the covariates in the same manner as in Section 4.2.1, by formulating the Poisson rate and the GPD parameters as functions of the covariate of interest, when in this case the covariate is two-dimensional. Note that for the rate of exceedance, we can use the same "binning" approach as in Section 4.2.1. In this case, we focus on the joint covariate domain and divide it into $k_b$ small subsets, such that each covariate pair $\{x_{1,t}, x_{2,t}\} \in \mathcal{B}_i$ if $x_{1,t} \in \mathcal{B}_{1,i}$, $x_{2,t} \in \mathcal{B}_{2,i}$, where $\mathcal{B}_{1,i}$ and $\mathcal{B}_{2,i}$ are individual bins for $x_{1,t}$ and $x_{2,t}$ respectively and $\mathcal{B}_i = \mathcal{B}_{1,i} \times \mathcal{B}_{2,i}$. Then, we can define $N_i^u$ as the number of threshold exceedances $Y_t | Y_t > u, \mathbf{X}_t \in \mathcal{B}_i$, where $\mathcal{B}_i$ is the $i^{th}$ bin, and we can model it as $N_i^u \sim \text{Poiss}(\phi(\mathbf{x}_t))$ for $i = 1, \ldots, k_b$ and $k_b$ is the total number of combined bins $\mathcal{B}_i$.

### 5.2.1 Covariate parameterisation

We consider three different forms of basis function in this chapter, corresponding to P-spline (Brezger and Lang, 2006; Eilers and Marx, 2010), and two novel approaches, a Voronoi tessellation formulation, and a 2-D BARBaR-f model using bivariate Von-Mises kernels. Further details and an overview of each formulation

are presented below.

## 5.2.2 Spline basis

There are different spline formulations one could use to represent a bivariate co-variate space (Ruppert et al., 2003). The most natural extension of the models presented in Section 4.2.2 involves expressing the parameter $\theta(\mathbf{x})$ in terms of an appropriate basis for the joint domain $\mathbb{D}$ of the covariates, where $\mathbb{D} = \mathbb{D}_1 \times \mathbb{D}_2$, and hence creating an appropriate P-spline basis for each of the covariates consid-ered. In our case, since we are interested in a seasonal covariate (day of the year) and a directional one (wave direction), we adopt two bases of periodic B-splines. These are evaluated at each of the $n$ observations, yielding an $n \times p_i$ basis matrix, where $p_1$ and $p_2$ represent the number of basis functions in the respective basis. Note that one could choose to first divide the covariate domain into $k_b$ bins so as to reduce the dimension of the spline basis. Then we can define a basis matrix for the two-dimensional domain $\mathbb{D}$ as a Kronecker product, denoted as $\otimes$, of the marginal basis matrices (Eilers and Marx, 1996), such that

$$B_\theta(\mathbf{x}) = B_\theta(x_1) \otimes B_\theta(x_2),$$

where $B_\theta(\mathbf{x})$ is an $k_b \times p$ matrix, $p = p_1 p_2$. Alternatively, one can use the original undivided domain, and obtain the combined basis as

$$B_\theta(\mathbf{x}) = \left(B_\theta(x_2) \otimes \mathbf{1}_{p_1}\right) \odot \left(\mathbf{1}_{p_2} \otimes B_\theta(x_2)\right),$$

where $B_\theta(\mathbf{x})$ is an $n \times p$ matrix, $p = p_1 p_2$, $\odot$ denotes the element-by-element multiplication of two matrices and $\mathbf{1}_{\tilde{p}} = (1, \ldots, 1)$ is a $1 \times \tilde{p}$ vector of ones, for $\tilde{p} = \{p_1, p_2\}$ (Eilers et al., 2006). In both cases, the model parameter $\theta(\mathbf{x})$ can then be expressed as in Eq. 2.3.2 for some $p \times 1$ vector of basis coefficients $\boldsymbol{\beta_\theta}$. As in previous chapters, we opt to use the original undivided domain in order to avoid potential information loss as well as issues with the arbitrariness of the binning procedure.

In order to implement the P-spline approach, we also need a roughness penalty matrix. Recall from Section 2.3.4 that, in the case of a one dimensional covariate, we can use a difference matrix $D_\theta$ to define a penalty matrix $Q_\theta = D'_\theta D_\theta$. For higher dimensions, we proceed by first obtaining the difference matrix for each of the marginal bases, in this case $D_{\theta,1}$ for $B_\theta(x_1)$ and $D_{\theta,2}$ for $B_\theta(x_2)$, where $D_{\theta,1}$ and $D_{\theta,2}$ are $n \times p_1$ and $n \times p_2$ matrices respectively. We can then construct each penalty matrix as

$$P_{\theta,1} = D_{\theta,1} \otimes I_{p_1},$$

$$P_{\theta,2} = I_{p_2} \otimes D_{\theta,2},$$

where $I_p$ denotes the $p \times p$ identity matrix, and the joint smoothness penalty is the Kronecker sum

$$P_\theta = \rho_{\theta,1} P_{\theta,1} + \rho_{\theta,2} P_{\theta,2} = \rho_{\theta,1} \left( D_{\theta,1} \otimes I_{p_1} \right) + \rho_{\theta,2} \left( I_{p_2} \otimes D_{\theta,2} \right), \quad (5.2.1)$$

for some roughness coefficients $\rho_{\theta,1}, \rho_{\theta,2} \geqslant 0$. It is worth noticing that this approach assumes a constant smoothness across each covariate domain, while the models

we propose allow varied degrees of smoothness to arise in different parts of the covariates domain.

### 5.2.3 Voronoi basis

We consider using a Voronoi tessellation of the covariate space as a basis. This approach results in a partitioning of the covariate space by selecting some points known as sites, so that for each site there is a corresponding sub-region consisting of all covariates that are closer to that site than to any other (Berg et al., 2008). Let $\mathbf{x} \in \mathbb{D}$, where $\mathbb{D}$ is some metric space where we can define some distance metric $\boldsymbol{R}$. We can then select $K$ sites $\boldsymbol{v_j}, j = 1, \ldots, K$, and define the corresponding $K$ regions $\mathbb{D}_j$ such that the region $\mathbb{D}_k$ is the set of all points $\mathbf{x} \in \mathbb{D}$ whose distance to $\boldsymbol{v_k}$ is not greater than their distance to any of the other sites $\boldsymbol{v_j}$, where $j \neq k$. In practice, we first need to define some distance metric $\boldsymbol{R}(\mathbf{x}, \tilde{D}) = \inf\{\boldsymbol{R}(\mathbf{x}, \boldsymbol{v}) \mid \boldsymbol{v} \in \tilde{\mathbb{D}}\}$ between the point $\mathbf{x}$ and the subset $\tilde{\mathbb{D}}$. Then, the Voronoi cell $\mathbb{D}_k$ of $\boldsymbol{v_k}$ is

$$\mathbb{D}_k = \{\mathbf{x} \in \mathbb{D} \mid \boldsymbol{R}x, \mathbb{D}_k) \leqslant \boldsymbol{R}(\mathbf{x}, \mathbb{D}_j) \text{ for all } j \neq k\},$$

where all the regions $\mathbb{D}_j$ are assumed to be disjoint.
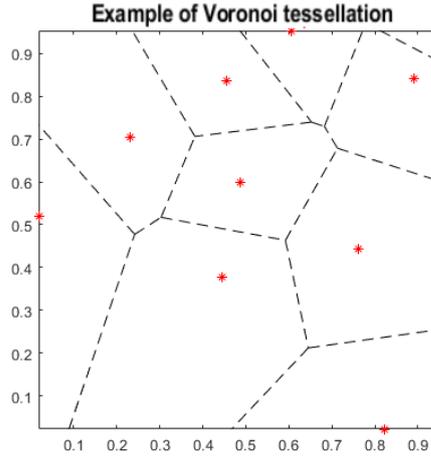
Figure 5.2.1: Example of a Voronoi tessellation of a $[0,1]$ covariate space, with cell edges shown by the dashed line and sites by the red $*$.

Figure 5.2.1 shows an example of such a partition , where for illustration purposes we choose $[0,1]$ as a covariate space. Here, we consider the case where $\mathbb{D}$ is a Euclidean plane, and we have a finite set of sites $\boldsymbol{v_1}, \ldots, \boldsymbol{v_K}$. In this case each site $\boldsymbol{v_k}$ is simply a point, and its corresponding Voronoi cell $\mathbb{D}_k$ consists of every point in the Euclidean plane whose distance to $\boldsymbol{v_k}$ is less than or equal to its distance to any other $\boldsymbol{v_j}$, $j \neq k$. Each such cell is a convex polygon, the line segments of the Voronoi diagram are all the points in the plane that are equidistant to the two nearest sites and the Voronoi vertices (nodes) are the points equidistant to three (or more) sites (Okabe et al., 2009).

Different choices for the distance metric $\boldsymbol{R}$ are possible, with a common example being the Euclidean distance, such that $\boldsymbol{R}(\mathbf{x}, \mathbb{D}_k) = \sqrt{(x_1 - v_{1,k})^2 + (x_2 - v_{2,k})^2}$. In case like ours, where one is interested in periodic covariates, a periodic distance formulation should be chosen. To do so, we can find the distance between two points on a plane in terms of their respective angle, since $x_1, x_2 \in [0, 2\pi]$, such

that

$$\boldsymbol{R}^*(\mathbf{x}, \mathbb{D}_k) = \arctan\left\{\frac{\sin(x_1 - v_{k,1})}{(x_1 - v_{k,1})}\right\} + \arctan\left\{\frac{\sin(x_2 - v_{k,2})}{\cos(x_2 - v_{k,2})}\right\}.$$

Following the same intuition as for the radial basis functions in Chapter 4, one can use a Voronoi tessellations to form a basis matrix $B_\theta(\mathbf{x})$, where each entry $b_{i,j}$ is just the distance contribution for component $j$ given some observation $\mathbf{x}$ such that

$$b_{i,j} = \boldsymbol{R}(\mathbf{x}_i, \mathbb{D}_j),$$

and a model parameter $\theta(\mathbf{x})$ can be expressed as in Eq. 2.3.2. Then, for some $p \times 1$ vector of basis coefficients $\boldsymbol{\beta_\theta}$, the resulting $\theta(\mathbf{x})$ is a piece-wise constant function on the covariates plane. We call this approach Voronoi tessellation basis regression (VTBR).

There are many algorithms to implement a Voronoi tessellation of the space, although they can be slow in implementation: we refer the reader to the monograph by Berg et al. (2008) for more details, as well as the surveys by Aurenhammer (1991) and Okabe et al. (2009). Many of these can be difficult to implement and hard to extend to higher dimensions (Reem, 2009). While boundaries may be difficult to compute, we are mostly interested in determining what cells a covariate value belong to. In our case, we aimed to find an approximation to the Voronoi partition that could be easily adapted to any desired combination of periodic and non-periodic domains, extendible to higher dimensions, easily computed and differentiable, in case we were interested in using derivative-based MCMC methods during inference. Hence we looked for a formulation that was similar, in mathe-

matical expression, to some known kernel densities, and whose behaviour similarly

captured the symmetrical decay of the density as points further from the site of

interest are considered. We also aimed to preserve the shape of the cells as well

as the option of having, as in an original Voronoi tessellation, each observation

contribute to a single cell only. First, we note that, by using an exponential func-

tion of a symmetrical distance metric, we can ensure the positivity of the resulting

outcome as well as obtain a function that is easily differentiable. Furthermore,

given that all points $\mathbf{x} \in [0, 2\pi)$ and the covariate domains are periodic, we note

that the distances $\boldsymbol{R}^*(\mathbf{x}, \mathbb{D}_k) \in [0, d^*]$, where $d^* = \sqrt{(\pi^2 + \pi^2)}$ is the maximum

distance between any two points on the space. Then, if we compute

$$\tilde{\boldsymbol{R}}(\mathbf{x}, \mathbb{D}_k) = \exp\left\{-\frac{\boldsymbol{R}^*(\mathbf{x}, \mathbb{D}_k)}{s}\right\}, \qquad (5.2.2)$$

for a small $s > 0$, e.g. $s = 10^{-4}$, $\tilde{\boldsymbol{R}}(\mathbf{x}, \mathbb{D}_k) \to 0$ for larger values of $\boldsymbol{R}^*(\mathbf{x}, \mathbb{D}_k)$ and

$\tilde{\boldsymbol{R}}^*(\mathbf{x}, \mathbb{D}_k) \to 1$ for smaller distances $\boldsymbol{R}^*(\mathbf{x}, \mathbb{D}_k)$. If for each point $\mathbf{x}$ we compute the

distance to each of the sites $\boldsymbol{v}_j$, $j = 1, \ldots, p$, we can then normalise each distance

by the sum of all the distances, that is

$$\boldsymbol{R}(\mathbf{x}, \mathbb{D}_k) = \frac{\tilde{\boldsymbol{R}}(\mathbf{x}, \mathbb{D}_k)}{\sum_{j=1}^{p} \tilde{\boldsymbol{R}}(\mathbf{x}, \mathbb{D}_j)}. \qquad (5.2.3)$$

Then, provided $s$ is small, e.g. $s < 10^{-3}$, all but the smallest distance will yield a

value $\boldsymbol{R}(\mathbf{x}, \mathbb{D}_k) = 0$, while for the closest site $\boldsymbol{v}_{k*}$, $\boldsymbol{R}(\mathbf{x}, \mathbb{D}_{k*}) = 1$. Note that larger

values for $s$ produce a smoother transition between subregions, where the metric

$\boldsymbol{R}(\mathbf{x}, \mathbb{D}_k)$ no longer yields binary 0-1 contributions. Instead, observations close to

the edges of a region will be the result of contributions of neighbouring regions,

yielding to smoother fuzzy edges.

We illustrate the effect of varying values of the decay parameter in Figure 5.2.2. Here we consider a random partition of a periodic domain into 6 cells, each with randomly chosen coefficients. While location and coefficients for each cell component remain unaltered, we investigate increasing values of the decay parameter.
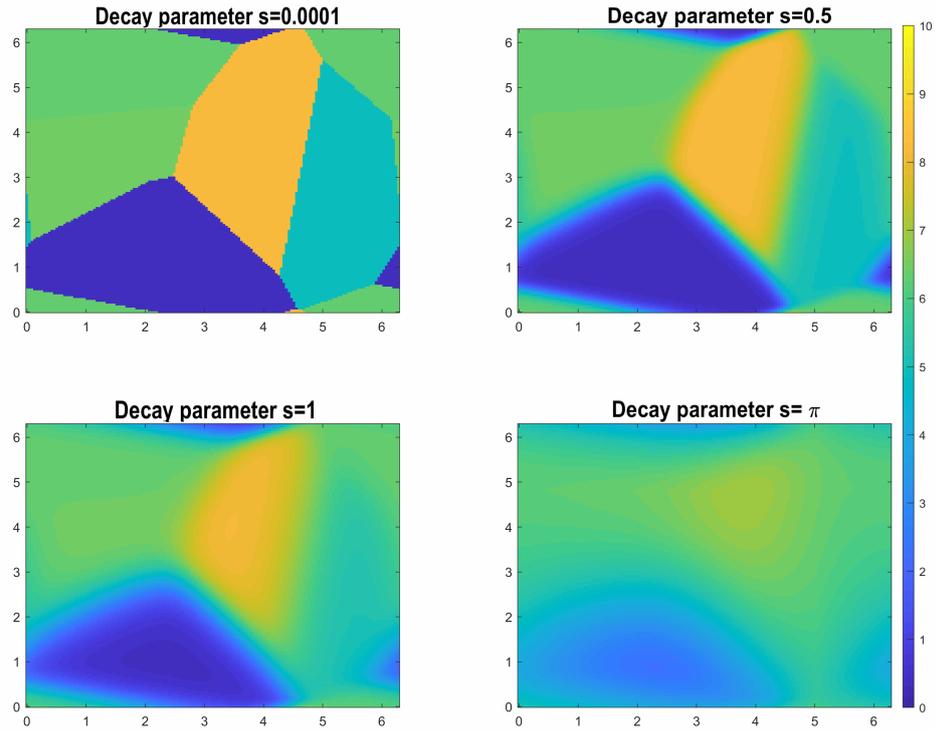


Figure 5.2.2: Partition of a periodic domain into 6 cells with fixed random locations and coefficients, and decay parameter (left to right, top to bottom) $s = 0.0001, 0.1, 1, \pi$.

For the work presented in this chapter, we obtain the tessellation for the VTBR model using the metric from Eq. 5.2.2, with a fixed very small value for the decay parameter. We also consider two further variations. For the first one, denoted as VTBR-smooth (VTBR-s), we let parameter estimates change smoothly between adjoining cell by allowing larger decay parameters. Finally, we consider the case

where contributions from all cells sum to one at all points by using the metric from Eq. 5.2.3: we denote the resulting model as VTBR-smooth-rescaled (VTBR-sr). The specific details for the models considered are presented in Section 5.3.1.

## 5.2.4 Bayesian adaptive radial basis functions in 2-D

We now consider an extension of the BARBaR-f formulation introduced in Section 4.2.2 for a two-dimensional covariate. As we have seen in Section 2.3.5, radial basis functions can be defined for spaces of dimensions higher than one when an appropriate symmetric distance function is used. We can model the bases for $\phi(\mathbf{x})$, $\xi(\mathbf{x})$ and $\nu(\mathbf{x})$ as vectors of $p$ kernels, where each of these kernels is defined on the two-dimensional covariate space, and represent a general parameter $\theta(x)$ as the linear combination of these kernels. For the case with two periodic covariates, we could have used an approximation to the bivariate wrapped normal, which would have been constructed similarly to Eq. 4.3.12. Nevertheless, following the same approach, 5 terms in each of the dimensions would have to be considered, for a total of 25 terms in the final computation. Due to the considerable increase in computational cost, we propose using the cosine variant of the bivariate Von-Mises kernels with density proportional to

$$f(x_1, x_2) \propto \exp\{\kappa_1(x_1 - m_1) + \kappa_2 \cos(x_2 - m_2) - \kappa_3 \cos(x_1 - m_1 - x_2 + m_2)\}$$

for covariates $x_1, x_2 \in [0, 2\pi]$, where $m_1, \kappa_1 > 0$ and $m_2, \kappa_2 > 0$ are the means and concentration for $x_1$ and $x_2$ respectively and $\kappa_3 < \frac{\kappa_1 \cdot \kappa_2}{\kappa_1 + \kappa_2}$ is related to their correlation. Figure 5.2.3 shows the effect of different values of the parameters

$\kappa_1, \kappa_2, \kappa_3$ for a single Von-Mises component with a fixed $\beta$ coefficient.
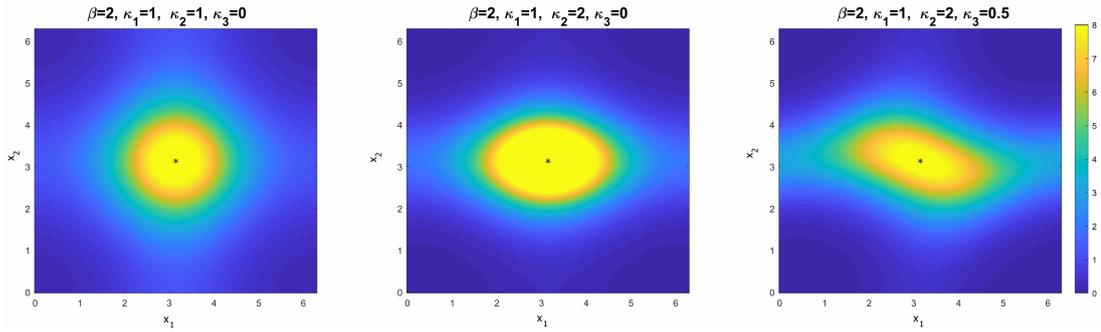


Figure 5.2.3: Single Von-Mises component, for $x_1, x_2 \in [0, 2\pi]$, with a fixed coefficient, for different values of the $\kappa_1, \kappa_2, \kappa_3$ parameters.

We refer the reader to the work by Mardia (2013) for a discussion on different variants of the bivariate Von Mises distribution.

## 5.3 Inference procedures and methods evaluation

In a similar way to Chapter 4, we use the generalised Pareto distribution (GPD) peaks over threshold (POT) model, as described in Section 4.2.1, and implement Bayesian methods for the estimation of the unknown parameters, following an approach broadly similar to Algorithm 2. In Section 5.3.1 we describe all the formulations used to represent the Poisson-GPD model parameters, while in Section 5.3.2 we illustrate the structure of our Bayesian model and we discuss the prior and posterior structures for different parameterisation approaches.

## 5.3.1   Investigated models

We used the formulations introduced in Section 5.2 to obtain the following models for inference:

Model 1 **P-spline** : penalised spline with 20 fixed knots in each covariate domain, for a total of 400 basis components. The knot locations were spaced evenly on the two-dimensional covariate domain. Note that this approach is used as a benchmark;

Model 2 **P-spline-l**: penalised spline with only 5 fixed knots unevenly spaced fixed knots in each covariate domain, assigned depending on the density of observations, with a total of 16 basis components;

Model 3 **VTBR**: Voronoi tessellation with 8 cells and sites initiated depending on the density of observations and updates as part of the inference;

Model 4 **VTBR-s**: Version of the Voronoi tessellation where the edge between cells is allowed to change smoothly, with the rate of this change depends on the unknown "decay" parameter $s$ in Eq. 5.2.2, and where each cell has a corresponding decay parameter $s_j$. The model is fitted using 6 cells, with sites initiated depending on the density of observations and the decay parameters initiated to some very small value, e.g. $s_j = 10^{-4}$ for all cells;

Model 5 **VTBR-sr** : similar to Model 4, we consider a tessellation with unknown

decay parameters. Each basis component is also rescaled by the sum of the

contributions from the other components: this insures that, at each given

observations, the contributions of all cells sum to 1, in a similar manner to

spline formulations,in an attempt to ensure stability in the computations.

The model is fitted using 6 cells, sites are initiated depending on the density

of observations and the decay parameters are initiated to some very small

value, e.g. $s_j = 10^{-4}$ for all cells;

Model 6 **BARBaR-f**: Bivariate Von Mises RBFs with 4 kernels, with locations initi-

ate again via a k-means algorithm and using $\Sigma_\theta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ as initial covariance

matrix for all model parameters;

Model 7 **BARBaR-fr**:  Bivariate Von Mises RBFs with 4 kernels, with locations

initiated depending on the density of observations and using $\Sigma_\theta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ as

initial covariance matrix for all model parameters.  Similarly to Model 5,

each basis component is also rescaled by the sum of the contributions from

the other components.  Given the additional rescaling step, which provides

a formulation similar to a B-spline basis, we decide to omit the baseline

coefficient.

Each approach scales differently with the number of components used for the

analysis.  Table 5.3.1 summarises the total number of unknown parameters that

need to be estimated during inference, hence providing some insight in model

| Model | Basis param. per comp. | No. hyperparam. | Tot. no. param. | Tot. no. param. in studies |
|:---:|:---:|:---:|:---:|:---:|
| P-spline | 1 | 2 | $\mathbf{p_1 p_2 + 2}$ | 402, for $p_1 = p_2 = 20$ |
| P-spline-l | 1 | 2 | $\mathbf{p_1 p_2 + 2}$ | 27, for $p = 5$ |
| VTBR | $1 + 2$ | 1 | $\mathbf{3p + 1}$ | 25, for $p = 8$ |
| VTBR-s | $1 + 2 + 1$ | 1 | $\mathbf{4p + 1}$ | 25, for $p = 6$ |
| VTBR-sr | $1 + 2 + 1$ | 1 | $\mathbf{4p + 1}$ | 25, for $p = 6$ |
| BARBaR-f | $1 + 2 + 3$ | 2 | $\mathbf{6p + 2}$ | 26, for $p = 4$ |
| BARBaR-fr | $1 + 2 + 3$ | 1 | $\mathbf{6p + 1}$ | 25, for $p = 4$ |

Table 5.3.1: Summary of the total number of unknown parameters for the models fitted in Sections 5.4 and 5.5. P-spline bases have $p_1$ and $p_2$ components for the two covariates domain respectively, and the other models have $p$ total components. The second column considers the basis formulation parameters required for each component listed by type, e.g. each BARBaR-f component is defined by 1 coefficient, 2 locations and 3 $\kappa$ parameters.

complexity. For both the simulation study and the hindcast analysis, we opted to choose the number of components for each approach in order to have a similar number of unknowns and model complexity. As we have pointed out in Section 5.1, the P-spline benchmark model requires a much larger number of parameters to be estimated.

## 5.3.2   Bayesian inference

Given the results obtained from one-dimensional models, in order to improve mixing we decide to introduce a hyperparameter $\lambda_\theta$ for the prior of the coefficients for all the non-spline based models, including a separate hyperparameter $\lambda_{0,\theta}$ for the baseline coefficient in the BARBaR-f models. Spline models, as before, benefit from the use of roughness coefficients $\rho_{\theta,1}, \rho_{\theta,2}$. Hence, following the approaches and notation from Sections 4.2.2 and 5.2, model fitting corresponds to estimating, for each of the distribution parameters, the unknown parameters:

- $\Omega_\theta = \{\boldsymbol{\beta_\theta}, \rho_{\theta,1}, \rho_{\theta,2}\}$ for both P-spline models;

- $\Omega_\theta = \{\boldsymbol{\beta_\theta}, \boldsymbol{v}_{\theta,1}, \boldsymbol{v}_{\theta,2},\}$ for the Voronoi tessellation VTBR model;

- $\Omega_\theta = \{\boldsymbol{\beta_\theta}, \boldsymbol{v}_{\theta,1}, \boldsymbol{v}_{\theta,2}, \boldsymbol{s}, \lambda_\theta\}$ for the adapted Voronoi tessellation models VTBR-s and VTBR-sr;

- $\Omega_\theta = \{\boldsymbol{\beta_\theta}, \boldsymbol{m}_{\theta,1}, \boldsymbol{m}_{\theta,1}, \boldsymbol{\kappa}_{\theta,1}, \boldsymbol{\kappa}_{\theta,2}, \boldsymbol{\kappa}_{\theta,3}, \lambda_{0,\theta}, \lambda_\theta\}$ for both the Von-Mises BARBaR-f models.

Recall from Section 4.3.1, we can express the joint posterior distribution for the Poisson-GPD model parameters in terms of the relevant conditional posterior distributions. For all the above models, we let $N_i^u$ be the number of threshold exceedances $Y_t | Y_t > u, X_t \in \mathcal{B}_i$, where $\mathcal{B}_i$ is the $i^{th}$ two-dimensional bin, and $N_i^u \sim \text{Poiss}\left(\phi\left(\mathbf{x}_t\right)\right)$ for $i = 1, \ldots, k_b$. These resulting independent threshold exceedances are

$$Y_t | Y_t > u \sim \text{GP}\left(\psi\left(\mathbf{x}_t\right), \xi\left(\mathbf{x}_t\right)\right),$$

and we use the notation $\theta\left(\mathbf{x}\right)$ to refer, more generally, to the Poisson rate, GPD scale and shape, as both will have the same basis model structure from Eq. 2.3.2.

Then for the P-spline models, we have

$$\boldsymbol{\beta_\theta} | \rho_{\theta,1}, \rho_{\theta,2} \approx \rho_\theta^{1/2} \exp\left(-\frac{\rho_\theta}{2} \boldsymbol{\beta_\theta}^{\text{T}} Q_\theta \boldsymbol{\beta_\theta}\right), \tag{5.3.1}$$

$$\rho_{\theta,1}, \rho_{\theta,2} \sim \text{Gamma}(10^{-3}, 10^{-3}),$$

where in this case $Q_\theta$ correspond to the matrix $P_\theta$ as defined in Eq. 5.2.1. This set up is similar to the case with a one-dimensional covariate, where now the

coefficients have a prior density proportional to the improper Gaussian density in Eq. 5.3.1, where $Q_\theta$ depends on penalty matrices on both covariates $P_{1,\theta}, P_{2,\theta}$ with corresponding roughness parameters $\rho_{\theta,1}, \rho_{\theta,2}$. These are given an uninformative Gamma prior distribution, which is conjugate with the prior Gaussian distribution for $\beta_\theta$.

For the Voronoi tessellation model, for each cell component $j = 1, \ldots, k$, we suggest the following prior distributions:

$$\beta_{\theta,j} | \lambda_\theta \sim \text{MVN}(0, \lambda_\theta I)$$

$$m_{\theta,j} \sim \text{Uniform}(d_l, d_r),$$

$$\gamma_\lambda \sim \text{Gamma}(g, h),$$

$$\lambda_\theta \sim \text{Gamma}(10^{-3}, 10^{-3}).$$

Here, $[d_l, d_r]$ denotes the range of the domain of the covariate and the coefficients $\beta_{\theta,j}$ are mutually independent of each other and of the other parameters, with a ridge-type penalty imposed by the normal prior and the corresponding smoothness hyperparameter is common across all the cell components. For the VTBR-s and VTBR-sr models, obtained from an adaptation of the Voronoi tessellation, we impose that the decay parameter $s \sim \text{Gamma}(g, h)$ to guarantee its positivity.

For the Von Mises BARBaR formulations, for each component $j = 1, \ldots, k$, we

suggest the following prior distributions:

$$\beta_{\theta,j}|\lambda_\theta \sim \text{MVN}(0, \lambda_\theta I)$$

$$m_{\theta,j} \sim \text{Uniform}(d_l, d_r),$$

$$\kappa_{\theta,1,j}, \kappa_{\theta,2,j} \sim \text{Gamma}(g, h),$$

$$\kappa_{\theta,3,j} \sim \text{N}(0, t),$$

$$\lambda_\theta \sim \text{Gamma}(10^{-3}, 10^{-3}).$$

Additionally, the BARBaR-f model formulation includes a baseline coefficient, for which we assume $\beta_{\theta,j}|\lambda_{0,\theta}l \sim \text{N}(0, \lambda_{0,\theta})$. In this case, a separate prior, and corresponding hyperparameter, are proposed for the baseline term, such that $\lambda_{0,\theta} \sim$ Gamma$(10^{-3}, 10^{-3})$. In fact, we expect the other coefficients to capture local variations, and the baseline to represent the average behaviour of the parameter being modelled. Hence, we can assume baseline and components' coefficients might have considerably different behaviours, so that separate priors would ensure better mixing of the MCMC and faster convergence.

### 5.3.3   MCMC inference algorithms

For inference, we follow the same general approach presented in Section 4.3.2 using a Metropolis-within-Gibbs MCMC algorithm. At each iteration of the MCMC, each of the parameters is sampled in turn conditionally on the values of others. The full conditional posterior distribution is Gamma by conjugacy for both the precision parameters $\rho_{\theta,1}, \rho_{\theta,2}$, in the spline models, and the penalty parameters

$\lambda_{0,\theta}, \lambda_{\theta}$, for the other models, and these parameters are sampled exactly in a Gibbs step. All the other parameters are estimated using a general Metropolis-Hastings (MH) scheme, since full conditional distributions are not available in closed form for their posteriors. Sampling is performed using the methods presented in Section 2.2.2, where the specific method depends on the basis model type. Here we consider the approaches for one of the statistical model parameters, which are applied to all model parameters in turns for each MCMC sweep.

**P-spline models**    Similarly to the work in Chapter 4, the coefficients of the spline basis are updated with a single Riemann manifold Metropolis-adjusted Langevin algorithm (mMALA) step. Then, a Gibbs step is used to update the two precision parameters $\rho_{\theta,1}, \rho_{\theta,2}$ given the current MCMC state. All knots locations are assumed to be fixed and remain unchanged during the inference.

**VTBR models**    For the Voronoi tessellation based model, we first update all the coefficient parameters using an mMALA step. Then, all sites locations, in each dimension, are considered one by one and updated using a simple RWMH, as this performed well in test in terms of mixing of the chain and convergence. Finally, the penalty parameter $\lambda_{\theta}$ is updated via Gibbs sampling. The same set up is used for the VTBR-s and the VTBR-sr models, although these include an additional stage before the update of the penalty parameter. This involves a joint RWMH step for all the decay parameters of the different cells. Again, the RW algorithm was preferred as it yielded good convergence and mixing, while resulting in lower computational costs than the mMALA.

**BARBaR models** The two models based on Von Mises kernels both follow the same approach. As for all other models, all the basis coefficient are updated via an mMALA step, while all location parameters are later updated, one at a time, with a RWMH step. Then, a single joint mMALA update is performed on the three covariance parameters for all the components. In this case, we opt for a gradient based method to reduce the number of MH steps to be performed and improve the MCMC mixing and convergence, as a RW approach tested poorly in this case. Then, using Gibbs sampling, the penalty parameters are updated. Note only one penalty parameter is updated for the BARBaR-fr model, which does not have a baseline term, while two penalty parameters are present and updated for the BARBaR-f formulation.

### 5.3.4 Return value distributions and comparison methods

To assess performance we first focus on comparison of the posterior estimates of the statistical model parameters. We then follow the same methodology introduced in Section 4.4.1 to obtain distributions of return values corresponding to long return periods. To apply the procedure summarised in Algorithm 3 when two covariates are present, it is sufficient to adapt the way we define bins, by dividing the joint covariate domain into $n_b$ small subsets $S_j$, $j = 1, \ldots, n_b$. Then each covariate pair $\{x_{1,t}, x_{2,t}\} \in S_j$ if $x_{1,t} \in \mathcal{S}_{1,J}$, $x_{2,t} \in \mathcal{S}_{2,j}$, where $\mathcal{S}_{1,J}$ and $\mathcal{S}_{2,j}$ are individual bins for $x_{1,t}$ and $x_{2,t}$ respectively and $\mathcal{S}_j = \mathcal{S}_{1,j} \times \mathcal{S}_{2,j}$.

We consider estimates of "omni-covariates" values, which in this case means concurrently omni-directional and omni-seasonal. We also inspect the model perfor-

mance by directional octant as well as for 12 seasonal sectors corresponding to 30-day periods.

For the case study in Section 5.4, we then compare all these estimated distributions with the corresponding values obtained from the original parameters used to simulate the datasets. As part of this final assessment, we also use the statistical tools described in Section 4.4.1, namely the Kolmogorov-Smirnov (KS) and the Kullback-Leibler (KL) criterion, to quantify the discrepancy between estimated and the "true" underlying distributions.

## 5.4 Case studies

We follow the same approach to simulate data as the one used by Jones et al. (2016) for the samples used in Chapter 4. We consider three cases, each with 100 samples of approximately 2000 observations, and assume they correspond to an observational period of $\mathcal{T}$ years. All the samples were constructed to reflect some features that we are likely to observe in hindcast data from various ocean basins (Randell et al., 2014). Furthermore, they cover a range of behaviours with increasingly complicated underlying models for the Poisson-GPD parameters. For example, in Case 1 the Poisson parameter is assumed to be constant, while in Case 2 and Case 3, we define $\phi(\mathbf{x}) = \tilde{\phi}(\mathbf{x}) \times 2000/c_\rho$, where $\tilde{\phi}(\mathbf{x})$ is defined below and we let $c_\rho = \int_0^{360} \tilde{\phi}(\mathbf{x})d\mathbf{x}$ in order to simulate 2000 observations.

**Case 1:** For extreme value threshold $u(\mathbf{x}) = 0$, we simulate 2000 observations with a uniform Poisson rate $\phi(\mathbf{x}) = 2000/360$ per degree covariate, and a low order

Fourier parameterisation of GPD shape $\xi(\mathbf{x}) = -0.2 + \frac{1}{10}\sin\left(x_1 - \frac{\pi}{6}\right)\sin(x_2)$ and

scale $\psi(\mathbf{x}) = 2.5 + \frac{1}{2}(\sin(x_1) + \cos(x_1 - 0.5) + \cos(x_1))$.

**Case 2:** For extreme value threshold $u(\mathbf{x}) = 0$ and the same Fourier param-

eterisation of GPD shape and scale as in Case 1, a non-uniform Poisson rate

$\phi(\mathbf{x}) = \frac{1}{2}\max\left(1.3 + 2\sin(x_1) + \sin(x_2) + 1.6, 0\right) \times 2000/c_\rho$.

**Case 3:** For extreme value threshold $u(\mathbf{x}) = 0$, a non-uniform Poisson rate

$$
\begin{aligned}
\phi(\mathbf{x}) \;=\; & 0.1 + 5\mathring{F}_2\left(\boldsymbol{x}, \begin{bmatrix} 0.5\pi \\ 1.6\pi \end{bmatrix}, \Sigma_{\rho,1}\right) + 30\mathring{F}_2\left(\boldsymbol{x}, \begin{bmatrix} 0.7\pi \\ 3.5 \end{bmatrix}, \Sigma_{\rho,2}\right) \\
& + \; 10\mathring{F}_2\left(\boldsymbol{x}, \begin{bmatrix} 1.5\pi \\ 0.3\pi \end{bmatrix}, \Sigma_{\rho,3}\right) \times \frac{2000}{c_\rho},
\end{aligned}
$$

where

$$
\Sigma_{\rho,1} = \begin{bmatrix} 0.3 & -0.15 \\ -0.15 & 0.35 \end{bmatrix}, \qquad \Sigma_{\rho,2} = \begin{bmatrix} 2.5 & 0 \\ 0 & 0.7 \end{bmatrix}, \qquad \Sigma_{\rho,3} = \begin{bmatrix} 1.5 & 0.3 \\ 0.3 & 0.4 \end{bmatrix},
$$

$$
\xi(\mathbf{x}) = -0.3 + 1.2\mathring{F}_2\left(\boldsymbol{x}, \begin{bmatrix} 0.6\pi \\ 3.5 \end{bmatrix}, \Sigma_{\xi,1}\right) + 2.2\mathring{F}_2\left(\boldsymbol{x}, \begin{bmatrix} 1.4\pi \\ 0.2\pi \end{bmatrix}, \Sigma_{\xi,2}\right)
$$

where

$$
\Sigma_{\xi,1} = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.5 \end{bmatrix}, \qquad \Sigma_{\xi,2} = \begin{bmatrix} 2 & 0.3 \\ 0.3 & 2.2 \end{bmatrix},
$$

and

$$\psi\left(\mathbf{x}\right) = 1.8 + 1.2\mathring{F}_2\left(\boldsymbol{x}, \begin{bmatrix} 0.3\pi \\ 1.4\pi \end{bmatrix}, \Sigma_{\psi,1}\right), \qquad \text{where } \Sigma_{\psi,1} = \begin{bmatrix} 1.5 & 0 \\ 0 & 1 \end{bmatrix}.$$

Here, we define $\mathring{F}_2(\cdot)$ to an approximation of a bivariate wrapped normal density, such that

$$\mathring{F}_2\left(\boldsymbol{x}, \boldsymbol{m}, \Sigma\right) = \frac{1}{2\pi\sqrt{|\Sigma|}} \sum_{j=-4\pi}^{6\pi} \sum_{k=-4\pi}^{6\pi} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{m} - \begin{bmatrix} 2\pi j \\ 2\pi k \end{bmatrix}\right)\Sigma^{-1}\left(\boldsymbol{x} - \boldsymbol{m} - \begin{bmatrix} 2\pi j \\ 2\pi k \end{bmatrix}\right)'\right\}.$$
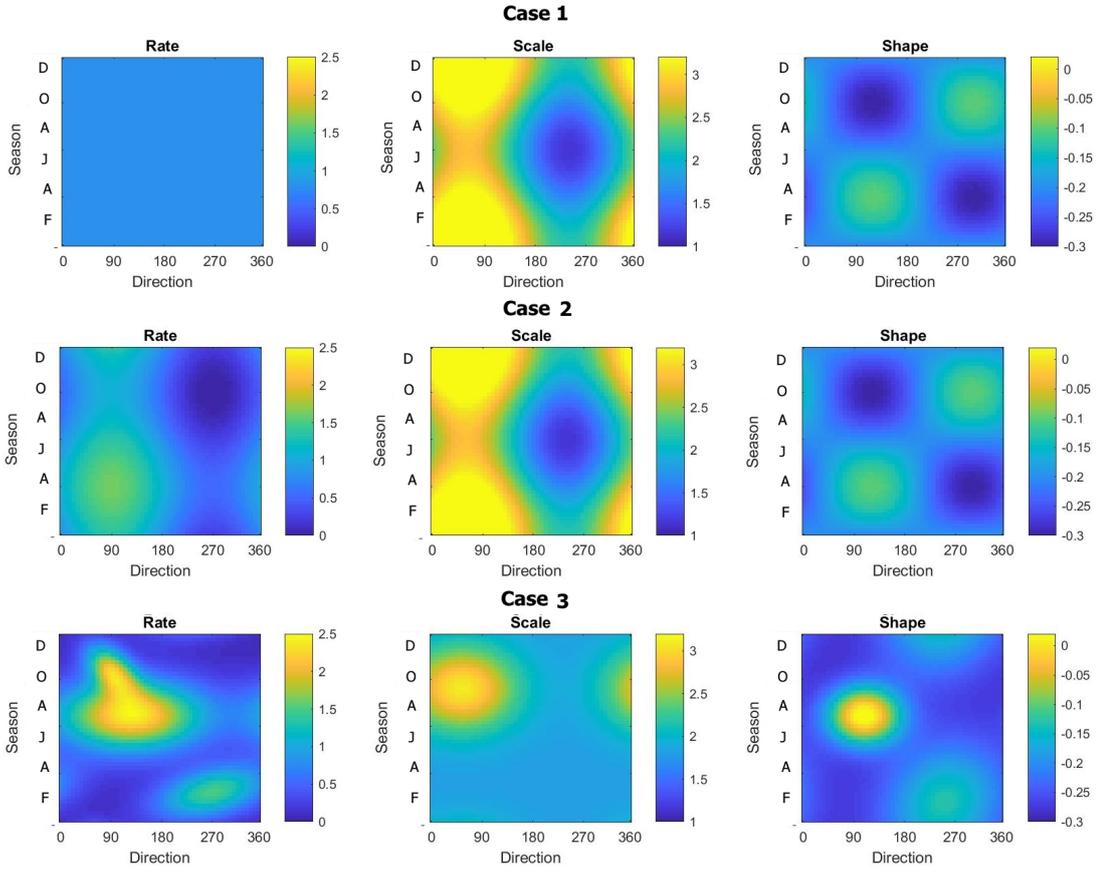


Figure 5.4.1: Parameter values used to simulate datasets for Case 1 (top), Case 2 (middle) and Case 3 (bottom). From left to right, panels show parameter variation of GPD shape $\xi\left(\mathbf{x}\right)$, scale $\psi\left(\mathbf{x}\right)$ and Poisson rate $\phi\left(\mathbf{x}\right)$ with direction and season for each case.

Note that, for plotting and analysis purposes, we will refer to the response as significant wave height Hs, and to the two covariate as "wave direction" (or direction)

and "season" for $x_1$ and $x_2$ respectively.

Figure 5.4.1 shows the resulting parameter variations with respect to the two co-variates for all 3 cases. Note that the parameter variation of the GPD shape $\xi(\mathbf{x})$ and scale $\psi(\mathbf{x})$ with direction $x$ are identical in Cases 1 and 2. The main and only difference between them lies in the Poisson rate $\phi(\mathbf{x})$, which is constant in Case 1 only. In Cases 2, the Poisson rate is assumed to vary with respect to both covariates, has lowest values in the same directional sector where the scale parameter is also low. It is worth noticing that the shape $\xi(\mathbf{x})$ is negative throughout. Case 3 exhibits more complicated behaviour in the Poisson rate, while the shape parameter includes some positive values in a small part of the covariates domain.
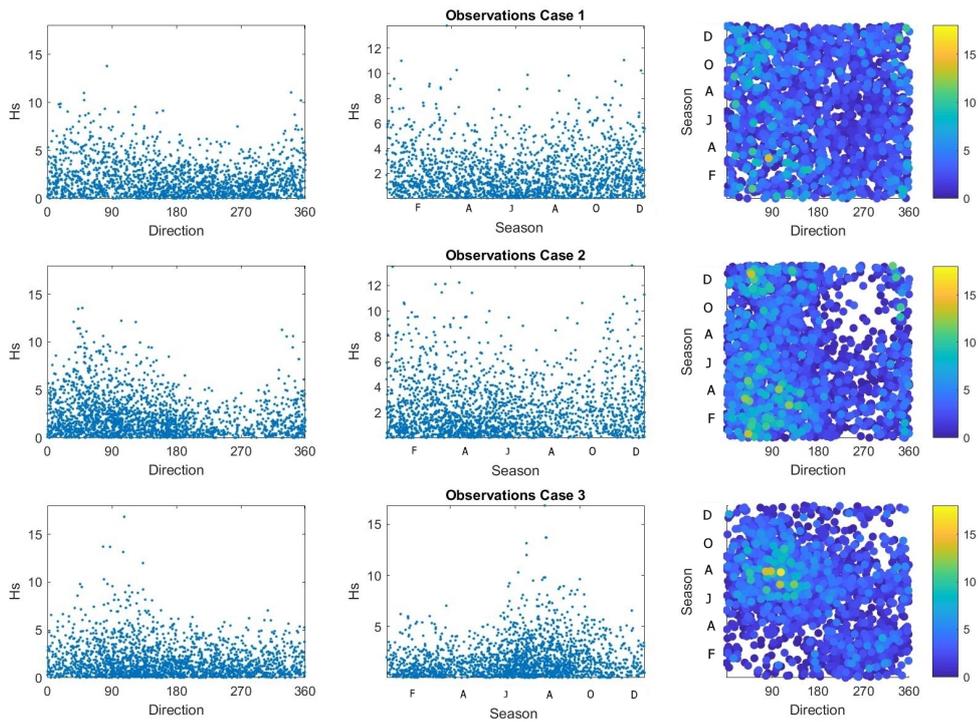


Figure 5.4.2: Illustrations of a sample realisation from each of Cases 1 (top), Case 2 (middle) and Case 3 (bottom). From left to right, panels show the response Hs values against direction, season and with respect to both covariates.

Figure 5.4.2 illustrates a typical sample realisations for each of the three cases. It is possible to notice the difference in size and variation of the exceedances, as

well as the subsets of the domain with very few observations, corresponding to the lowest values of the Poisson rate parameter.

### 5.4.1 Inference and tests

For each case, the extreme value threshold $u$ was fixed at zero during inference, while both the GPD shape $\xi$ and scale $\psi$ of exceedance size vary as a function of the covariate $x$. As before in Chapter 4, we use a constant model for the Poisson rate $\phi$ of threshold exceedance Case 1, while this is covariate dependent in Cases 2 and 3. We proceed to perform inference using the methodology and models described in Section 5.3. For all cases, we obtain parameter estimates based on MCMC run lengths of 60000 in total, with the first 30000 samples removed as burn-in. Only 1000 draws, one every 30, are stored, in order to circumvent issues with memory storage and the internal memory available for computations in Matlab.

After a first inspection, we notice a similar performance for models across the different cases. This is especially true for Case 1 and 2, where the GPD parameters, used for simulation, have the same form. Furthermore, although we move from a constant threshold exceedance rate in Case 1 to a covariate-dependent one in Case 2, we opted to propose a functional form which produced a limited difference in magnitude between the larger and smaller rate observed across the covariate domain. Hence, we decide to focus on Case 3 specifically, as this is the one characterised by the most complicated combined behaviour for all three underlying model parameters. We include results and a brief analysis for the first two cases in Appendix 5.A for completeness.
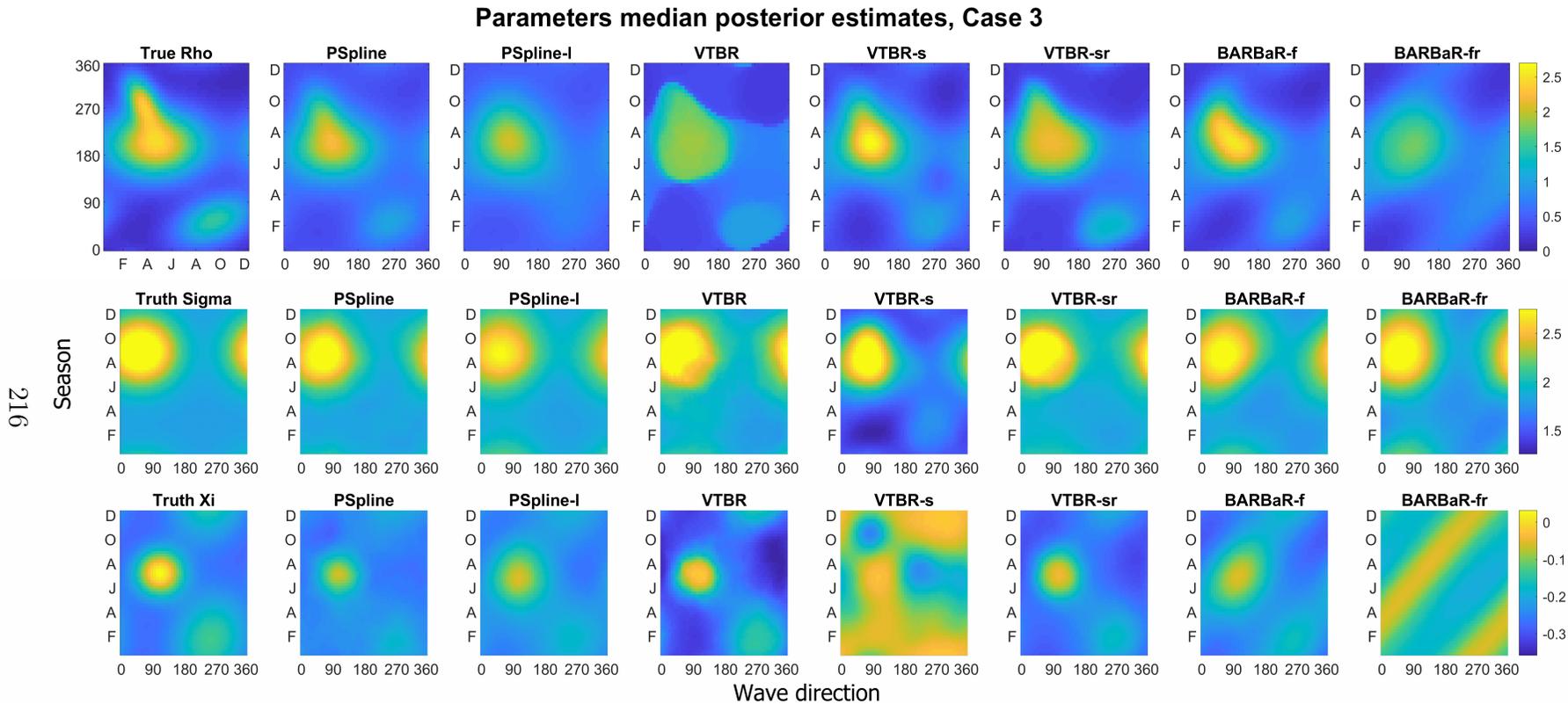
Figure 5.4.3: Parameter estimates for rate of occurrence $\phi(\mathbf{x})$ of the exceedances (upper), the GPD scale $\psi(\mathbf{x})$ (middle) and GPD shape $\xi(\mathbf{x})$ (lower) for simulated samples from Case 3, for Models 1-7. Each panel illustrates the posterior median estimate, while the leftmost panels show the true parameters used during the simulation.

**Interquantile ranges of parameters posterior estimates, Case 3**
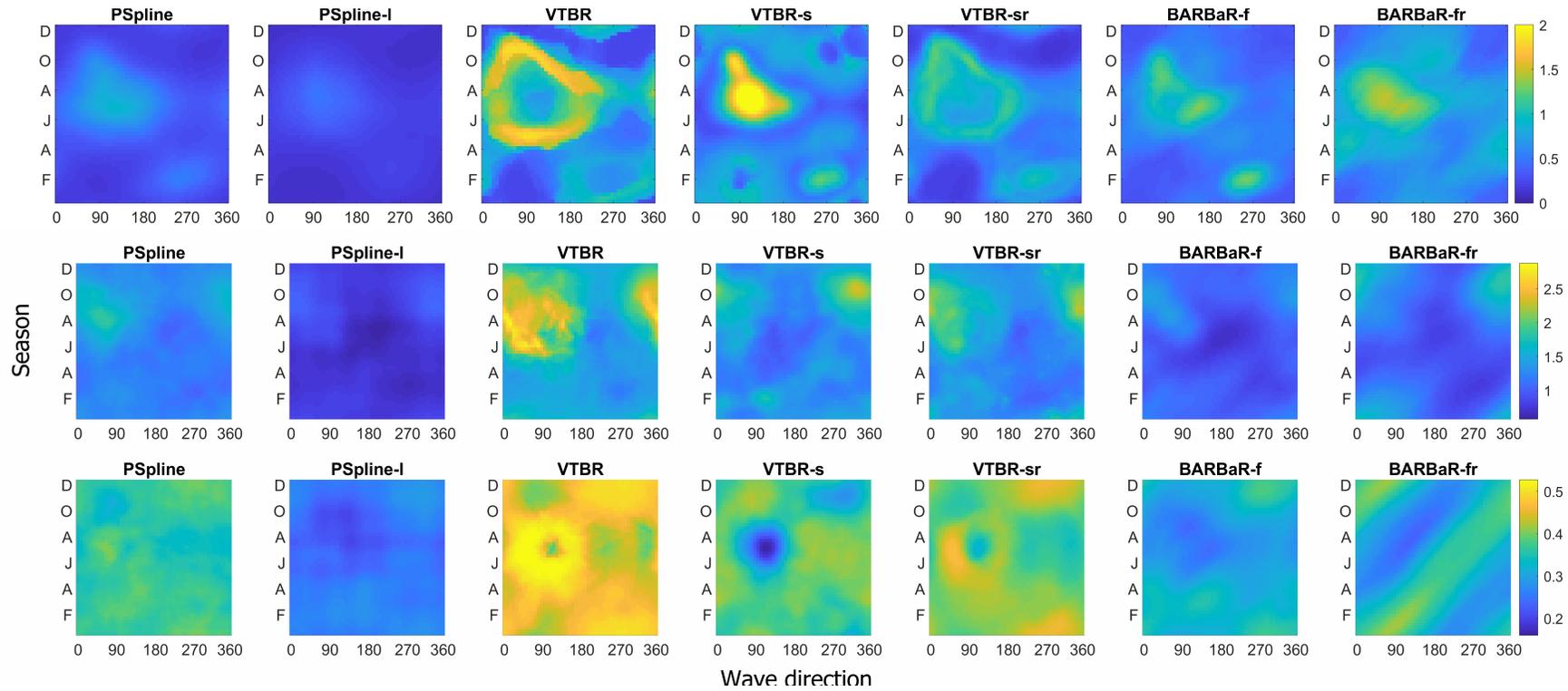
Figure 5.4.4: 95% interquantile ranges for the posterior estimates of the rate of occurrence $\phi\left(\mathbf{x}\right)$ of the exceedances (upper), the GPD scale $\psi\left(\mathbf{x}\right)$ (middle) and GPD shape $\xi\left(\mathbf{x}\right)$ (lower) for simulated samples from Case 3, for Models 1-7.

We first inspect the median parameter estimates over the sample realisations for the three model parameters across models 1-7, shown in Figure 5.4.3. It is also essential to consider the uncertainty that characterises each of the models, so we produce plots showing the range between the upper and lower 0.25 quantile, which are shown in Figure 5.4.4.

Overall, the models seem to be able to detect the underlying patters present in the three model parameters. It is worth noticing a couple of factors. First, despite detecting the correct shape for the function of the Poisson rate on the covariate space, all models obtain lower median estimates than the true values. A similar performance can be noticed across most models for the GPD shape parameter, while both the VTBR-s and the BARBaR-fr model struggle more and yield median estimates that are higher than the truth. These models also required more careful tuning and additional testing, although they all eventually showed good mixing of the chain.

When we consider the interquantile ranges for the estimates, all models seem to show similar pattern, although the models based on the Voronoi formulations show more uncertainty and overall wider interquantile ranges for both the shape and the scale parameter, as well as some localised potential overestimates in the Poisson rate. It is also worth noticing that the P-spline-l model yields the narrowest interquantile ranges out of all the formulations.
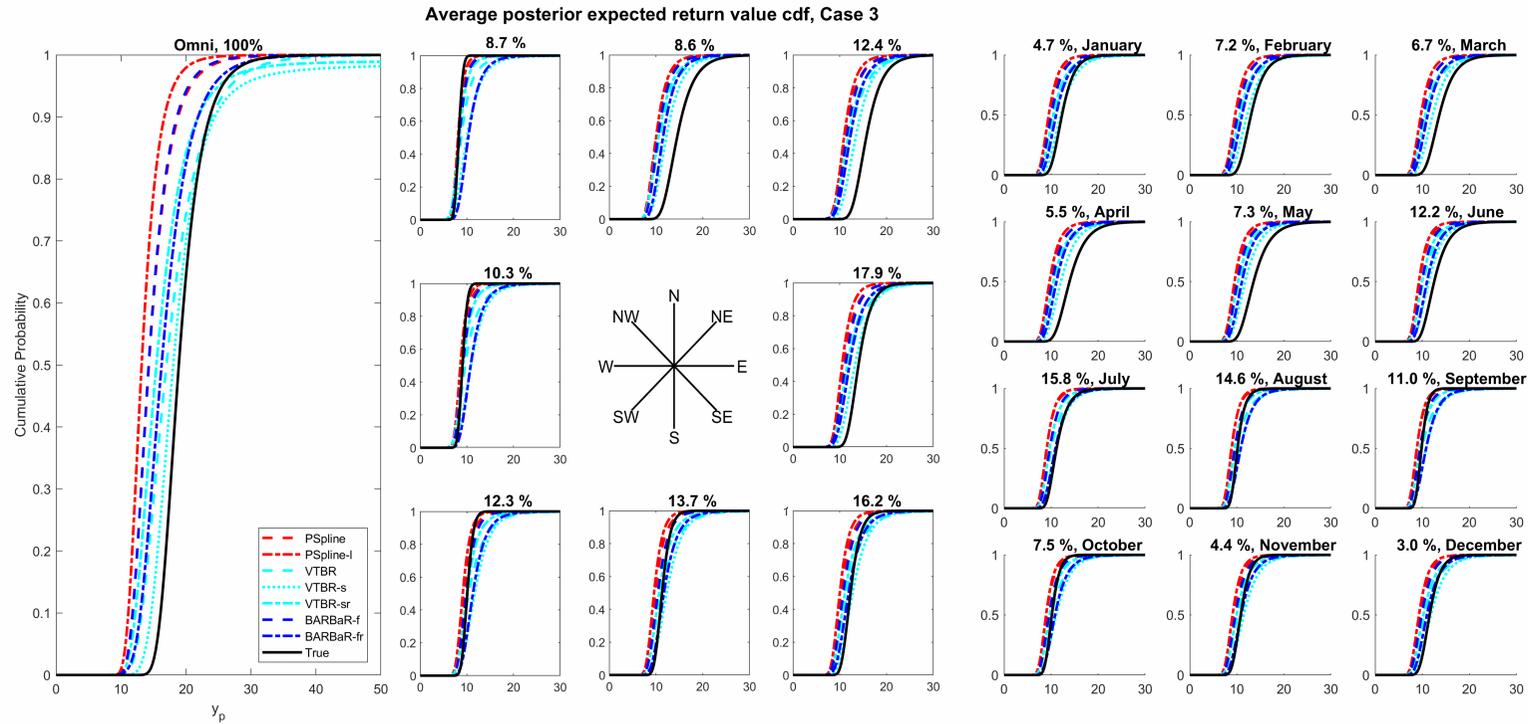
Figure 5.4.5: Average posterior expected return value cdf for simulated samples from Case 3, corresponding to a return period of ten times the period of the original sample. The panel show, from left to right, the omnidirectional return value distribution (left), the corresponding directional estimates (middle) and the seasonal estimates split by month (right). The title for each panel gives the expected percentage of individuals in that directional sector. The true return value distribution is given in solid black.

Since the scale and shape parameter of the GPD are negatively correlated, it is necessary to consider return values in order to properly assess inference performance of the models. We proceed to consider the posterior cumulative distribution functions (cdf) of return values for a return period of ten times the period of the original sample. These are shown in Figure 5.4.5. In terms of the omni-covariate cdf, all models yield lower return levels as compared to the estimates from the true underlying parameters. The Voronoi based models incur in some issues for higher probabilities, which are likely due to the wider interquantile ranges visible in the parameter estimates. The benchmark and the BARBaR-f model yield an almost identical omni-covariate cdf, although some small differences are noticeable in sectoral cdf's.

When considering directional and seasonal results, again we notice return levels lower than the truth are obtained, and all models seems to perform similarly, showing seasonal and directional variability.

We then employ the Kolmogorov-Smirnov (KS) and the Kullback-Leibler (KL) criterion, as defined in Section 4.4.3, to compare the distributions of return values statistically and consider how they perform across different realisation samples. Figures 5.4.6 and 5.4.7 summarise the characteristics of the distributions for KS and KL divergence criteria respectively. We note that all models perform similarly, especially in terms of maximum divergence, as shown by the KS test values. The results are consistent across models both for the omni-covariate estimates and in each of the directional and seasonal sub-sectors, with higher median values and wider ranges for sectors with fewer observations.
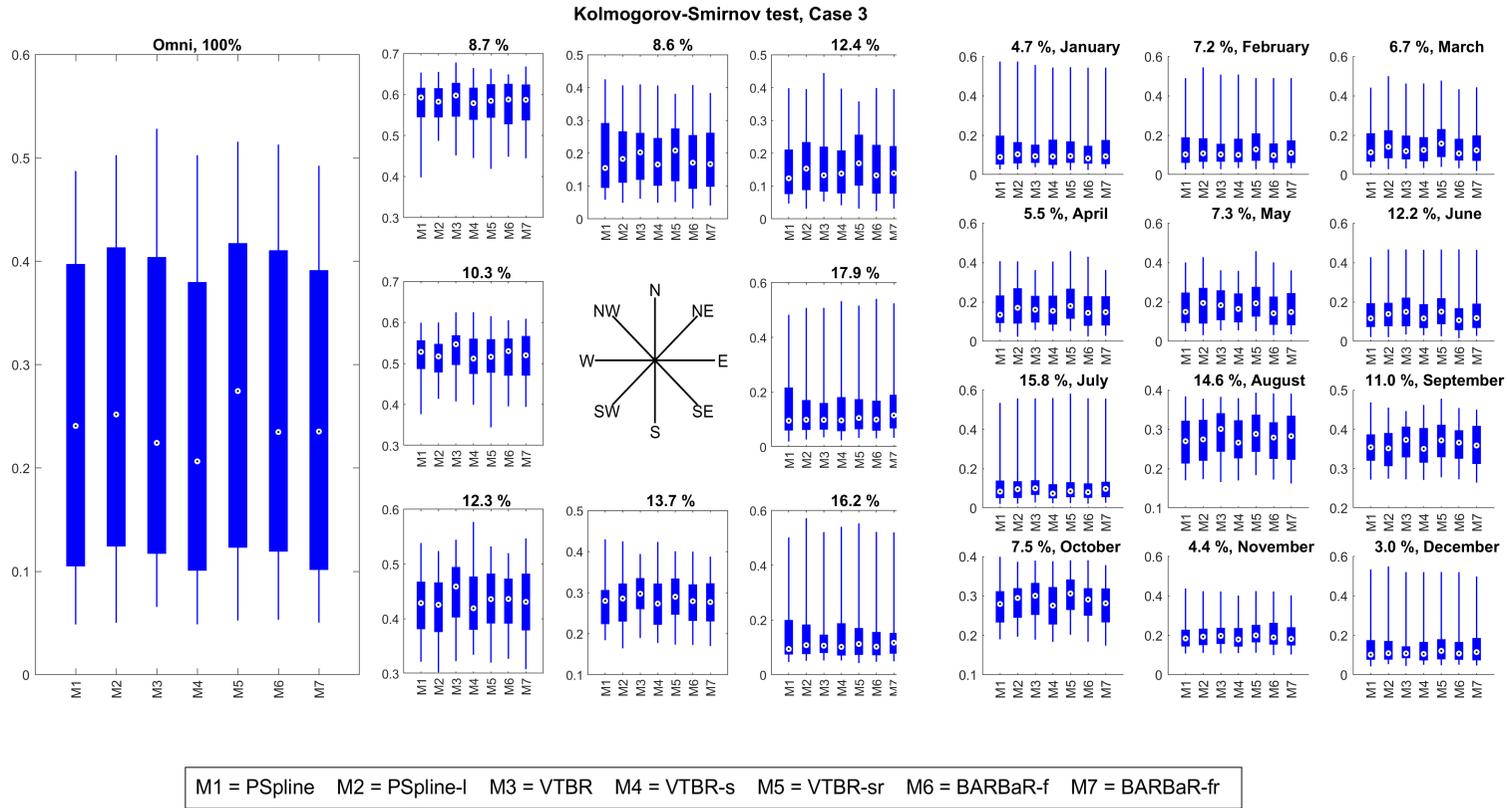
Figure 5.4.6: Box-whisker comparison of samples of the Kolmogorov-Smirnov (KS) divergence criterion between omnidirectional (left) posterior expected return value cdf's (corresponding to a return period of ten times that the original sample), by directional sector (middle) and seasonal monthly sector (right), estimated under samples from the true return value distribution and those estimated under models of each of 100 sample realisations for Case 3. The sample of the KS divergence criterion are summarised by the median (white disc with black central dot), the interquartile range (blue rectangular box, with vertical lines showing the 2.5%, 97.5% interval).
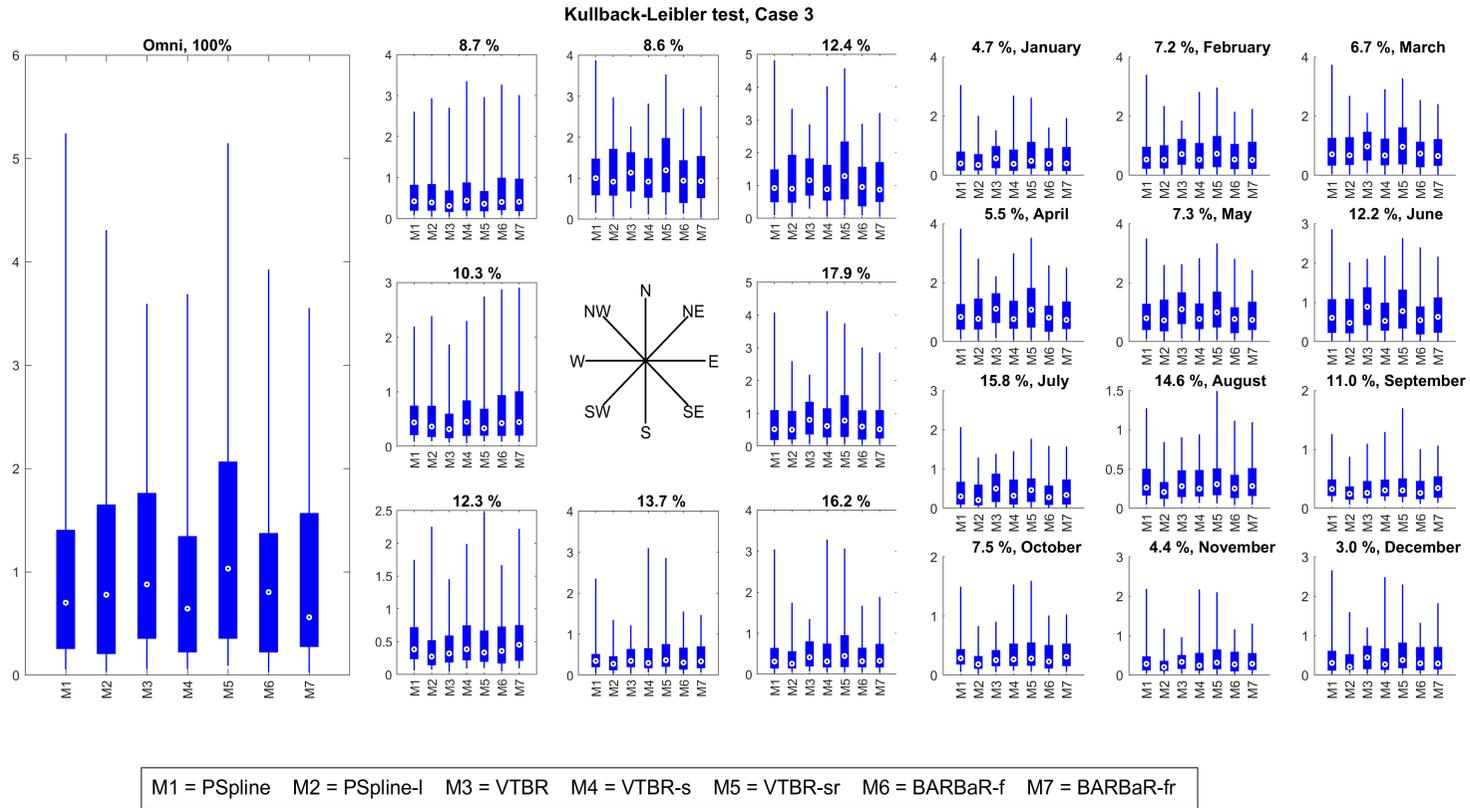
Figure 5.4.7: Same as Figure 5.4.6, here for the Kullback-Leibler (KL) criterion for Case 3.

Overall, no model seems to be performing consistently better than the others, and performance seems to match the one of the benchmark P-spline model. Some additional variation can be seen in the KL criterion estimates, which consider the total distance of the posterior expected cdf's from the true one. Here the VTBR-sr model yields the highest median omnidirectional divergence values and widest 75% interquantile ranges, with similar results from the directional and seasonal sectors. This is in agreement with the average expected posterior cdf results. No other model seems to show consistently better better performance for all directional, seasonal and omni-covariate tests.

## 5.5   Hindcast data study

While in Section 5.4 we tested and assessed the models on some simulated case studies, we now proceed to apply them to the datasets introduced in Section 4.1. In both cases, we apply the models to storm peak significant wave heights Hs, and include corresponding wave direction and season (given by day of the year) as covariates. Note that, unlike previous chapters, we decided to consider both datasets in this case. In fact, due to the distinct physical conditions that affect these distant ocean basins, observations exhibit different covariate dependent behaviours in these two locations. For both hindcasts, we consider the same models for parameter formulations as in Section 5.4 and apply the methodology described in Section 5.3 for inference. As for the case studies in Section 5.4, results are based on run lengths of 60000 run lengths, with the first 30000 samples removed as burn-in and a total of 1000 draws, one every 30, being stored after that.

## 5.5.1 Northern North Sea hindcast
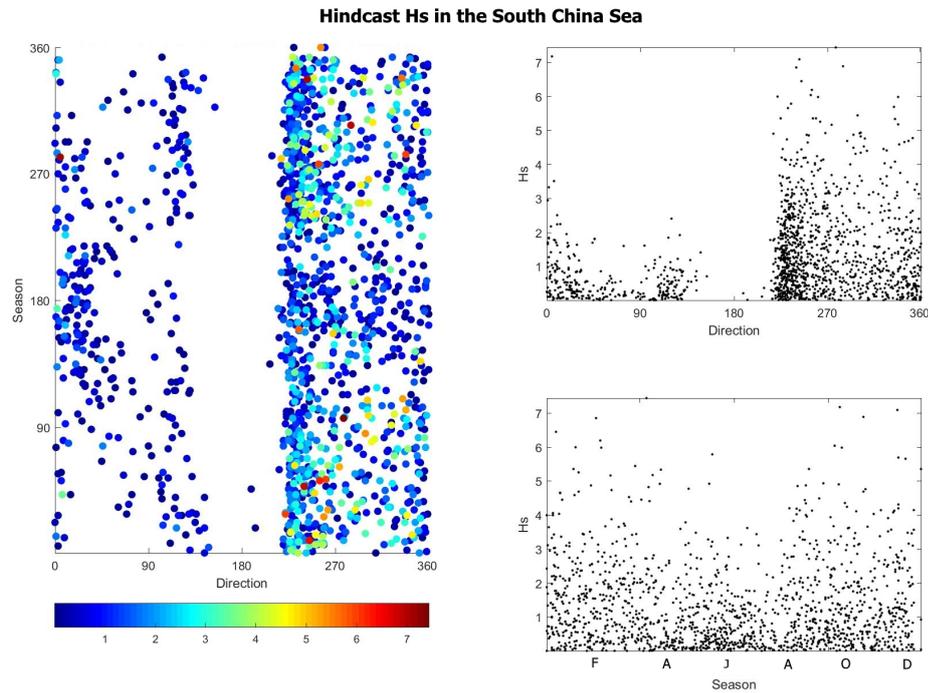


Figure 5.5.1: Storm peak significant wave height Hs hindcast (in meters) for a location in the Northern North Sea, with corresponding wave direction (in degrees, with 0 corresponding to waves approaching from the North) and season (in day of the year), both increasing clockwise.

We know from Section 4.5, we know that extreme sea states have a directional component, and waves are affected by different factors, such as land shadows from land masses and sea surface fetches. In the North Sea in particular, the longest available fetches correspond to the Norwegian Sea to the North, the Atlantic Ocean to the West, and the North Sea to the South. Extreme sea states are also partly shielded by Scandinavia to the East and the British Isles to the South-West. Extreme sea states also exhibit seasonal variation, and are dominated by winter storms, which are more intense and numerous, originate in the Atlantic Ocean and propagate eastwards across the northern part of the North Sea. It is worth noticing that, in contrast with the high variability from directional values we observed in Chapter

4, the seasonal variation present in the North Sea is less marked. For both the frequency of threshold exceedances and their size, the magnitude of changes with respect to season is lower, which should be reflected in the Poisson-GPD model parameters.

Before we can proceed to the inference stage for all the models discussed, the extreme value threshold needs to be estimated. Given the non-stationarity present, it is advisable to consider covariate-dependent thresholds, and we perform some exploratory analysis, which can be found in Appendix 5.B.1, to aid our choice. We decide to proceed with a threshold obtained from an non-exceedance probability of 0.7, which yields 978 threshold exceedances. Figure 5.5.1 shows the corresponding threshold exceedances and entire dataset, and by visual inspection we can recognise the directional-seasonal behaviour described above.

We first inspect the parameter estimates for all model parameters across Models 1-7, which are shown in Figure 5.5.2. All models seem to detect the same underlying patterns across all the model parameters, with only minimal differences in the range of values of the median estimates. When we consider the threshold exceedance rate, the Voronoi based models yield median estimates with a wider range of value, while the VTBR-sr and the BARBaR-fr seem to capture some inconsistent patterns which might be indicative of unreliable model fitting.
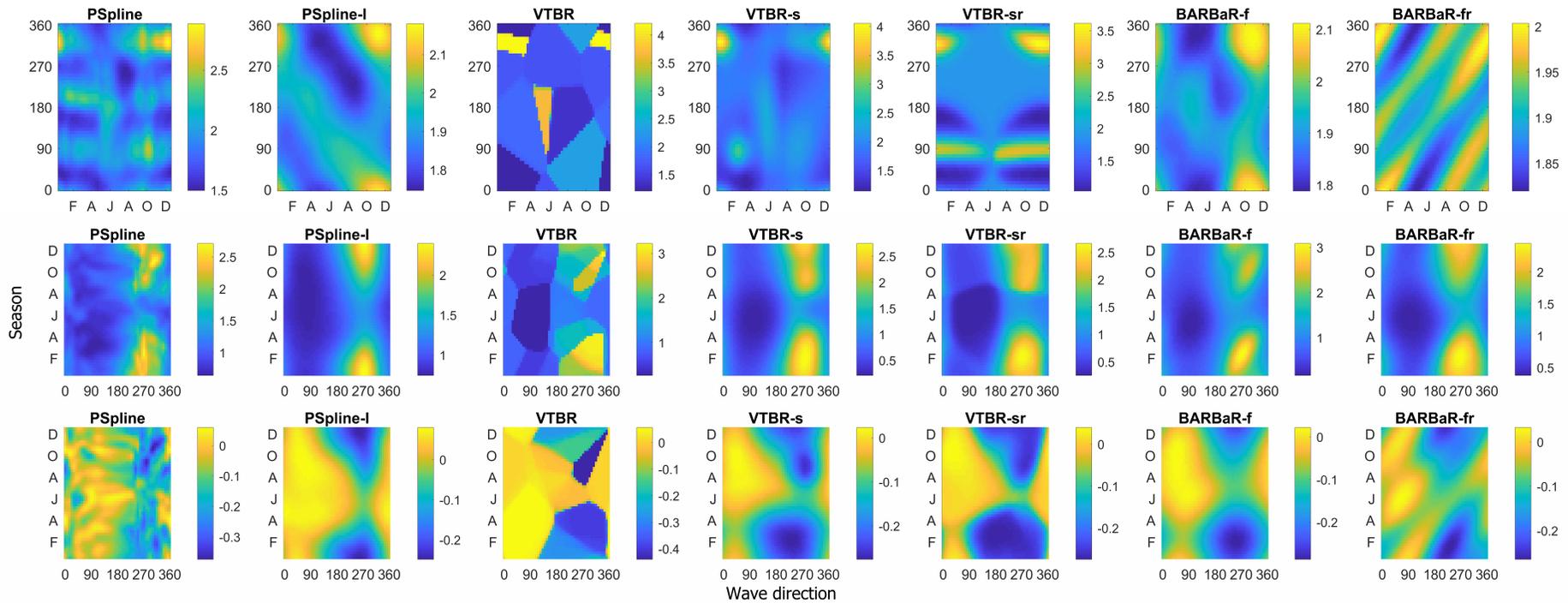
Figure 5.5.2: Parameter estimates for rate of occurrence $\phi(\mathbf{x})$ of the exceedances (upper), the GPD scale $\psi(\mathbf{x})$ (middle) and GPD shape $\xi(\mathbf{x})$ (lower) for the North Sea hindcast dataset shown in Figure 4.1.1, for Models 1-7. Each panel illustrates the posterior median estimate. Different color scales are used depending on the model, in order to show both the pattern detected and the range of values estimated.
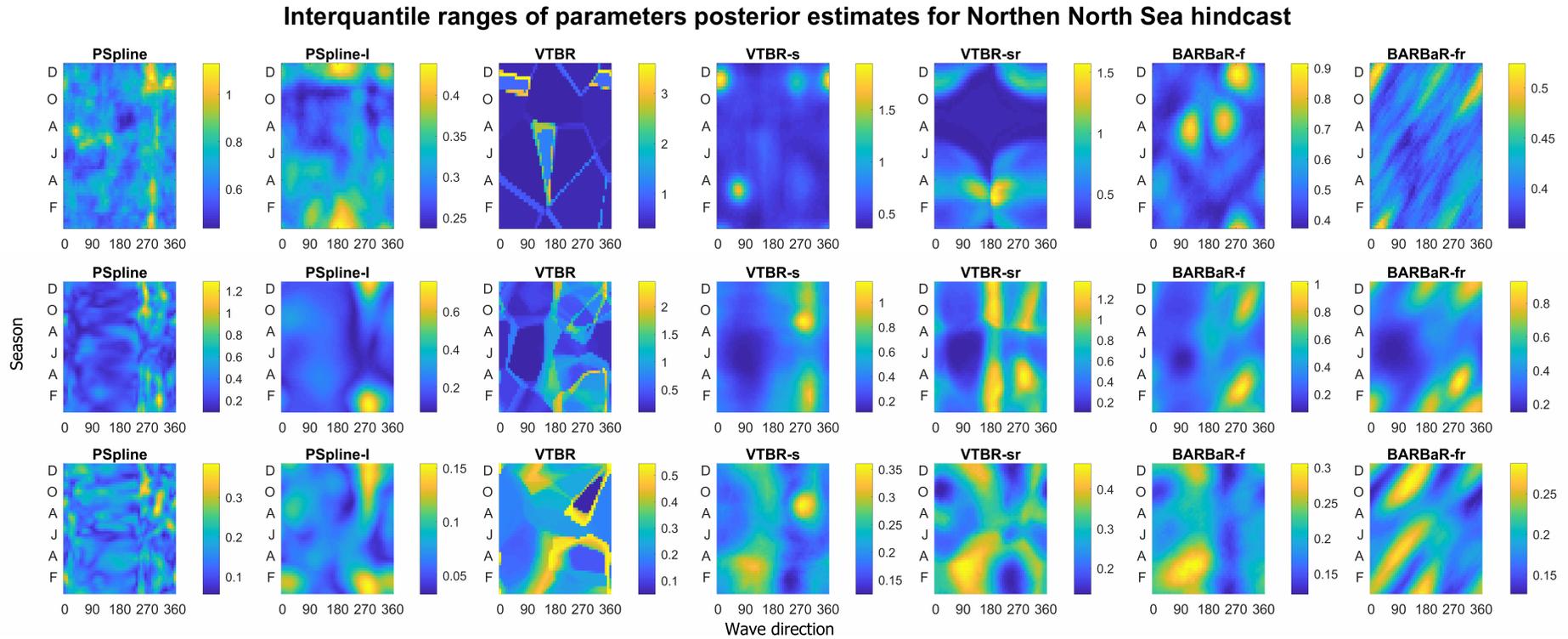
Figure 5.5.3: 95% interquantile ranges for the posterior estimates of the rate of occurrence $\phi(\mathbf{x})$ of the exceedances (upper), the GPD scale $\psi(\mathbf{x})$ (middle) and GPD shape $\xi(\mathbf{x})$ (lower) for the North Sea hindcast dataset shown in Figure 4.1.1, for Models 1-7. Different color scales are used depending on the model, in order to show both the pattern detected and the range of values estimated.

Again similar patterns emerge across all models when we consider the 95% interquartile range for the estimates, although the Voronoi based formulations yield, as before, wider intervals. Visually, we can immediately notice that the estimates from the P-spline model with the high number of component are less smooth, as could be expected. It is also possible to notice the fragmented nature of the estimates from the Voronoi based models. This was not noticeable in the case studies as results from the chains of 100 simulated samples were pulled together to obtain the final median estimate, hence producing a much smoother surface, while in this case only 1000 posterior estimates from the MCMC chain contribute to the final plots.

If we inspect the patterns detected across models, we can see how the rate of threshold exceedance is largest for winter storms from approximately a southern (180°) or northern (360°) direction. The estimates for the GPD scale show a considerable degree of variation, with the largest values corresponding to winter storms from a West-North West direction ($250° - 360°$). In term of the GPD shape, while we notice that this shows considerably large uncertainty, we can also detect some directional variability, again mainly concentrated in the winter months.

We proceed to consider the posterior expected cdf's of return values, corresponding to a return period of ten times the period of the original sample, which corresponds to approximately 500 years. These are shown in Figure 5.5.4. All models seem to produce very similar estimates both as omni-covariate and when focusing on specific directional and seasonal sectors. In fact, due to the scale and shape parameters of the generalised Pareto distribution being negatively correlated, different values of the two parameters can yield the same return value estimates.
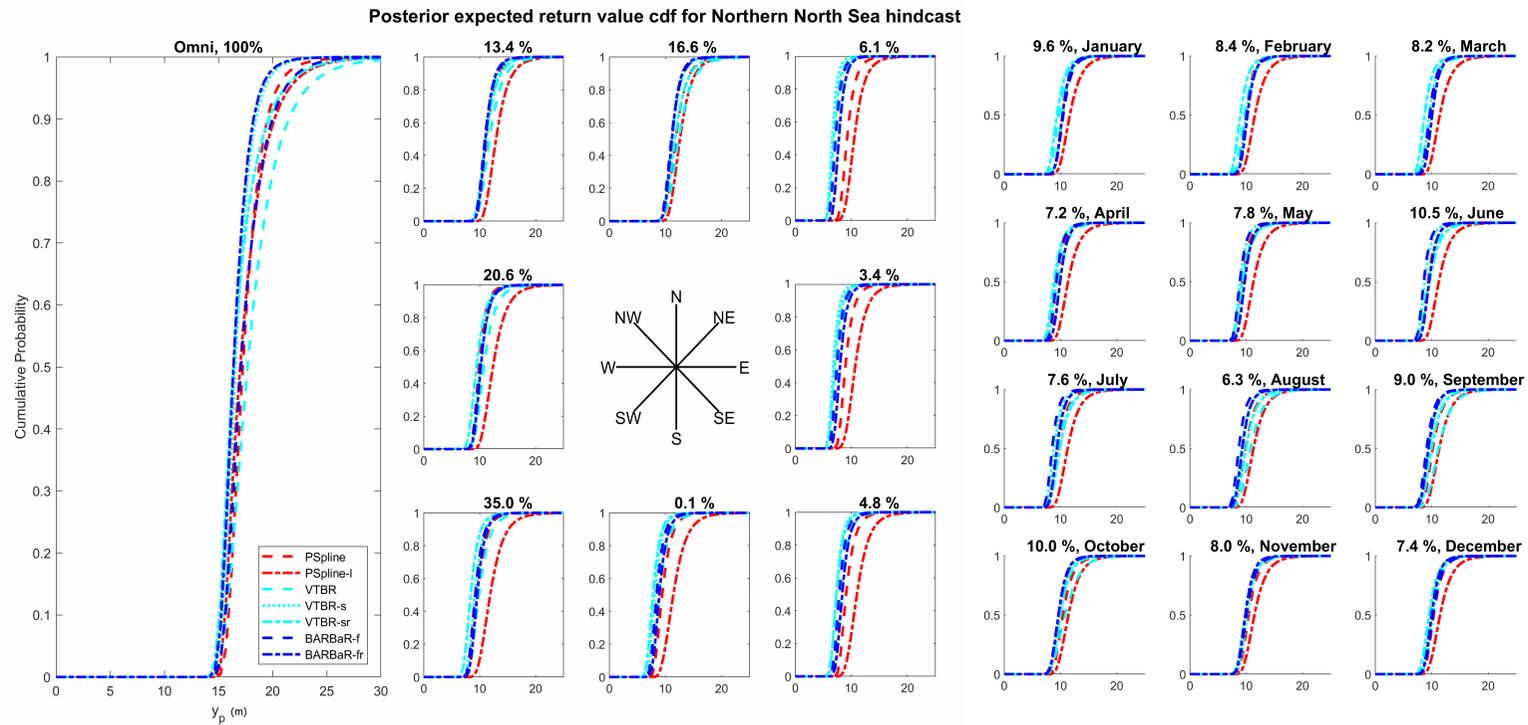
Figure 5.5.4: Posterior expected return value cdf for the North Sea hindcast dataset shown in Figure 4.1.1, corresponding to a return period of ten times the period of the original NNS sample. The panel show, from left to right, the omnidirectional return value distribution (left), the corresponding directional estimates (middle) and the seasonal estimates split by month (right). The title for each panel gives the expected percentage of individuals in that directional sector.

When considering estimates by sectors, we can notice some clear seasonal and directional variation, and can see that the most severe storms come from the North-West in winter months. Models seem to yield very similar estimates, with higher agreement in the sectors with highest number of observations, corresponding to waves coming from the South-West and North.

### 5.5.2 South China Sea hindcast
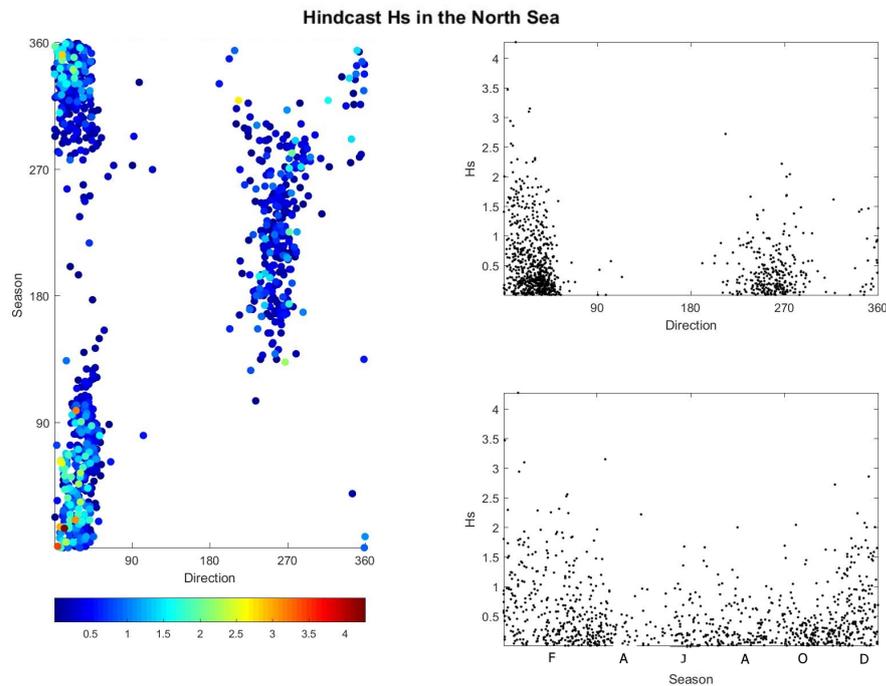


Figure 5.5.5: Storm peak significant wave height Hs hindcast (in meters) for a location in the South China Sea, with corresponding wave direction (in degrees, with 0 corresponding to waves approaching from the North) and season (in day of the year), both increasing clockwise.

The sample consists of hindcast observations between July 1956 and June 2007 for a location in the Makassar Strait between the islands of Borneo and Sulawesi in Indonesia. As mentioned in Section 3.4 when hindcast from a neighbouring site was considered, the main climatic features for the whole region are monsoonal,

with occasional typhoons passing through it. Southwest monsoons occur between July and September and northeast monsoons between December and March.

It is worth noticing some differences with the Northern North Sea data from Section 5.5.1. First, here the sea states present are less severe, with the largest value of Hs in the sample falling just over 4m high, compared to some Hs observations exceeding 7m in the North Sea. Secondly, the seasonality of the sea states is more marked here, and strongly correlated to the directional component of the waves considered, as visible from Figure 5.5.5. We can expect this variability to affect the rate and size of exceedances, and hence affect the Poisson-GPD model parameters.

As before for the North Sea hindcast, we perform some exploratory analysis to choose a suitable threshold (see Appendix 5.B.1 for further details). We decide to proceed with a threshold obtained from an non-exceedance probability of 0.65, which yields 1469 threshold exceedances. These exceedances, as well as the entire dataset, are shown in Figure 5.5.5.

It is worth noticing the South China Sea hindcast is the most complicated dataset we consider in this Chapter, due to the observations being sparse over the covariate domain and the underlying physical process showing a strong correlation between the two covariates. It is then unsurprising that here the models show the most disparity in performance that we have observed so far.
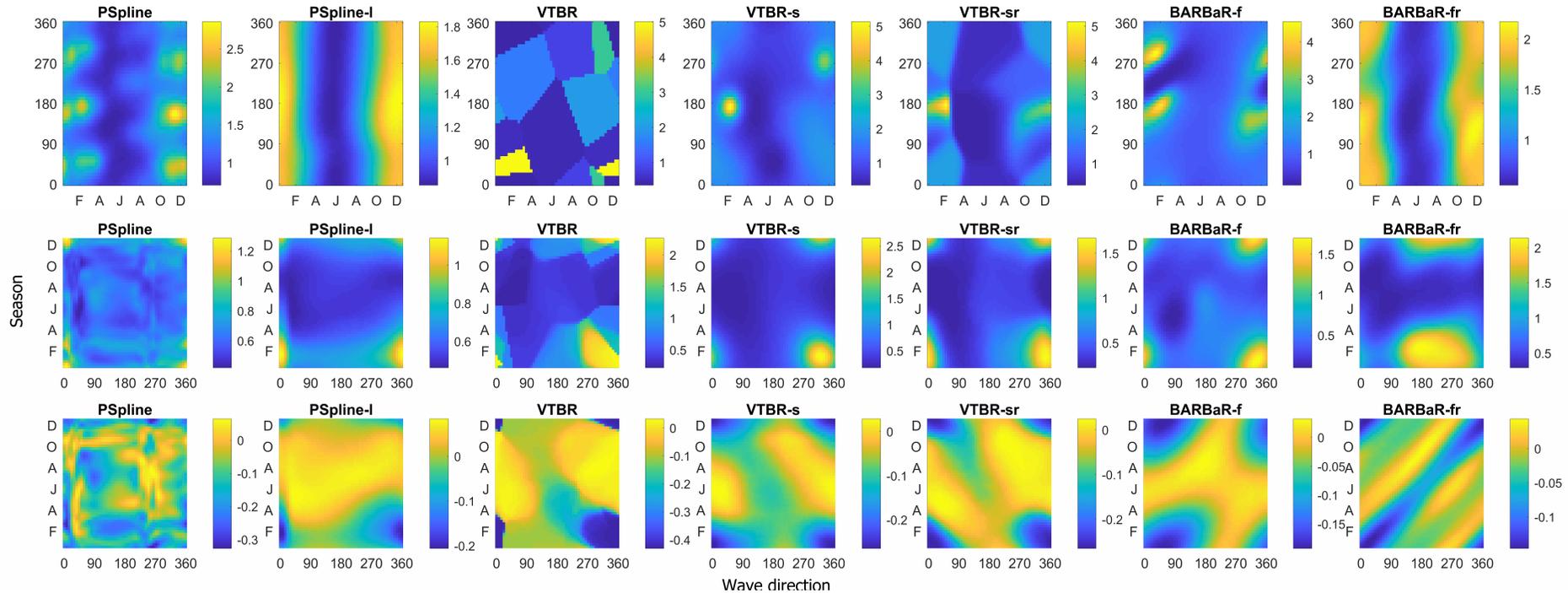
Figure 5.5.6: Parameter estimates for rate of occurrence $\phi(\mathbf{x})$ of the exceedances (upper), the GPD scale $\psi(\mathbf{x})$ (middle) and GPD shape $\xi(\mathbf{x})$ (lower) for the South China Sea hindcast dataset shown in Figure 5.5.5, for Models 1-7. Each panel illustrates the posterior median estimate. Different color scales are used depending on the model, in order to show both the pattern detected and the range of values estimated.
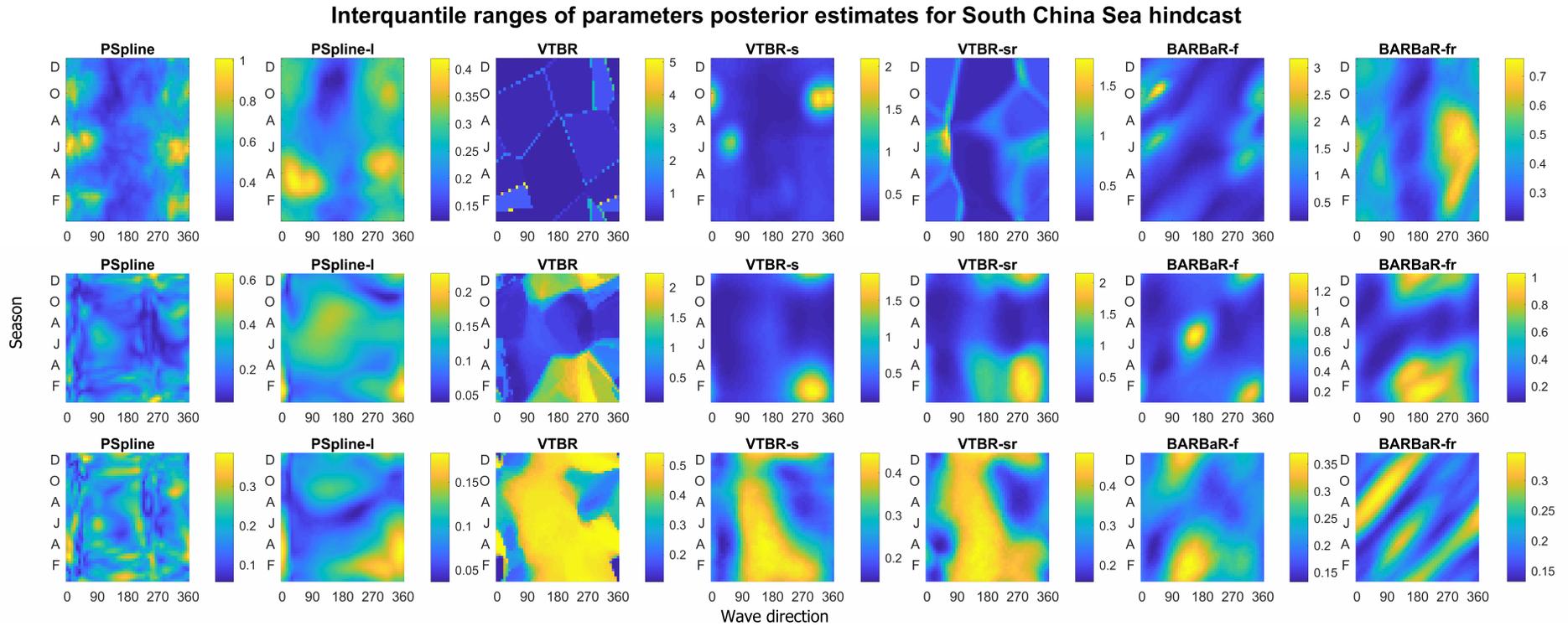
Figure 5.5.7: 95% interquantile ranges for the posterior estimates of the rate of occurrence $\phi(\mathbf{x})$ of the exceedances (upper), the GPD scale $\psi(\mathbf{x})$ (middle) and GPD shape $\xi(\mathbf{x})$ (lower) for the South China Sea hindcast dataset, for Models 1-7. Different color scales are used depending on the model, in order to show both the pattern detected and the range of values estimated.

First, we inspect the parameter estimates for all model parameters across Models 1-7, which are shown in Figure 5.5.2. All models seem to detect the some underlying patterns across all the model parameters, although these patterns seem to be perceived slightly differently across formulations, as some highlight local behaviour more, as for example the P-spline model assume constant smoothness. This is especially visible for the threshold exceedance rate, where the Voronoi based models yield median estimates with the largest range.

Again similar patterns emerge across all models when we consider the 95% interquartile range for the estimates, with the spline models exhibiting the smallest variation. As for previous cases, Voronoi based formulations yield wider intervals, with some very localised areas of wider uncertainty in the VTBR estimate, localised along what looks like cell edges, for the Poisson rate. As for the Northern North Sea hindcast,the P-spline and the VTBR model produce, respectively, "wiggly" and fragmented estimates.

We can compare what we know about the physical processes affecting the hindcast location with the patterns detected in the fitted models. The exceedance rate detects the areas with little to no observations, while the GPD scale shows localised behaviour, with maxima corresponding to the north-east monsoon.

We proceed to consider the posterior cumulative distribution functions of return values, corresponding to a return period of ten times the period of the original sample. This is, again, approximately 500 years. These are shown in Figure 5.5.4. Let us first consider the omni-directional omni-seasonal cdf estimates. Here most models seem to produce broadly similar estimates, which seem to agree with the results from the P-spline benchmark model. The spline based models yield

generally lower estimates than the others, which agrees with lower median values observed for the GPD scale. The BARBaR-fr model deviates considerably from the others, and shows overestimates of the return levels, likely due to a combination of high rate, scale and positive shape for large areas of the domain. While the effect in the single directional and seasonal sectors is limited, this issue is inflated by the joint contributions in the omni-directional omni-seasonal estimates. Models differ in estimates across most sectors, but they show higher agreement in the sectors with highest number of observations. Despite these differences, we can notice for all of them some clear seasonal and directional variation, yielding higher estimates for the monsoonal directional-seasonal sectors.
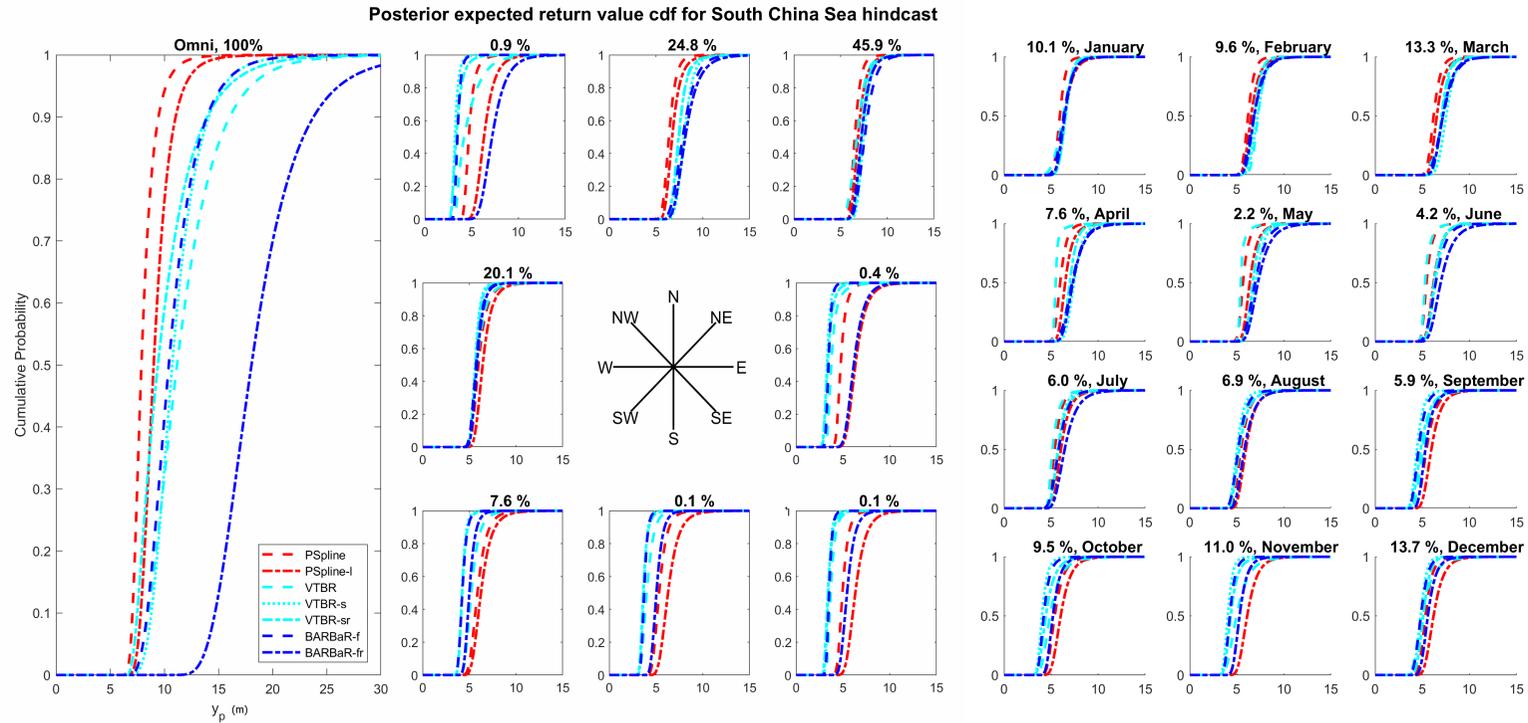
Figure 5.5.8: Posterior expected return value cdf for the South China Sea hindcast dataset shown in Figure 5.5.5, corresponding to a return period of ten times the period of the original sample. The panel show, from left to right, the omnidirectional return value distribution (left), the corresponding directional estimates (middle) and the seasonal estimates split by month (right). The title for each panel gives the expected percentage of individuals in that directional sector.

## 5.6 Discussion and concluding remarks

Extreme occurrences of environmental phenomena have potentially devastating consequences associated with heavy costs, pollution and risk to human lives. Extreme value analysis provides the ideal framework for modelling and predicting rare phenomena with very low probabilities of occurrence, although various issues arise from a requirement to produce such long-range extrapolations based on data that is, by definition, sparse. Environmental processes are often very complex and observations tend to be non-homogeneous in space, time and with respect to covariates. It then becomes essential to build models that adequately include these factors, as ignoring or failing to capture the effects of covariates adequately leads to unreliable and unrealistic analysis.

The need for approaches that yield reliable and efficient analysis for non-homogeneous extreme data has been the driving theme for this thesis. We have used peaks over threshold methods, which examine observations above an arbitrary large value to characterise the behaviour at the tail of the distribution. In Chapter 3, we focused on two models from the literature, namely the generalised Pareto distribution and the non-homogeneous Poisson point process formulation. These models are theoretically equivalent and both require adaptations to account for covariate effects. We have then considered the implementation and consequences of adaptations to include covariate effects, and showed how each method has advantages and limitations.

The rest of the thesis has focused on issues that arise once a suitable extreme value model has been chosen. We have concentrated specifically on extreme character-

istics of the ocean environment, which are often characterised by multiple sources of non-stationarity (such as wave or wind direction, season, fetch, longitude and latitude). Amongst the available methods, we have opted to capture sample non-homogeneity by representing the parameters of the statistical model as functions of covariates. One then has to consider a variety of issues, as different parameterisations are available from the literature. In oceanographic applications, it is often necessary to consider models that are complex enough to be practically useful, such as semi-parametric and non-parametric models, without being too complex. Furthermore, the framework for covariate-dependent modelling should be, in general, flexible and scalable, while producing reliable results. More specifically, we are ideally interested in an approach which can be applied to samples from neighbouring or distant locations without requiring extensive *ad hoc* changes. A model should also be computational efficient and stable, characteristics that become even more important when multiple covariates are included in the analysis.

We first started to investigate these issues in Chapter 4, where we have focused on modelling non-stationary series with one-dimensional covariates. We have introduced novel methodology based on radial basis functions and compared its performance to existing models, in particular the standard P-spline approach, which was used as a benchmark. We assessed model performance using case studies and later applied the models to a hindcast significant wave height dataset. The case study highlighted important observations on ease of implementation, stability and potential for further refinements. When applied to the hindcast data, the models have shown promising results, at least matching the performance of the benchmark model for both omnidirectional and directional summary results.

In this chapter, we have considered the case where two covariates are present, as it introduces considerable complications, from a modelling, computational and prediction perspective, and provides a stepping stone towards higher-dimensional models. In order to yield reliable results, the formulation needs to appropriately capture not only the behaviour associated with each covariate, but also any potential interaction between them. While a variety of parametric and non-parametric formulations provide flexible approaches in 1-D, extending these models to a two-dimensional covariate domain is not straightforward.

The computational cost involved also increases considerably when a new covariate is included, so efficiently and scalability are essential. For example, while non-parametric regression can already lead to a large number of correlated parameters when one covariate is considered, this issue is exacerbated for multiple covariates and is more marked when a gridded approach to a two-dimensional covariate domain is used, such as for P-splines. We introduced novel methodology, by extending the approach based on radial basis functions from Chapter 4 as well as considering partition models to represent the interaction with the covariates domain. The aim was to obtain some models with the potential of being more efficient and stable than the P-spline gridded approach.

We have again used a simulation study to test and assess the performance of the models considered, and later applied them to two hindcast significant wave height datasets with different characteristics. From the simulation study, we see that some of the models perform similarly to the benchmark P-spline model. The basic versions of the new models, the VTBR and the BARBaR-f, seem particularly promising, although some issues with uncertainty and stability emerged which can

provide interesting avenues for further investigation. We also found it necessary to introduce, for all models, an additional hyperparameter for the coefficients prior to imposing a varying level of smoothness, in a similar way as for the spline formulations. The P-spline approach with a low number of components often performed better than the benchmark, which is likely a consequence of the aforementioned issues with parameter correlation in bases with high numbers of components.

Applying the models to hindcast data shed valuable insight on their performance and opportunities for improvement. The characteristics of the first dataset from the Northern North Sea are relatively well described by the models. While data are sparse for a small portion of the directional domain, changes in season are much smoother. Here all models yield good and consistent estimates, and perform well in terms of quality of inference. The South China Sea hindcast provided a more challenging application. In this case, large portions of the covariates domain have little to no observations, and the underlying oceanographic phenomena have strong joint directional-seasonal characteristics. These issues proved challenging for all models, and careful tuning was necessary.

Furthermore, since the scale and shape of the GPD are negatively correlated, problems with approaches yielding over- and under-estimates can only be properly detected when we start to consider return values. In particular, despite differences in the parameter estimates, most models still provided similar summary results. All model formulations were able to capture the directional and seasonal variation present in both the simulated and hindcast data, with good agreement also for different directional-seasonal sectors.

In conclusion, the inspection of extreme value models for non-stationary series and the deployment of different formulations to capture the underlying non-homogeneity has yielded some interesting results and provided clear areas for further research. The models considered show promising results and manage to represent the covariate-dependent features in one- and two-dimensions. However, model implementation showed some additional issues when compared to the penalised spline approach. This is to be expected as the latter is a well-established framework which has been widely researched in the literature and developed over a number of years. However, the novel formulations for two-dimensional covariates are considerably more parsimonious, as clearly shown by Table 5.3.1.

In this chapter, we implemented the standard penalised spline model with a medium-sided basis (with approximately half the total components used, for example, by Randell et al. (2015a)), which already required noticeably longer computational time for inference in comparison to other approaches. Overall, although the results were not as stable as one would hope, their encouraging performance in modelling non-homogeneity indicates that there is scope of improvement and the prospects of computational saving warrant further study.

# Appendix

## 5.A  Inference and tests for Cases 1 and 2

While we decided to focus on the results from Case 3, we consider here the results
and performances from the first two cases. As mentioned in Section 5.4.1, we use
an extreme value threshold $u$ fixed at zero during inference, which corresponds
to the value used to simulate the cases realisations. The results show below are
obtained via Bayesian inference, with MCMC run lengths of 60000 in total,the first
30000 samples removed of which were discarded as burn-in. As before in Chapter
4, we use a constant model for the Poisson rate $\phi$ of threshold exceedance Case 1,
with the same resulting chains being later used to produce posterior expected cdf
estimates for all Models 1-7, while the rate is modelled as covariate dependent in
Cases 2, although the changes are limited in range.

We first inspect the median parameter estimates, as well as interquantile ranges,
over the sample realisations for the three model parameters in Case 1 across models
1-7, shown in Figure 5.A.2. Visual inspection suggests that estimates are of similar
quality across models: most differences emerge in the estimation of the GPD shape
parameter. As far as the latter is concerned, the VTBR-s and the BARBaR-fr

seem to struggle the most with the estimation, albeit producing estimates which seem to capture a similar pattern to the one detected by the other models. These issues are confirmed by the range between the upper and lower quantile for the shape parameter $\xi(\mathbf{x})$ for most models. Purely in terms of parameter estimates, the Voronoi-based formulations seem to show more uncertainty and overall wider interquantile ranges for both the shape and the scale parameter.

It is worth remembering that, for Case 2, the scale and shape parameters were chosen to have the same form as in Case 1. The main distinguishing feature lies in the threshold exceedance rate, and consequently in the fact that the observations are no longer uniformly scattered on the covariates domain. Nevertheless, the presence of areas of the domain with sparser data seems to have little to no effect on the estimation of the GPD parameters, which yield remarkably similar results to case one in both median and uncertainty estimates. This is probably due to the fact that, although the rate changes over the domain, we proposed here a case where the difference in magnitude between the larger and smaller rate observed is fairly contained. As far as the Poisson rate is concerned, again all models seem to obtain consistent median estimates. Nevertheless, it is worth noticing that the spline formulations yield narrower interquantile ranges, while the BARBaR models shows the highest level of uncertainty with estimates of over a unit higher than the ones from other models in part of the domain.
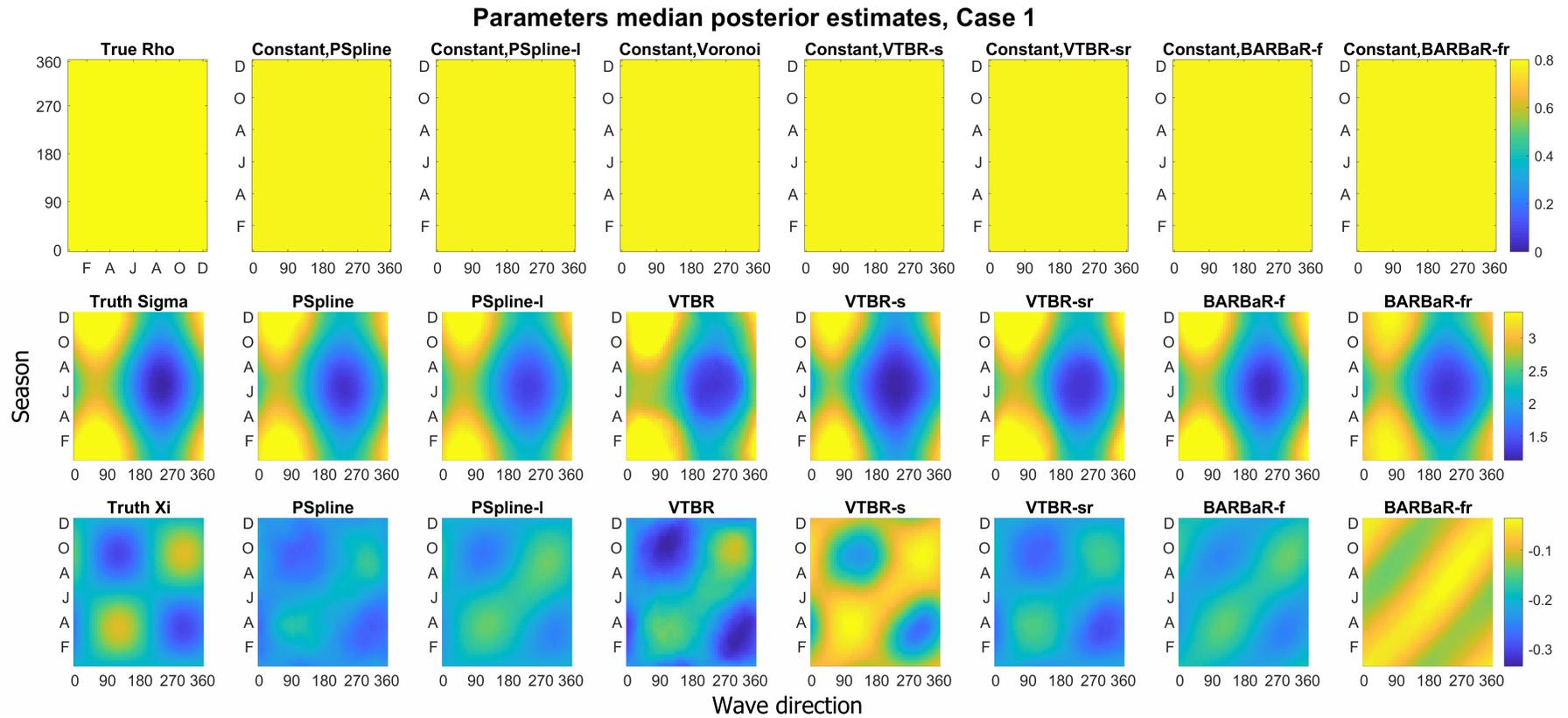
(a) Case 1, median posterior estimates.

**Interquantile ranges of parameters posterior estimates, Case 1**
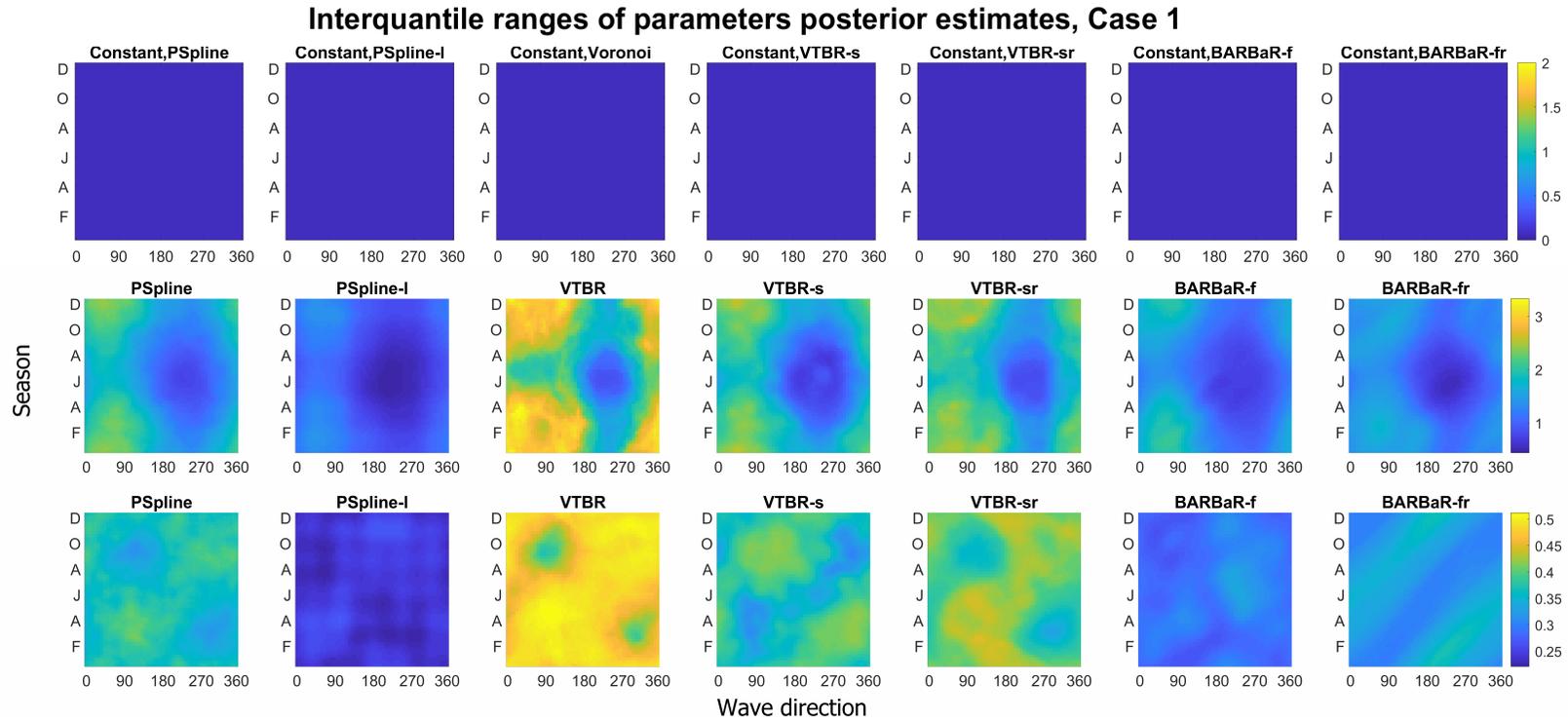
(a) Case 1, posterior estimates 95%interquantile ranges.

Figure 5.A.2: Parameter estimates (Figure 5.A.1a) and 95% interquantile ranges (Figure 5.A.2a) for rate of occurrence $\phi(\mathbf{x})$ of the exceedances (upper), the GPD scale $\psi(\mathbf{x})$ (middle) and GPD shape $\xi(\mathbf{x})$ (lower) for Case 1, for Models 1-7. Each panel illustrates the posterior median estimate.
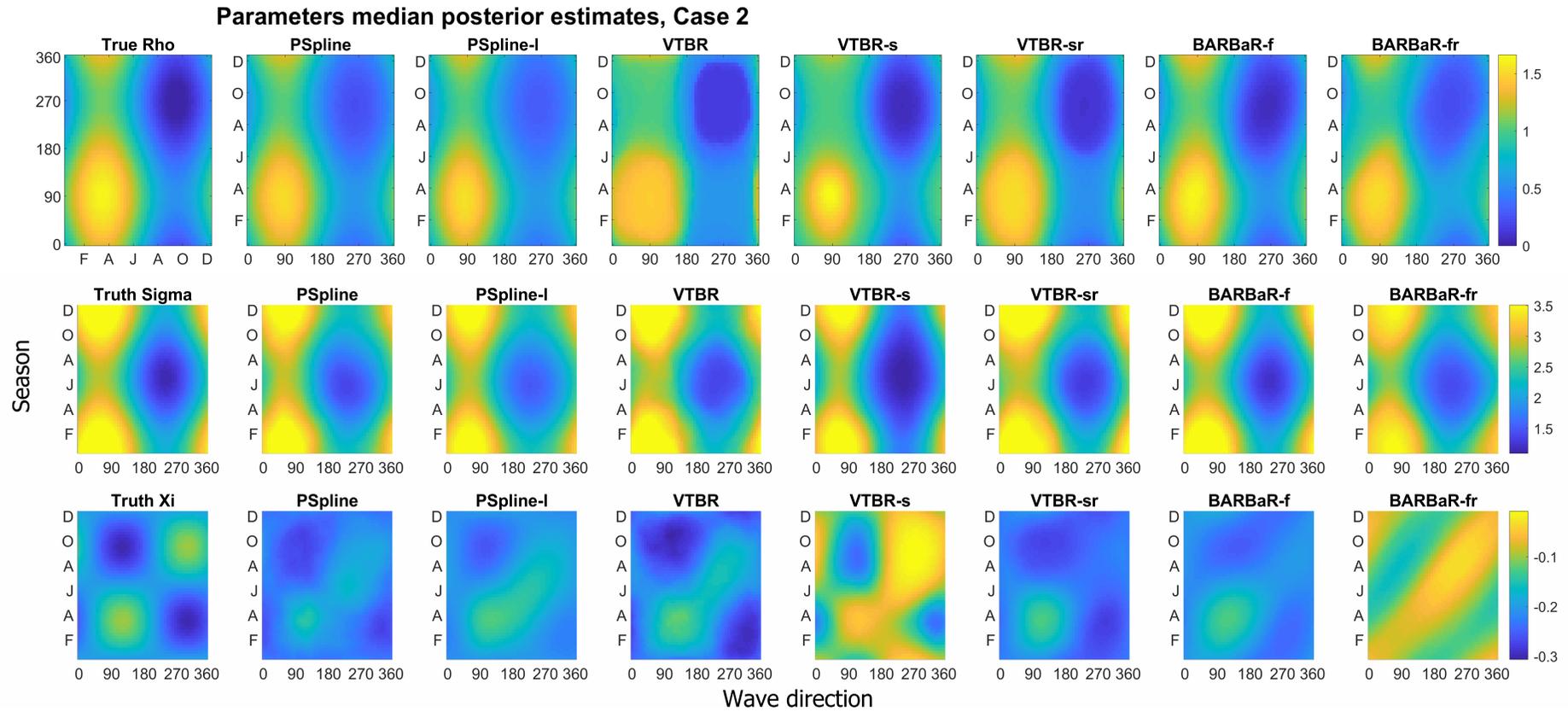
(a) Case 2, median posterior estimates.

**Interquantile ranges of parameters posterior estimates, Case 2**

(a) Case 2, posterior estimates 95% interquantile ranges.
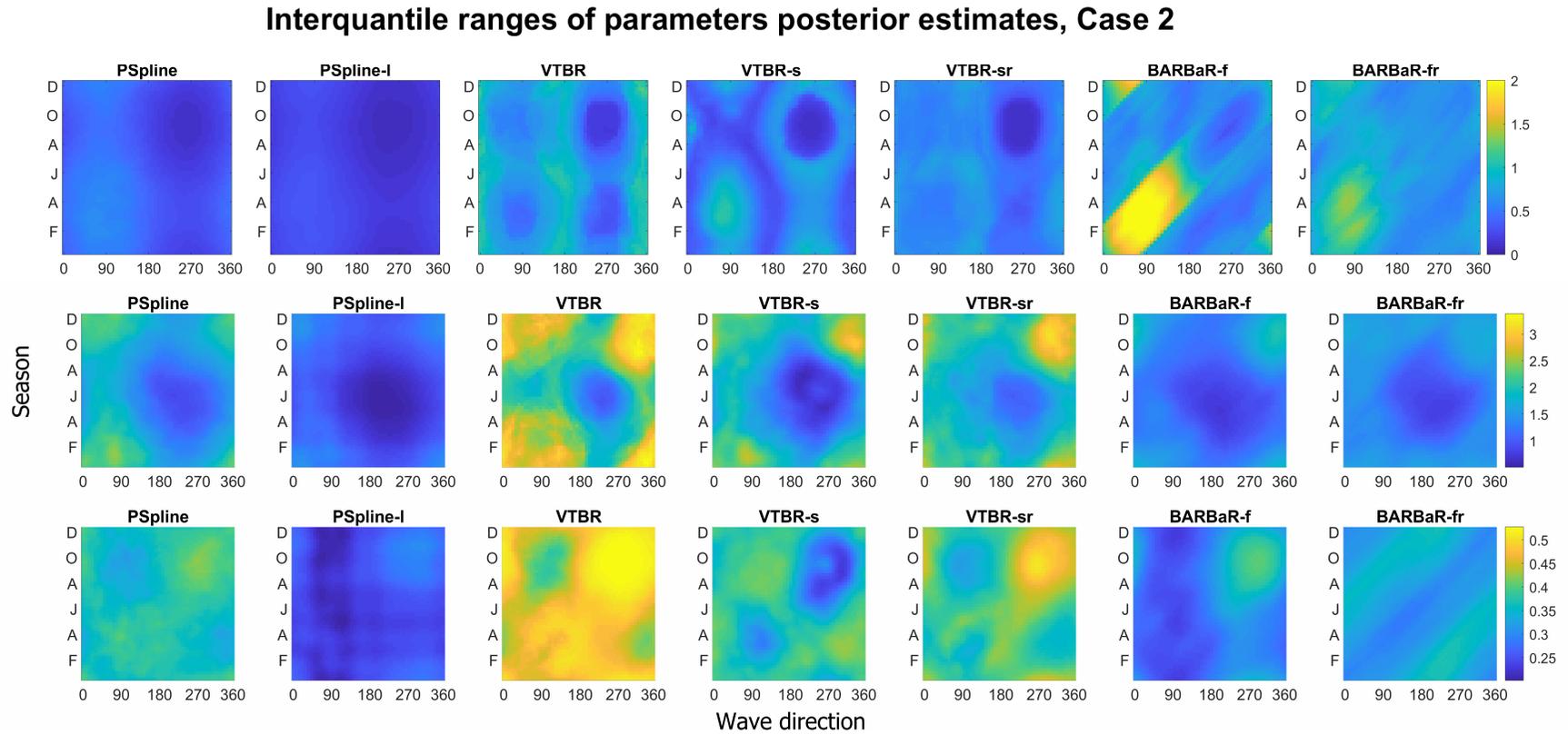
Figure 5.A.4: Same as Figure 5.A.2, here for Case 2.

We proceed to consider the posterior cumulative distribution functions of return values, corresponding to a return period of ten times the period of the original sample. These are shown in Figure 5.A.6 for Case 1 and 2. Here we notice that omnidirectionally, the benchmark P-spline model produces the closest estimates to the ones from the true underlying model. The additional uncertainty detected for the Voronoi-based models, as well as in the BARBaR-fr estimates for the shape parameter, result in overestimates of the return values.

Overall, the standard BARBaR-f model and the low-dimensional spline approach produce the closest estimates to both the benchmark and the truth. Similar patterns are visible if we consider the results by directional octant or seasonal interval, as can be expected by the uniform distribution of data over the covariate domain and the corresponding constant rate of exceedance.

For Case 2, again the two spline models, as well as the BARBaR-f approach, yield the closest estimates to the cdf obtained from the underlying true parameters. Similar results can be observed for Case 2, which is consistent with the findings about the parameter estimates. Again, the limited range of values of the covariate-dependent rate of exceedance ensures that little difference is noticeable across directional and seasonal sectors. The most noteworthy difference lies in the performance of the VTBR-s and VTBR-sr models, which show some issues for higher probabilities, which can be explained as the result of wider confidence intervals across all three parameter estimates, as well as a consequence of some ill-fitting results for a limited number of realisations.

**Average posterior expected return value cdf, Case 1**
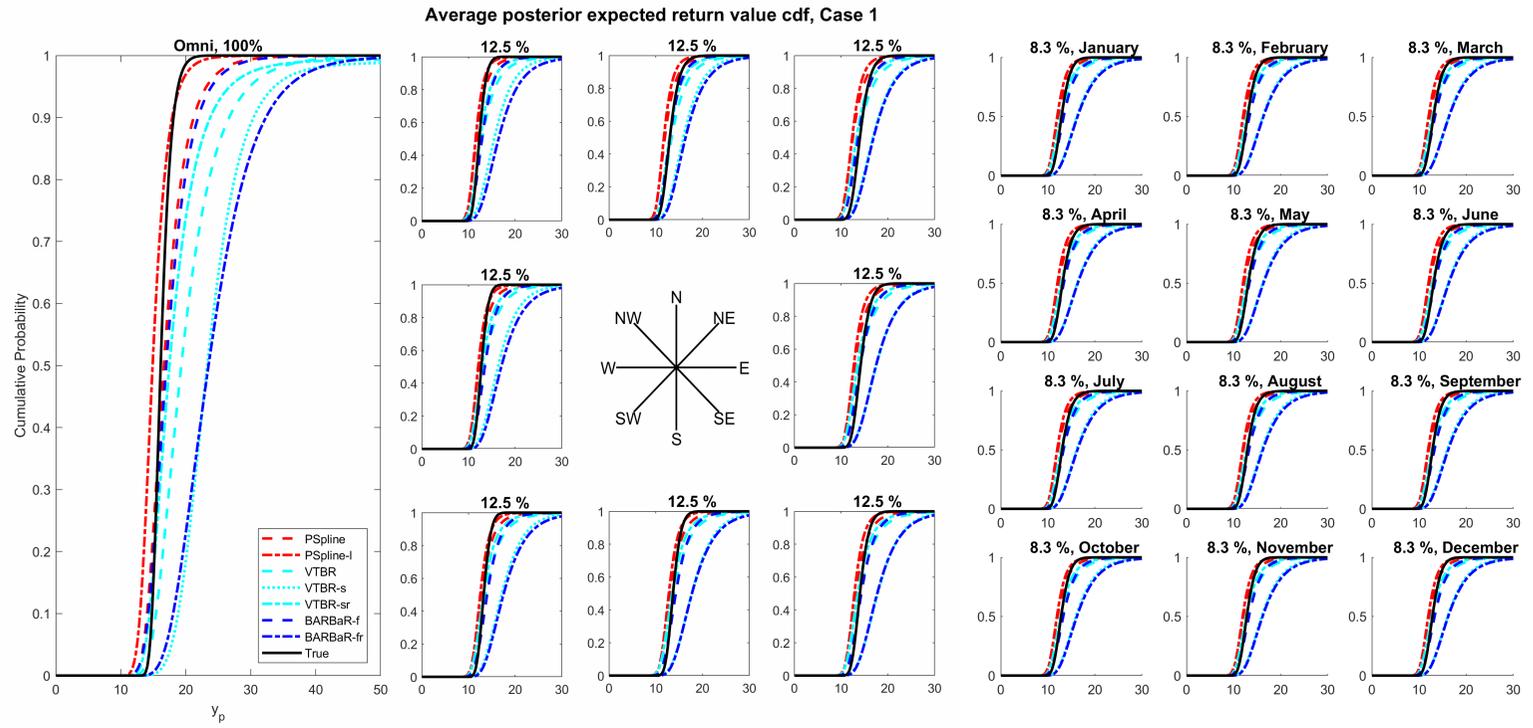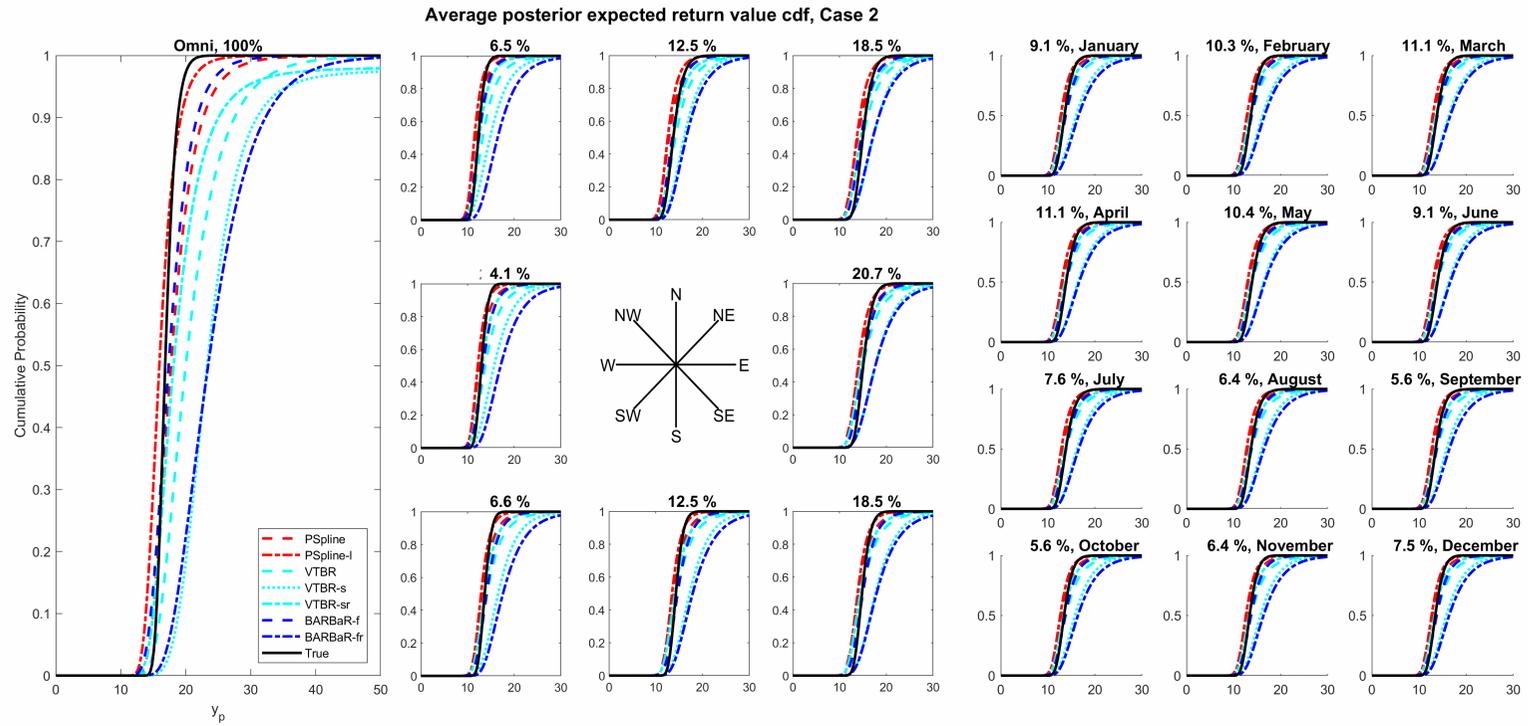
(a) Case 1

(a) Case 2

Figure 5.A.6: Average posterior expected return value cdf for simulated samples from Case 1 (Figure 5.A.5a) and 2 (Figure 5.A.6a), corresponding to a return period of ten times the period of the original sample. The panel show, from left to right, the omnidirectional return value distribution (left), the corresponding directional estimates (middle) and the seasonal estimates split by month (right). The title for each panel gives the expected percentage of individuals in that directional sector. The true return value distribution is given in solid black.

Figures 5.A.8 and 5.A.10 summarise, for both Case 1 and Case 2, the characteristics of the distributions for KS and KL divergence criteria respectively. In both cases, we note that all models perform similarly, especially in terms of maximum divergence, as shown by the KS test values. The results are consistent across models both for the omni-covariate estimates and in each of the directional and seasonal sub-sector. Some variation can be detected in the KL criterion estimates, which consider the total distance of the posterior expected cdf's from the true one.

In Case 1, the Voronoi-based models show, both in the omni-covariate and sectoral values, marginally wider interquantile ranges and hence seem to yield in more varied results across different realisations, which is consistent with the uncertainty previously detected in the parameter estimates. Nevertheless, in Case 2, none of the models seems to show consistently better or worse KL criterion values across all directional, seasonal or omni-covariate cdf estimates.
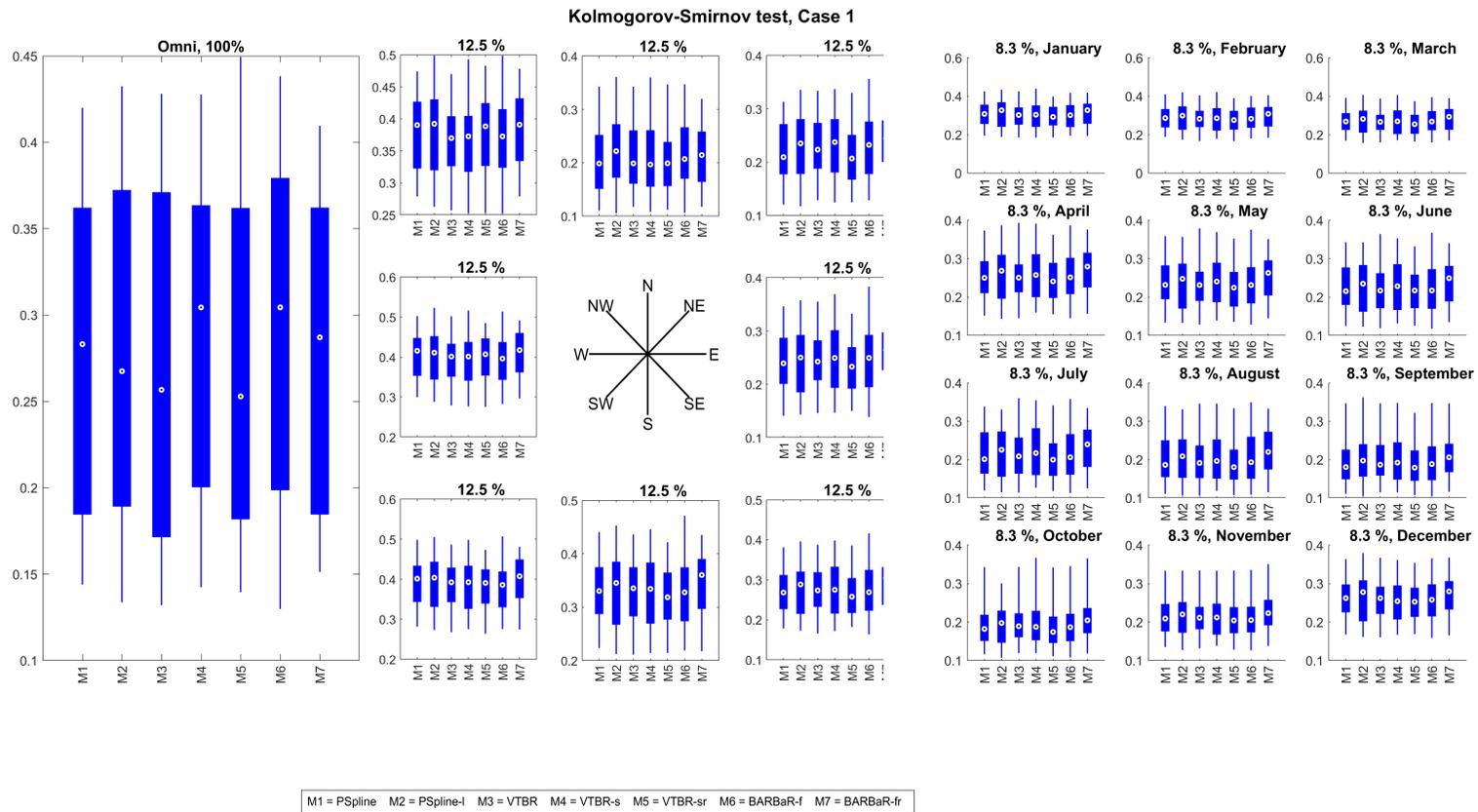
(a) Case 1
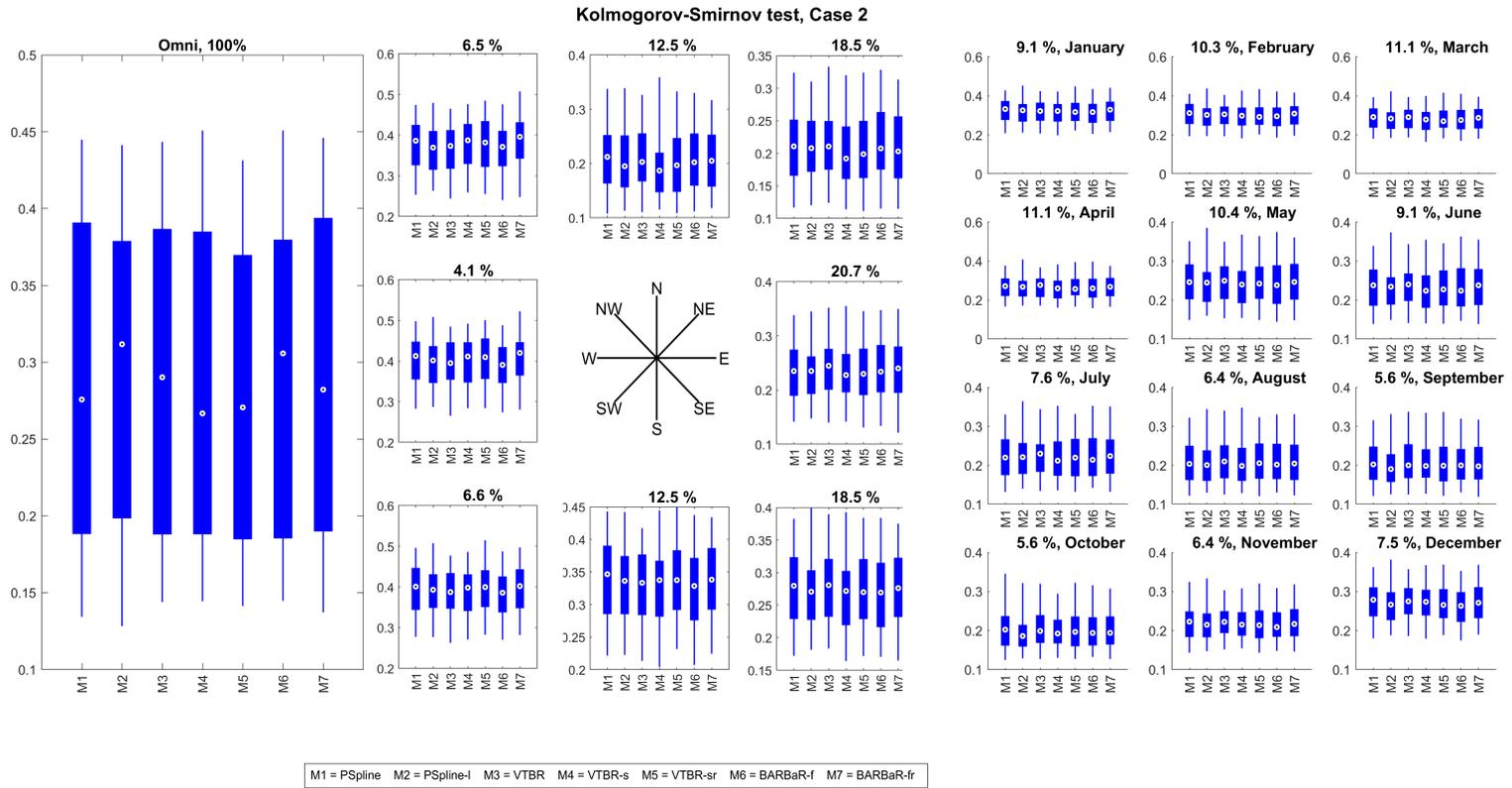
(a) Case 2

Figure 5.A.8: Box-whisker comparison of samples of the Kolmogorov-Smirnov (KS) divergence criterion between omnidirectional (left) posterior expected return value cdf's (corresponding to a return period of ten times that the original sample), by directional sector (middle) and seasonal monthly sector (right), estimated under samples from the true return value distribution and those estimated under models of each of 100 sample realisations for Case 1 (Figure 5.A.7a) and 2 (Figure 5.A.8a). The sample of the KS divergence criterion are summarised by the median (white disc with black central dot), the interquartile range (blue rectangular box, with vertical lines showing the 2.5%, 97.5% interval)

**Kullback-Leibler test, Case 1**



M1 = PSpline     M2 = PSpline-l     M3 = VTBR     M4 = VTBR-s     M5 = VTBR-sr     M6 = BARBaR-f     M7 = BARBaR-fr
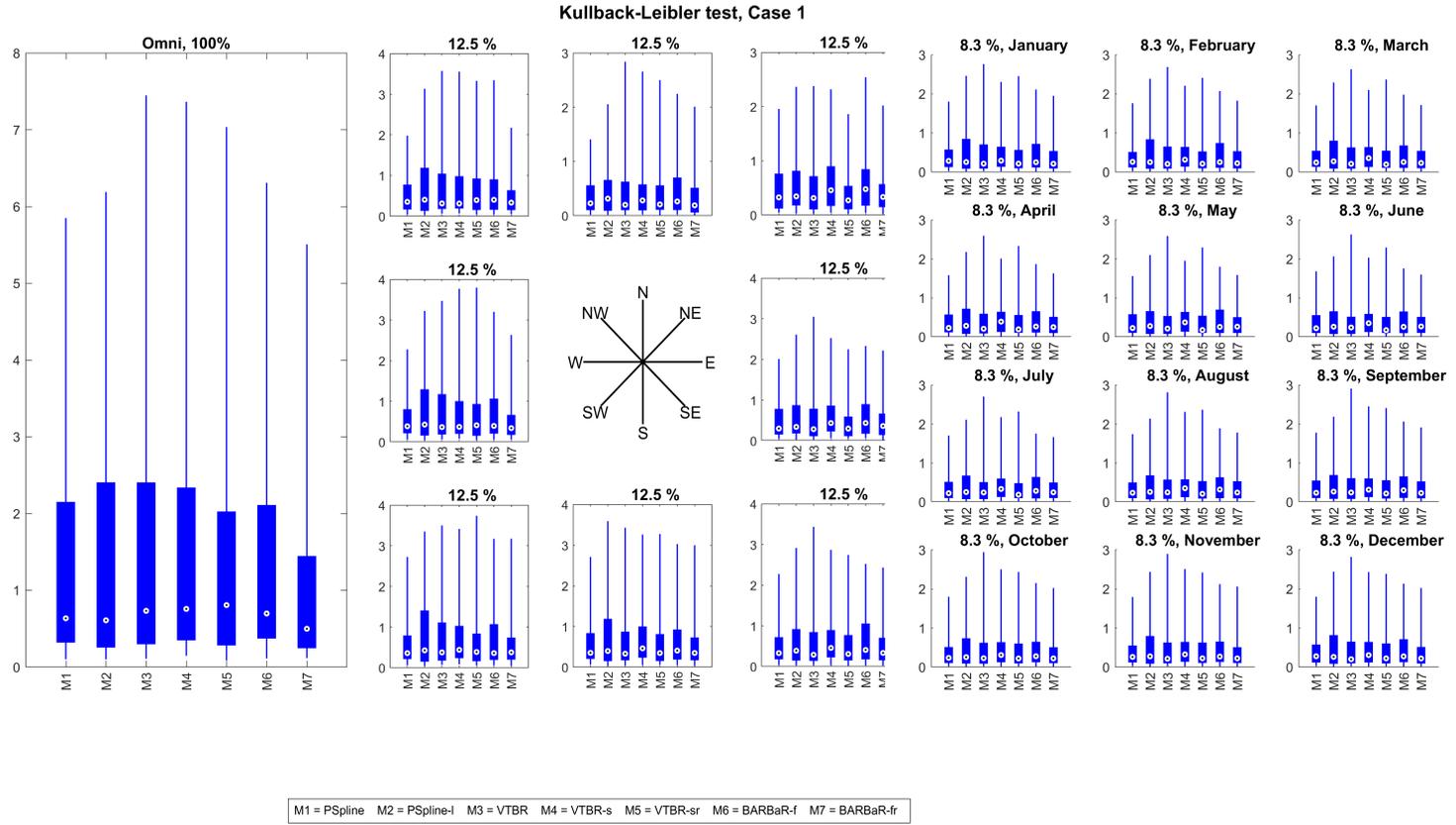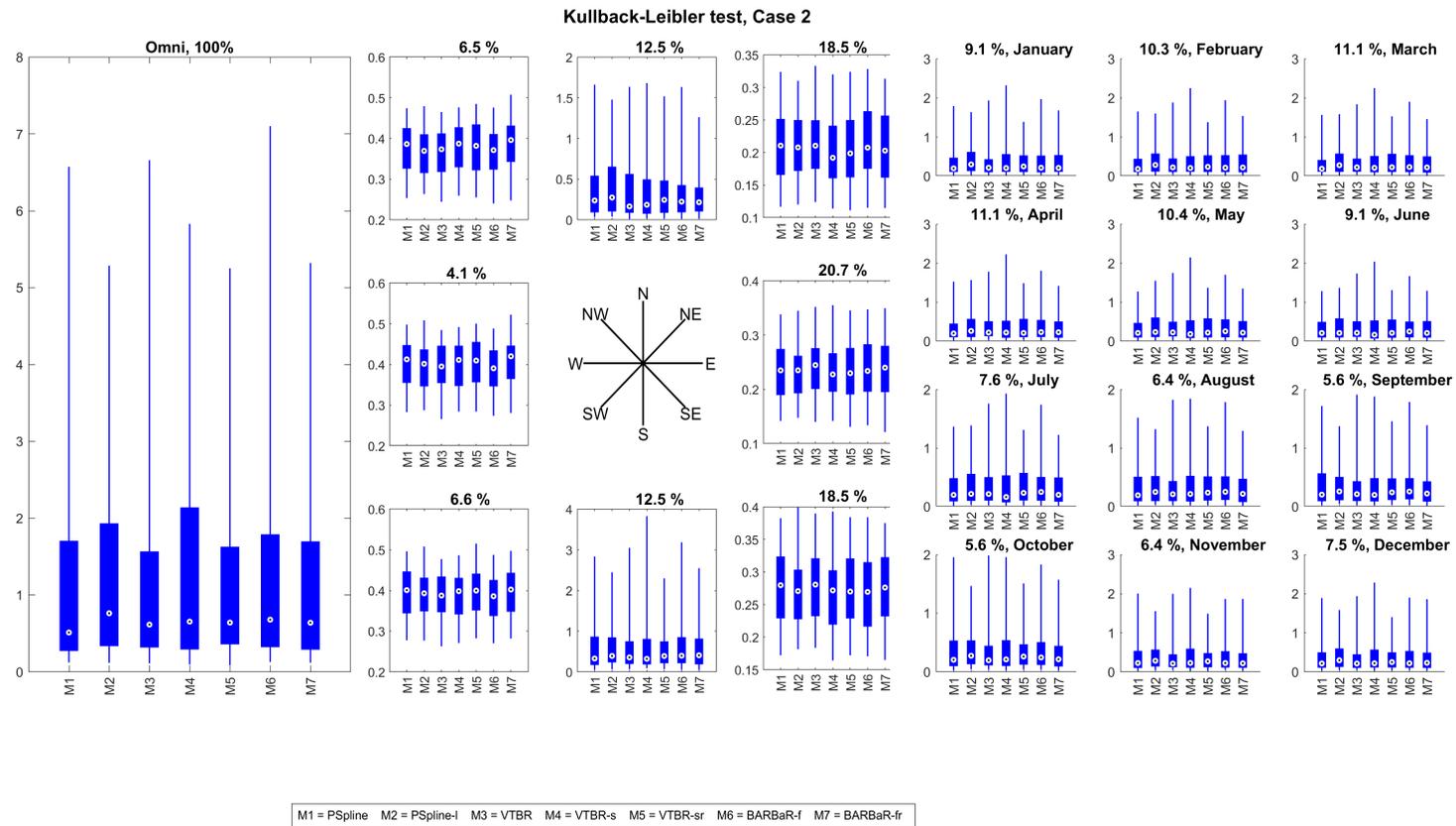
(a) Case 1

(a) Case 2

Figure 5.A.10: Same as Figure 5.A.8, here for the Kullback-Leibler (KL) criterion.

## 5.B  Threshold diagnostics for hindcast datasets

Given the presence of strong non-stationarity, it is again advisable to consider covariate-dependent thresholds for the two hindcast datasets. Here we decide to show results corresponding to a non-exceedance probability varying between 0.6 and 0.8. Higher and lower values were also considered initially, although these were excluded following the same criterion illustrated below. The smaller range of non-exceedance probabilities extracted allows us to produce clearer and more informative plots, which can focus on the most interesting values.

We proceed with the same aim as for previous chapters, looking for a threshold that is low enough to allow as much data to be preserved as possible, while still ensuring we are only considering extreme data and yielding stable and reliable results. Nevertheless, the more covariates are considered, the harder to visualise diagnostic plot becomes. Furthermore, multiple covariates affect both the efficacy and the interpretability of the methods used in previous chapters. Here, we then found it useful to consider stability of fit by comparing the median of the posterior expected return value cdf across different return periods for the desired thresholds. For consistency, these estimates were all achieved using the same model, namely the spline Model 1 from Section 5.2.2. Figure 5.B.1 shows these cdf estimates, with shaded areas corresponding to the interquantile ranges for the lowest (0.6) and highest (0.8) non-exceedance probability.

For both datasets, we can notice that the lowest threshold shown (0.5 exceedance probability) yields considerably different results from the remaining ones. This is indicative of a value that is too low, such that data that is not extreme is included
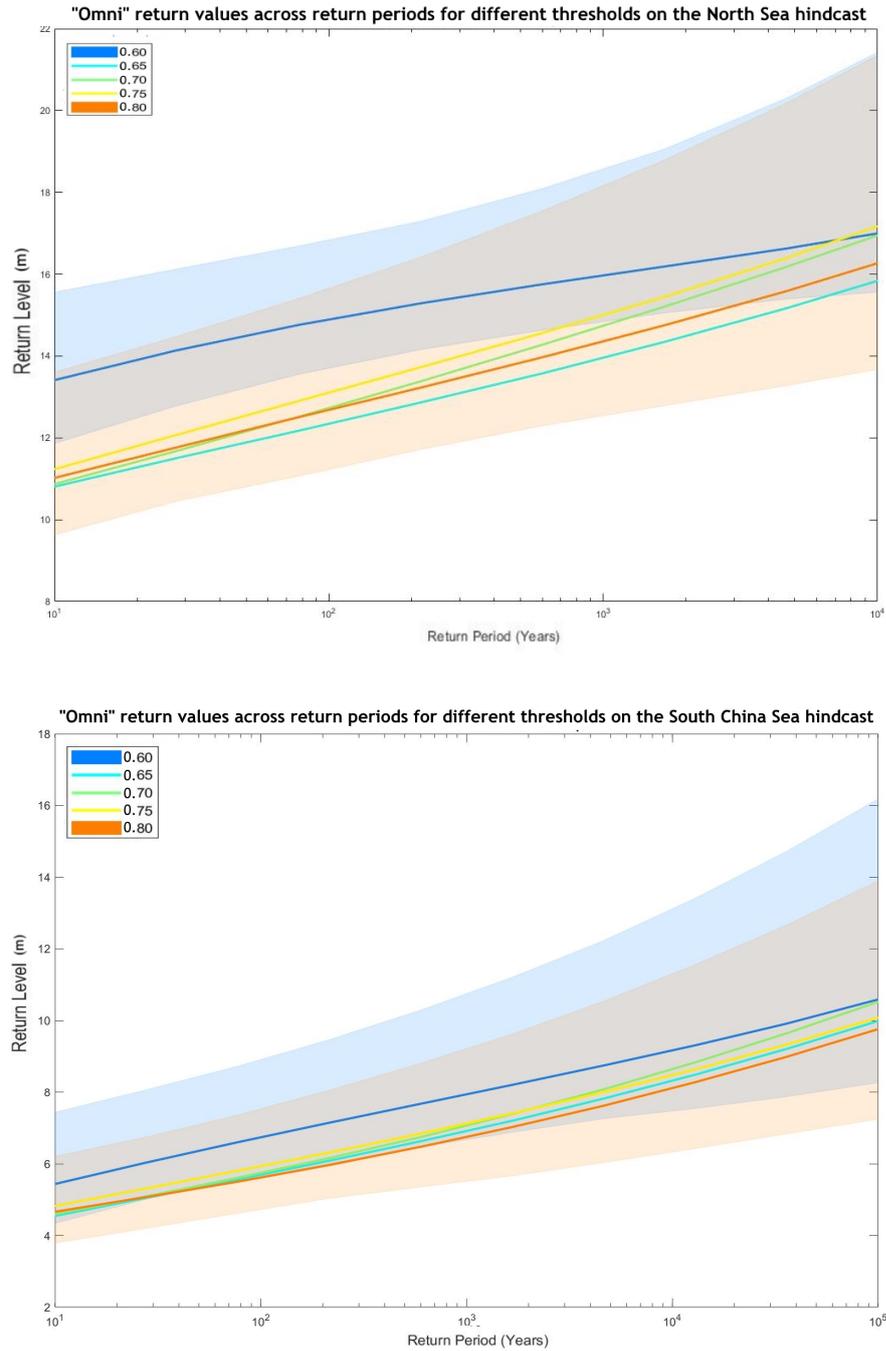
Figure 5.B.1: Omnidirectional posterior expected return median return value estimates for the North Sea hindcast HS (top) and South China Sea (bottom) introduced in Section 1.2 across increasing return periods (pictured in log-scale for readability). Here, Model 1 from Section 5.3.1 is used to fit the exceedances of thresholds obtained from non-exceedance probability ranging between 0.6 and 0.8, with 95% interquantile ranges shaded for the highest and lowest value.

in the inference. On the other hand, the highest values of non-exceedance proba-bility yield more uncertainty, characterised by wider interquantile ranges. We then proceed to consider the remaining values. As expected, results are more similar for these and a choice is not as obvious and remains arbitrary (see Section 2.1.2 and Coles, 2001). We decide to focus on values that yield the closest estimates, while still preserving as much of the original data as possible. Hence, we proceed to use thresholds obtained with non-exceedance probabilities of 0.7 and 0.65 for the Northern North Sea and the South China Sea datasets respectively.

# Chapter 6

# Further work

In this section, we present a few possible areas where the models proposed could be improved and extended, along with some suggestions on ways this could be achieved.

## 6.1 Bayesian inference and dimension-jumping models

Efficiency of prior and proposal distributions for MCMC algorithms are essential. First, one could consider different choices of prior distributions for the parameters of the models considered in Chapter 5. In a similar approach to the one used for P-spline models, we have already introduced some hyperparameters for the prior of coefficients in two-dimensional covariate models. In our case, this has allowed the imposition of a varying level of smoothness through a ridge penalty

term. Alternative formulations for this term are possible, for example based on a LASSO penalty or the approach proposed by Denison et al. (1998b).

Similar considerations are possible for the remaining formulation parameters. For example, we could consider adapting the approach recommended by Richardson and Green (1997) and introduce a hyperprior for the $\kappa$ parameters in the bivariate Von-Mises densities for the BARBaR models, as well as the decay parameter in the adapted Voronoi-based models.

Further studies could also investigate other aspects of the inferece. For example, one might want to systematically investigate the sensitivity of each of the models to the choice of starting values for the MCMC, as well as assess whether any differences emerge in the way sample size affects the various formulations.

A second extension would consider dimension-jumping models. In Chapter 4, we applied reversible-jump techniques for a one-dimensional covariate, both for spline and BARBaR based models. A first area for further development would focus on the BARS reversible-jump proposals, especially the approach used to update coefficients introduced in Section 4.3.3. It is worth remembering that, given the application to a complicated likelihood surface such as the generalised Pareto one, it is particularly important to propose candidate parameter values which aid the convergence of the MCMC to its stationary distribution. In our study, the presence of considerable variability between results for different sample realisations suggests that there is scope for refining the methodology.

The application of reversible-jump methodology to two-dimensional covariate domains would be natural. The challenge with adaptive spline methods lies in stable and efficient implementation of the RJ MCMC scheme and would require care-

ful adaptations of the proposals used for models with a single covariate.  The

work of Denison et al. (1998b) and Zhou and Shen (2001) would provide a useful

starting point for multi-dimensional BARS formulations, while the approaches by

Heikkinen and Arjas (1998) and Bodin and Sambridge (2009) would be helpful

for adaptive Voronoi tessellation models.  In order to implement reversible jump

for BARBaR models in two dimensions, one could start from the work by Zhang

et al. (2004) and Dellaportas and Papageorgiou (2006) on dimension changing ap-

proaches for multivariate mixtures of Gaussian densities.

For all the formulations considered, one would hope that the ability of model com-

ponents to adjust location and extent and increase or decrease in number would

regulate the smoothness of the curve estimated, allowing varying levels of rough-

ness in different areas of the covariate domain and reducing the need for global

penalty procedures.

## 6.2  Study of computational efficiency

In order to shine further light on the performance of the models proposed, one

could develop a study to test and compare model efficiency. As a way to evaluate

this, we could examine the effective sample size (ESS, Geyer 1992), which provides

a measure of the equivalent number of independent iterations that an MCMC

represents, defined as

$$ESS = \frac{n_{\text{iter}}}{1 + 2\sum_{k=1}^{\infty} c_k},$$

where $c_k$ is the autocorrelation of the MCMC chain at lag $k$, and $n_{\text{iter}}$ is the actual

chain length. Note that, in practice, the sum of the autocorrelations is truncated

when the autocorrelation $c_k$ drops beneath a certain level. The comparison of effective sample sizes for different models, given the same samples and inference methods, can provide some indication of relative computational efficiency. A different approach would consider computational run-times, although this requires the use of the same hardware resources for all chains.

While the work in this thesis focused on quality of inference for the proposed models, a study of this type for the approaches presented in Chapters 4 and 5 could provide useful information on their computational efficiency and ideas for possible improvements.

## 6.3    Model extension and threshold selection

The methodology proposed in Chapters 4 and 5 could be used for estimation of other parameters. For example, typical results of an extreme value analysis on a sample of data are generally not invariant to measurement scale. Introducing a measurement scale parameter may provide better estimation of return values (e.g. Wadsworth et al. 2010, Reeve et al. 2012). Let $Y \in \Omega$ be a random variable, independent and identically distributed according to a distribution function $F_Y$ defined over the domain $\Omega$. When measurement scale is taken into consideration, $Y$ is rescaled to $Y^* \in \Omega$ using the Box-Cox transformation with parameter $\zeta \in \mathbb{R}$, such that the transformed variable becomes

$$Y^* = \frac{Y^\zeta - 1}{\zeta},$$

$Y^* = \log(Y)$ for the case $\zeta = 0$, with the latter case being defined only for positive domains $\Omega^+$ for which $Y > 0$. Extending this to a continuous covariate-dependent model may provide additional benefits. It is worth noticing that, even in a stationary framework, there is confounding between the GEV shape parameter and the Box-Cox parameter $\zeta$. It is therefore possible that some simplifying assumption would be needed in order to perform inference for a covariate-dependent parameterisation (Wadsworth et al. 2010).

We acknowledged that using covariates in the threshold is essential to fitting covariate-dependent extreme value models. For this thesis, we opted to use the same covariate-dependent threshold for all models, in order to focus on the Poisson-GPD parameters estimation. While this was obtained using a P-spline formulation, one could investigate proposing a threshold using, in each case, the same formulation as for the other model parameters.

Since applications have shown that threshold specification in extreme value analysis is problematic, there is also scope for different approaches to threshold selection. Some authors prefer *whole-sample models* in which an extreme value tail model for threshold exceedances is combined with a model for non-exceedances. Since likelihood for such a model is defined for the whole sample, the threshold can be estimated as part of inference. Recent work in Shell (Randell et al. 2015b) and elsewhere has shown that piecewise models and mixture models provide a useful and computationally tractable approach to whole-sample inference. It is therefore desirable to consider whole-sample inference in the current research. Due to the considerable uncertainty in the choice of threshold, it is also often important to perform model estimation over multiple plausible thresholds (Randell et al. 2015a).

## 6.4 Models for multi-dimensional covariates

Extreme characteristics of the ocean environment vary with covariates such as direction, season and location. Incorporating this variation effectively and systematically within extreme value models is essential for methods to be useful. Raghupathi et al. (2016) show how spatial effects, captured by the longitude and latitude of a hindcast location, affect storm severity. It is then natural to consider 3-dimensional spatio-seasonal and 4-dimensional spatio-directional-seasonal covariate models. Hence one could investigate possible extensions of the models introduced to higher-dimensional domains.

Extending the models proposed to higher dimension can be relatively straightforward in form. For P-splines, one can easily use the approach from Section 5.2.2 to obtain a overall basis which is a linear combination of suitable bases over the domain of each covariate. As far as the BARBaR- and the Voronoi-based models are concerned, the adaptation requires the use of distance metrics defined on higher-dimensional domains. Although obtaining a suitable basis might be simple, computational efficiency and stability becomes even more essential with the introduction of additional covariates. It is then necessary to consider parameterisations that are complex enough to be practically useful, while still being computationally sustainable and yielding stable results. In order to implement analysis with higher-dimensional covariates, one should explore and resolve some of the statistical and computational issues that have emerged. The studies outlined in this section could provide useful insight and guidance for the development and implementation of such models.

# Bibliography

Aurenhammer, F. (1991). Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405.

Behrens, C. N., Lopes, H. F., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227–244.

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

Berg, M. d., Cheong, O., Kreveld, M. v., and Overmars, M. (2008). *Computational geometry: algorithms and applications*. Springer-Verlag TELOS.

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics*, 9(1):122–140.

Bodin, T. and Sambridge, M. (2009). Seismic tomography with the reversible jump algorithm. *Geophysical Journal International*, 178(3):1411–1436.

Boldi, M.-O. and Davison, A. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B*, 69(2):217–229.

Bottolo, L., Consonni, G., Dellaportas, P., and Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes*, 6(1):27–36.

Breivik, Ø., Reistad, M., and Haakenstad, H. (2007). A high-resolution hindcast study for the North Sea, the Norwegian Sea and the Barents Sea. In *10th International Workshop on Wave Hindcasting and Forecasting*.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 50:967–991.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: Series D*, 47(1):69–100.

Buhmann, M. D. (2003). *Radial basis functions: theory and implementations*, volume 12. Cambridge Uiversity Press.

Cabras, S., Castellanos, M. E., and Gamerman, D. (2011). A default Bayesian approach for regression on extremes. *Statistical Modelling*, 11(6):557–580.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Chavez-Demoulin, V. (1999). *Two problems in environmental statistics: capture-recapture analysis and smooth extremal models*. PhD Thesis. Department of Mathematics, Swiss Federal Institute of Technology, Lausanne.

Chavez-Demoulin, V. and Davison, A. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C*, 54:207–222.

Chavez-Demoulin, V. and Davison, A. (2012). Modelling time series extremes. *Revstat*, 10:109–133.

Coles, S. G, D. M. J. (1999). Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23.

Coles, S. (2001). *An introduction to statistical modelling of extreme values*. Springer.

Coles, S. and Walshaw, D. (1994). Directional modelling of extreme wind speeds. *Applied Statistics*, 43:139–157.

Coles, S. G. and Powell, E. A. (1996). Bayesian methods in extreme value modelling: a review and new developments. *International Statistics Review*, 64:119–136.

Coles, S. G. and Simiu, E. (2003). Estimating uncertainty in the extreme value analysis of data generated by a hurricane simulation model. *Journal of Engineering Mechanics*, 129:1288.

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

Coles, S. G., Tawn, J. A., and Smith, R. L. (1994). A seasonal Markov model for extremely low temperatures. *Environmetrics*, 5(3):221–239.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B*, 49:1–39.

Davison, A. and Smith, R. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B*, 52(3):393–442.

Davison, A. C. and Ramesh, N. I. (2000). Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society: Series B*, 62(1):191–208.

de Boor, C. (2001). *A practical guide to splines.* Springer-Verlag, New York (Applied Mathematical Sciences Revised Edition).

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

de Zea Bermudez, P. and Turkman, M. A. (2003). Bayesian approach to parameter estimation of the generalized Pareto distribution. *Test*, 12(1):259–277.

Dellaportas, P. and Forster, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 86(3):615–633.

Dellaportas, P. and Papageorgiou, I. (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing*, 16(1):57–68.

Denison, D. and Holmes, C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics*, 57(1):143–149.

Denison, D., Mallick, B., and Smith, A. (1998a). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society: Series B*, 60(2):333–350.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998b). Automatic Bayesian curve-fitting. *Journal of the Royal Statistical Society: Series B*, 60:330–50.

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika*, 88:1055–1071.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. CRC press.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, pages 89–102.

Eilers, P. H. C., Currie, I. D., and Durban, M. (2006). Fast and compact smoothing on multi-dimensional grids. *Computational Statistics and Data Analysis*, 50:61–76.

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

El Adlouni, S. and Ouarda, T. B. (2009). Joint Bayesian model selection and parameter estimation of the generalized extreme value model with covariates using birth-death Markov chain Monte Carlo. *Water Resources Research*, 45(6).

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, volume 33. Springer.

Ferro, C. A. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B*, 65(2):545–556.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distributions of the largest or smallest member pf a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24:180–190.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Taylor & Francis.

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science*, 7:473–483.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, 1:19.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 73:123–214.

Girosi, F. and Poggio, T. (1990). Networks and the best approximation property. *Biological cybernetics*, 63(3):169–176.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalised Linear Models*. Chapman & Hall.

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo. *Oxford Statistical Science Series*, pages 179–198.

Gunther, H., Hasselmann, S., and Janssen, P. (1992). *The WAM Model, Cycle 4*. DKRZ.

Hall, P. and Tajvidi, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science*, pages 153–167.

Hammersley, J. (2013). *Monte carlo methods*. Springer Science & Business Media.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning. Data mining, inference and prediction.* Springer.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Heikkinen, J. and Arjas, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, 25(3):435–450.

Heikkinen, J. and Arjas, E. (1999). Modeling a Poisson forest in variable elevations: a nonparametric Bayesian approach. *Biometrics*, 55(3):738–745.

Higham, N. J. (2002). *Accuracy and stability of numerical algorithms*, volume 80. Siam.

Hosking, J. R. (1998). *L-Moments.* Wiley Online Library.

Hosking, J. R., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261.

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B*, 52(1):105–124.

Jonathan, P. and Ewans, K. (2011). Modeling the seasonality of extreme waves in the gulf of mexico. *Journal of Offshore Mechanics and Arctic Engineering*, 133(2):021104.

Jonathan, P., Ewans, K. C., and Forristall, G. Z. (2008). Statistical estimation of

extreme ocean environments: The requirement for modelling directionality and other covariate effects. *Ocean Engineering*, 35:1211–1225.

Jones, M., Randell, D., Ewans, K., and Jonathan, P. (2016). Statistics of extreme ocean environments: non-stationary inference for directionality and other covariate effects. *Ocean Engineering*, 119:30–46.

Kalos, M. H. and Whitlock, P. A. (2009). *Monte carlo methods*. John Wiley & Sons.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21.

Leadbetter, M. and Rootzen, H. (1988). Extremal theory for stochastic processes. *The Annals of Probability*, pages 431–478.

Leadbetter, R., Lindgren, G., and H., R. (1983). *Extremes and Related Properties of Random Sequences and Processes*. New York.

MacDonald, A., Scarrott, C., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011). A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157.

MacKay, D. J. (1998). Introduction to monte carlo methods. In *Learning in graphical models*, pages 175–204. Springer.

Mardia, K. V. (2013). Statistical approaches to three key challenges in protein

structural bioinformatics. *Journal of the Royal Statistical Society: Series C*, 62(3):487–514.

Mardia, K. V. and Jupp, P. E. (2009). *Directional statistics*, volume 494. John Wiley & Sons.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models. 37.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computational machine. *Journal of Chemical Phusics*, pages 1087–1091.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Nobile, A. and Green, P. J. (2000). Bayesian analysis of factorial experiments by mixture modelling. *Biometrika*, 87(1):15–35.

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.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. (2009). *Spatial tessellations: concepts and applications of Voronoi diagrams*, volume 501. John Wiley & Sons.

Ouarda, T. and El-Adlouni, S. (2011). Bayesian nonstationary frequency analysis of hydrological variables 1. *Journal of the American Water Resources Association*, 47(3):496–505.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, pages 1065–1076.

Pauli, F. and Coles, S. (2001). Penalised likelihood inference in extreme value theory. *J. Appl. Statist.*, 28:547–560.

Perez-Cruz, F. (2008). Kullback-Leibler divergence estimation of continuous distributions. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 1666–1670.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131.

Powell, M. J. (1987). Radial basis functions for multivariable interpolation: a review. *Algorithms for approximation.*

Raghupathi, L., Randell, D., Ewans, K., and Jonathan, P. (2016). Fast computation of large scale marginal extremes with multi-dimensional covariates. *Computational Statistics and Data Analysis*, 95:243258.

Randell, D., Feld, G., Ewans, K., and Jonathan, P. (2015a). Distributions of return values for ocean wave characteristics using directional-seasonal extreme value analysis. *Environmetrics*, 26:442–450.

Randell, D., Turnbull, K., Ewans, K., and Jonathan, P. (2015b). Bayesian inference for non-stationary marginal extremes. *(Submitted to Environmetrics August 2015, draft at www.lancs.ac.uk/∼jonathan).*

Randell, D., Zanini, E., Vogel, M., Ewans, K., and Jonathan, P. (2014). Omnidirectional return values for storm severity from directional extreme value models: the effect of physical environment and sample size. *Proceedings of 33nd Inter-*

*national Conference on Ocean, Offshore and Arctic Engineering, San Fransisco*, OMAE2014-23156.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning.* MIT Press.

Reem, D. (2009). An algorithm for computing Voronoi diagrams of general generators in general normed spaces. In *2009 Sixth International Symposium on Voronoi Diagrams*, pages 144–152. IEEE.

Reeve, D. T., Randell, D., Ewans, K. C., and Jonathan, P. (2012). Accommodating measurement scale uncertainty in extreme value analysis of North Sea storm severity. *Ocean Engineering*, 53:164–176.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B*, 59(4):731–792.

Roberts, G. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4:337–358.

Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1):110–120.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B*, 60(1):255–268.

Roberts, G. O., Rosenthal, J. S., et al. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367.

Robinson, M. E. and Tawn, J. A. (1997). Statistics for extreme sea currents. *Applied Statistics*, 46:183–205.

Rosenblatt, M. et al. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

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

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.

SEAFINE (2015). *South East Asia Meteorological and Oceanographic Hindcast (SEAMOS) Fine Grid Hindcast Study Joint Industry Project*. Oceanweather Inc.

Sharef, E., Strawderman, R. L., Ruppert, D., Cowen, M., Halasyamani, L., et al. (2010). Bayesian adaptive B-spline estimation in proportional hazards frailty models. *Electronic journal of statistics*, 4:606–642.

Sharkey, P. and Tawn, J. A. (2017). A Poisson process reparameterisation for Bayesian inference for extremes. *Extremes*, 20(2):239–263.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B*, pages 1–52.

Smith, R. (1987). Approximations in Extreme Value Theory. *Technical Report No. 205*, University of North Carolina.

Smith, R. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4(4):367–393.

Sriram, K., Shi, P., and Ghosh, P. (2015). A Bayesian quantile regression model for insurance company costs data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Tierney, L. (1998). A note on Metropolis-Hastings kernels for general state spaces. *Annals of Applied Probability*, pages 1–9.

Undén, P., Rontu, L., Jarvinen, H., Lynch, P., Calvo Sánchez, F. J., Cats, G., Cuxart, J., Eerola, K., Fortelius, C., Garcia-Moya, J. A., et al. (2002). HIRLAM-5 scientific documentation.

Uppala, S. M., Kållberg, P., Simmons, A., Andrae, U., Bechtold, V. d., Fiorino, M., Gibson, J., Haseler, J., Hernandez, A., Kelly, G., et al. (2005). The ERA-40 re-analysis. *Quarterly Journal of the royal meteorological society*, 131(612):2961–3012.

Wadsworth, J. L. and Tawn, J. A. (2012). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *Journal of the Royal Statistical Society: Series B*, 74(3):543–567.

Wadsworth, J. L., Tawn, J. A., and Jonathan, P. (2010). Accounting for choice of measurement scale in extreme value modelling. *Annals of Applied Statistics*, 4:1558–1578.

Wallstrom, G., Liebner, J., and Kass, R. E. (2008). An implementation of Bayesian adaptive regression splines (BARS) in C with S and R wrappers. *Journal of Statistical Software*, 26(1):1.

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S., and Girolami, M. (2014). Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics and Probability Letters*, 91(2002):14–19.

Yee, T. W. and Stephenson, A. (2007). Vector generalized linear and additive extreme value models. *Extremes*, 10:1–19.

Zhang, Z., Chan, K. L., Wu, Y., and Chen, C. (2004). Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm. *Statistics and Computing*, 14(4):343–355.

Zhou, S. and Shen, X. (2001). Spatially adaptive regression splines and accurate knot selection schemes. *Journal of the American Statistical Association*, 96(453):247–259.