Prior distributions for Bayesian inference about extremes

Jeremy Gye Colman

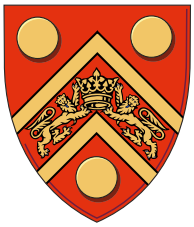
Submitted for the degree of Doctor of Philosophy

School of Mathematics and Statistics

Submitted April 2024

Supervisor: Professor Jeremy Oakley

University of Sheffield



I dedicate this thesis to my dear wife, Jill, without whose encouragement the work would not have started and certainly not have finished.

I thank my supervisor, Professor Jeremy Oakley, for the thoroughly helpful way he has conducted his supervisory rôle. I invariably left supervisory meetings feeling happier and more strongly motivated than I was at the start of them.

I also thank him for his willingness at the start to take me on as a postgraduate researcher of nearly three times the usual age of such workers and after over 50 years absence from academic work.

During my research I have had in turn two advisors, Professor Tim Heaton and Dr Kostas Triantafyllopoulos. Both have been very encouraging and helpful but I am pleased to say that I have not needed to call on them to deal with any problems in my relationship with my supervisor.

Finally, I thank my old friend Professor Rod Rainey for bringing to me the original problem to do with extreme waves in the North Sea that started me on this fascinating line of research. I thank him also for acting in a rôle that we certainly did not foresee at the start: that of the expert whom I interviewed to elicit his opinions about wave heights.

Summary

Extreme events although very rare can have very large, and very unwelcome effects. To obtain some reassurance that people are highly protected from such harm some understanding is needed as to the size and likelihood of extremes. But because extreme events are very rare, so is direct evidence of their likelihood. A particular extreme may never have been observed. And yet, we still wish to understand its likelihood so that we can protect ourselves from its effects.

A very beautiful theorem of probability theory appears to offer a way of assessing the probability of the occurrence of events well beyond the data that have been observed. Although its preconditions set an ideal that rarely if ever has been met, a wider theory has been developed that is directly applicable to practical problems. But there is still a catch: the forecasts of extremes are subject to huge uncertainty, sometimes rendering them practically useless.

One way of tightening the estimates of extremes that has been increasingly recognised over the past 30 years or so is to use the Bayesian paradigm for inference about extremes to bring into account informative priors, especially those obtained by elicitation from experts. That way specific observations of the quantity of interest are supplemented by other data.

Elicitation is a complex process that needs to be conducted according to strict protocols to yield from experts both valid estimates of quantities and of the uncertainty of those estimates. Typically only a small number of quantiles can be produced for each quantity. But for Bayesian inference the priors need to be full probability distributions.

This thesis examines options for stretching, so to speak, a small number of elicited quantiles into a full distribution. It points to the disadvantages of fitting quantiles to a parametric textbook distribution. It presents a new way of obtaining a wide variety of distributions defined as Gaussian processes over series of knots. It also considers the use of minimal distributions in the sense that a bare minimum of assumptions is made. Finally it reports the results of an elicitation exercise based on these principles.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Extreme events are rare, by definition, but they can give rise to extreme consequences 1
that are highly unwelcome, not to say catastrophic. There is a worldwide interest in understanding the size and likelihood of extreme events. The motivation for wishing to study extremes is to provide a high degree of protection against some calamity. To do that necessarily requires statistical methods: typically what is needed is the estimation of quantiles.

In this thesis we mainly consider the univariate case such as the occurrence of extremes 2
at a particular location. We discuss bringing in relevant information to supplement specific, maybe site-specific, observations, and some of that information will be derived from experience at other sites. We do not consider formal multivariate models although some of our conclusions will be relevant to those models.

Formally, the interest is in quantiles of the form x_p, known as the $1/p$ years **return** 3
level satisfying $F(x_p) < 1 - p$, where p is a very small annual probability and F is a typically unknown distribution function. The period $1/p$ is known as the **return period**. For example, in considering how high the sea dikes of Holland need to be, the Netherlands government set $p = 0.0001$ (de Haan, 1990). Any attempt to estimate the corresponding x_p must clearly involve extrapolation well beyond any historical data.

That in itself rules out use of methods based solely on the empirical distribution of the variable of interest.

4 Although the focus of this thesis is univariate extreme value theory it is worth referring to the practical applications of the theory. As the Netherlands example above shows, one class of application is to give assurance that a given protection against a catastrophe is highly unlikely to be overwhelmed. In addition to coastal sea surges, extreme value modelling is used in connection with all kinds of extreme weather events such as rainfall, flooding, and deep ocean structures such as oil rigs. The theory has been applied to the study of earthquakes, and to the maintenance and operation of nuclear installations. In the world of finance extreme value modelling is central to insurance companies' management of re-insurance risk. It has also been applied in the wider context of portfolio management.

§ 1.1 Extreme value theory

5 The core theory, known as **Extreme Value Theory (EVT)**, that can be applied to provide such estimates of extreme behaviour is a branch of the theory of probability that is concerned with extremes of very long sequences of random variables, $\{X_n\}$. In classical EVT it is assumed that the $\{X_n\}$ are independent and identically distributed (and therefore form stationary sequences). By contrast, in practical statistical applications it is common for the available data to exhibit dependency, non-stationarity and for the sequences of data to be relatively short. Developments of the theory have dealt successively with relaxing the classical conditions to enable modelling to address practical problems relating to extreme behaviour.

6 One of the principal results of EVT is that in a wide range of circumstances there is a class of limiting distributions that apply asymptotically to extremes irrespective of the underlying, but unknown, distributions of the data. Even the extremes of the observational data are likely to be far away from the asymptote, so the key to making use of the theory in practice therefore lies in finding ways to process the data in such a way that the conditions of the extreme value theorems apply reasonably closely. The major issue in this task is to determine in each case the threshold above which EVT might be considered to apply closely enough. Below that threshold it is accepted that observations will not necessarily, indeed will very likely not, be governed by the probability distributions from EVT and that such observations will not be directly relevant to estimating extreme behaviour. There is therefore a continuing tension between on the one hand ensuring that only relevant data are used and, on the other hand, minimising the variance of the results by using as much data as possible. Many

of the elaborations of EVT have been driven by the need to find ways of resolving that tension.

Whatever the means, the result in any case is that the data to which EVT is to be applied 7
are likely to be disappointingly meagre. It therefore makes sense to use legitimate means to add relevant information to the analysis. Bayesian inference, with its requirement for prior distributions that could encode expert knowledge to supplement the information contained in the observations, is a perhaps obvious choice. Nevertheless, it has in practice come surprisingly late to the world of extremes. Even now it is not unusual for engineers to work with point estimates of 10000 year return levels, and the relevant international standard for offshore oil and gas installations (International Organization for Standardization, 2015) refers to the *possibility* [my emphasis] of assessing the uncertainty of such estimates (using confidence intervals - a non-Bayesian concept) but does not demand any such assessment.

In work I undertook in collaboration with a prominent hydrodynamicist before beginning 8
the research reported in this thesis, Bayesian methods were used to examine extreme wave heights at a specific deep sea location (Rainey and Colman, 2014). That work showed how using priors that encoded even quite basic expert knowledge could substantially reduce the uncertainty in estimates of return levels. It was (independently) consistent with earlier findings (Egozcue, Pawlowsky-Glahn and Ortego, 2005).

The work presented in this thesis is based on the point of view that the Bayesian 9
approach using informative priors is likely to provide better quality inference about extremes than other methods. On that assumption, the immediate question is how to construct appropriate informative priors. There is a substantial literature on the elicitation of expert opinion for use in Bayesian inference quite generally. It is seen as a means of moving away from inappropriately non-informative or minimally informative priors. The idea is that experts in the subject matter of the inference can bring into the inference relevant sector knowledge that is independent of the source of observational data being analysed.

§ 1.2 Contribution to the field

This thesis examines options for the construction of informative priors based on expert 10
elicitation. We show that using informative priors can be expected to reduce quantified uncertainty in the resultant estimates, but that established extreme value methods are

subject to unquantified and, we argue, unrecognised uncertainty arising from failure to make full enough use of prior knowledge. In particular, one further source of uncertainty arises from what might be characterised as "inappropriately informative priors", that is, the use of reasonable seeming but arbitrary assumptions or restrictions in setting up the Bayesian model.

11 We develop a method for defining priors using a Gaussian process (GP) model. To do so we show how a PDF and its corresponding CDF can both be constructed by constraining GPs to meet the specific requirements of density and distribution functions.

12 We show, further, how the distribution function can be constrained to comply with quantiles elicited from experts. We show that by using the GP model there is no limit to the number of probability functions that are all compliant with the experts' data. The implication is a full analysis of uncertainty in the forecasts of extremes needs to take into account the wide range of compliant distributions that might be used as the basis for priors.

13 Although our work here is exclusively in the field of inference about extremes, our methods for quantifying uncertainty are applicable more widely, in any application in which priors had hitherto assumed a particular parametric form in the absence of clear evidence of that form's applicability to the particular case.

14 The GP-based priors can be computationally demanding limiting their use in face-to-face elicitation. To seek a much quicker method we examine the use of a minimal PDF and show how it can practicably be used in an elicitation.

§ 1.3 Structure of the thesis

15 The thesis is structured as follows:

- Chapter 2 is a review of the literature on EVT

- Chapter 3 discusses the use of Bayesian methods and particularly the case for using informative priors in Bayesian inference about extremes.

- Chapter 4 discusses the use of elicitation as a means to provide informative priors relevant to extremes;

- Chapter 5 describes the approach used in an important paper (Coles and Tawn (1996)) that applied Bayesian methods including priors derived from elicitation. We show a worked example of that approach using similar data.

- In Chapter 6 we review how previous authors have used Gaussian processes to generate wide ranges of prior distributions to reflect information elicited from experts. We show that the methods used there have some disadvantages, and we then suggest that some methods developed in a different context could be developed to attack our problem.

- Chapter 7 presents the new method that we have developed accordingly, showing how an unlimited number of valid priors can be constructed to encapsulate fully information elicited from an expert.

- Chapter 8 applies the methods developed in Chapter 7 to an example in which the modelling is restricted to observations above a high threshold. Those observations are assumed to be suitable for modelling using EVT. The case chosen for illustration is that of rainfall data that heavily overlaps that used in Coles and Tawn (1996) but is not exactly the same.

- Chapter 9 examines a much simpler approach than the method described in the previous chapter based on fitting a density defined on a small number of knots to tertiles and to a mode. It is much faster to compute and is therefore easier to use in face-to-face elicitation.

- In Chapter 10, we report an experiment in using elicitation to examine wave heights at a specific location. We report the results of using priors based on a specific parametric distribution and compare them with those using the simpler prior constructed as described in Chapter 9.

Chapter 2

Extreme Value Theory

This chapter summarises the principal theorems of EVT drawing attention to the sources 16
of bias and uncertainty to which they give rise when applied to statistical problems. Much of the theory was originally developed at a time when the use of Bayesian methods was uncommon. In recent years the development of Bayesian methods in the context of extremes has been more marked. In this chapter we refer to the literature on Bayesian methods but our major discussion of Bayesian methods is in Chapter 3.

A major source of both bias and uncertainty is the fact that however large the number of 17
observations of the phenomenon of interest, the set of those directly relevant to inference about extremes is unavoidably small. Datasets recording observations for as long as 100 years are very rare, yet there is a demand for estimates of 1000- and 10000-year events. It is not uncommon, or surprising, for there to be literally no actual observations of 1000-year or higher extremes in a given dataset. And even if the observations include such extremes, as sometimes they might, the observer cannot easily determine that such extremes are present in the data.

Bayesian inference, using informative **priors**, offers a way to bring into the inference 18
expert knowledge beyond that contained in the observational data. It is to be expected that the results, in the form of posterior distributions (**posteriors**) will be tighter than inference based only on sparse data. We refer in this chapter to an influential

paper, (Coles and Tawn, 1996), which we examine in more detail in Chapter 5 and use its methods to construct a worked example using similar data. We discuss there a comparison between Bayesian and frequentist methods that suggests that the former bring advantages particularly in respect of forecasts of extremes beyond the range of the observational data. Likewise we postpone to Chapter 4 discussion of the literature on the use of elicitation of prior knowledge from experts both in the context of extreme value estimates and more widely.

§ 2.1 Distribution of maxima

19 Classical EVT considers the behaviour of the maxima of sequences of independent identically distributed (iid) random variables $(X_1, \ldots, X_n)$ each subject to the same distribution function, $F(\cdot)$. It examines whether and, if so, under what conditions the distribution of $M_n = \max\{X_1, \ldots, X_n\}$ might converge as $n \to \infty$.

20 It is straightforward to write down the distribution function of M_n:

$$P(M_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x) = F^n(x). \tag{2.1}$$

As it stands, however, there are two problems with this representation. First, it leads to degenerate functions in the limit: for all x, as $n \to \infty$, $F^n(x) \to 0$, unless $F(x) = 1$. Secondly, the scale and location of the F^n may both vary as n increases, rendering comparisons difficult. A normalisation of scale and location potentially deals with both of these problems. Accordingly, EVT examines the conditions under which there exist sequences of real numbers $\{a_n > 0\}$ and $\{b_n\}$ such that $(M_n - b_n)/a_n$ has a non-degenerate limit distribution as $n \to \infty$, so that

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \tag{2.2}$$

(Gnedenko, 1943).

Definition 1. *Such a function G is known as an* ***extreme value distribution****.*

21 The central theorem of classical EVT is as follows:

Theorem 2. *(Fisher and Tippett, 1928) The class of extreme value distributions is* $G_\xi(\alpha x + \beta)$ *with* $\alpha > 0, \beta$ *real, where*

$$G_\xi(x) = \exp\left(-(1+\xi x)^{-1/\xi}\right), \quad 1 + \xi x > 0. \tag{2.3}$$

Definition 3. G_ξ *is known as* ***the Generalised Extreme Value (GEV) distribution*** *with* ***extreme value index*** ξ *(also known as the* ***shape parameter*** *of the distribution).*

Definition 4. *The underlying distribution,* F*, is said to belong to the* ***maximum domain of attraction (MDA)*** *of* G_ξ*, written* $F \in MDA(G_\xi)$*.*

MDAs are closed under affine transformations. That has the consequence that in 22
fitting the asymptotic distribution, G, to the data it is not necessary to make separate estimates of either the $\{a_n > 0\}$ and $\{b_n\}$ which are in effect bundled into the estimates of the three parameters of the distribution.

The class of extreme value distributions therefore forms a simple one-parameter family 23
apart from scale and location parameters. Including scale and location, the density of the generalised extreme value distribution is as follows:

$$G_\xi(x) = \exp\left[-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right], \quad 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0. \tag{2.4}$$

Although the single Equation 2.4 completely describes a single distribution function, 24
historically extreme value distributions were seen as belonging to three classes depending on the value of the extreme value index ξ as follows:

(i) For $\xi > 0$, use $G_\xi((x-1)/\xi)$ and set $\gamma = 1/\xi > 0$ then

$$\Phi_\gamma(x) := \begin{cases} 0, & \text{if } x \le 0; \\ \exp(-x^{-\gamma}), & \text{if } x > 0 \end{cases} \tag{2.5}$$

is usually called the **Fréchet** distribution class. It is unbounded above in support.

(ii) With $\xi < 0$ use $G_\xi(-(1+x)/\xi)$ and set $\gamma = -1/\xi$ then

$$\Psi_\gamma(x) := \begin{cases} \exp(-(-x)^\gamma), & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases} \tag{2.6}$$

is usually known in the context of EVT as the **Weibull** distribution class. (Properly, it is the reverse of the Weibull distribution as used in, for example, survival time analysis.) The support for the Weibull distribution has a finite upper bound, as does that of all distributions in its MDA. For that reason the Weibull distribution has been commonly used to model extremes when it is known e.g. for physical reasons that the phenomenon being modelled must be bounded.

(iii) With $\xi = 0$,

$$\Lambda(x) := \exp(-e^{-x}) \tag{2.7}$$

is called the **Gumbel** distribution. Many commonly used distributions, such as normal, log-normal and gamma, are in the MDA of the Gumbel distribution, including distributions with finite upper bounds and those which are unbounded in support. For these reasons it has been common to model extreme values using the Gumbel distribution alone. Apart from concern about the weakness of that argument for restricting modelling to just one of the three feasible classes, Cohen points out that even if the data are known to be from a distribution in the MDA of the Gumbel distribution, a quicker and better approximation may be derived by modelling with either of the other classes and that therefore a preference for modelling with a Gumbel distribution may be misguided (Cohen, 1982) .

25 The three classes of extreme value distribution are shown as densities in a standard form in Figure 2.1.

$$\begin{aligned} \text{Fréchet}: \quad & \Phi(x) = \begin{cases} 0, & x \leq 0 \\ \exp(-x), & x > 0 \end{cases} \\ \text{Weibull}: \quad & \Psi(x) = \begin{cases} \exp(x), & x \leq 0 \\ 1, & x > 0 \end{cases} \\ \text{Gumbel}: \quad & \Lambda(x) = \exp(e^{-x}) \quad x \in \mathbb{R}. \end{aligned}$$

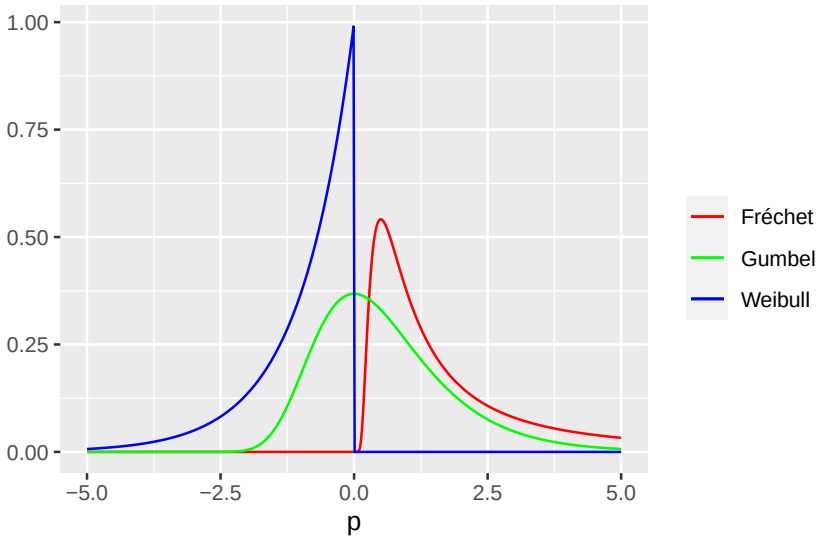


Figure 2.1: The three extreme value distributions as densities in standard form

Gnedenko established necessary and sufficient conditions (which we do not state here) 26
for a distribution function to belong to each class. Their force is that the behaviour of a distribution function away from its right-hand tail is irrelevant to whether it is in an MDA, and, if so, in which (Gnedenko, 1943). Note that this theoretical point is not the same as the more practical point made above relating to the exclusion from the analysis, in particular cases, of data that are unrelated to the physical processes underlying extremal behaviour.

The irrelevance of non-extreme data to the behaviour of the right-hand tail appears to 27
justify analysing extremes by looking solely at observations that lie above some high threshold. We shall see in Section 2.5 that such an approach is well-established in practice. Nevertheless, the difficulty of defining just how high the threshold needs to be has led to a number of approaches in which the whole dataset of observations (or, sometimes, all observations above a **low** threshold) is modelled using mixture models, as we shall see in Subsection 2.7.3.

§ 2.2 Analysis of block maxima

Historically the application of the theory outlined above was based on the idea that 28
assuming that an asymptotic distribution, G, say, exists then it might be reasonable to model the situation on the basis that distribution of the maxima of long blocks of observations, $M_{k,n} = \max\{X_{k,1}, \ldots, X_{k,n}\}$, where $k = 1, \ldots, r$ and n is large, actually

is G (Gumbel, 1958). In practice, it is often convenient, especially if the data show seasonal variability, to take the blocks as whole years.

29 Although this approach is well established, it is subject to serious, even obvious, problems, including, in particular:

(i) The assumption that the distribution of the observed block maxima is close to the asymptotic distribution may be questionable because the rate of convergence of the $F^n(a_n x + b_n)$ to the limiting form can be extremely slow (Hall, 1979);

(ii) by taking only one datapoint per year the number of observations used may be relatively small, leading to large sampling error in the results;

(iii) for the same reason, vast amounts of potentially informative data are unused (Thom, 1954); and, specifically,

(iv) in a year with a particularly high maximum no account will be taken of secondary maxima in that year, even if they are higher than the annual maxima in other years. That is an important omission as it potentially introduces a downward bias of large, but unknown, size into the model.

§ 2.3 Distribution of set of highest values

30 One way of improving on the block maxima method is based on the observation that the information about the tail that is provided by r block maxima cannot be greater, and may well be less, than that provided by the r highest order statistics of the same $\{X_i |\, i = 1 \ldots N\}$. If F is in the MDA of the extreme value distribution G, then the limiting distribution of the r largest $\{X_i |\, i = 1 \ldots N\}$ can be derived straightforwardly in terms of G, see, for example, Embrechts et al. (2013).

31 In addition to the idea that the r largest observations might provide more information than r block maxima, a second idea is to use the s highest observations from each block, yielding rs observations. That approach literally multiplies the number of data points

being used, addressing one of the problems of the block maxima approach. But it also potentially creates or increases another problem: that the data points may no longer reasonably be regarded as being drawn from an iid sequence, as the theory requires. In many practical applications, high observations close in time of the variable of interest have a common cause (for example, a single storm at sea will create very many high waves). If that happens several of the r highest in a given period may be strongly correlated. To address that problem of clustering a number of approaches have been developed based on the idea of identifying clusters and treating each cluster separately, perhaps by choosing a single observation from each cluster that is presumed to be independent of its counterpart in the other clusters.

§ 2.4 Making adjustment for dependence

As in all of the classical EVT, it is assumed that the $\{X_i\}$ are independent and so the 32
exceedances used need to be independent too. In many practical applications, however, the raw exceedances are known to be dependent (for example, wave heights during a single storm (see Section 2.3)). A common adjustment to deal with that situation is to examine the data for the existence of clusters of high values and to use only the maximum of each cluster (the **peak**) for what is therefore known as the **Peaks Over Threshold (POT)** method. Clusters may be identified by inspection, by knowledge of the physical circumstances (for example, in marine applications, when storms were known to have occurred), or by the application of some rule which might define a cluster, for example, as a period in which the observed variable has exceeded some threshold for a certain number of consecutive observations and then fallen below that threshold for a further stated number of consecutive observations (Davison and Smith, 1990).

Just as with the block maxima method, the use in the POT method of just one 33
observation in each cluster makes no use of whatever information is contained in all the other exceedances in that cluster, leading to unnecessarily high variance in its estimates. Fawcett and Walshaw argue that, even worse, it may introduce large biases, particularly into the estimates of return levels that are often of importance in practical applications. To overcome the problem of biased estimates of return levels, Fawcett and Walshaw (2007) describe a method to adjusting the standard errors of the parameters of the Generalised Pareto Distribution (GPD) along the following lines:

(i) the set of **all** exceedances is modelled with the GPD, ignoring dependence between exceedances, to produce maximum likelihood estimates of the parameters, σ and ξ, of the GPD and of the information matrix $\mathbf{H}$;

(ii) the covariance matrix, still ignoring dependence, is then approximately equal to $\mathbf{H}^{-1}$;

(iii) an adjustment as proposed in Smith (1990) is then made replacing that approximation with $\mathbf{H}^{-1}\mathbf{V}\mathbf{H}^{-1}$ where $\mathbf{V}$ is the covariance matrix of the likelihood gradient vector. $\mathbf{V}$ can be estimated as the empirical covariance matrix of the annual components of the log-likelihood sum.

34 In Fawcett and Walshaw (2006) the same authors propose another method for capturing the information from all exceedances whilst dealing with the issues raised by dependence. Examining hourly gust maximum wind speeds at a location in the Pennines, they find marked seasonal variation. They adopt a modelling approach based on 12 seasons corresponding to the calendar months and fit a GPD to each such season. In addition to seasonal variation they find that there is substantial short-term serial correlation in the hourly figures. To deal with that, they examine Markov chain models. They find that a second-order Markov chain model performs better than a first-order model in the analysis of clustering behaviour, but that if the quantity of interest is return level estimation then the simpler first order model would be adequate.

§ 2.5 Exceedances over a high threshold

35 As an alternative to looking at maxima or order statistics as such, hydrologists for many years have used empirical study of exceedances, that is $\{X_i | X_i > u\}$, where u is a threshold chosen empirically. Theoretical work leading to more soundly based statistical methods was undertaken by, for example, Todorovic and Rousselle (1971) and Todorovic and Zelenhasic (1970). Subsequently, and relevantly to the work described in this thesis, analysis based on thresholds has formed the basis of many applications of Bayesian methods to the study of extremes.

36 A major advance over the empirical work of hydrologists was the proof by Pickands that if the underlying distribution F of the $\{X_i\}$ was in the MDA of an extreme value distribution G, then for a sufficiently high threshold u, the distribution of the exceedances $\{X_i | X_i > u\}$ would be in the family he identified as the **GPD** with the parameters directly related to the corresponding G, and conversely (Pickands, 1975).That idea was developed into a modelling framework by Davison and Smith that has been widely used especially in hydrology (Davison and Smith, 1990). It depends

on the following theorem, sometimes known as the Second Fundamental Theorem of EVT

Theorem 5. *(Embrechts et al., 2013, Theorem 3.4.5) The following assertions are equivalent:*

(i)
$$F \in MDA(G_\xi); \tag{2.8}$$

(ii) *There exists a positive measurable function* $a(\cdot)$ *such that for* $1+\xi x > 0$,
$$\lim_{u \uparrow x_F} \frac{\overline{F}(u + a(u)x)}{\overline{F}(u)} = \begin{cases} (1+\xi x)^{-1/\xi}, & \text{if } \xi \neq 0 \\ e^{-x}, & \text{if } \xi = 0, \end{cases} \tag{2.9}$$
where $x_F = \inf(x : F(x) = 1)$ *and* $\overline{F}(\cdot) = 1 - F(\cdot)$*;*

(iii) *for* $x, y > 0, y \neq 1$,
$$\lim_{s \to \infty} \frac{U(sx) - U(s)}{U(sy) - U(y)} = \begin{cases} \dfrac{x^\xi - 1}{y^\xi - 1} & \text{if } \xi \neq 0, \\ \dfrac{\log x}{\log y} & \text{if } \xi = 0, \end{cases} \tag{2.10}$$
where $U(t) = F^{-1}(1 - t^{-1})$.

The right hand side of Equation (2.9) defines the tail of a GPD. In its most general form the GPD is given by
$$G_{\xi,\sigma}(x) = 1 - \left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-1/\xi}, \tag{2.11}$$
where μ is the location parameter and σ is the scale parameter.

The left hand side of equation (2.9) is equivalent to

$$\lim_{u \uparrow x_F} P\left(\frac{X-u}{a(u)} > x \mid X > u\right) \tag{2.12}$$

that is, the limiting probability that X exceeds x, given that X exceeds u. So the theorem gives the asymptotic distribution of scaled exceedances over a given (high) threshold, u, with $a(u)$ the scaling factor.

37 Although the question as to how high to set u is crucial, it applies equally to the subject of Section 2.6, so we will deal with it below in the following Section 2.7.

§ 2.6 Poisson point process

38 Closely related to the POT method is the idea of modelling the occurrence of extremes as a point process: the timing of occurrences is modelled as a Poisson point process (PPP) and the magnitude by an extreme value distribution. It turns out that this idea leads to a representation of extremes from which the results of the classical theory may all be derived.

39 Suppose Equation(2.2) holds. Then $n\{1 - F(a_n x + b_n)\} \to -\log G(x)$. If $X_1, \ldots, X_n$ is a random sample from F then let $Y_{n,i} = \frac{X_i - b_n}{a_n}$. A point process P_n can then be defined on $\mathbb{R}^2$ with points at $(\frac{i}{n+1}, Y_{n,i})$. It can then be shown that in a subset of $\mathbb{R}^2$ (corresponding to values of X over a high threshold, u) P_n converges as $n \to \infty$ to a process P with an intensity function

$$\lambda(x) = \frac{1}{\sigma}\left\{1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right\}_+^{-\frac{\xi+1}{\xi}}, \tag{2.13}$$

where $z_+ = \max(z, 0)$, which is precisely the GEV distribution of Equation(2.3) with scale and location parameters included (Smith, 1989).

40 Figure 2.2 illustrates P_n of a lognormal density for $n = 10^4$.

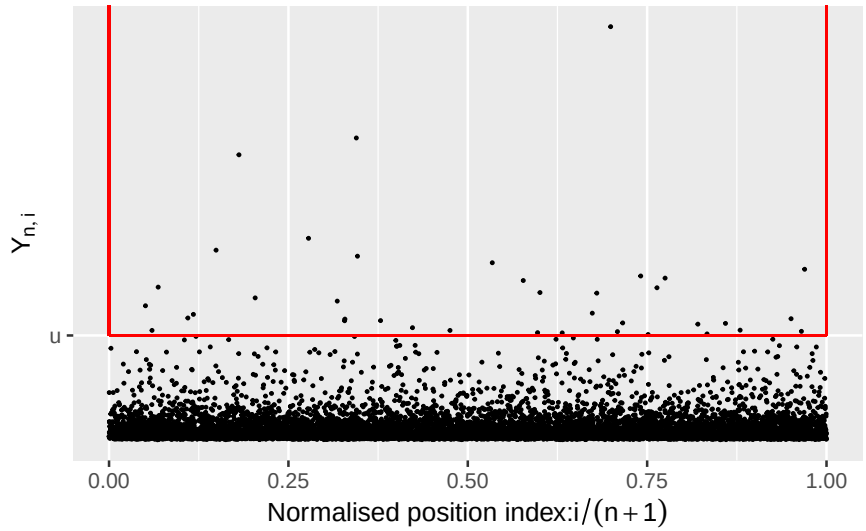


Figure 2.2: Poisson point process: $\{P_n : n = 10000\}$ lies within the red lines

§ 2.7 Setting the height of the threshold

There is a bias/variance trade-off in setting the height of the threshold u whether for the POT or the PPP approach: too low and the extreme value theorem does not apply; too high and there may be so few exceedances that the variance of the results may be impracticably high. A wide variety of approaches to that question have been reported which can be grouped into the categories described below (adapted from Scarrott and MacDonald (2012)): 41

(i) judgement aided by graphical diagnostics of the data;

(ii) rules of thumb, such as setting u to the 90% quantile of the observations;

(iii) mixture models to draw on non-extreme information to account for uncertainty in the choice of threshold.

2.7.1 Graphical diagnoses

42 Data visualisation can help inform judgment on threshold selection. Many graphical methods have been used. We shall see below (Chapter 5) an example of the use of a graphical method. But there are problems in relying on such approaches because they depend on personal interpretation of the plot, and some plots can be open to a wide range of interpretations.

43 One common approach is to choose u by inspection of the plot of estimates of $e(u) := \mathrm{E}(X - u \mid X > u)$ which is known as the **mean excess function** of X. It can be shown that if the observations are truly iid, for $u < x_F$, $e(u)$ is linear in u for all sufficiently high values of u. A starting point for the POT analysis can then be chosen as a value of u above which the plot appears to be linear. In practice the judgement as to where that point lies can be open to considerable uncertainty and therefore subjectivity. As can be seen in Figure 2.3, which is a plot of mean excess significant rainfall at a location in south-west England. (This figure is used in Chapter 5 as part of the detailed description of Coles and Tawn (1996).) It should be noted that

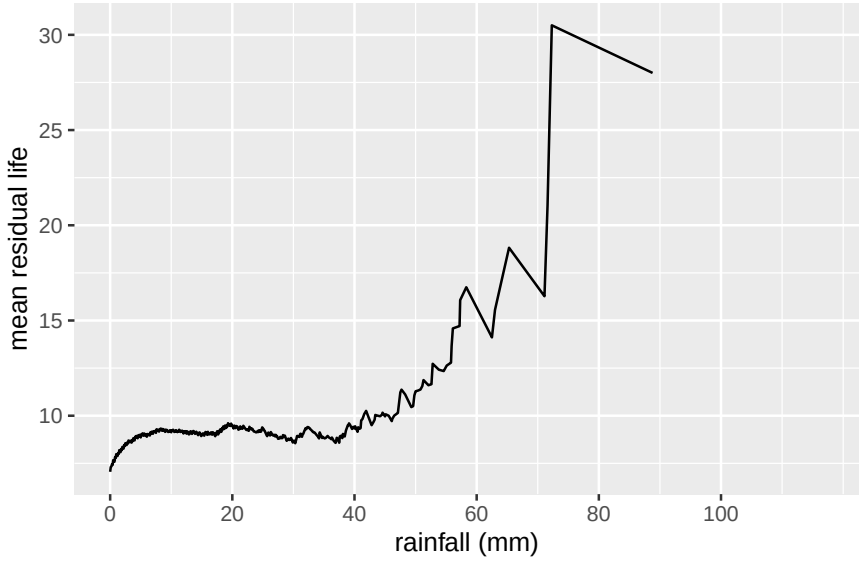


Figure 2.3: Example of a mean residual life plot

Figure 2.3 has been produced using R code written as part of the work on this thesis, and therefore without using functions supplied as part of an R package. Such functions often include error bars automatically calculated based on frequentist confidence interval calculations. We do not use them as they do not fit with the Bayesian ethos of this thesis.

There are several reasons for this effect: 44

(i) since u is high, the number of observations $\{X|X > u\}$ will generally be low giving a high variance to estimates of $\mathrm{E}(X - u|X > u)$;

(ii) the data points are unlikely to be *exactly* iid; and

(iii) may include the effects of more than one generative process.

These problems with using graphical methods are compounded if the inherent uncertainty 45
in the choice of u is ignored in subsequent analysis. At the very least it is necessary to test the eventual results for their sensitivity to the choice of u.

2.7.2 Rules of thumb

The idea of the rules of thumb in this context is that if the distribution of observations 46
lies in the domain of attraction of an extreme value distribution then, given a sample $\{X_1, \ldots, X_n\}$ of size n, there is a function of n that determines the order statistic k such that $X_{(1)} \geq \cdots \geq X_{(k)} \geq \cdots \geq X_{(n)}$ to define the threshold to be used. For that idea to work well some well-founded assumption needs to be made as to the extreme value distribution applying to the observations. In practice, rules are sometimes used, such as the simple statistic that k should correspond to the 90% quantile of the $\{X_i\}$ (DuMouchel (1983)), despite a lack of theoretical justification. And even when there is a theoretical justification the choice of k may not be straightforward. Graphical methods may be needed to explore the range of options.

2.7.3 Mixture models

As shown above, there appear to be good reasons when analysing extremes for ignoring 47
non-extreme data, including the fact that non-extreme events are often the result of physically different processes from those generating extremes, and that classical extreme value theory does not apply to non-extreme events. In practice the point at which events become, or cease to be, extreme may be far from obvious. As we mentioned in Section 2.1, a number of approaches have therefore been developed in which all observations (or perhaps above a **low** threshold) are modelled using a mixed distribution. That distribution would be chosen so that for high observations (that is, for the **tail**) some kind of extreme value distribution would dominate. And for lower observations (the **bulk**) some other kind of distribution would dominate.

48 Although these approaches have much in common, it is convenient here to examine them under three headings:

(i) parametric bulk models;

(ii) semi-parametric bulk models; and

(iii) nonparametric bulk models.

These headings are roughly drawn, different authors having differing views on which category particular models should inhabit.

2.7.4 Parametric bulk models

49 The idea here is that the mixture comprises a parametric distribution for the bulk below the threshold, and a tail model, the GPD, for example, for those above the threshold, which is itself a parameter to be estimated. That avoids the subjectivity involved in using graphical methods to choose the value of the threshold. If a Bayesian approach is being used, it also permits multiple choices of threshold to be handled straightforwardly.

50 Behrens et al. (2004) is an example of such a Bayesian approach. They use a gamma distribution to define the priors in the cases described in the paper and find that when, instead, a Weibull distribution is used as the basis for the priors the inferences about extreme values were unaffected. They recognise that a more thorough study of the effect of other distributions below the threshold would be helpful.

51 An example of an approach that was not Bayesian, although it did depend on simulation, is Frigessi et al. (2002). It was an early attempt to define a mixture model with a continuous transition between bulk and tail models (assuming that there was a lower bound on the support). Instead of defining a threshold with one bulk distribution modelling to the left and a tail distribution modelling to the right, both distributions are defined on the whole support, the relative weights attached to each being determined by a transition function that varies continuously. The Weibull distribution was suggested as the basis for the leftmost modelling, in part because it has a light tail. The rightmost

part would be modelled by the GPD. The transition function, $p(\cdot)$, could be a CDF. The approach could be seen as a generalisation of that in Behrens et al. (2004) with the transition function a simple step

$$p(x, \theta) = \begin{cases} 0, & \text{if } x < \theta; \\ 1, & \text{if } x \geq \theta, \end{cases} \tag{2.14}$$

where θ is a parameter defining the threshold.

Scarrott and MacDonald (2012) point out that the definition of left and right distributions 52
over the whole range may in practice give too much weight

(i) to the tail distribution in the bulk because the GPD reaches its pole at zero, and

(ii) to the bulk distribution in the tail because the weight given to the bulk distribution tends to zero only asymptotically.

2.7.5 Semiparametric bulk models

Carreau and Bengio (2009) construct a hybrid Pareto model for asymmetric fat-tailed 53
data. It comprises a mixture of m hybrid Pareto distributions that are themselves mixture distributions. Each of these components is built by stitching a GPD tail (with parameters α, β and ξ) to a Gaussian (with parameters μ and σ), so that at their junction the resulting density and its derivative are continuous. Those two constraints reduce the number of free parameters to three, chosen as ξ, μ and σ. The remaining two parameters, α and β, are then functions of the three free parameters.

Using a mixture of m hybrid Paretos does not require a threshold to be either set 54
or estimated. It has been shown that the tail behaviour of the mixture of m hybrid Pareto distributions is determined by the component that has the heaviest tail (the **tail dominant** distribution) (Kang and Serfozo (1999)). So, if a threshold is required to be stated it is suggested that the junction between the Gaussian and GPD in the tail dominant distribution be chosen. The other components in the mixture of hybrid Paretos were shown to out-perform a mixture of Gaussians and one hybrid Pareto at modelling the target distribution away from the tail.

do Nascimento et al. (2012) propose a mixture comprising a bulk distribution for 55

observations below the threshold, u, made up of a weighted mixture of k gamma densities and a tail distribution, for observations above u as a single GPD. Inference is performed using Markov chain Monte Carlo (MCMC) methods. The resulting predictive density gets naturally smoothed out taking into account all possible values of the threshold irrespective of any discontinuities at the threshold, u, of any individual mixture density.

2.7.6 Nonparametric bulk models

56 Tancredi et al. (2006) argue that if the aim is to give a single estimate of the return level corresponding to a given return period, and to attach a measure of uncertainty to the estimate, then the threshold, $u(x)$, should depend on the data, x, and is part of the estimation procedure. They model the whole set of observations (above a low threshold) with a mixed distribution comprising a set of uniform distributions for the bulk of the observations below the threshold $u = u(x)$, and an extreme value distribution for the tail. The arrangements are illustrated in Figure 2.4. The image shows that the bulk distribution is a mix of uniform distributions whilst the tail is an extreme value distribution. The bulk distribution is therefore nonparametric and its CDF is a piecewise linear approximation. The widths and the number of uniform distributions are allowed to vary. That bulk distribution is then combined with a PPP distribution for the tail (thus avoiding the awkwardness of the dependence of the GPD scale parameter).

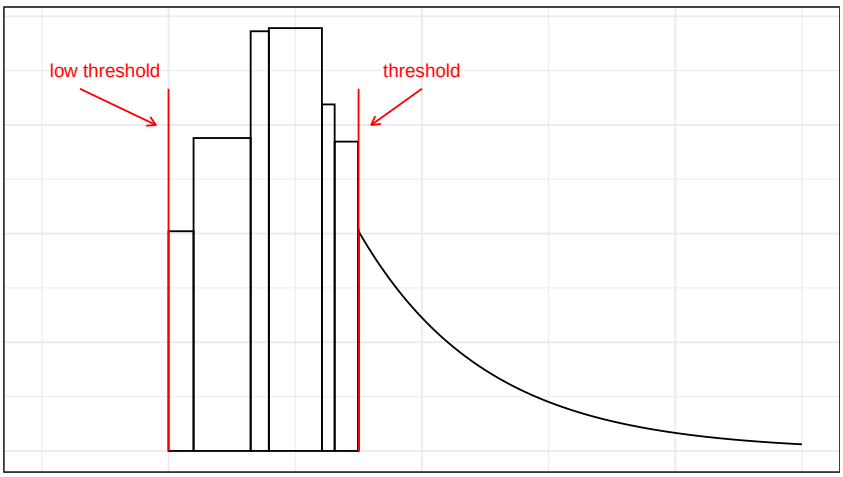


Figure 2.4: Example of mixture model proposed by Tancredi et al. (2006)

57 MacDonald et al. (2011) also propose a model in which the threshold is a parameter to be estimated rather than a constant to be determined in advance. Below the threshold,

observations are assumed to follow a non-parametric density. Above it either a GPD or a PPP applies. The authors point to the potentially simpler computation of their proposed model compared with that of Tancredi et al. (2006). As with the method proposed by do Nascimento et al. (2012), the possible discontinuity at the threshold in any individual mixture model density does not translate to a discontinuity in the estimated posterior predictive density obtained through MCMC sampling: any individual discontinuities are, in effect, integrated out.

The mixture model in MacDonald et al. (2011) is compared with the then more usual 58
model with a threshold determined in advance. It is found, unsurprisingly, that the credibility intervals with the mixture model are wider than the credibility intervals with the threshold fixed. The comparison also shows that the return levels are estimated more robustly with the mixture model compared to the fixed threshold approach, despite the fact that the thresholds are sensitive to the choice of prior. The paper concludes that the mixture model gives a more reliable uncertainty assessment to attach to the return level point estimate.

Northrop et al. (2017) use Bayesian cross-validation to address the bias-variance trade- 59
off in choosing the threshold in the context of prediction of future extremes. Rather than choosing a single threshold, the idea is to treat different thresholds as providing competing models for extreme prediction, and rank each model for its predictive ability. Predictions of extremes are then obtained by averaging over all the models, weighted by their individual ranks.

§ 2.8 Dependent Sequences

Various authors have examined the scope for relaxing the condition that the sequence 60
$X_1, \ldots, X_n$ of random variables with common distribution F must be independent. Loynes showed that if the assumption of independence is replaced by the requirement that the sequence $\{X_i\}$ be a stationary stochastic process then similar results to the classical theory would hold at least if X_i and X_j are *nearly* independent when $|j - i|$ is large (Loynes, 1965). The concept of being "nearly independent" is formally defined as the sequence $\{X_i\}$ being uniformly mixing:

Definition 6. *Let $\{X_n\}$ be a sequence of independent iid random variables and let $g(k)$ be a function such that $g(k) \to 0$ as $k \to \infty$. Suppose $A \in \mathfrak{B}(X_1, \ldots, X_m)$ and*

$B \in \mathfrak{B}(X_{m+k+1}, \ldots, X_{m+k+n})$, *for some* m, *then the sequence* $\{X_n\}$ *is* ***uniformly mixing*** *if*

$$|P(A \cap B) - P(A)P(B)| < g(k) \tag{2.15}$$

where $\mathfrak{B}$ *denotes the* σ*-field generated by the random variables indicated.*

61 Considering the case in which the $\{X_i\}$ are a stationary normal sequence, Berman showed that the distribution of the maxima would be in the same MDA as that of a sequence of independent normal random variables provided certain simple correlation restrictions on the sequence were satisfied (Berman, 1964).

62 Leadbetter et al. (1983) generalised these results by defining a weaker mixing condition than that used by Loynes (1965), denoted as $D(u_n)$, that makes precise the loose concept of the stationary sequence $\{X_i\}$ being "not too dependent". The definition of $D(u_n)$ relates to the joint distribution functions $F_{i_1 \ldots i_n}(x_{i_1}, \ldots, x_{i_n})$ of sets of random variables $\xi_{i_1}, \ldots, \xi_{i_n}$. For brevity, we write $F_{i_1 \ldots i_n}(u)$ for $F_{i_1 \ldots i_n}(u, \ldots, u)$.

63 **Definition 7.** *The condition* $D(u_n)$ *will then be said to hold if for any integers* $i_1 < \cdots < i_p$ *and* $j_1 < \cdots < j_{p'}$ *for which* $j_1 - i_p \geq l$, *and any real* u,

$$|F_{i_1,\ldots,i_p,j_1,\ldots,j_{p'}}(u_n) - F_{i_1,\ldots,i_p}(u_n)F_{j_1,\ldots,j_{p'}}(u_n)| \leq \alpha_{n,l},$$

where $\alpha_{n,l_n} \to 0$ *as* $n \to \infty$ *for some sequence* $l_n = o(n)$.

The following theorem, stronger than that in (Berman, 1964), was then proved, where $M_n = \max(\xi_1, \ldots, \xi_n)$:

Theorem 8. *(Leadbetter et al., 1983, Theorem 3.3.3) Let* $\{X_i\}$ *be a stationary sequence and* $a_n > 0$ *and* b_n *given constants such that* $P\{a_n(M_n - b_n) \leq x\}$ *converges to a non-degenerate distribution function* $G(x)$. *Suppose that* $D(u_n)$ *is satisfied for* $u_n = \frac{x}{a_n} + b_n$ *for each real* x. *Then* $G(x)$ *has one of the three standard extreme value forms (as described in Section 2.1).*

§ 2.9 Trends and seasonality

As noted above, classical EVT does not apply if the data are subject to a trend 64
or to seasonal variation about a stationary position. Various approaches have been described aimed at overcoming such problems in individual cases. There are well known techniques for de-trending and for smoothing data. What they have in common is that they introduce into the eventual estimates further uncertainty deriving from uncertainty about the choices of de-trending or smoothing technique that are made. Any thorough quantification of the uncertainty in the estimates needs to account for that added by such adjustments.

§ 2.10 Conclusions

We have seen that classical EVT was derived to fit an ideal world. We have described 65
subsequent developments that extend its validity to more realistic scenarios but as is usual in statistical modelling it cannot be expected that even these later models will describe the real world exactly. When these models are used to forecast extreme events we must expect them to be very uncertain.

Chapter 3

Bayesian methods: choice of priors

The main theme of this thesis is the use of the Bayesian paradigm in applying the theorems of extreme value theory summarised in Chapter 2. Bayesian inference depends upon making use of prior information: it requires the analyst to express prior information in the form of a probability distribution. The word "prior" does not necessarily mean earlier in time: Jaynes (2003, Section(4.1)) states that "any additional information beyond the immediate data ... of the current problem is by definition prior information". 66

Although Bayesian methods have been known for very many years, until the late twentieth century their use was severely constrained by computational problems because purely analytical methods were usually not feasible for the problems arising in practice, so that computationally heavy numerical methods were required. That situation has been transformed by two developments: first, the enormous increase in the speed, capacity and availability of computers, and secondly, by the development of theoretical advances in simulation models and specialised software enabling users to write the simulation models that Bayesian methods require. 67

In this chapter we consider the choice of priors and why it is especially important to 68

use informative priors when estimating extremes.

§ 3.1 Importance of priors in analysis of extremes

69 In many applications of Bayesian methods, the volume of data is so great that the information it contains swamps any reasonable prior the analyst may choose, so that very flat, uninformative priors may be used. But, as noted above, in extreme value problems the data that is most relevant to estimating extremes is usually very scarce. For example, in my earlier work on extreme wave heights referred to in Chapter 1, and summarised below, the dataset comprised over 350000 observations of wave heights made over a total of 45 years at the location in question. That meant that in the method of annual maxima (see Section 2.2) only 45 of them could be used. Even using the POT method (see Section 2.4), fewer than 200 observations were realistically available for use. And within these very sparse datasets the observations that were most relevant to estimating extremes were inevitably the even sparser set of extremes of the observations. (Rainey and Colman, 2014)

70 In such cases the choice of prior distribution is likely to have a great effect on the results of the analysis, and the justification for using non-informative priors becomes highly questionable. The formulation of priors also takes on great importance as it provides the means for bringing into account relevant information that is not present in the direct experimental data. Sources of knowledge independent of the data are frequently available, such as known physical constraints, understanding of the generative mechanism of the quantities of interest, and information from other locations (Coles and Tawn, 1996). It therefore makes sense to use as much valid prior information as can be obtained. An important specific source of information that goes beyond well-established knowledge is the judgement of experts.

§ 3.2 Non-informative priors

71 Many of the papers that examine Bayesian approaches to extremes focus on modelling issues rather than the particular data or results, and for simplicity use non-informative priors. Engelund and Rackwitz (1992) consider the effect on inference of various choices of non-informative priors, conclude that some well known non-informative priors can give rise to nonsensical results in inference about extremes and propose some minimally informative priors to avoid that problem, but the priors they suggest are not based on the specific science of the phenomenon of interest. In parallel, in the related field of survival analysis, several authors described schemes for eliciting priors for parameters

in the relevant distributions: Berger and Sun (1993); Singpurwalla (1988); Singpurwalla and Song (1989).

The focus of Smith and Naylor (1987) is also on a modelling issue, namely, a comparison 72
between the Bayesian and maximum likelihood approaches for estimation of a Weibull distribution. In their Bayesian approach, two priors are used, one at least of which might have been based on relevant scientific knowledge. Although the authors advocate trying to incorporate any available physical information in priors, they make clear that their two priors are not intended to represent real physical information.

§ 3.3 Minimally informative priors

The idea of basing priors on very general scientific knowledge is also described by 73
de Zea Bermudez et al. (2001), who propose constructing separate models depending on each of the three classes of extreme value distribution based on what they describe as a very careful elicitation of the extreme value index ξ (de Zea Bermudez et al., 2001; de Zea Bermudez and Turkman, 2003). By that they mean expert prior knowledge of the likely lightness or heaviness of the tail of the underlying distribution. Similar approaches are taken by, for example, Cooley (2005), Ortego et al. (2010), and Renard et al. (2006).

In 2014 I was approached by a leading hydrodynamicist, Professor Rod Rainey FREng, 74
with a problem about forecasting extreme wave heights at a specific location in the North Sea in the context of the design of oil rigs. One of the regulatory requirements that must be met before the design for an ocean oil rig can be approved is that it must be shown to be safe in the event of 10000-year waves, that is, waves of a height that might be expected to occur in any year with a probability lower than 1 in 10000. One major risk that faces an oil rig is that the crest of a wave might catch the deck of the rig and knock the rig over. It follows that the estimate of the 10000-year wave determines the minimum height of the deck, and is a very strong influence on the cost of the whole rig.

Rainey had been provided with a report by a consultant that offered a variety of ways 75
of calculating the 10000-year wave height as a single figure at a specific location west of Shetland and then took the arithmetic mean of the results again presented as a single figure. Rainey doubted the validity of some of the methods the consultant used. He was not convinced that it was reasonable for those results to be taken into the mean.

He was looking for a more rational approach. That led us to consider ways of applying EVT to wave heights at the west of Shetland location.

76 In ocean hydrology a standard approach to forecasting the height of extreme wave crests is to begin by forecasting what is known as the **significant wave height**, denoted by the symbol H_s. Significant wave height was originally defined as the mean of the highest one-third of the waves, as measured from the trough to the crest of the waves but is now formally defined as $H_s = 4\sqrt{m_0}$ where m_0 is the variance of the wave heights observed during a given period. In practice wave heights are not measured directly but are calculated from measurements of the vertical acceleration of the measuring device. Significant wave height is thus a statistical description of individual wave heights rather than measurement of any individual wave during the period of measurement.

77 The length of the period over which variance is measured may vary according to the use to be made of the data. In the offshore oil industry a period of 20 minutes is used. For other uses longer periods may be used. The length of the period has an effect on the height and volatility of the H_s measurements. In a short period high individual waves have a greater impact on the variance than in a long period. So a short period will give rise to more volatile values of H_s than a long period and will permit higher values of H_s to be attained.

78 In our elicitation exercise discussed in Chapter 10, and for reasons explained there, we shall not be using forecasts of individual wave heights. For completeness, however, we record here that where such forecasts are required there is a second stage that makes use of a probability distribution of individual wave heights and wave crests for given significant wave heights. The current practice in the oil industry is to use the Forristall distribution which is based on Stokes's second order theory. Its distribution function is:

$$F(h) = 1 - \exp\left[-2.263\left(\frac{h}{H_s}\right)^{2.126}\right] \tag{3.1}$$

where h denotes individual wave height. The distribution is a refinement of the Rayleigh distribution which is derived from theoretical analysis (Longuet-Higgins, 1953):

$$F(h) = 1 - \exp\left[-2\left(\frac{h}{H_s}\right)^{2}\right] \tag{3.2}$$

and was obtained with the parameters estimated from measurements of hurricanes in the Gulf of Mexico (Forristall, 1978). In a study of the Northern North Sea, Kvingedal et al. (2018) state that the Forristall distributions for individual wave height and crest

heights in long-crested seas "are considered to provide the most appropriate descriptions" of these measured wave data.

Looking now at the specific west Shetland location, we noticed that although the 79
data available about significant wave heights there comprised some 350000 data points covering 45 years, only a very small number of these were relevant to forecasting extremes. Our first model was an analysis of annual maxima as described in Section 2.2 using just the 45 corresponding data points. We fitted the generalised extreme value distribution, Equation 2.4, using (MCMC) methods with the modelling language OpenBUGS, obtaining posterior densities for the three parameters: the location, μ, the size, σ, and the shape, ξ.

Initially we had thought that we lacked prior information about all three parameters and 80
therefore based our simulation on very flat priors for each of them. When we calculated the return levels implied by our posterior estimates of the parameters, however, we realised immediately that the results were impossibly pessimistic. For instance, the forecast 1000-year significant wave height, H_s, fell in a 95 % credible range that extended as far as over 40 metres. Holliday et al. (2006) suggest that the highest value of H_s ever recorded by scientific instruments (near Rockall, west of Scotland) was 18.5 metres and figures of around 18 metres have been recorded in severe hurricanes.

A simple way to prevent the forecast extreme waves from reaching physically implausible 81
values was to note that the shape parameter ξ yields unbounded forecasts if $\xi > 0$. We therefore re-ran the simulation with the prior $\xi \leq 0$. The results of that simulation were no longer physically implausible.

Our work was intended only to show that a simple Bayesian model was capable of 82
yielding plausible estimates of extreme values. It would, of course, be possible to add complications to the model. For instance, wave heights are always constrained by the depth of the water. We could therefore reasonably have used a prior that limited the estimated extremes to the depth of the sea at the location in question, 135 metres. We could also take into account such aspects of dependance, seasonality, clusters and thresholds as described above. To do so would not lead to an especially complex MCMC model.

Another, earlier, example of the use of more detailed priors in a study of extreme coastal 83
waves is given by Egozcue et al. (2005). Modelling the occurrence of storms as PPP, and

maximum wave height exceedances per storm with the generalised Pareto distribution, they develop a joint prior for the shape parameter ξ and the scale parameter β of the Pareto distribution to give effect to constraints on wave height that they consider, *a priori*, to apply. These constraints relate to:

(i) known bound to wave height (depth of the sea at the specific location),

(ii) the probabilities of maximum wave height exceeding a given high level, falling below a given low level, and attaining a given middle level;

(iii) the rate of decline of the density of excesses. That is obtained by setting a minimum value for the extreme value index ξ.

§ 3.4 Conclusion

84 The cases just described show that when using Bayesian methods to forecast extremes, even simply constructed semi-informative priors can improve the tightness of the results. The question that obviously arises is how to develop more fully informative priors in such circumstances. In the next chapter we describe how elicitation of information from experts can help with that task.

Chapter 4

Use of expert elicitation in Bayesian analysis of extremes

We have seen that in Bayesian analysis of extremes the choice of prior can be expected 85
to be very important and that informative priors can encapsulate important information not available in the data being analysed. Sometimes there exists relevant empirical evidence that can be used to produce informative priors, but where that is not existent or where its scope is limited, expert judgement may be usable. A formalised, documented procedure (Colson and Cooke, 2018) by which the judgement of an expert or experts is obtained is known as **elicitation**.[1] In this chapter we use the term to refer specifically to elicitation in a form that provides probability distributions that may be used as priors in Bayesian analysis of extremes. The broad idea is that the expert is questioned by a person, whom we describe as the **analyst**, with the aim of producing numerical opinions which the analyst can then use to produce a full probability distribution over the sample space X, say $\{f(x) : x \in X\}$, that can be used as a prior. We shall see that there are various differing protocols for managing the elicitation process.

§ 4.1 Relevant literature

Despite its apparently restrictive title, the 278-page European Food Safety Agency 86

[1] For simplicity we may refer freely in this thesis to 'expert' in the singular or to 'experts' in the plural without implying either that there should be only one expert or that multiple experts are required.

(2014) is a very full review of the theory and practical application of elicitation to risk assessment in sectors not restricted to food and feed safety. It sees elicitation as one of the means of providing a quantitative assessment of risk in the areas where judgement is required rather than just factual knowledge. The experts used need to be skilled in estimation but also able to give realistic judgements as to the accuracy of their estimates. Elicitation therefore involves uncertainty expressed as probabilities.

87 The paper goes on to describe in detail three specific but differing approaches to elicitation:

(i) the Sheffield protocol (Oakley and O'Hagan, 2019). The paper draws attention specifically to the behavioural methods used for aggregating potentially differing opinions at a meeting of the experts;

(ii) the Cooke protocol (Cooke (1991)), in which the experts do not provide their opinions collectively, and do not meet to do so, their opinions being aggregated mathematically based on their individual performance with trial or seed questions;

(iii) a Delphi protocol (Helmer, 1968), in which written opinions are sought remotely and then revised in the light of feedback about the opinions of the other experts.

88 In Bojke et al. (2021), and Bojke et al. (2022), the focus of the work was to develop standardised principles, referred to as a **reference case**, for the use of elicitation in a specific context, health care decision-making (HCD). The earlier unpublished paper sets out in greater detail the material described in the later paper. To do that, the paper reviews 16 existing guidelines on elicitation, five of which were generic, the others being specific to particular domains. The paper reports some experiments to test various aspects of the guidelines. A set of nine principles for judging the suitability of choices available for elicitation is developed, as follows:

(i) Transparency. Making available details of how an elicitation has been conducted is argued to improve the validity of expert judgements and to permit peer assessment of elicitation.

(ii) Fitness for purpose. The information that is elicited should be suitable for use in further analysis.

(iii) Consistency, but respecting the constraints of the decision-making context. In the specific case of HCD elicitation will potentially be used by a wide variety of types of user, some with considerable resources others with much less. The reference case for HCD needs to be usable by all such user bodies.

(iv) Reflecting uncertainty at the individual expert level. Uncertainty of knowledge of individual experts should be specifically elicited.

(v) Recognising and acting on biases. Elicitation tasks need to be designed and conducted to minimise the effect of the many heuristics and biases that are known to be used when experts are asked for complex judgements.

(vi) Suitability for experts who possess substantive skills, who are less likely to possess normative skills. The elicitation task needs to be suitable for experts who are not necessarily skilled in probability and statistics.

(vii) Recognising where adaptive skills are required. The elicitation task should recognise where experts need to adapt their specialised knowledge to related areas in which they may not be expert.

(viii) Recognising and act on between-expert variation. Where different experts give different judgements, the reasons for such differences should be understood and their effect on the eventual decision be explored.

(ix) Promoting high performance. Different experts may bring different levels of subject-matter expertise, and different levels of skills in, for example, probability and statistics. Elicitation needs to be able to encourage experts with differing levels of skills to best express their beliefs about a quantity of interest.

89 Examining existing guidelines in elicitation, the paper concludes that they contain many choices for which there is no empirical support, and where the principles do not present a clear basis for preferring one set of choices over another. In consequence, it is argued that the specific constraints of the chosen application (in this case HCD) may best determine the choices to be made.

90 O'Hagan et al. (2006) is an extensive combination of literature review, commentary and synthesis about the field of elicitation. It is aimed at a diverse target audience including decision-makers as well as researchers in many fields. It collects a series of findings about best practice in elicitation. The Sheffield Elicitation Framework (SHELF) (https://shelf.sites.sheffield.ac.uk), which is a specific protocol for conducting elicitations, was developed in parallel with this book. In our own work using elicitation described in Chapter 10 we shall draw very much on the methods O'Hagan et al. (2006) describe.

§ 4.2 Difficulties in elicitation

91 As shown by the references cited above, a great deal of attention has been paid to the development of protocols for elicitation: the process of elicitation is far from straightforward. Many authors have pointed to difficulties from a number of sources that appear to be inherent in the process:

(i) **heuristics and biases**: the effect of irrelevant or partially relevant influences on the judgements elicited from the experts;

(ii) **imprecision and (dramatically) incomplete information**: the desired output $\{f(x) : x \in X\}$ is a function defined at an uncountable infinity of points in a sample space X, such that:

$$\begin{aligned} &0 \leq f(x) \leq 1, \forall x \in X \\ &\int_{x \in X} f(x) = 1. \end{aligned} \tag{4.1}$$

and yet only a small number of judgements can be elicited from the experts;

(iii) uncertain **calibration**: how far an individual expert's opinions can be validated as reflecting reality when their opinions are about matters that are unknown.

4.2.1 Heuristics and biases

Throughout the whole process of elicitation major risks arise from the unnoticed use 92
by the experts of instinctive processes, known as heuristics, used by people making judgements under uncertainty. O'Hagan et al. (2006, Section 3.4) discuss the effect of three heuristics:

(i) **availability** - judgements are based on the instances that come most easily to mind

The risk is that the judgements ignore the effect of relevant instances that did not come to mind;

(ii) **representativeness** - judgements about an event A are based on those relating to another event B according to the similarity between A and B.

The risk is that A is also influenced by events that are not similar to it, so ignoring such events would give rise to erroneous judgements;

(iii) **anchor-and-adjustment** - judgements are based on an initial estimate (known as an anchor) that is then adjusted up or down.

The risk is that the anchor is given excessive weight and the adjustments made are insufficient. Experiments have shown that such an effect can arise when the anchor is not a relevant estimate, and is even known by the expert not to be.

93 Experts have been shown to be at risk of the effects of all these heuristics. There is some evidence that very experienced experts are able to avoid them when making judgements in their own area of expertise, but in an elicitation project the experts are being asked to make judgements not so much about matters in which they are expert but rather about probabilities relating to such matters. If the experts have little or no previous experience in probability and statistics, some training might well be desirable. And even then, it may be that specific thought would need to be given to structuring the elicitation in a way that experts lacking such skills can still contribute effectively, as Bojke et al. (2022) point out (see above) going on to suggest that questions should be posed in a manner consistent with how experts express their knowledge. For example, they suggest that elicitation tasks should specify observable quantities, such as probabilities (expressed as proportions or frequencies), rather than more complex quantities such as odds ratios or variances.

94 The elicitation process needs to be structured paying attention to these biases that can occur in judgement under uncertainty. The facilitator, and, often, the experts need to be aware of them and take them into account throughout the facilitation exercise.

4.2.2 Imprecision and incomplete information

95 The desired output from elicitation is a complete probability distribution. And yet it can only be obtained from a (small) finite number of judgements by the experts, and each expert will inevitably be limited in the precision with which they can express the numbers that they state. (We leave for discussion below the issue of calibration, so here we are assuming that the experts' judgements are all well calibrated.)

96 Although some authors (Lindley et al., 1979; O'Hagan, 1988) see the expert having a "true" probability distribution that elicitation cannot, however, ever fully obtain, others (such as, for example, Winkler (1967); O'Hagan et al. (2006)) regard the experts' assessments as necessarily uncertain, so that an expert might well be satisfied that each of a variety of different distributions represented their view. In practice, whichever view is taken it is clear "that the expert's elicited probabilities must be treated as imprecise"(O'Hagan et al., 2006). There does not appear to be consensus on the best way of dealing with this issue.

97 As to the number of judgements the expert may be asked to make to define their probability distribution, the practical limit arises through tiredness and the risk that the expert becomes anchored to previously given judgements. Beyond some point, eliciting more judgements provides no extra information and may even introduce biases. In practice very small numbers of judgements are commonly used. For example, in the

case considered in Coles and Tawn (1996), which is described in more detail in Chapter 5, the expert provided twelve quantified judgements relating to quantiles used in the analysis, and of these only six judgements were used directly.

4.2.3 Calibration

There is a very large literature on calibration, the question of how closely a source's 98
probability judgements correspond with the relative frequency with which events are observed. A **source** may be an individual, a group of individuals, or some kind of prediction model. In an elicitation the expert is generally being asked to give judgements about matters that are unknown, so it is difficult to check the calibration of such judgements. But an expert's calibration can be tested by, for example, using calibration questions, which are items from the expert's field that are, however, uncertain to the expert (i.e., the expert does not know the true values or does not have the true values readily accessible), but are either known to the analysts at the time of the elicitation or will be known during the analysis period.

A perfectly calibrated source would make judgements that coincided with the frequencies 99
of occurrence of the events in question. **Over-confidence** would result in judged probabilities being higher than the corresponding actual frequencies of occurrence; **under-confidence** would be the reverse. Calibration might be compromised without any consistent pattern off over- or under-confidence: O'Hagan et al. (2006) define one such pattern, where the judged probability of low-frequency events is even lower, and the judged frequency of high-frequency events is even higher, as **over-extremity**. They identify another potential defect in a source's calibration as poor **discrimination**, that is a tendency for the source to assign similar probability judgements to events with different frequencies of occurrence.

§ 4.3 Processes for elicitation

We have referred in Section 4.1 to the numerous elicitation protocols examined in the 100
literature. One such model is described in(O'Hagan et al., 2006, Section 2.2.2). It has been developed into a very full package for elicitation, SHELF (Oakley and O'Hagan, 2019). It proceeds through the following stages:

(i) Background and preparation

It is not necessarily straightforward to identify the variables for which expert assessment is needed. The choice will depend on the statistical model that is to be used and on the structure of the data. At this stage too the statistician or facilitator needs to acquire the technical expertise needed to communicate with the experts. Background briefing for the experts needs to be created during this first stage.

(ii) Identify and recruit experts

Experts used in elicitation projects may need to be more than available individuals who have a high degree of subject-matter knowledge. If the results of the elicitation exercise are to be convincing to third parties, evidence of and reputation for expertise will be needed. High reputation alone is not necessarily enough: O'Hagan et al. (2006, Section 3.3.1) point out the important distinction between a professional and an expert. It is possible that not all professionals employed in a particular subject area, especially very senior professionals, do in fact possess the degree of expertise required for a particular elicitation.

(iii) Motivating and training the experts

We noted above that it is likely that some at least of the experts will not be highly familiar with expressing their opinions in probabilistic terms. In addition to explaining to the experts how the process of elicitation will work, therefore, it may well be necessary to train the experts in probability, in how their judgements could be undermined by common biases and shortcuts in thinking. Practice elicitation sessions might well be helpful.

(iv) Structuring and decomposition

The variables that are the subject of the elicitation do not exist in isolation; there is some kind of structure that relates them to each other and to other variables that may be well known. The experts themselves are likely to have opinions about such interrelationships. Identifying such structure is relevant to reviewing with the experts the evidence on which they will draw in making their judgements.

(v) The elicitation

A cyclical process is used: from the experts' judgments a probability distribution is fitted and presented to the experts. If they do not consider it to be adequate the process is repeated. For a function to be a valid probability distribution function or probability density it must be highly constrained. As well as complying with the experts' judgements, it must also satisfy Equations 4.1. It will be capable of being described by numerous descriptive measures known generally as **summaries**, such as mean, variance, quantiles, moments etc. And in many cases, it will be smooth and well-behaved (having only one or two modes, for example). We have already noted, however, that in the specific context of extreme value analysis, results can be highly sensitive to the specific distribution function that is chosen to represent the expert's judgements.

§ 4.4 Conclusion

We conclude that the effective use of elicitation requires the use of a well-designed protocol. Without it, the results are likely to be fatally compromised in numerous ways. But, as to the choice of protocol, numerous options have been presented in the literature with some authors (eg Bojke et al. (2022) finding the evidential basis for choosing between them rather lacking. 101

Chapter 5

Coles and Tawn on elicitation

Coles and Tawn (1996) is an influential paper describing an approach to expert elicitation 102
in predicting extreme rainfall at a specific location in south-west England. In this chapter
we discuss their approach, and illustrate it by working through their methods with an
example data set.

§ 5.1 Constructing the Coles and Tawn approach

The problem to be addressed is that when studying extreme phenomena, that is, 103
phenomena that might occur on average as rarely as once in 1000 or 10000 years, relevant
observational data will be very scarce. The basic idea is that properly elicited expert
opinion, independent of the observational data, can be taken into account alongside the
observational data and thus increase the precision and reduce the uncertainty pertaining
to forecasts of extreme behaviour.

The observational data used had been recorded at one rain gauge site spanning the 104
period 1932-88. Although there were known to be some seasonal characteristics, the
temporal dependence at extreme levels had been found to be extremely weak. Expert
information was provided by Dr Duncan Reed, an acknowledged expert in hydrological
science with specialised knowledge of rainfall in the region. Computationally, an MCMC
method was used.

105 Working backwards, we can describe the Bayesian mechanism Coles and Tawn used for taking elicited expert opinion into account as follows:

(i) forecasts of extremes are derived from posterior samples from a parametric probabilistic model;

(ii) the parameters of the model are fitted to a subset of the observational data that excludes observations judged too low to be relevant to extreme behaviour;

(iii) during the fitting process parameters are drawn as samples from prior distributions defined by hyperparameters;

(iv) the hyperparameters are fitted to quantiles elicited from the expert.

106 It follows that as analysts Coles and Tawn needed to make a series of decisions:

(i) what parametric model to choose;

(ii) how to exclude irrelevant observations;

(iii) how to define the prior distributions;

(iv) how to convert the expert's quantiles into parameters of the prior distributions.

§ 5.2 The choice of parametric model

Coles and Tawn examined the merits of choices made in earlier work. They recognised 107
that the annual maxima approach (see Section 2.2) would not be an efficient use of the data. They also saw a disadvantage in the GPD model as used by Pickands (see Section 2.5). That method used the potentially helpful relation, for all thresholds $\tilde{u} > u$:

$$p(X_i - \tilde{u} \leq y | X_i > u) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)_+^{1/\xi} \tag{5.1}$$

where $\tilde{\sigma} = \sigma + \xi(\tilde{u} - \mu)$. Coles and Tawn found, however, that using the GPD model seemed restrictive since the scale parameter $\tilde{\sigma}$ is dependent on the choice of $\tilde{u}$ so a single uninformative prior cannot be chosen that is valid at all thresholds.

They therefore worked explicitly with the point process model (see Section 2.6) in which 108
the parameters would be independent of the threshold level. That model was based on a point process characterisation of extremes $\{X_i : i = 1, \ldots, n\}$ over a high threshold u. Accordingly the set $\{X_i : i = 1, \ldots, n\}$ viewed on the interval $[u, \infty)$ is approximately a non-homogeneous PPP with intensity function:

$$\lambda(x) = \frac{1}{\sigma}\left\{1 + \xi\left(\frac{x - \mu}{\sigma}\right)\right\}_+ \tag{5.2}$$

where $z_+ = \max(z, 0)$ and where μ, σ, ξ are the location, size and shape parameters respectively of the GEV distribution.

5.2.1 Excluding irrelevant observations

We have just seen that the PPP model depends on the selection of a high threshold, 109
u, below which the observational data is discarded. The analysis therefore required a decision on the level to be used for u. Coles and Tawn followed the method described above (see Subsection 2.7.1), namely inspection of the empirical plot of mean residual life (MRL). Although the exact dataset used by Coles and Tawn cannot be traced, a nearly identical set has been provided as data in the R package `evdbayes`. That dataset covers almost the same period as used by Coles and Tawn but contains slightly fewer observations: 19,667 rather than 19,725. The `evdbayes` dataset enables us to produce a very similar MRL plot, see Figure 2.3. It shows clearly the complication caused by increased variability at large values of rainfall. Nevertheless, Coles and Tawn discerned evidence of linearity above a threshold of about 40mm. They proceeded on that basis but examined later the sensitivity of the results to other choices of u.

5.2.2 Defining the prior distributions

110 As can be seen in Equation 5.2, the three parameters of the PPP are: (i) location, μ, (ii) size, σ, and (iii) shape, ξ. Coles and Tawn argued, however, that it would not be satisfactory to use these parameters directly because there would be likely to be a negative dependence between σ and ξ. That dependence arises from the fact that the weight of the tail of the distribution would be increased by increasing either σ or ξ.

111 To avoid such dependence, use was to be made of quantiles (in this context known as return levels):

$$q_i = \mu + \frac{\sigma\left[\{-\log(1-p_i)\}^{-\xi} - 1\right]}{\xi} \tag{5.3}$$

for three values of $p : p_1 > p_2 > p_3$. But since the q_i are not independent the authors worked with the differences

$$\begin{aligned} \tilde{q}_1 &= q_1 - e_1 \\ \tilde{q}_2 &= q_2 - q_1 \\ \tilde{q}_3 &= q_3 - q_2, \end{aligned} \tag{5.4}$$

where e_1 is the physical lower limit of the process. In this case, rainfall, it is obvious that $e_1 = 0$.

112 We note that if we combine Equations 5.3 and 5.4, we obtain:

$$\begin{aligned} \tilde{q}_1 &= \mu + \frac{\sigma\left[\{-\log(1-p_1)\}^{-\xi} - 1\right]}{\xi} \\ \tilde{q}_2 &= \frac{\sigma}{\xi}\left[\{\log(1-p_1)\}^{-\xi}\} - \{\log(1-p_2)\}^{-\xi}\}\right] \\ \tilde{q}_2 &= \frac{\sigma}{\xi}\left[\{\log(1-p_2)\}^{-\xi}\} - \{\log(1-p_3)\}^{-\xi}\}\right], \end{aligned} \tag{5.5}$$

It follows that the priors for the size parameter, σ, and the shape parameter, ξ, are determined by those for $\tilde{q}_2$ and $\tilde{q}_3$, and that the prior for the location parameter, μ, is determined by that for $\tilde{q}_1$, conditional on the values of σ and ξ.

113 As to the prior distribution to which the elicited quantities were to be fitted, Coles and Tawn chose the gamma distribution for each $\tilde{q}_i$ with the PDF $f_i(\cdot)$ being given by

$$f_i(x) = \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} x^{\alpha_i - 1} e^{-\beta_i x} \quad x \in (0, \infty) \tag{5.6}$$

where $\Gamma(\cdot)$ is the gamma function. They assumed that the three distributions were independent.

The gamma distribution has support on the positive real line but no further detailed reasons were given for that choice of distribution. Because the elicitation would be working with the differences $\tilde{q}_i : i = 1, 2, 3$, the implied distributions of the underlying q_i would be one gamma distribution, the sum of two, and the sum of three gamma distributions respectively, a situation the authors describe as having a sense of arbitrariness. 114

5.2.3 Converting the expert's quantiles into parameters of the Poisson point process

The expert was to be asked for his estimated median and 90% quantiles of each of the $\tilde{q}_i$, from which the values of α_i and β_i could then be calculated. The joint prior for the q_{p_i} is then to be obtained from Equations 5.4 and 5.6 as: 115

$$\begin{aligned} f(q_{p_1}, q_{p_2}, q_{p_3}) &\propto \tilde{q}_{p_1}^{\alpha_1 - 1} \exp(-\beta_1 \tilde{q}_{p_1}) \prod_{i=2}^{3} \tilde{q}_{p_i} \exp\{-\beta_i \tilde{q}_{p_i}\} \\ &= q_{p_1}^{\alpha_1 - 1} \exp(-\beta_1 q_{p_1}) \prod_{i=2}^{3} (q_{p_i} - q_{p_{i-1}}) \exp\{-\beta_i (q_{p_i} - q_{p_{i-1}})\} \end{aligned} \tag{5.6}$$

on $0 \leq q_{p_1} \leq q_{p_2} \leq q_{p_3}$.

The posterior, $\pi((\mu, \sigma, \xi)|x)$, is derived using Equation 5.3 and the Jacobian of the transformation $(q_{p_1}, q_{p_2}, q_{p_3}) \mapsto (\mu, \sigma, \xi)$. 116

Explicit analytical calculation of the marginal distributions of $\pi((\mu, \sigma, \xi)|x)$ being impossible, a Markov chain Monte Carlo method based on the Gibbs sampler was used. It successively updated the individual parameters μ, σ, and ξ conditionally on the current values of the other parameters. 117

§ 5.3 Results

The expert took account of the general rainfall climate over the region containing the specific location as well as particular characteristics of the site, such as altitude and average annual rainfall. He did not, however, refer to the data available from the site. 118

p_i	Median (mm)	90% quantile	α_i	β_i
0.1	59	72	38.9	0.67
0.01	43	70	7.1	0.16
0.001	100	120	47.0	0.39

Table 5.1: Elicited prior medians and 90% quantiles for distributions of $\tilde{q}_i$ with associated gamma parameters for the prior distribution

The expert's three pairs of quantiles were fitted to gamma distributions, yielding parameters $\alpha_i, \beta_i : i = 1, 2, 3$. The results were given as shown in Table 5.1.

119 The Gibbs sampler converged and after a burn-in of 2000 iterations provided material for analysis based on the subsequent 8000 iterations. The results are presented graphically in two ways:

(i) Showing a range of estimates of the return level for each period of years together with 2.5% and 97.5% quantiles and the medians. These ranges are based on the fact that for each draw of the three parameters, μ, σ, ξ, the implied return level can be calculated for any chosen period of years. Figure 5.1 shows the resulting distributions for each return period from 10 to 10000 years.

(ii) Univariate marginals are shown (Figure 5.2 of the prior (red) and posterior (black) distributions of each GEV parameter.[1]

120 The authors examine the relative merits of their Bayesian approach and frequentist methods. They compare their results for the quantity q_p, namely the $1 - p$ quantile of the annual maximum distribution to a version of q_p calculated from the Maximum likelihood estimate (MLE)s. They find that the Bayesian approach gives both a more sharply rising curve and one that passes more closely to the empirical return levels of the observations than the MLE method. They suggest that the shape of the MLE curve may be over-influenced by less extreme data.

121 Comparison of the results with those obtained with a set of uninformative priors (independent Gaussian or log-Gaussian priors with large variance) showed that within

[1] Both Figures are derived using the `evdbayes` dataset rather than the slightly different dataset used in the original paper.

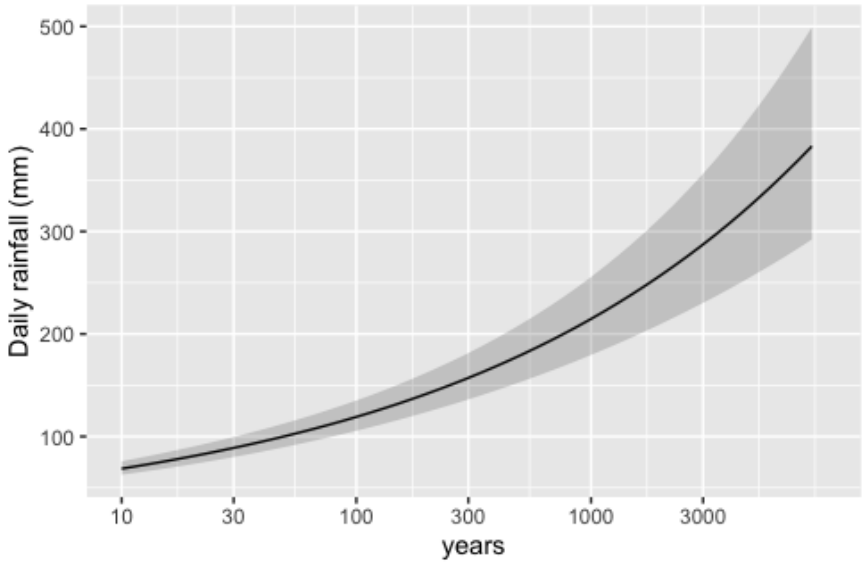


Figure 5.1: Implied return levels by return period: median and 2.5% and 97.5% quantiles

the range of the data the two posteriors were very similar, whereas, for longer return periods the informative priors produce posteriors with higher means and much lower variance than with the uninformative priors (such priors would give results that would be consistent with frequentist methods). We show a similar comparison graphically below in Section 5.5. The conclusion was that use of informative, but not site-specific, prior knowledge led to improved precision of the estimates beyond the range of the data.

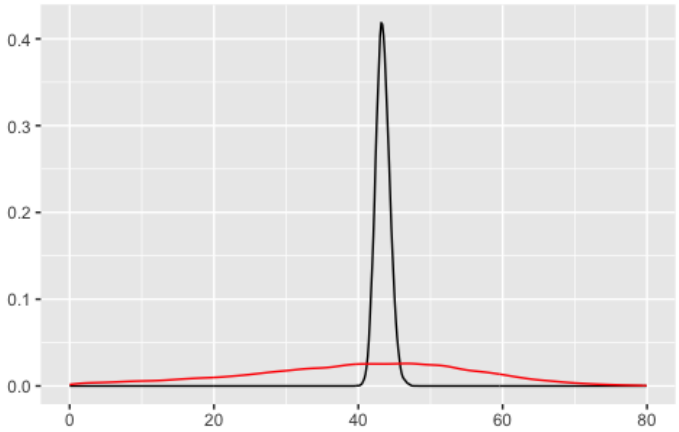
§ 5.4 Discussion

Coles and Tawn (1996) was one of the earliest papers in which informative priors based on elicitation were used to analyse extremes. There are a number of aspects of the work which might nowadays be handled differently: 122

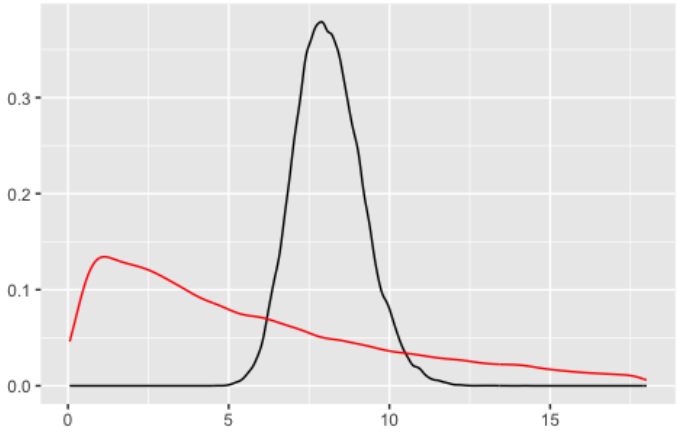
(i) Because it obtains priors for the three parameters by a one-to-one transformation of variables

$$(q_1, q_2, q_3) \mapsto (\mu, \sigma, \xi) \tag{5.7}$$

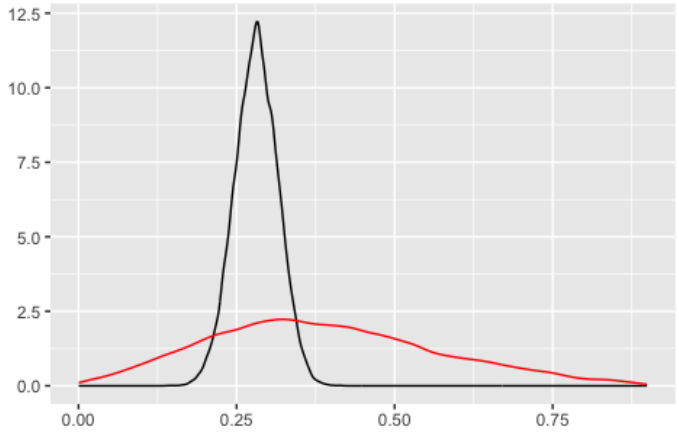
the method is restricted to making use of only three pairs of elicited quantiles. Although six pairs of quantiles were elicited, three of them could not be used in



(a) Location, μ



(b) Scale, σ



(c) Shape, ξ

Figure 5.2: Univariate marginals by parameter:
Posteriors: black. Priors: red

estimation, but only as a cross-check. And the choice of which three of the six pairs to use in estimation would have affected the results.

(ii) Working with differences between quantiles ostensibly requires the expert to estimate the median and 90% quantile of *differences* between quantiles of the underlying distribution without estimating the quantiles themselves. That seems likely to be a difficult task without risking the required independence of the three quantities.

(iii) the reasonable-seeming choice of the gamma distribution to model the differences in quantiles is nevertheless arbitrary.

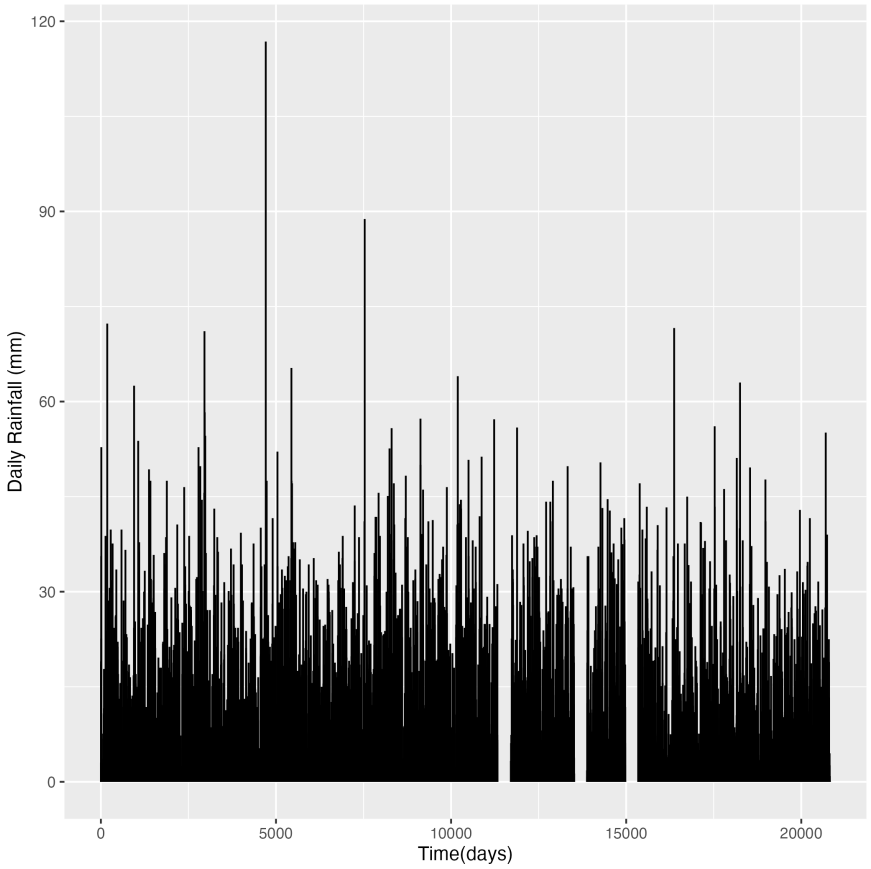
(iv) the effects of the assumption that a gamma distribution is appropriate are not examined.

The analysis concludes that with high probability the shape parameter, ξ, is positive. 123
Using the GEV distribution, such a value of ξ implies that some at least of the return levels for daily rainfall implied by the Hamiltonian Monte Carlo (HMC) simulation might be unbounded above. While noting that fact, the paper does not discuss the realism of such a finding. There are, nevertheless, some arguments in favour of permitting the parameter ξ to exceed zero, particularly in a study of rainfall, as we shall discuss in more detail below (see Section 8.1).

§ 5.5 Worked example of the Coles and Tawn approach

To present a fuller understanding of the methods used by Coles and Tawn we apply 124
them to an artificial data set as set out in the following paragraphs. It is convenient at this point to describe the approach we have taken to computation both in this chapter and throughout this thesis. All the work has been done using the languages R (R Core Team, 2024) and Stan (Stan Development Team, 2024). R is a language and environment for statistical computing and graphics and includes a very large number of packages of software for specific tasks. Stan is a platform for statistical modelling and high-performance statistical computation. In our work we have primarily used the facilities Stan provides for MCMC modelling, making use of interfaces between Stan and R that have been provided as R packages (initially `rstan` and, later, `cmdstanr`). Although we are aware that R packages exist for the analysis of extremes, we have

Figure 5.3: A rainfall dataset



developed our own R code for our analysis. Likewise, we have created Stan code specific to extremes rather than using any pre-written code that may exist. (At the outset of this research there was little or no such code. We have not particularly followed any subsequent development that there might have been.)

125 It is convenient to use the rainfall dataset in the R package `evdbayes`. The full dataset is shown in Figure 5.3. Because that dataset is remarkably similar to that used by Coles and Tawn, it seems reasonable to use the same expert data that were elicited by Coles and Tawn, comprising medians and 90% quantiles $\tilde{q_1}, \tilde{q_2}, \tilde{q_3}$ of the differences between rainfall corresponding to three probabilities, namely 0.1, 0.01 and 0.001. For each probability we fit the data to a gamma distribution with results shown in Table 5.2. The three pairs of parameters of the gamma distributions define the elicited prior, $\pi((\mu, \sigma, \xi)|x)$.

p_i	Median (mm)	90% quantile	α_i	β_i
0.1	59	72	39.0	0.66
0.01	43	70	6.1	0.13
0.001	100	120	46.7	0.46

Table 5.2: Elicited prior medians and 90% quantiles for distributions of $\tilde{q}_i$ with associated gamma parameters for the prior distribution

It is now necessary to choose a high threshold to restrict the analysis to observations 126
for which the GEV model is most likely to be appropriate. Coles and Tawn did so by inspection of the MRL plot. The plot for the dataset we are using was shown above as Figure 2.3. As noted there, it is certainly not obvious what the choice of threshold should be. The aim is to find a high threshold above which the plot is reasonably linear (until the sparsity of observations gives rise to large variances). By inspection the lowest high threshold appears to be around 37mm and possibly a figure as high as 50mm might be chosen, corresponding, respectively, to 132 and 26 observations. We will work initially with a threshold of 40 (86 observations) and then explore the effect of using lower or higher figures for the threshold.

We are now able to carry out an MCMC simulation to estimate the posterior distributions 127
based on the given priors, the threshold and the observational data. The simulation is performed using code we have written in the language Stan (Stan Development Team, 2024) and processed using the R package `cmdstanr`.

Before examining the results of the simulation we can carry out a cross-check on the 128
simulation using the additional quantiles elicited from the expert. These correspond to 30-, 300- and 3000-year return periods. The results are shown in Table 5.3. As was observed in Coles and Tawn (1996)[Section 3], the 30-year and 300-year derived prior values are very close to those elicited from the expert but there is a discrepancy in the 3000-year quantiles.

Table 5.3: Comparison of elicited and derived prior estimates for statistics of return values associated with a range of return periods

Return period	Elicited prior values(mm)		Derived prior values (mm)	
(years)	Median	90% quantile	Median	90% quantile
30	75	95	76.7	95.7
300	140	170	140.2	175.8
3000	260	300	296.2	354.1

129 As it happens, the discrepancy is not so large in our case as in the original paper: their derived prior values at 3000 years were as high as 350.7 and 419.3. Nevertheless a conflict between the opinion of Dr Reed, the expert, and the simulated results requires further examination. The expert's response as reported in Coles and Tawn (1996)[Section 3] was that he doubted the validity of extrapolations based on the GEV model at very extreme levels. Amplifying this view in recent correspondence with us Dr Reed stated that his doubts related not just to the GEV model but to any model, and not only at very extreme levels. He commented that he would not trust "a single-site analysis beyond return periods of say twice the record length. 30 years fine, 300 years not. 300 years: you simply must do some kind of pooling to exploit data from other sites. 3000 years a lot of pooling" (Reed, 2023). Dr Reed's recent statement confirms that the difference between his opinions of the 3000-year rainfall and the results of the modelling exercise is attributable to his disbelief in any likelihood model, including the specific model used by Coles and Tawn, rather than any incoherency in the prior. We therefore proceed, as did they, on the basis of the original prior but maintain caution about the validity of the analysis at extreme levels.

130 The univariate marginals of prior and posterior distributions of each parameter were shown above in Figure 5.2. The posterior distributions are clearly much more precise than the priors but consistent with them.

131 The simulation proceeds in four independent chains with different starting points. There were 1000 warmup iterations in each chain and 1000 post-warmup iterations making two sets of 4000 each in total. At each iteration of the simulation in each chain, a triplet of the GEV parameters, μ, σ, ξ, is generated defining an extreme value distribution.

132 The results of the simulation are summarised in Table 5.4. That table includes values of two diagnostics: $\hat{R}$, which is a measure of the convergence achieved in the simulation; and $n_{\texttt{eff}}$ which is a measure of the effective sample size taking account of the autocorrelation between successive iterations. The unit in which $n_{\texttt{eff}}$ is measured is the number of effective samples. It may be compared with the actual number of samples taken in the simulation, in this case 4000. It can be shown that in a simulation with n (effective) iterations $\lim_{n\to\infty} \hat{R} = 1$, so the values of $\hat{R}$ give some comfort that reasonable convergence has been achieved, and that is supported by inspection of the trace plots for the three parameters μ, σ, ξ (Figure 5.4). As to autocorrelation, inspection of the autocorrelation plots of each parameter give rise to no concern, as can be seen, for example, in the plot shown as Figure 5.5.

Parameter	$\hat{R}$	n_{eff}	mean	sd	2.5%	50%	97.5%
qtilde1	1.007	1178	69.478	3.355	63.457	69.379	76.692
qtilde2	1.008	1038	50.093	4.828	41.122	50.000	59.847
qtilde3	1.004	1190	96.677	13.569	72.103	95.896	125.578
mu	1.003	1920	43.367	0.968	41.559	43.332	45.373
nu	1.002	1476	1.926	0.149	1.647	1.922	2.236
xi	1.002	1486	0.283	0.034	0.217	0.284	0.350
sigma	1.005	1344	8.063	1.046	6.232	8.023	10.271
log-posterior	1.003	1066	-19.275	1.228	-22.527	-18.950	-17.872

Table 5.4: Summarised results of the simulation using u = 40

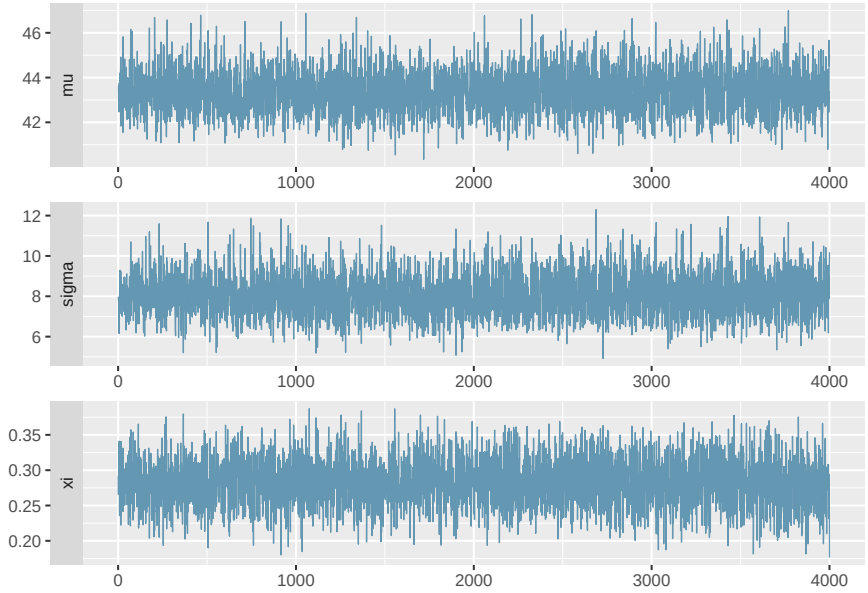


Figure 5.4: Traceplot plot for GPD parameters

The range of forecasts of return values that flow from the individual extreme value distributions is shown graphically in Figure 5.6. In each graph, we summarise the resulting distributions in terms of estimated quantiles of rainfall for a full range of return periods. We also show in red the empirical return values of the data. The effect of using elicited priors is clearly shown by comparing Figure 5.6a with the equivalent produced using uninformative priors, Figure 5.6b. It is clear that the use of elicited priors vastly improves the precision of the estimates of extreme return levels. It is also notable that the mean forecast is consistently higher in the elicited case than in the uninformative case, as can be seen in Figure 5.7 where the two sets of results are presented together. 133

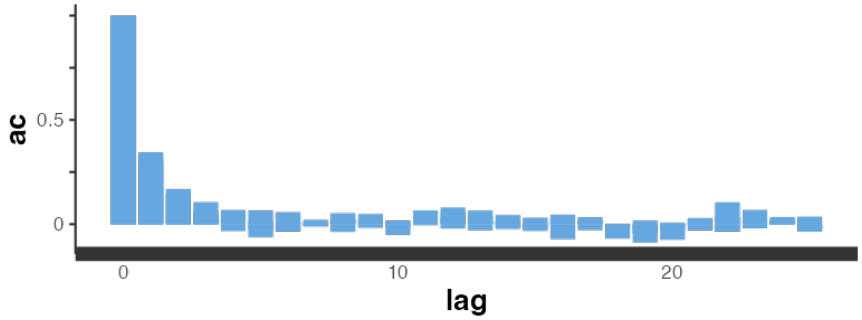
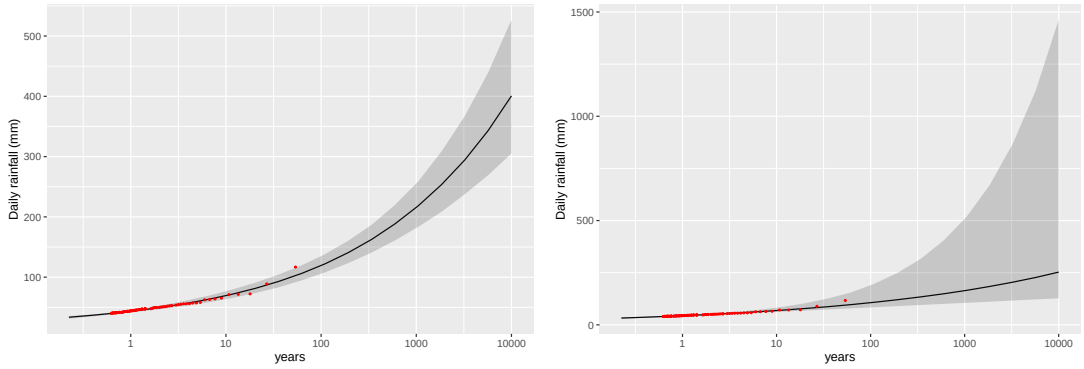


Figure 5.5: Autocorrelation plot for parameter μ



(a) Elicited priors

(b) Uninformative priors

Figure 5.6: Forecast return levels for range of periods with different priors:
Key: medians (black), and 2.5% to 97.5% quantile range (grey), empirical return periods of data (red)

134 Coles and Tawn (1996) tested the 30-, 300- and 3000-year quantiles that they had elicited from their expert. We can do the same: Figure 5.8 shows that the expert's quantiles are slightly lower than the return levels generated by the model.

135 To test the sensitivity of the results to the choice of high threshold, u, we additionally ran the simulation for the cases $u = 35$ and $u = 50$. In Table 5.5 the results of the simulations are presented showing for each threshold an indication of the spread of the estimates of extreme values for three different periods. The results for $u = 35$ and $u = 40$ are very similar at all return periods. The notable feature of the results for $u = 50$ are the greater spread of the figures. That is an unsurprising consequence of the small size of the sample of observations exceeding 50mm.

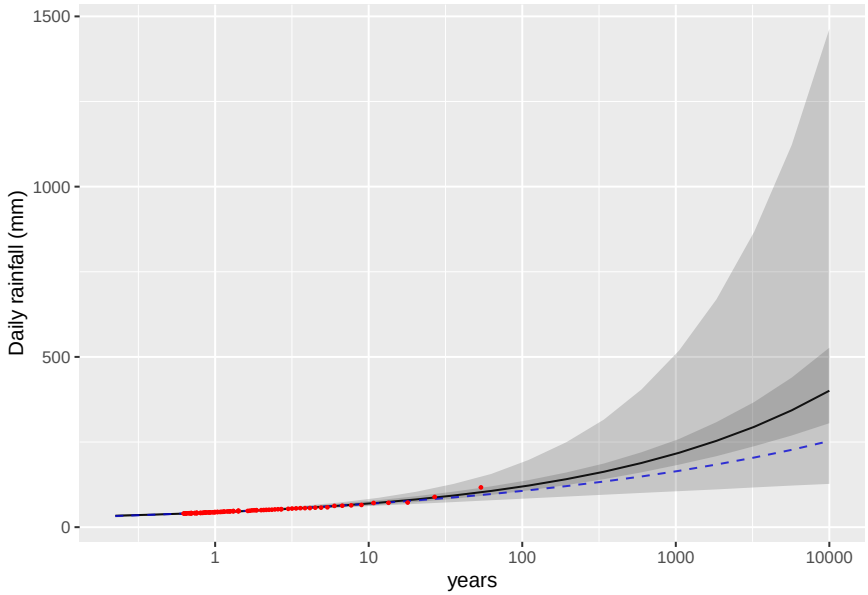


Figure 5.7: Comparison of results using elicited and uninformative priors
Key: Elicited priors: black line and dark shading. Uninformative priors: dashed blue line and light shading. Empirical return levels of data: red dots.

§ 5.6 Practical use of posterior distributions

In our reproduction, above, of the Coles and Tawn methodology we have followed the 136
authors in not addressing how the results might be used in practice. The posterior distribution of their forecasts of a particular extreme value, the 10000-year return level, for instance, does not immediately enable an engineer to decide a specific physical structure that would resist such an extreme: ideally a single figure would need to be derived. The Bayesian approach can produce such a figure. If the quantity of interest is a random variable X whose value depends on a parameter θ, then the quantity

$$p(X) = \int p(X|\theta)p(\theta) \tag{5.8}$$

describes the distribution of X taking account of uncertainty about the values of θ and the uncertainty about the value of X once θ is known. It follows that the expected posterior predictive value,

$$\mathbb{E}(X) = \int xp(X = x)dx, \tag{5.9}$$

is a point estimate summarising the combined effect of all the uncertainty underlying the forecast.

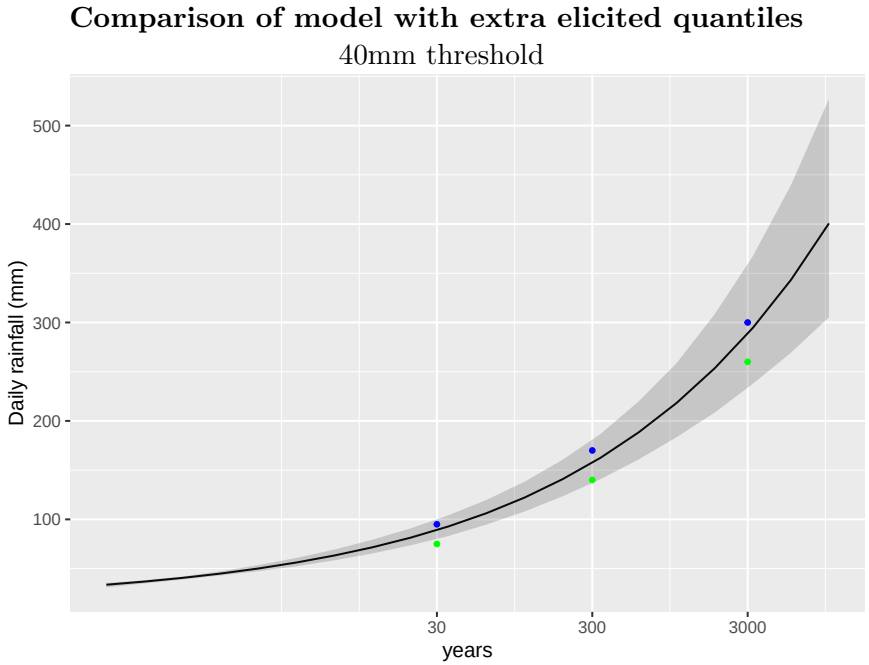


Figure 5.8: Testing additional elicited quantiles using the 40mm model.
blue = 90% quantiles, green = median quantiles

137 Although the point estimate yielded by Equation 5.9 is the best summary of the Bayesian modelling taking into account all uncertainties, care needs to be taken in its use. It would not make sense to calculate the figure until being satisfied that the posterior predictive distribution does indeed take into account all relevant uncertainty and that it excludes irrelevant material. It would be possible to calculate an expectation based on the predictions in Figure 5.6b, and such a point estimate would hide that fact that the underlying posterior failed to make use of relevant information, namely that underlying the informative priors that produce the much tighter estimates in Figure 5.6a.

§ 5.7 Conclusions

138 We have described in some detail the work of Coles and Tawn (1996). We have drawn attention in Section 5.4 to some respects in which the approach taken might be questioned, nevertheless we have shown (using a very similar dataset) that it is reproducible and, of course, using modern software its computation is very fast.

Threshold (mm)	110-year			1050-year			1000-year		
	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
35	113	126	141	182	218	260	293	383	500
40	108	122	138	184	218	259	305	401	527
50	98	115	135	178	213	255	309	421	570

Table 5.5: Effect of changing the threshold on forecast extreme return values

Chapter 6

Gaussian process models for prior distributions from elicited quantiles

We have seen (for example, in Figure 5.6) that information elicited by an analyst from an 139
expert *can* (depending on the skill of the expert) lead to improved inference particularly in circumstances where the relevant data points are sparse, as is generally the case in extreme value statistics. To be used in Bayesian inference, the information the expert provides must be put into the form of a probability distribution, and we note that to insist that that distribution take any particular parametric form, a gamma distribution, or lognormal, say, may well be an unwarranted assumption. We therefore consider how non-parametric methods may be used.

A non-trivial problem that arises when using non-parametric methods for this purpose is 140
the need for the resultant probability functions not only to match the quantiles elicited from the expert but also to satisfy constraints (such as monotonicity) that apply across the distributions' whole support. Any successful method must provide a solution to that problem.

There is a literature dating over many years on density estimation, both parametric and 141

non-parametric and both Bayesian and frequentist. That literature addresses a problem that is, however, subtly different from ours in that its aim is to identify the best fit of density to possibly extensive data whereas for our purposes we wish to understand the full range of densities that are compatible with typically sparse data. The literature on the emulation of complex models is also relevant despite being aimed at a rather more general problem.

142 This chapter argues that:

(i) to obtain a proper understanding of the uncertainty arising from the necessarily limited information that the expert can provide, it is necessary to use for inference a full range of distributions that would be compatible with the expert's information;

(ii) a GP model provides a very flexible and comprehensive non-parametric approach to the elicitation problem in general; but,

(iii) that flexibility of the GP model can give rise to realisations that meet the constraints but are not valid distributions, because, for example, they give rise to negative "probabilities". That reduces, potentially severely, the efficiency of the modelling process.

§ 6.1 The need to generate a range of probability distributions

143 It is important to note that we use the plural word "distributions". We do so because where the expert has supplied quantiles, there are countless probability distributions that could be fitted. As Gosling points out: "We must also take into account all the distributions that we believe to be consistent with the judgements provided by the expert"(Gosling, 2005, Section 2.5). To account fully for the uncertainty in the resulting estimates of extremes it is necessary to consider a full range of feasible distributions; to choose just one distribution will lead to assessments of uncertainty that are subject to biases of which the size and direction are both unknown.

144 In contrast to the view that a full range of distributions should be taken into account, however, some authors, notably Jaynes (2003), argue that the choice of prior distribution

should always be as conservative as possible and they express conservatism in terms of the entropy, $H[p]$, of the probability density, p:

$$H[p] = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx \tag{6.1}$$

(in the univariate case). From that point of view, the most conservative prior distribution is that which maximises $H[p]$ subject to whatever constraints apply. For example, if we wish to find the univariate density, p, that maximises $H[p]$ subject to being constrained to a given variance, σ^2, and given mean μ, it must satisfy three constraints, as follows:

$$\begin{aligned} \int_{-\infty}^{\infty} p(x)\, dx &= 1 \\ \int_{-\infty}^{\infty} xp(x)\, dx &= \mu \quad \mu \in \mathbb{R}, \\ \int_{-\infty}^{\infty} x^2 p(x)\, dx &= \sigma^2 \quad \sigma \in \mathbb{R}. \end{aligned} \tag{6.2}$$

Using straightforward calculus of variations it is easy to show that the only function $p(x)$ that satisfies all three constraints is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \left\{ \frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\}, \tag{6.3}$$

namely the normal density.

It is not always possible to derive a maximum entropy distribution: where the constraints 145
comprise quantiles of a distribution function with unbounded support, the maximum entropy distribution is indeterminate. With bounded support, however, where, for example, $X \sim [a, b]$ where a, b are both finite and the given quantiles $q_i : i = 1, \ldots, n$ are at points v_i, the PDF, $f(x) : x \in [a, b]$, of the maximum entropy distribution can be stated, as follows:
let

$$p_i \equiv P(v_{i-1} < X \leq v_i) = q_i - q_{i-1} \quad i = 1, \ldots, n$$

where $v_0 = a$ and $v_{n+1} = b$. Then

$$f(x) = \sum_{i=1}^{n+1} \frac{q_i - q_{i-1}}{v_i - v_{i-1}} \mathbf{1}_{(v_{i-1}, v_i]}(x) \tag{6.4}$$

where for convenience we have set $q_0 = a$ and $q_{n+1} = b$ and the operator $\mathbf{1}(x)$ is defined

thus:

$$\mathbf{1}_{(a,b)}(x) = \begin{cases} 1 & \text{if } x \in (a,b) \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

146 We do not pursue this method any further here. We note that the explicit assumption in the use of the maximum entropy distribution is that the given $q_i : i = 0, \ldots, n+1$ comprise the totality of information about the situation. In the context of expert elicitation with a very small number of quantiles, we judge that experts are unlikely to agree that their beliefs about the distribution are truly so limited. In particular, the sharp drops in the value of $f(x)$ as $x \downarrow a$ and $x \uparrow b$ seem likely to be unwelcome in many cases.

§ 6.2 Gaussian process models

147 In this section we describe briefly what GPs are and review the literature on their use in defining priors in the context of elicitation.

148 GP models have been used, sometimes under quite different names and varying definitions, for a wide range of purposes. As Rasmussen and Williams (2006, Preface) make clear, GP models are mathematically equivalent to many well known models and in the statistics community (alone) have been discussed many times. We use the following definition:

Definition 9. *A **GP** is a collection of random variables, any finite number of which have a joint Gaussian distribution.(Rasmussen and Williams, 2006)*

149 A GP is therefore a generalisation of the Gaussian probability distribution whereby a sample (usually known in this context as a **realisation**) drawn from a GP is a function (or, more precisely, comprises the values of that function at a finite number of points).

150 In the general context the basic idea is that the problem to be solved is one in which inferences are to be drawn about a given function by combining *data* (in the form of the values taken by the function at a finite number of points) with *prior beliefs*. In the

Bayesian paradigm the prior beliefs are represented as probability distributions and the inferences are posterior distributions.

The GP approach to defining priors and extracting posterior distributions is capable of great flexibility. That very flexibility does, however, present some potential difficulties. Constraints on a function, such that it must be non-negative or monotone can be difficult to model with a GP (O'Hagan and Forster, 1994, paragraph 13.50). That is precisely the situation in the typical case we face here, as we seek to fit a bounded, non-negative, monotone non-decreasing function (all of which a CDF must be) to small number of quantiles supplied by the expert, whilst ensuring that it attains its lower bound of zero and its upper bound of exactly unity. 151

Typically the expert's prior expectations will also impose further constraints that arise from qualitative aspects such as whether the density could be multimodal, or how smooth a function it might be. (We shall assume for the purposes of illustration, and not unrealistically, that the expert has asserted that the density is **unimodal** and very smooth). So each simulated density, $f(\cdot)$, and corresponding distribution function, $F(\cdot)$, say, must comply with a number of constraints: 152

(i) $f(\cdot)$ must be a valid probability density function:

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1 \qquad f(x) \geq 0 \quad \forall x \in \mathbb{R} \tag{6.6}$$

(ii) equivalently, $F(\cdot)$ must be non-negative, monotonic non-decreasing, $F(-\infty) = 0$ and $F(\infty) = 1$

(iii) the quantiles of $F(\cdot)$ must comply with those defined by the expert:

$$F(x_i) = y_i \quad i = 1, \ldots, n \tag{6.7}$$

(iv) or, equivalently,

$$\int_{-\infty}^{x_i} f(x)\,dx = y_i \quad i = 1, 2, \ldots, n \tag{6.8}$$

(v) $f(\cdot)$ must be unimodal: $\exists x_0 \in \mathbb{R}$ with all the following properties:

$$\begin{aligned} x \neq x_0 &\Rightarrow f(x_0) > f(x) \\ x_0 < x_1 < x_2 &\Rightarrow f(x_0) > f(x_1) \geq f(x_2) \\ x_1 < x_2 < x_0 &\Rightarrow f(x_0) > f(x_2) \geq f(x_1) \end{aligned} \tag{6.9}$$

We shall see below that the definition of "unimodal" can be widened slightly to include the **plateau** case: $\exists x_0 < x_1 \in \mathbb{R}$ with all the following properties:

$$\begin{aligned} x \in [x_0, x_1] &\Rightarrow f(x) = f(x_0) \\ x_1 < x_2 < x_3 &\Rightarrow f(x_1) > f(x_2) \geq f(x_3) \\ x_2 < x_3 < x_0 &\Rightarrow f(x_0) > f(x_3) \geq f(x_2) \end{aligned} \tag{6.10}$$

It can be seen that some of the constraints apply to the density alone, and others apply (with obvious modification) to the corresponding distribution function as well..

153 In the following subsections we review the literature on using non-parametric methods based on GPs in prior elicitations and examine how authors have dealt with the problem of constraining the GP so that its realisations are valid probability distributions and densities.

6.2.1 Oakley and O'Hagan (2007)

154 Oakley and O'Hagan (2007) address an essentially similar problem to that considered here, namely a non-parametric approach to construct a distribution from a small number, n, of statements, $\{(x_i, y_i), i = 1, \ldots, n\}$, that have been elicited by an analyst from an expert. Formally, they seek a density function $f(\theta)$ representing the analyst's beliefs about the expert's density function consistently with data D, which are taken to include the requirement $\int_{-\infty}^{\infty} f(\theta)\,d\theta = 1$.

155 They assume that the analyst's prior beliefs about $f(\theta)$ can be represented by a GP with a mean function and covariance function that are modelled hierarchically in terms of a vector α of hyperparameters. For the mean function, they choose $E\{f(\theta|\alpha\} = g(\theta|u)$

that is a member of a suitable parametric family with parameters $u \in \alpha$, arguing that the GP model is sufficiently flexible that a parametric choice of mean function allows the true $f(\theta)$ to have any form at all.

As to the choice of covariance function, they argue that the analyst would expect the variance of $f(x)$ to be smaller where $f(x)$ is small. To reflect that expectation, a covariance function (denoted a **scaled stationary covariance function**) of the form 156

$$\mathrm{Cov}\left(f(\theta), f(\varphi)|\alpha\right) = g(\theta|u)g(\varphi|u)\,\sigma^2 \exp\left\{-\frac{1}{2vb^*}(\theta-\varphi)^2\right\} \tag{6.11}$$

is chosen.

Oakley and O'Hagan work primarily in terms of the density function and therefore to incorporate the quantile constraints they evaluate covariances between percentiles of the distribution function F in terms of the mean and covariance functions of the GP they constructed for the density as described below in Subsection 6.2.6. In the case they choose to exemplify, in which $g(\theta|u) \sim \mathcal{N}(\mu, v)$, that is the normal density with mean μ and variance v, the covariances involving percentiles can be given in closed form. 157

An informative prior is set for b^* to reflect the prior belief that the ratio $f(\theta)/g(\theta)$ should not fluctuate excessively but neither should it be expected to be constant. Minimally informative priors are set for the other hyperparameters. 158

The posterior for the density function of $f(\cdot)$ can then be derived, as follows. Let the data comprise a vector D with mean H and variance-covariance matrix $\sigma^2 A$, and if the covariance between D and $f(\theta)$ is $\sigma^2 t$, where A and $t(\theta)$ are functions of m, v and b^*. (Note that the notation $t(\theta)$ is not intended to refer to the t- distribution.) Then the joint distribution of D and any finite set of points on the function $f(\theta)$ is multivariate normal, as is the distribution of $f(\theta)|D, \mu, v, b, \sigma^2$, which is a GP defined by: 159

$$\begin{aligned} \mathrm{E}\left(f(\theta)|\alpha\right) &= g(\theta) + t(\theta)^T A^{-1}(D-H) \\ \mathrm{Cov}\left(f(\theta), f(\varphi)|\alpha\right) &= \sigma^2\left(g(\theta|u)g(\varphi|u)c(\theta,\varphi) - t(\theta)^T A^{-1} t(\varphi)\right) \end{aligned} \tag{6.12}$$

The posterior distribution on α follows from the multivariate normal likelihood for $D|\alpha$, and, by integrating out σ^2 the joint posterior of the parameters μ, v, b^* can be found:

$$p(\mu, v, b^*|D) \propto \frac{1}{vb^*\sigma^{n+2}}|A|^{-1/2}\exp\left(-\frac{1}{2}(\log b^*)^2\right) \tag{6.13}$$

Using MCMC methods, tripartite samples of values of $\{\mu, v, b^*\}$ may be drawn from this joint posterior. Finally, each tripartite sample can then be used to sample the density function, $f(x)$, from the posterior, $p\left(f(\cdot)|D, \mu, v, b^*\right)$, at a finite number of values of x.

160 There are two practical difficulties in applying Oakley and O'Hagan's approach as it stands to the rainfall case considered by Coles and Tawn. The first is that the natural choice of mean function $m(\cdot|u)$ would be the density of the gamma distribution fitted to the given quantiles, but in that case the covariances involving the percentiles cannot be given in closed form. The second difficulty is the low acceptance rate required to ensure that $f(x) \geq 0 \, \forall x$, so the process can become very inefficient.

161 In addition, Gosling et al. (2007) identify two further difficulties:

(i) the use of a normal distribution as the mean function of the GP might understate the analyst's uncertainty about the tail of the density function; and

(ii) in some circumstances, when the underlying distribution perfectly matches the expert's judgements, the method might yield results implying that the analyst knows the true density with no uncertainty at all.

6.2.2 Gosling, Oakley, and O'Hagan (2007)

162 In Gosling et al. (2007) the the authors follow an approach that is very similar to that in Oakley and O'Hagan (2007) in that the facilitator's beliefs about the density function $f(\theta)$ are represented by a GP that is updated in the light of information elicited from the expert [1]. But to address the problems they see with the method described in the earlier paper, they base the GP on an underlying t-distribution rather than the normal distribution.

163 The facilitator's objective is to develop beliefs about the expert's density function $f(\theta)$, assumed to be smooth and infinitely differentiable everywhere. The facilitator represents $f(\theta)$ as a GP and updates that representation in the light of information elicited from the expert. The initial form of the GP encompasses the facilitator's prior

[1] For convenience the facilitator is denoted as male and the expert as female

beliefs and is defined by a prior expectation function, a prior covariance and a correlation function.

In the following paragraphs we describe: 164

(i) the construction of priors for the GP;

(ii) how the priors are updated in the light of judgements from the expert about quantiles of the distribution $f(\cdot)$;

(iii) making use of information about derivatives;

(iv) incorporating sign information.

The facilitator's prior expectation of $f(\theta)$ is of the form 165

$$\mathrm{E}\left[f(\theta)|m, v, d, b^*, \sigma^2\right] = t(\theta|m.v, d) \tag{6.14}$$

where the hyperparameters comprise, first, the location, scale and degrees of freedom parameters $\mathbf{w} = (m, v, d)$ of the t-distribution, and, secondly, the quantities σ^2, b^* which respectively indicate the closeness of the simulated functions to the underlying t-function and the smoothness of the simulated functions.

The facilitator's prior covariance between $f(\theta)$ and $f(\varphi)$ is given by: 166

$$\mathrm{Cov}\left[f(\theta), f(\varphi)|\mathbf{w}, b^*, \sigma^2\right] = \sigma^2 t(\varphi|\mathbf{w})t(\theta|\mathbf{w})c(\theta, \varphi|\mathbf{w}, b^*) \tag{6.15}$$

in which the correlation function $c(\cdot, \cdot)$ takes the value 1 at $\theta = \varphi$ and is a decreasing function of $|\theta - \varphi|$. The paper uses

$$c(\theta, \varphi|\mathbf{w}, b^*) = \exp\left\{-\frac{1}{2vb^*}(\theta - \varphi)^2\right\} \tag{6.16}$$

167 Before the GP thus constructed can be updated in the light of information elicited from the expert, priors need to be defined for the five hyperparameters. The paper deals with priors for m, v and d separately from those for b^* and σ^2.

168 As regards m, v and d, the facilitator has only vague prior beliefs about m and v. He believes that d needs to cover a range up to a near-normal density but he has no grounds for preferring any point on that range over any other. A uniform distribution is therefore chosen for d leading to the following priors:

$$\begin{aligned} d &\sim \mathcal{U}_{[0,40]} \\ p(m, v, d) &\propto v^{-1} \end{aligned} \tag{6.17}$$

169 The priors σ^2 and b^* need to take into account beliefs the facilitator has about $f(\cdot)$. He will also wish to make choices that improve the efficiency of the sampling process. These criteria can be expressed in graphical terms in the plane defined by σ^2 and b^*, such that sampling will be concentrated on a central region of the plane. That region is defined by four lines, as shown in Figure 6.1. The joint prior density of (b^*, σ^2) is shown as contours in the figure and has been calculated by reference to the regions defined by the lines. The regions have the following characteristics:

(i) The $f(\cdot)$ derived from draws of (b^*, σ^2) from within the region will not focus too much on $f(\cdot)$ that are very close to the underlying t-density. The lower sloping line marks the top of a region in which the maximum absolute difference between the $f(\cdot)$ and the underlying t-density is less than 0.25.

(ii) The $f(\cdot)$ will not include too many multimodal cases. The left vertical line marks the leftmost boundary of a region in which fewer than 20% of draws of $f(\cdot)$ will have three or more modes.

(iii) The $f(\cdot)$ will not include too many cases which are not valid densities as a result of $f(\theta) < 0$ for some θ. The upper sloping line marks the lower boundary of a region in which the proportion of points that are negative exceeds 0.1.

(iv) The correlation matrix will be invertible without numerical problems. The numerical problems arise more frequently to the right of the vertical line $b^* = 5$

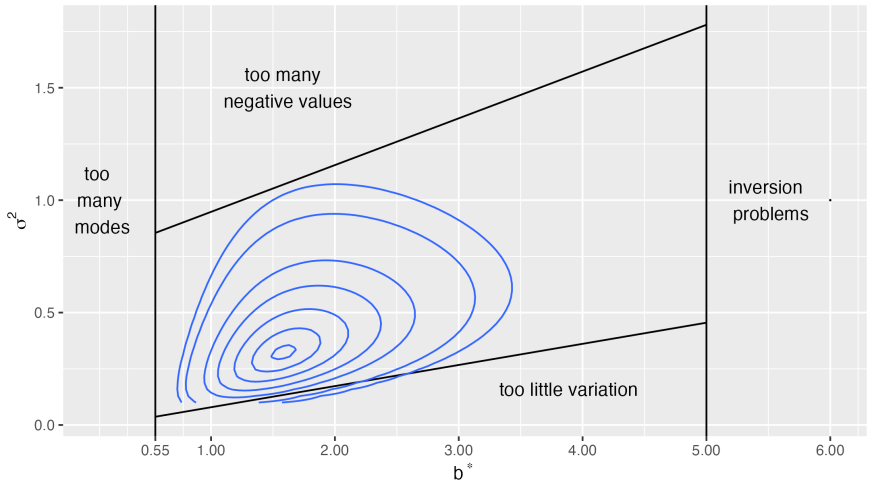


Figure 6.1: Joint prior density of (b^*, σ^2) calculated from linear criteria

If the expert expresses her judgements in terms of quantiles of the distribution for θ, 170 the vector of judgements D is of the form:

$$\begin{aligned} D^T &= \left(\int_{x_0}^{x_1} f(x)dx, \ldots, \int_{x_{n-1}}^{x_n} f(x)dx \right) \\ &= (P_{x_0,x_1}, \ldots, P_{x_{n-1},x_n}) \end{aligned} \tag{6.18}$$

The expectation of D is then

$$\begin{aligned} \mathrm{E}[D|\mathbf{w}]^T &= \left(\int_{x_0}^{x_1} g(x|\mathbf{w})dx, \ldots, \int_{x_{n-1}}^{x_n} g(x|\mathbf{w})dx \right) \\ &= H^T \end{aligned} \tag{6.19}$$

The facilitator's beliefs about $f(\cdot)$ now conditional on D as well as on $\mathbf{w}, b^*, \sigma^2$ is also a GP constructed using

$$\begin{aligned} \mathrm{Cov}(f(\theta), P_{x_i,x_j}|\mathbf{w}, b^*, \sigma^2) &= \sigma^2 g(\theta|\mathbf{w}) \int_{x_i}^{x_j} g(x|\mathbf{w}) c(\theta, x|b^*, v)dx \\ \mathrm{Cov}(P_{x_i,x_j}, P_{y_i,y_j}) &= \sigma^2 \int_{y_i}^{y_j} \int_{x_i}^{x_j} g(x|\mathbf{w}) g(y|\mathbf{w}) c(x, y|b^*, v)dx\, dy \end{aligned} \tag{6.20}$$

giving rise to a mean

$$\mathrm{E}[f(\theta|D, \mathbf{w}, b^*, \sigma^2] = g(\theta|\mathbf{w}) + \mathbf{t}(\theta|\mathbf{w}, b^*, \sigma^2)^T A^{-1}(D - H). \tag{6.21}$$

and a covariance

$$\mathrm{Cov}(f(\theta), f(\varphi)|D, \mathbf{w}, b^*, \sigma^2) = \sigma^2(g(\theta|\mathbf{w})g(\varphi|\mathbf{w})c(\theta, \varphi|b^*, v) - \mathbf{t}(\theta|\mathbf{w}, b^*, \sigma^2)^T A^{-1}\mathbf{t}(\varphi|\mathbf{w}, b^*, \sigma^2)^T, \quad (6.22)$$

where, in each equation,

$$\mathbf{t}(\theta|\mathbf{w}, b^*, \sigma^2)^T = \left(\mathrm{Cov}(f(\theta), P_{x_0,x_1}|\mathbf{w}, b^*, \sigma^2), \ldots, \mathrm{Cov}(f(\theta), P_{x_{n-1},x_n}|\mathbf{w}, b^*, \sigma^2)\right). \quad (6.23)$$

171 Although it is not regarded as feasible to ask the expert to give values of the derivative of $f(\cdot)$ for different values of θ, the authors see scope for asking the expert for some simple judgements. For example, the expert could be asked to give the mode M of the distribution, where $df(M)/d\theta = 0$. The paper shows how the data vector D given in Equation 6.18 is changed to

$$\begin{aligned} D^T &= \left(\int_{x_0}^{x_1} f(x)dx, \ldots, \int_{x_{n-1}}^{x_n} f(x)dx\right) \\ &= (P_{x_0,x_1}, \ldots, P_{x_{n-1},x_n}, 0), \end{aligned} \quad (6.24)$$

the facilitator's prior expectation Equation 6.19 becomes

$$\begin{aligned} \mathrm{E}[D^T|\mathbf{w}] &= H^T \\ &= \left(\int_{x_0}^{x_1} g(x|\mathbf{w})dx, \ldots, \int_{x_{n-1}}^{x_n} g(x|\mathbf{w})dx, \frac{dg(M|\mathbf{w}}{d\theta}\right). \end{aligned} \quad (6.25)$$

It is then straightforward to construct a GP incorporating this information.

172 Similarly, it is shown how the covariances and the facilitator's prior expectation may be amended to permit information from the expert about the sign of certain quantities to be incorporated in the analysis. That is helpful when the expert has been asked to give the mode of the distribution because without the condition

$$\frac{d^2f(M)}{d\theta^2} < 0 \quad (6.26)$$

modes cannot be distinguished from anti-modes. The paper shows how information about the sign of the second derivative can be included in the construction of the GP.

Generalising from the idea of using information from the expert about the sign of the second derivative of $f(\cdot)$, the paper then shows how to incorporate sign judgements for any order of derivative of $f(\cdot)$ and for $f(\cdot)$ itself. The latter facility helps address one of the problems with using a GP to represent a probability density, namely that realisations of the GP may not be valid densities because they include negative values of $f(\theta)$ for some θ. The remedy, that prevents the posterior density being negative "too often" is to include the judgement that $f(\theta_i) > 0$ for a set of values $\{\theta_i : i = 1, \ldots, m\}$. 173

6.2.3 Finite-dimensional Gaussian approximation

The method described above (Subsection 6.2.2) reduces but does not eliminate the disadvantage (shared with the method in Subsection 6.2.1) that realisations of the GPs to which they give rise may not all be valid probability density or distribution functions. That means that a further stage of rejection is required to obtain valid results. Such a stage is a source of inefficiency in using the methods that it would be good, if possible, to avoid. We now describe some methods that were devised in a rather different context but which can be modified for our purposes. These methods bring the advantage that GPs may be produced such that all of them satisfy the requirements of consistency with probability densities or distributions. 174

The context of the work we describe here arises particularly in the construction of **emulators** for computer experiments. Such an emulator is a statistical surrogate that is very much cheaper to run than some large scale computer model and where the differences between the emulator's outputs and those of the large model can be estimated statistically. The methods presented in the papers we describe below have in common the use of GPs defined at a finite set of knots to approximate the modelling of a given function that is in principle known but in practice prohibitively expensive to calculate, and that is subject to known linear inequality constraints. 175

Maatouk and Bay (2017) consider the case in which the function to be approximated is $f : [0, 1]^d \to \mathbb{R}$, where d is the number of dimensions, and we are given the value of f at n distinct locations: 176

$$f(\mathbf{x}^{(i)}) = y_i \quad i = 1, ..., n. \tag{6.27}$$

(For consistency with subsequent notation we have changed Maatouk and Bay (2017) indexing generally to start at unity rather than zero, as in the paper.)

Let $(Y(\mathbf{x}))_{\mathbf{x}\in[0,1]^d}$ be a zero-mean GP with covariance function K, and let $C^0([0, 1]^d)$ be the space of continuous functions on $[0, 1]^d$. Then let the $C \subset C^0$ be the subspace of continuous functions that comply with the given set of linear inequality constraints.The 177

interpolation conditions and inequality constraints to which Y is subject are, therefore:

$$Y\left(\mathbf{x}^{(i)}\right) = y_i$$
$$Y \in C$$

The aim is to derive approximations to the infinite-dimensional GP Y of the form:

$$Y^N(\mathbf{x}) := \sum_{j=1}^{N} \xi_j \varphi_j(\mathbf{x}) \quad \mathbf{x} \in \mathbb{R}^d, \tag{6.28}$$

where

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^T \tag{6.29}$$

is a realisation of a d-dimensional zero-mean GP with covariance matrix $\mathbf{\Gamma}^N$, and $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_N)^T$ is a vector of basis functions constructed as described below By this construction, Y^N is a zero-mean GP with covariance function

$$K_N(\mathbf{x}, \mathbf{x}') = \boldsymbol{\varphi}(\mathbf{x}) \Gamma \boldsymbol{\varphi}(\mathbf{x}'). \tag{6.30}$$

178 With such a structure the task of simulating an infinite-dimensional GP is reduced to the simulation of a finite-dimensional $\boldsymbol{\xi}$ subject to the **interpolation conditions**, $Y_N(\mathbf{x}^{(i)}) \equiv \sum_{j=1}^{N} \xi_j \varphi_j(\mathbf{x}^{(i)}) = y_i \, (i = 1, \ldots, n)$, and to $Y_N \in \mathcal{E}$, where $\mathcal{E}$ denotes the desired set of **inequality conditions**.

179 With the sequence of knots, $\{u_i : 0 < u_1 < \cdots < u_N < 1\}$ where $u_N < \inf(x : F(x) = 1)$ we can then define for $j \in [2, N-1]$ a sequence of individual basis functions $\varphi_j(x) = h_j(x)$ each derived from the "witch's hat" function

$$h(x) = \begin{cases} 1 - |x| & |x| < 1 \\ 0 & \text{otherwise,} \end{cases} \tag{6.31}$$

and with support on the range $[u_{j-1}, u_{j+1}]$.

180 It is shown that Y^N is then indeed a finite-dimensional GP that converges uniformly pathwise to Y with unit probability as $N \to \infty$. Similarly, it is then shown that monotonicity constraints can be applied, in the one-dimensional case with which we are concerned, by using basis functions of the form $\varphi_j(x) := \int_0^x h_j(t)dt$ where $x \in [0, 1]$.

López-Lopera et al. (2017) generalise the approach in Maatouk and Bay (2017) to permit a wider range of constraints to be applied. The aim is to provide a method for calculating a finite-dimensional GP of the form 181

$$Y_m(x) = \sum_{j=1}^{m} \xi_j \varphi_j(x), \quad \text{s.t.} \begin{cases} Y_m(x_i) = y_i & i = 1, \ldots, N \\ Y_m \in \mathcal{E}. \end{cases} \tag{6.32}$$

Focussing on the case in which the inequality constraints may be expressed in the matrix form $\boldsymbol{l} \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u}$, the situation is expressed as follows:

$$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}) \quad \text{s.t.} \begin{cases} \boldsymbol{\Phi\xi} = \boldsymbol{y} & \text{(interpolation conditions)} \\ \boldsymbol{l} \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u} & \text{(inequality conditions).} \end{cases} \tag{6.33}$$

The authors note that provided the number, q, of constraints is not exceeded by the number, m, of knots then $\boldsymbol{\Lambda}$ is injective and that $\boldsymbol{\Lambda\xi} = \boldsymbol{\eta}$ has a unique solution when $\boldsymbol{\eta}$ is in the image space of $\boldsymbol{\Lambda}$.

An algorithm is presented for sampling from the finite-dimensional GP with linear inequality constraints. We describe a slightly amended version of that algorithm below (Algorithm 1) in the direct context of the problem we address in this thesis but, in short, the posterior distribution of Equation (6.33) is obtained from 182

$$\boldsymbol{\xi} | \{\boldsymbol{\Phi\xi} = \boldsymbol{y}, l \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u}\} \sim \mathcal{TN}(\boldsymbol{\Lambda\mu}, \boldsymbol{\Lambda\Sigma\Lambda^T}, \boldsymbol{l}, \boldsymbol{u}) \tag{6.34}$$

where

$$\boldsymbol{\mu} = \boldsymbol{\Gamma\Phi}^T[\boldsymbol{\Phi\Gamma\Phi}^T]^{-1}\boldsymbol{y}, \quad \text{and} \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma\Phi}^T[\boldsymbol{\Phi\Gamma\Phi}^T]^{-1}\boldsymbol{\Phi\Gamma} \tag{6.35}$$

and the notation $\mathcal{TN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{l}, \boldsymbol{u})$ denotes a truncated multivariate normal distribution with mean $\boldsymbol{\mu}$, variance matrix $\boldsymbol{\Sigma}$, lower limit vector $\boldsymbol{l}$, and upper limit vector $\boldsymbol{u}$.

6.2.4 Notation

We will use the notation adopted in Rasmussen and Williams (2006). 183

A real GP, $f(\mathbf{x})$, is determined by its mean function, $m(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d$, and its covariance function $k(\mathbf{x}, \boldsymbol{x'})$: 184

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], \\ k(\mathbf{x}, \boldsymbol{x'}) &= \mathbb{E}[(f(\mathbf{x}) - m(\boldsymbol{x'}))(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))] \end{aligned} \tag{6.36}$$

written

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \boldsymbol{x'})). \tag{6.37}$$

185 By varying the choice of covariance function, and, in particular, by making use of the fact that linear combinations of positive-definite functions are positive-definite, an extremely wide range of functions can be modelled as GPs (Rasmussen and Williams, 2006, ch.4).

186 If we have two vectors $X = (\mathbf{x}_i : i = 1, \ldots, m)^T$, $X' = (\boldsymbol{x'}_j : j = 1, \ldots, n)^T$ then we write

$$K(X, X') = \begin{pmatrix} k(\mathbf{x}_1, \boldsymbol{x'}_1) & \ldots & k(\mathbf{x}_1, \boldsymbol{x'}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_m, \boldsymbol{x'}_1) & \ldots & k(\mathbf{x}_m, \boldsymbol{x'}_n) \end{pmatrix} \tag{6.38}$$

187 We have **training data** comprising a set of pairs $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$. The $\{\mathbf{x}_i\}$ are denoted the **design inputs** and written X. The $\{y_i\}$, written $\boldsymbol{f}$, are the corresponding **training outputs**. For clarity of exposition in this chapter, we present the noise-free case, so $y_i = f(\mathbf{x}_i) \, \forall i$.

6.2.5 Representing an unknown function as a Gaussian process

188 The definition of a GP refers to the joint distribution of a *finite* number of random variables. We represent a single function by treating each of those random variables as the value taken by the function at a single point. So a single draw from their joint distribution forms a (finite) set of value pairs for a single function. Our task is, therefore, to find the conditional distribution of $\boldsymbol{f}_*$, namely of the values of $f(\cdot)$ at n_* **test** points $X_* = \{\boldsymbol{x}_{*j} : j = 1, \ldots, n_*\}$, that is to find $\boldsymbol{f}_* | X_*, X, \boldsymbol{f}$.

Theorem 10. *(Rasmussen and Williams, 2006, Equation 2.19) The distribution of*

$$\boldsymbol{f}_* | X_*, X, \boldsymbol{f}. \tag{6.39}$$

is multivariate normal with mean

$$\bar{\boldsymbol{f}}_* = K(X_*, X) K(X, X)^{-1} \boldsymbol{f} \tag{6.40}$$

and variance

$$\boldsymbol{\Sigma} = K(X_*, X_*) - K(X_*, X) K(X, X)^{-1} K(X, X_*) \tag{6.41}$$

Proof. The joint distribution of $\boldsymbol{f}$ and $\boldsymbol{f_*}$ is given by:

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f_*} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{6.42}$$

corresponding to a joint probability distribution of the form

$$p(\boldsymbol{y}) = (2\pi)^{-\frac{n+n_*}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y}-\boldsymbol{\mu})\right\} \tag{6.43}$$

where

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f_*} \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \mathbf{0} \\ \bar{\boldsymbol{f}}_* \end{bmatrix} \tag{6.44}$$

and

$$\Sigma = \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \tag{6.45}$$

For ease of reading we write

$$\begin{aligned} K(X,X) &= A_{11} \\ K(X,X_*) &= A_{12} \\ K(X_*,X) &= A_{21} \\ K(X_*,X_*) &= A_{22}, \end{aligned} \tag{6.46}$$

then we have

$$\Sigma = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

and we write

$$\Sigma^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

so that

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_n & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{n_*} \end{bmatrix}$$

whence

$$\begin{aligned} B_{11}A_{11} + B_{12}A_{21} &= \boldsymbol{I}_n \\ B_{11}A_{12} + B_{12}A_{22} &= \mathbf{0}. \\ B_{21}A_{12} + B_{22}A_{22} &= \boldsymbol{I}_{n_*} \\ B_{21}A_{11} + B_{22}A_{21} &= \mathbf{0}. \end{aligned} \tag{6.47}$$

Now, the joint probability distribution (Equation 6.43) may be written in terms of the

two components of $\boldsymbol{y}$, namely $\boldsymbol{f}$, and $\boldsymbol{f_*}$ and the exponential term would become:

$$-\frac{1}{2}\left\{(\boldsymbol{f_*}-\bar{\boldsymbol{f}}_*)^T B_{22}(\boldsymbol{f_*}-\bar{\boldsymbol{f}}_*)+2(\boldsymbol{f_*}-\bar{\boldsymbol{f}}_*)^T B_{12}\boldsymbol{f}+\boldsymbol{f}^T B_{11}\boldsymbol{f}\right\} \tag{6.48}$$

which is seen to be a quadratic form in $\boldsymbol{f_*}$ given $\boldsymbol{f}$. Completing the square, we see that the exponential term in the probability distribution of $\boldsymbol{f_*}|\boldsymbol{f}$ is proportional to the form

$$\begin{aligned}-\frac{1}{2}\left\{(\boldsymbol{f_*}-\bar{\boldsymbol{f}}_*)^T B_{22}(\boldsymbol{f_*}-\bar{\boldsymbol{f}}_*)+2{\boldsymbol{f_*}}^T B_{21}\boldsymbol{f}\right\} = \\ -\frac{1}{2}\left\{\boldsymbol{f_*}-\bar{\boldsymbol{f}}_*+B_{22}^{-1}B_{21}\boldsymbol{f}\right\}^T B_{22}\left\{\boldsymbol{f_*}-\bar{\boldsymbol{f}}_*+B_{22}^{-1}B_{21}\boldsymbol{f}\right\}+K\end{aligned} \tag{6.49}$$

where K denotes a constant not involving $\boldsymbol{f_*}$. The density of $\boldsymbol{f_*}|\boldsymbol{f}$ is therefore:

$$\begin{aligned}\boldsymbol{f_*}|\boldsymbol{f} &\sim \mathcal{N}(\bar{\boldsymbol{f}}_*-B_{22}^{-1}B_{21}\boldsymbol{f}, B_{22}) \\ &= \mathcal{N}(\bar{\boldsymbol{f}}_*+A_{21}A_{11}\boldsymbol{f}, A_{11}-A_{12}A_{22}^{-1}A_{21})\end{aligned} \tag{6.50}$$

Returning to our original notation, we have thus shown that $\boldsymbol{f_*}|\boldsymbol{f}$ has a multivariate normal distribution with mean

$$\bar{\boldsymbol{f}}_*+K(X,X_*)K(X,X)^{-1}\boldsymbol{f}$$

and variance

$$K(X^*,X^*)-K(X^*,X)K(X,X)^{-1}K(X,X^*)$$

□

189 It follows that the corresponding density is

$$p(\boldsymbol{y}|X_*,X,\boldsymbol{f})=(2\pi)^{-\frac{n_*}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\bar{\boldsymbol{f}}_*)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\bar{\boldsymbol{f}}_*)\right\}. \tag{6.51}$$

6.2.6 Integrals or derivatives as training inputs

190 In arriving at Theorem 10 we assumed that our data comprise a set of training outputs $\{y_i\}$ taken by the function f at the design inputs $\{\mathbf{x}_i\}$. In some contexts, including, in particular, the case where f is a probability density or distribution, the data might instead comprise or include derivatives or integrals of f at the design inputs. Both cases can generally be dealt with as GPs by modifying the covariances and means in equations 6.40, and 6.41.

191 In the case of differentiation, suppose that our training data comprise n sets of value

pairs

$$\boldsymbol{f'}_j = \{(x_j^i, \omega_j^i)\} \quad j = 1, \ldots, n,\ i = 1, \ldots, d, \tag{6.52}$$

each $\boldsymbol{f'}_j$ corresponding to the n points of the j^{th} partial derivative of the function f, that is:

$$\omega_j^i = \left.\frac{\partial f(\mathbf{x})}{\partial x_j}\right|_{\mathbf{x}=x_j^i} \quad i = 1, \ldots, d \tag{6.53}$$

Then, provided the covariance function $k(\cdot, \cdot)$ is at least twice differentiable, the covariances between function values and partial derivatives and between partial derivatives and partial derivatives are: 192

$$\mathrm{Cov}(\omega_j^i, y_k) = \frac{\partial}{\partial x_j} k(\mathbf{x}_i, \mathbf{x}_k)\,, \quad \mathrm{Cov}(\omega_j^i, \omega_l^m) = \frac{\partial^2}{\partial x_j \partial x_l} k(\mathbf{x}_i, \mathbf{x}_m). \tag{6.54}$$

And provided the mean function $m(\cdot)$ is differentiable

$$\mathbb{E}\left[\frac{\partial f}{\partial x_i}\right] = \frac{\partial m}{\partial x_i} \quad i = 1, \ldots, d. \tag{6.55}$$

Similarly, where the training outputs comprise integrals (and, for illustration, showing only the univariate case) 193

$$\boldsymbol{F}_j = \left\{ \left(x_j, \int_{-\infty}^{x_j} f(\theta)\, d\theta \right) \right\} \quad j = 1, \ldots, n \tag{6.56}$$

then the covariances are

$$\begin{aligned} \mathrm{Cov}\left(\int_{-\infty}^{x_i} f(\theta)\, d\theta, y_j \right) &= \int_{-\infty}^{x_i} k(\varphi, \theta)\, d\varphi \\ \mathrm{Cov}\left(\int_{-\infty}^{x_i} f(\theta)\, d\theta, \int_{-\infty}^{x_j} f(\theta)\, d\theta \right) &= \int_{-\infty}^{x_i} \int_{-\infty}^{x_j} k(\varphi, \theta)\, d\varphi\, d\theta \end{aligned} \tag{6.57}$$

and

$$\mathbb{E}\left[\int_{-\infty}^{x_i} f(\theta)\, d\theta \right] = \int_{-\infty}^{x_i} m(x)\, dx \quad i = 1, \ldots, n. \tag{6.58}$$

See Oakley and O'Hagan (2007, Appendix) for an example of these integral derivations in a particular case.

§ 6.3 Conclusions

194 We have described approaches that have been taken to using GPs to model probability distributions or densities based on elicitation from experts for use as priors in Bayesian analysis. Although the methods proposed are workable they are subject to some inefficiencies. In the next chapter we present an approach to overcome these problems.

Chapter 7

Efficient simulation of probability functions from Gaussian processes

We have seen that earlier work on using non-parametric means to incorporate the results 195
of expert elicitation into Bayesian inference is potentially inefficient in that functions are simulated that cannot possibly be appropriate probability densities or distributions. In this chapter we present a new method that is a development of that work that also draws on some of the ideas presented in the previous chapter to avoid the potential inefficiencies. In the next chapter we apply this new method to the generation of valid solutions of the rainfall problem examined by Coles and Tawn (1996).

§ 7.1 Preliminaries

We assume that we are given a number of quantiles (that may be as small as one or 196
two) and wish to simulate a full range of unimodal probability densities $f(x)$ that are consistent with those quantiles. We also seek densities that are infinitely differentiable everywhere except at no more than a finite number of points. In what follows we will loosely use the term "infinitely differentiable" to include such cases. Following Maatouk and Bay (2017), we may also assume for convenience and without loss of generality that the densities have support only on the interval $[0, 1]$ (transforming the data inputs

correspondingly). It is also convenient to proceed by simulating the corresponding distribution functions, $F(x)$, such that $F'(x) = f(x)$, and then to derive densities by differentiation.

197 Let the given quantiles be as follows:

$$F(x_i) = y_i : \quad i = 1, \ldots, n \tag{7.1}$$

which we can write as a vector of cases as

$$F(\mathbf{x}) = \mathbf{y}. \tag{7.2}$$

We then also have to satisfy the following for $F(x)$ to be a valid distribution function:

$$F(0) = 0, \quad F(1) = 1, \quad x < x' \Rightarrow F(x) \leq F(x') \quad x, x' \in [0, 1]. \tag{7.3}$$

It follows that the requirements, $F(\mathbf{x}) = \boldsymbol{y}$, $F(0) = 0$, $F(1) = 1$, are the interpolation constraints.

198 Our task then is to simulate $F(x)$ subject to those constraints and to the further requirements stemming from our assumptions that $F(x)$ must be infinitely differentiable and that $f(x)$ must be unimodal.We describe below an approach in which all the conditions are met by an appropriate choice of finite-dimensional parameters.

§ 7.2 Finite Linear Sums of Basis functions

199 We create a zero-mean[1] finite-dimensional GP with a given covariance function such that each realisation, $\boldsymbol{\xi} = (\xi_0, \ldots, \xi_N)$, may be used with sets of basis functions $\{h_j(\cdot) : j = 0, \ldots, N\}, \{\varphi_j(\cdot) : j = 0, \ldots, N\}$ to define two functions, as follows:

$$f(x) = \sum_{j=0}^{N} \xi_j h_j(x)$$

$$F(x) = \sum_{j=0}^{N} \xi_j \varphi_j(x)$$

[1]Setting the mean of the GP to be zero implies nothing about the mean of any realisation of the GP.

where f is a valid PDF and F is the corresponding CDF and where all the interpolation constraints are complied with. The choice of covariance function does not affect the results described in this section but in making use of this structure we shall use the following covariance function:

$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\theta^2}\right) \tag{7.4}$$

with parameters σ, and θ to be chosen.

Although we use exactly the same methods as Maatouk and Bay (2017) and López- 200
Lopera et al. (2017), to construct a GP, each realisation of which is a set of parameters, $\boldsymbol{\xi}$, defining a function[2], it is important to understand that we are not seeking to approximate some unknown function. Each realisation of our GP generates a pair comprising a valid PDF and a valid CDF that are consistent with the interpolation constraints.

To achieve the smoothness in the PDF required by our assumptions about the probability 201
density we replace as basis function the simple witch's hat function as defined by equation (6.31) with the following smooth and infinitely differentiable function:

$$h(x) = \frac{1}{2} - \frac{1}{2}\cos(\pi(1 + x)). \tag{7.5}$$

We can then define the individual basis functions centred on a sequence of knots, 202
$\{u_i : 0 < u_1 < \cdots < u_N < 1\}$ where $u_N < \inf(x : F(x) = 1)$ and for $j \in [2, N-1]$

$$h_j(x) = \begin{cases} 0 & x < u_{j-1} \\ \frac{1}{2}\left[1 - \cos\left(\pi \frac{x - u_{j-1}}{u_j - u_{j-1}}\right)\right] & x \in [u_{j-1}, u_j] \\ \frac{1}{2}\left[1 + \cos\left(\pi \frac{x - u_j}{u_{j+1} - u_j}\right)\right] & x \in [u_j, u_{j+1}] \\ 0 & x > u_{j+1} \end{cases} \tag{7.6}$$

with obvious adjustments for the cases $j = 1$ and $j = N$.

It is important to note that these new basis functions satisfy what is in Maatouk and 203

[2]Actually in our method it defines two functions

Bay (2017) the crucial requirement

$$h_j(u_i) = \delta_{ij},$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{7.7}$$

and that then the following two theorems stated in Maatouk and Bay (2017) remain true:

Theorem 11. *$F_N(x) := \sum_{j=0}^{N} F'(u_j)h_j(x)$ converges uniformly pathwise to $F(x)$ when N tends to infinity.*

As to the requirement for monotonicity, we have:

Theorem 12. *With a choice of basis function $\varphi_j(x) = \int_0^x h_j(t)\,dt$ the approximation $F_N(x) := \sum_{j=0}^{N} \xi_j \varphi_j(x)$ is monotonic increasing if and only if the parameters $\{\xi_j\}$ are all non-negative.*

204 In that event $\xi_i = F'(u_i) : i = 0, \ldots, N$. In consequence, we see that the vector $\boldsymbol{\xi}$ may be viewed as a realisation of the GP $(F'(x))$, namely the derivative of a GP with covariance function K. It follows that the covariance matrix, $\boldsymbol{\Gamma}$ of $\boldsymbol{\xi}$ viewed as a GP is given by:

$$\boldsymbol{\Gamma}_{ij} = Cov(\xi_i, \xi_j) = \frac{\partial^2 K}{\partial \xi_i \, \partial \xi_j}. \tag{7.8}$$

205 The proofs of these propositions do not depend on the exact form of the hat functions used in the generation of the PDF, but only on the requirement that $h_j(u_i) = \delta_{ij}$.

§ 7.3 Differentiability, Monotonicity and Unimodality

We now show that in satisfying the interpolation constraints we also obtain well-behaved distribution functions and densities. Note that by construction we have also ensured that $F(0) = 0$. In the following, we will assume that $F(1) = 1$ is included in the interpolation constraint $F(\mathbf{x}) = \boldsymbol{y}$. It follows immediately from Theorem 12 and from the definition of $\varphi_j(\cdot)$ that $F_N(x) := \sum_{j=0}^{N} \xi_j \varphi_j(x)$ will be monotonic increasing and everywhere infinitely differentiable provided $\xi_j \geq 0 : j = 0, \ldots, N$. 206

We turn now to the conditions for the unimodality of $f(x)$. Because $h_j(u_i) = \delta_{ij}$, $f(u_i) = \xi_i$, and then, since $f(x)$ is monotonic between any two adjacent knots, the condition for $f(u_i)$ to be a local mode is 207

$$\xi_{i-1} < \xi_i > \xi_{i+1}$$

The condition for $f(x)$ to possess a single mode over $[0, 1]$ at $x = u_i$ is

$$\xi_0 \leq \xi_1 \leq \cdots \leq \xi_{i-1} < \xi_i > \xi_{i+1} \geq \cdots \geq \xi_N. \tag{7.9}$$

If, further,

$$0 \leq \xi_0 \leq \xi_1 \leq \cdots \leq \xi_{i-1} < \xi_i > \xi_{i+1} \geq \cdots \geq \xi_N \geq 0 \tag{7.10}$$

the conditions of both Equation(7.3) and Equation(7.9) are satisfied, so that $F(x)$ will then be monotonic and $f(x)$ will be unimodal.

It will be seen below that we can test for the feasibility of a single mode of $f(x)$ at each knot. If there is a mode at u_i, say, we are not too concerned if there is a plateau rather than a peak, that is, if $f(u_{i-1}) = f(u_i)$ or $f(u_i) = f(u_{i+1})$. We can therefore relax the strict inequalities in equation 7.10 to obtain 208

$$0 \leq \xi_0 \leq \xi_1 \leq \cdots \leq \xi_{i-1} \leq \xi_i \geq \xi_{i+1} \geq \cdots \geq \xi_N \geq 0, \tag{7.11}$$

which is a condition that would also permit plateaux below the peak density.

It is helpful to express the above in terms of matrices. We can define the matrix $\boldsymbol{\Phi}_{i,j} = \varphi_j(x_i)$ and then write: 209

$$y_i = \sum_{j=0}^{N} \xi_j \varphi_j(x_i) \quad i = 1, \ldots, n$$

as

$$\boldsymbol{\Phi\xi} = \mathbf{x}.$$

The conditions on $\boldsymbol{\xi}$ in Equation(7.11) may be expressed in the form

$$\boldsymbol{l} \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u}$$

where $\boldsymbol{l}$ and $\boldsymbol{u}$ are vectors of lower and upper bounds. In what follows we shall sometimes write these inequality conditions in the form:

$$\begin{aligned} \boldsymbol{\Lambda\xi} - \boldsymbol{l} &\geq \mathbf{0} \\ -\boldsymbol{\Lambda\xi} + \boldsymbol{u} &\geq \mathbf{0}. \end{aligned} \tag{7.12}$$

210 It follows, comparably to Equation(6.33), that we can restate our problem as that of sampling from

$$\boldsymbol{\xi}|\{\boldsymbol{\Phi\xi}, \boldsymbol{l} \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u}\}, \quad \text{where } \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \Gamma). \tag{7.13}$$

§ 7.4 The Sampling Algorithm

211 As noted above (Subsection 6.2.3), we slightly adapt the sampling algorithm from López-Lopera et al. (2017)[Algorithm 1] as set out in (our) Algorithm 1:

Algorithm 1: Sampling from GP with linear inequality constraints

Procedure: Sampling from $\boldsymbol{\xi}|\{\boldsymbol{\Phi\xi} = \boldsymbol{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u}, \}$ where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$;

Input: $\boldsymbol{y}, \boldsymbol{\Gamma} \in \mathbf{R}^{n\times N}, \boldsymbol{\Phi} \in \mathbf{R}^{n\times N}, \mathcal{C}$;

Compute conditional mean and covariance of $\boldsymbol{\xi}|\{\boldsymbol{\Phi\xi} = \boldsymbol{y}\}$:

$$\boldsymbol{\mu} = \boldsymbol{\Gamma\Phi}^T(\boldsymbol{\Phi\Gamma\Phi}^T)^{-1}$$
$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma\Phi}^T(\boldsymbol{\Phi\Gamma\Phi}^T)^{-1}\boldsymbol{\Phi\Gamma}$$

Find start for Monte Carlo simulation by solving:

$$\boldsymbol{\mu}_\xi^* = \min_{\boldsymbol{\xi}\in\mathbf{R}^N} \{\boldsymbol{\xi^T\Gamma}^{-1}\boldsymbol{\xi}|\boldsymbol{\Phi\xi} = \boldsymbol{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u}\}$$

Start is then solution of $\boldsymbol{\nu_\xi}^* = \boldsymbol{\Lambda\mu_\xi}^*$;

Sample from the truncated multinormal distribution

$$\boldsymbol{\Lambda\xi}|\{\boldsymbol{\Phi\xi} = \boldsymbol{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u}\} \sim \mathcal{TN}(\boldsymbol{\Lambda\mu}, \boldsymbol{\Lambda\Sigma\Lambda}^T, \boldsymbol{l}, \boldsymbol{u})$$

Set $\boldsymbol{\eta} = \boldsymbol{\Lambda\xi}$;

Result: $\boldsymbol{\xi}$ is solution of $\boldsymbol{\eta} = \boldsymbol{\Lambda\xi}$

§ 7.5 Hamiltonian Monte Carlo

López-Lopera et al. (2017) review and test various methods of sampling from a truncated multivariate normal distribution in the context of their work. Their tests showed that an approach based on HMC that had been introduced by Pakman and Paninski (2014) was clearly the fastest method, and in terms of a measure of effective sample size was also clearly the most efficient when applied to a test case of monotonicity. 212

As introduced by Metropolis et al. (1953), MCMC in its original form was a simple random walk of a parameter $\mathbf{x} \in \mathbb{R}^N$ where N denotes the number of coordinates. A well-known problem emerges if the constraints impose high correlations on the coordinates. Then, an extremely large number of steps may well be required to explore the probability space of the target distribution effectively. That problem arises because if the random walk is composed of **large** steps then the acceptance probability tends to be low, but when the steps are **small** then successive states of the system tend to be strongly correlated. 213

HMC provides a way of improving upon the simple random walk. The idea is to derive 214

proposed states in terms of the vector $\mathbf{x} = (x_1, \ldots, x_N)$ through random draws of a second parameter vector $\boldsymbol{S} = (s_1, \ldots, s_N)$ in such a way that small steps in $\boldsymbol{S}$ need not lead to small steps in $\mathbf{x}$. A state for $\mathbf{x}$ proposed in this way can therefore be distant from the current state but nevertheless have a high probability of acceptance (Neal, 2011). The relationship between $\mathbf{x}$ and $\boldsymbol{S}$ is defined by visualising the sampling process as the unforced movement of a notional (Newtonian) physical system where $\mathbf{x}$ describes the position and $\boldsymbol{S}$ the momentum of its N components. The total energy of the system, the **Hamiltonian** $H(\boldsymbol{S}, \mathbf{x})$, is conserved:

$$H(\boldsymbol{S}, \mathbf{x}) = U(\mathbf{x}) + K(\boldsymbol{S}) = \text{constant} \tag{7.14}$$

where $U(\mathbf{x})$ denotes the potential energy of the system and $K(\boldsymbol{S})$ its kinetic energy. Since $\mathbf{x}$ and $\boldsymbol{S}$ are independent and H is constant, we can derive samples of $\mathbf{x}$ by sampling $\boldsymbol{S}$. The case we are interested in is where $\mathbf{x}$ is a draw from the desired posterior distribution so we can set

$$U(\mathbf{x}) = -\log\left(\pi(\mathbf{x})L(\mathbf{x}|D)\right) \tag{7.15}$$

where the $\pi(\mathbf{x})$ is the prior and $L(\mathbf{x}|D)$ the likelihood given the data D. We are free to choose the form of the kinetic energy. A common choice is

$$K(\boldsymbol{S}) = \sum_{i=1}^{N} \frac{s_i^2}{2m_i} \tag{7.16}$$

where each m_i is the chosen variance of the corresponding s_i.

215 The HMC algorithm proceeds in two steps:

(i) draw new values of $\boldsymbol{S}$ from its distribution, Equation 7.16. This step sets a new value for $H(\boldsymbol{S}, \mathbf{x})$ which can be substantially different from its previous value.

(ii) make a Metropolis update starting from the new $\boldsymbol{S}$-state and the corresponding $\mathbf{x}$-state. In this second step the value of $H(\boldsymbol{S}, \mathbf{x})$ is conserved (or nearly so, as a result of rounding).

216 The Pakman and Paninski (2014) sampling algorithm is for linear and quadratic constraints and therefore is certainly applicable to our situation in which the constraints

are all linear. Without loss of generality (Li and Ghosh 2015) it is sufficient to consider sampling from

$$\log p(\mathbf{x}) = -\frac{1}{2}\mathbf{x}\cdot\mathbf{x} + \text{constant} \tag{7.17}$$

subject to

$$\boldsymbol{F}_j \cdot \mathbf{x} + g_j \geq 0 \quad j = 1, \ldots, n. \tag{7.18}$$

In the Pakman and Paninski (2014) approach the notional physical system may be 217
thought of as a particle moving in N-dimensional space so that $\mathbf{x} = (x_1, \ldots, x_N)$ are the physical coordinates of its location. The Hamiltonian is given by

$$H = \frac{1}{2}\mathbf{x}\cdot\mathbf{x} + \frac{1}{2}\boldsymbol{S}\cdot\boldsymbol{S} \tag{7.19}$$

such that the joint distribution is $p(\mathbf{x}, \boldsymbol{S}) = \exp(-H)$.

The solution of the equations of motion following from Equation(7.19) is 218

$$x_i(t) = a_i \sin t + b_i \cos t \quad i = 1, \ldots, N \tag{7.20}$$

where the constants a_i, b_i can be expressed in terms of the initial conditions as

$$\begin{aligned} a_i &= s_i(0) \\ b_i &= x_i(0) \end{aligned} \tag{7.21}$$

The two stages of the HMC are then:
1. Sample $\boldsymbol{S}$ from $\mathcal{N}(\mathbf{0}, \mathbb{I}_N)$;
2. Use this $\boldsymbol{S}$ and the last position vector $\mathbf{x}$ as initial conditions and let the particle move for a time T after which the position and momentum have values $\mathbf{x}^*$ and $\boldsymbol{S}^*$.
Pakman and Paninski (2014) show that the value $\mathbf{x}^*$ belongs to a Markov chain with equilibrium distribution $p(\mathbf{x})$.

If before expiry of the time T the particle hits a wall, that is if one of the m constraints 219
becomes active, it is reflected elastically, its new momentum vector being calculated accordingly.

After a time T the position of the particle, and its position $\mathbf{x}^*$ recorded as a sample 220

from the target distribution, that position becomes the starting point for the next cycle beginning again with Step 1.

§ 7.6 Finding a Starting Point for Sampling

221 To start sampling from $\mathcal{TN}(\boldsymbol{\Lambda\mu}, \boldsymbol{\Lambda\Sigma\Lambda}, \boldsymbol{l}, \boldsymbol{u})$ it is necessary to use a point that lies in the feasible space. Such a point may be found by solving a quadratic problem

$$\mu^*_{\boldsymbol{\xi}} = \min_{\boldsymbol{\xi}\in\mathbb{R}^N} \{\boldsymbol{\xi\Gamma\xi}^{-1} | \boldsymbol{\Phi\xi}, \boldsymbol{l} \leq \Lambda\boldsymbol{\xi} \leq \boldsymbol{u}\}, \tag{7.22}$$

and then using $\boldsymbol{\Lambda}\mu^*_{\boldsymbol{\xi}}$ as the starting point(López-Lopera et al., 2017).

222 Another option is to use Phase-I of the Phase-I/Phase-II approach to solving linear programming problems as described in Nocedal and Wright (1999, Chapter 16). That proceeds by defining a new linear program as described in the following paragraphs.

223 The interpolation and inequality constraints may be written:

$$\begin{aligned} a_i^T\eta &= b_i \quad i \in \mathcal{E} \\ a_i^T\eta &\geq b_i \quad i \in \mathcal{I} \end{aligned} \tag{7.23}$$

where the b_i correspond to the values of $\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{l}$ collectively and the a_i are the corresponding rows of $\boldsymbol{\Phi}, -\boldsymbol{\Lambda}, \boldsymbol{\Lambda}$ (with the latter matrix appearing twice because we are treating $\boldsymbol{l}$ and $\boldsymbol{u}$ separately as in Equation(7.12)).

224 We then define a linear program

$$\begin{aligned} \min_{\eta,z} \quad & \mathbb{I}_k z \\ a_i^T\eta + \gamma_i z &= b_i \quad i \in \mathcal{E} \\ a_i^T\eta + \gamma_i z &\geq b_i \quad i \in \mathcal{I} \\ z &\geq 0 \end{aligned} \tag{7.24}$$

where

$$\gamma_i = \begin{cases} -\text{sign}(a_i^T - b_i) & i \in \mathcal{E} \\ 1 & i \in \mathcal{I} \end{cases} \tag{7.25}$$

This linear program has an easily found feasible point from which it can always be solved. Loosely speaking, the new variable z measures how far any given η is from feasibility in the original problem. If η is feasible for the original problem then $(\eta, 0)$ is optimal in the new problem.

It is a helpful consequence of using this method to find a starting point, that in the 225
event that no feasible starting point exists, then the expense of Monte Carlo simulation is avoided. That contrasts with other methods in which unsatisfactory GPs are rejected only after they have been simulated.

It is also necessary to select the parameter T that governs the intervals at which the 226
position of the particle is taken. As part of the research presented in this thesis, the Algorithm 1 from López-Lopera et al. (2017), modified by the use of the Phase-I program, and including Pakman and Paninski's HMC algorithm, was encoded in an R package. In testing that package, it was observed that T needs to be chosen with some care. In particular, it must not be too short in relation to the interior dimensions of the feasible space. In that case, successive positions will tend to be correlated and HMC will fail to overcome the very problem which it was intended to solve. If T is too long, then the simulation will take an unnecessarily long time. Between the extremes of too long and too short, the precise choice of T does not appear to affect the validity of the result. There is scope for further research to examine in more detail how best to choose T. In the absence of such research it was sufficient in our work to choose T empirically.

§ 7.7 Sampling

The process described above generates a vector $\boldsymbol{\eta} = \boldsymbol{\Lambda}\boldsymbol{\xi}$. The final, straightforward, 227
stage is to solve the linear system $\boldsymbol{\eta} = \boldsymbol{\Lambda}\boldsymbol{\xi}$ to obtain a draw of $\boldsymbol{\xi}$ as desired.

The desired portfolio of probability distribution functions that represent a full range of 228
those that satisfy the model's constraints can now be generated easily: after the usual warm up phase of the HMC, each successive draw of $\boldsymbol{\xi}$ will yield a compliant probability distribution function.

§ 7.8 Choice of knots

229 We have shown above that by sampling it is possible to generate limitless examples of probability distribution functions and densities all satisfying the necessary constraints, including, of course, compliance with the opinions elicited from the expert. Each of these functions is defined as a realisation of a GP defined on a finite series of knots. Before we can proceed with a practical demonstration of the method it is necessary to decide both on the number, N, of knots and on their location.

230 In the context of the problem addressed by López-Lopera et al. (2017) and Maatouk and Bay (2017), it is clear that N should be large because their aim is to obtain a good approximation to a target function. (One can envisage tests to determine how large N needs to be to achieve a desired degree of closeness to the approximation). But as noted above in Section 7.2, in our case we are not seeking to approximate a target function, merely to generate compliant functions and there are certainly practical disadvantages in working with large values of N, especially in the context of a live elicitation session in which long computation times would be unacceptable. It is, of course, true that the larger the value of N the greater is the coverage of the space of consistent distributions. So, in practice it would be wise to choose a value of N that is as large as possible without rendering computation times excessive.

231 An additional feature of using a large number of knots is that the plots of the resultant densities will be smoother the larger the value of N. (The plots of the distribution functions will be smooth anyway, as a result of integration.) It is worth considering, however, whether this feature, smooth densities, is a real advantage in practice or not, and that will depend upon its influence or otherwise on the experts.

232 As to the location of the knots, there is a clear advantage in placing knots at the points corresponding to each of the quantities elicited from the experts: the minimum, the maximum, the tertiles and the mode, as it might be, for example. Additional knots could, in principle, be placed anywhere without destroying the validity of the resultant densities and distribution functions.

§ 7.9 Conclusions

233 We have presented a method using GPs for generating priors based on elicited quantities. It aims to overcome the potential problems with earlier approaches, specifically it ensures

that all the functions that are generated are valid probability densities or distributions. Through the use of a number of parameters the resulting priors are very tunable to the specific circumstances in which the model is to be used. Since each prior is generated as the result of a draw, and there is no theoretical limit to the number of draws that can be made, the method described is capable of furnishing a very wide range of priors.

Chapter 8

Results of applying the Gaussian Process model to rainfall data

In this chapter we set out the rationale for and summarise the construction of the GP model described in Chapters 6 and 7 so that it can be applied to rainfall data. 234

In outline there are three stages: 235

(i) to define each of the three pairs of GPs to be used in terms of their knots, their covariance functions and other parameters;

(ii) to construct an HMC model that can define priors for the three quantities $\tilde{q}_1, \tilde{q}_2$ and $\tilde{q}_3$ as GPs; and

(iii) using those priors to carry out an MCMC simulation to fit a generalised PPP distribution to the observations of daily rainfall that exceed a high threshold.

We consider these in turn.

§ 8.1 Defining the Gaussian process model

236 It should be noted that the simulation model provides a very wide range of choices. In addition to the number and location of knots, the user needs to choose the covariance function and its parameters. In addition there are parameters that govern the operation of the HMC simulation. All these parameters may interfere with each other and some combinations of the parameters do not give rise to workable models in the sense that no valid set of weights $\boldsymbol{\xi}$ exists.

237 Our approach to obtaining workable models has been to examine a large number of combinations of the parameters. For each such combination, we test the workability of the potential model by seeing if a starting point for the HMC simulation exists. We thus generate a portfolio of workable models comprising sets of models for each of the three elicited quantiles.

238 Each prior is defined on a set of knots. Two GPs are then constructed, one for generating the PDFs and the other for the corresponding CDFs such that any realisation of the CDF GP provides a CDF that fits the elicited quantiles of the three $\tilde{q}_i$. In this case all the $\tilde{q}_i$ are non-negative. As to whether we can define a maximum value, we follow Coles and Tawn (1996) in not setting any such figure. Part of the justification for that assumption is specific to rainfall. There is evidence that exceptionally extreme rainfall can occur at individual locations, certainly as much as three times any historical local precedent. See, for example, Coles et al. (2003), which attempts an extreme value analysis of a catastrophic rainfall event in Venezuela in 1999. In addition, the option of setting the constraint $\xi < 0$, which would certainly rule out unbounded results, might, however, overweight upper bounds that were too low, and thus lead to biased results.

239 Use could be made of more sophisticated methods of constructing the priors. In 3.3 we saw priors being defined to reflect prior opinions on aspects of the distribution of wave heights. The concept of **probable maximum precipitation (PMP)**, defined by the World Meteorological Organisation as"the greatest depth of precipitation for a given duration meteorologically possible for a design watershed or a given storm area at a particular location at a particular time of year", could be reflected in the priors. A

complication is that PMP is defined for an area not for a specific point location, which is the basis we have used throughout our work here.

One disadvantage of not defining a maximum value is that it is computationally 240
inconvenient to work with variables that are unbounded. To avoid that problem, we will transform all the rainfall figures into the range $[0, 1)$ using $t : x \mapsto \frac{x}{x+1}$, constructing the model to work with those transformed data, and then reversing the transformation at the end of the simulation.

8.1.1 Covariance function

We work with the well known exponential covariance function: 241

$$k(x, x') = \sigma^2 \left[\exp\left(-\frac{(x - x')^2}{2\theta^2} \right) \right]. \tag{8.1}$$

Any covariance function will generate realisations of the GPs that are probability 242
distributions fitting the elicited data. The only requirement is that they need to be acceptable to the expert. By appropriate choice of the two parameters, θ and σ the exponential covariance function is capable of producing a very wide range of distributions. It therefore seems a reasonably appropriate choice.

Table 8.1: Elicited data: $\tilde{q}_i : i = 1 : 3$

Probability	Median	90% quantile	Mode
0.100	0.59	0.72	0.631
0.010	0.43	0.70	0.510
0.001	1.00	1.20	1.075

8.1.2 Rainfall data

As before, we use the dataset obtained from the R package `evdbayes` which is nearly 243
identical to that used in Coles and Tawn (1996), as noted in Subsection 5.2.1. We decided that that close similarity enabled us to use the same quantiles as elicited from the expert in that earlier work. In addition, as our method requires an elicited mode, we chose one empirically for each of the three priors. The elicited data, including our chosen modes, are set out in Table 8.1.

The empirical choice of mode in the case, as here, of unimodality is not completely 244
straightforward as the sets of constraints that need to be satisfied depend on the position of the mode relative to the quantiles being used. An example of the different conditions that might apply depending on the relative position of the mode can be seen in Chapter 9.

In the case described in this chapter we decided that the mode would be located strictly between the two quantiles. To use our GP approach whilst avoiding that complication would require Equation 7.11 to be dropped, with a corresponding amendment to the matrix $\mathbf{\Lambda}$. The sampling method could then proceed. Such an approach would, however, generate complications of its own if the expert held the view that the density was unimodal.)

8.1.3 Knots

245 We described in Section 7.8 the considerations to be kept in view when deciding on the number and location of the knots. In addition to those points, we found in practice that computation time, unsurprisingly, increases with the number of knots. That led us to the idea of minimal knots, discussed in Chapter 9. But in this chapter we try to use as many knots as possible subject to computation time remaining feasible for a face-to-face elicitation meeting. The potential advantage of using more knots is that the resulting PDFs will have greater flexibility than in the minimal knots approach to fit the elicited data. The PDFs might also appear smoother than piecewise linear PDFs when displayed graphically but, as we shall see, they will still look rather different from textbook parametric density functions.

246 It seemed convenient to define the knots to include in each case the two quantiles and the mode we had chosen. Those numbers are transformed to d_1, d_2 and w, respectively, to lie in $[0, 1)$. The additional knots were defined as follows:

(i) for each prior, choose a number p for the total number of knots.

(ii) choose numbers p_1, p_2, p_3 such that $p = p_1 + p_2 + p_3$ and that the knots t are defined thus (expressed as R code)

```
t = c(
      seq(1/p, d1 - 1/p, length.out = p1),
      seq(d1, d2, length.out = p2),
      seq(d2 + 1/p, 0.75, length.out = p3)
      )
```

The choice of the parameters p_1, p_2, p_3, θ and σ is very wide but is not unconstrained. It 247
appears possible to choose the parameters in such a way that computation fails. Step 4 of the sampling algorithm (Subsection 7.4) requires the solution of

$$\boldsymbol{\mu}_{\xi}^{*} = \min_{\boldsymbol{\xi} \in \mathbf{R}^N} \{\boldsymbol{\xi}^{\boldsymbol{T}} \boldsymbol{\Gamma}^{-1} \boldsymbol{\xi} | \boldsymbol{\Phi}\boldsymbol{\xi} = \boldsymbol{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\} \tag{8.2}$$

which cannot be found if $\boldsymbol{\Gamma^{-1}}$ is reported as not being positive definite[1]. That calculation also appears to be very sensitive to the choice of which of the p_i is the mode of the density. The chosen values of the p_i are shown in Table 8.2.

Table 8.2: Numbers of knots for each simulation and location of mode

Probability	p_1	p_2	p_3	w
0.100	6	4	4	8
0.010	6	4	4	8
0.001	18	6	8	21

The parameters having been selected, the next steps are to construct the matrices used 248
in the sampling algorithm, namely Φ, Λ, Γ and Σ, and then to find an initial value as the starting point for the HMC process.

§ 8.2 Hamiltonian Monte Carlo outputs

Each of the three runs of the HMC draws 100 samples from the truncated multi-normal 249
distribution set out in Equation 6.34 but repeated here for convenience:

$$\boldsymbol{\Lambda}\boldsymbol{\xi} | \{\boldsymbol{\Phi}\boldsymbol{\xi} = \boldsymbol{y}, \boldsymbol{l} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\} \sim \mathcal{TN}(\boldsymbol{\Lambda}\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^{T}, \boldsymbol{l}, \boldsymbol{u}). \tag{8.3}$$

Each of the samples drawn, $\boldsymbol{\xi}$, comprises a set of weights equal in number to the 250
number, N, of basis functions, $h_j(x)$, each of which corresponds to a knot. As shown in Chapter 7, each sample defines a CDF:

$$F(x) := \sum_{j=0}^{N} \xi_j \varphi_j(x) \tag{8.4}$$

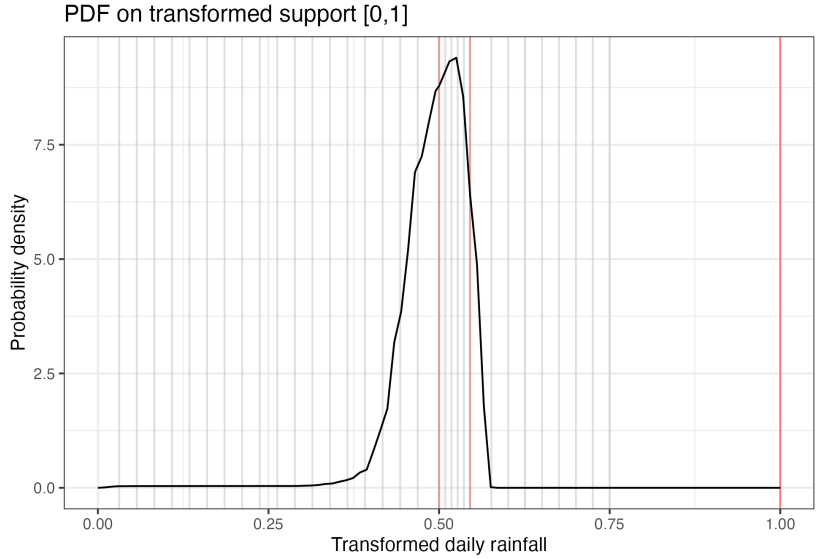
and a PDF

$$f(x) := \sum_{j=0}^{N} \xi_j h_j(x) \tag{8.5}$$

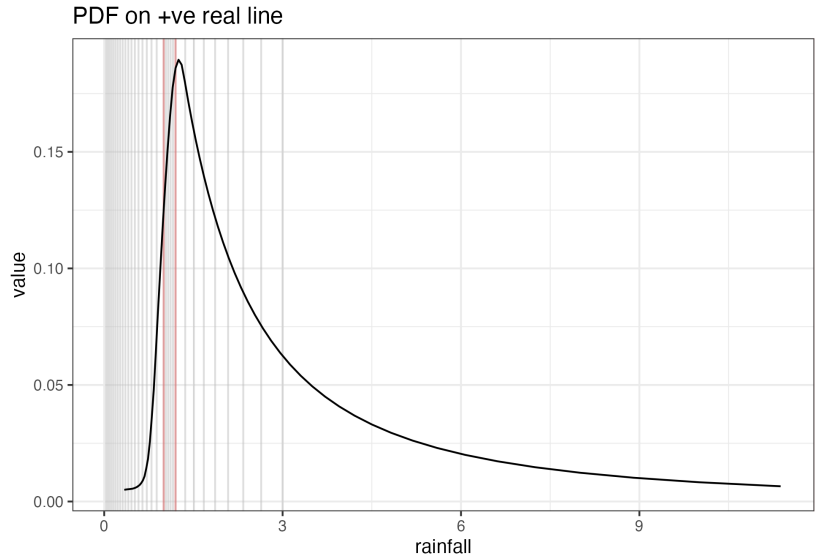
where $\varphi_j(x) = \int_0^x h_j(t)\, dt$.

[1] $\boldsymbol{\Gamma^{-1}}$ is defined mathematically to be positive-definite but due to rounding errors computationally may not be.

251 We show in Figure 8.1 and Figure 8.2 some examples of the pairs of PDFs and CDFs that emerge from the simulation.



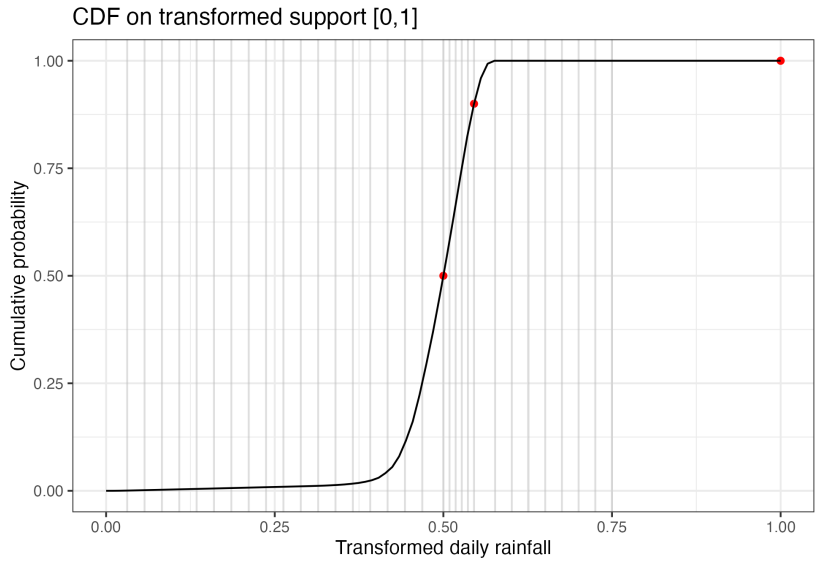
(a) PDF on transformed space



(b) pdf on untransformed space

252 We show a diagnostic summary of the run of the Stan program in Table 8.3. As before, the values of $\hat{R}$ and of $n_{\texttt{eff}}$ give some assurance that the simulation has converged.

253 The return level graph at Figure 8.3 shows that the GP-based priors have led to results that are consistent with those based on gamma distributions shown in Figure 5.8.

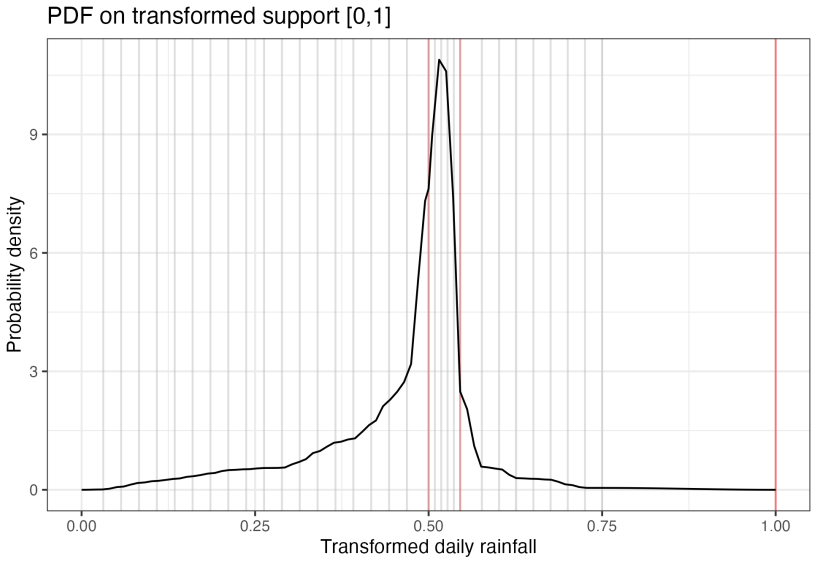


(c) cdf on transformed space

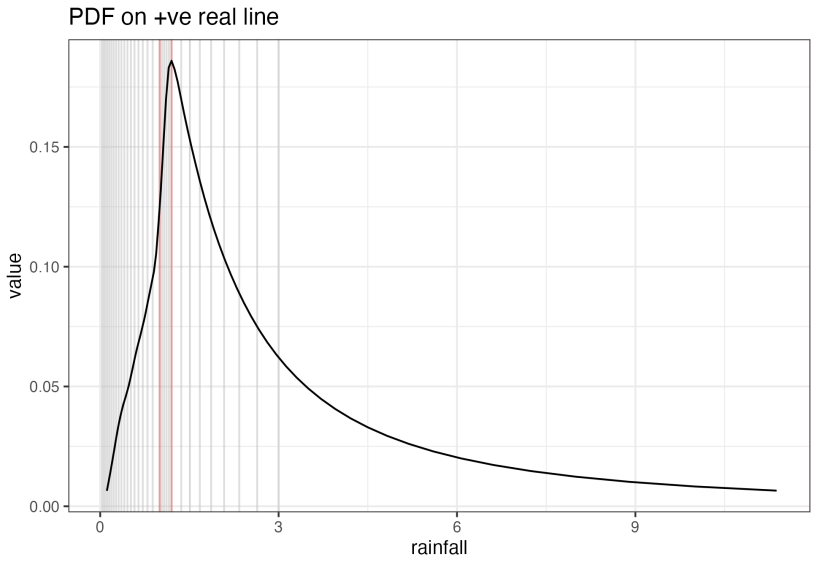
Figure 8.1: Example 1 of probability distribution and density resulting from the simulation. The red dots show the elicited tertiles and the point at which the cumulative density reaches unity.

Table 8.3: Diagnostic summary of results of simulation

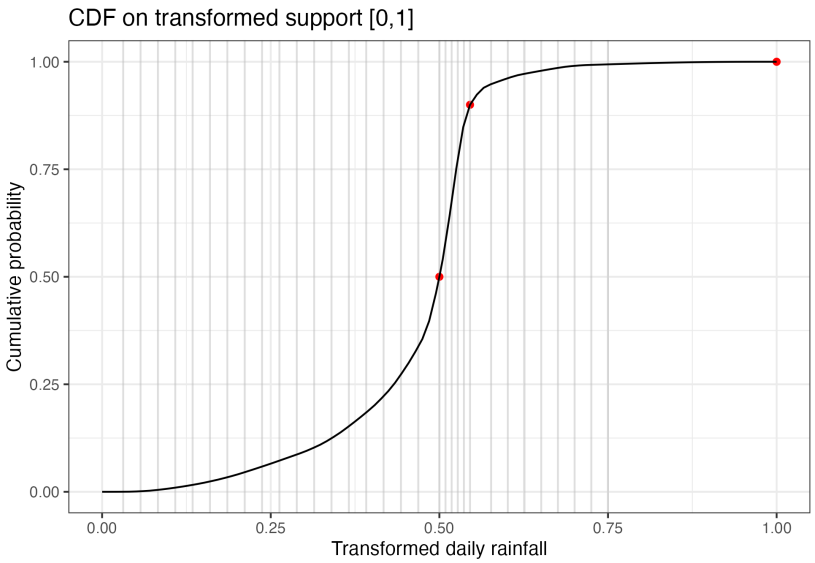
Parameter	$\hat{R}$	n_{eff}	mean	sd	2.5%	50%	97.5%
log-posterior	1.001	1731	75.974	1.276	72.659	76.343	77.376
qtilde1	1.001	1896	0.674	0.024	0.627	0.676	0.718
qtilde2	1.002	1417	0.484	0.042	0.410	0.481	0.575
qtilde3	1.001	1719	0.959	0.134	0.743	0.945	1.267
mu	1.001	3609	0.429	0.008	0.414	0.429	0.445
xi	1.001	2328	0.295	0.032	0.237	0.294	0.358
sigma	1.001	2380	0.075	0.008	0.059	0.075	0.090



(a) PDF on transformed space



(b) pdf on untransformed space



(c) cdf on transformed space

Figure 8.2: Example 2 of probability distribution and density resulting from the simulation. The red dots show the elicited tertiles and the point at which the cumulative density reaches unity.

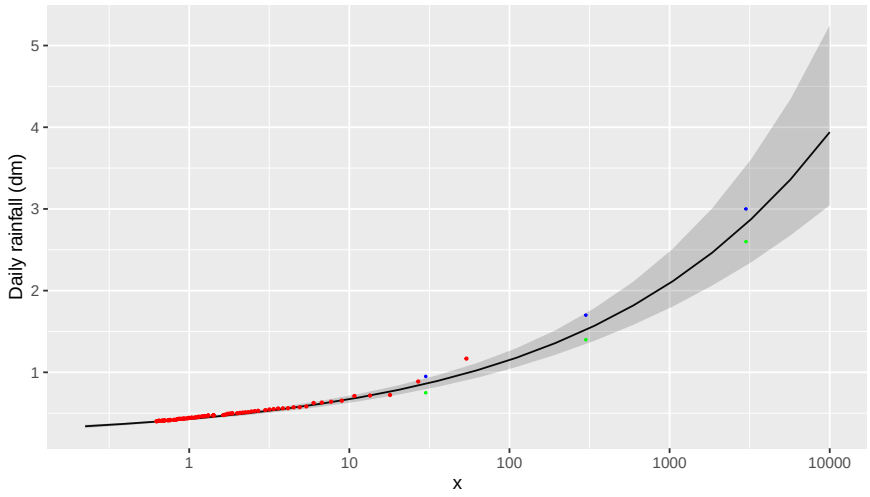


Figure 8.3: Implied return levels using Gaussian process priors compared with additional elicited quantiles
Key: red = observational data; black = median; grey shading = 95% credible range; blue = 90% additional quantiles, green = median additional quantiles

Chapter 9

Minimal knots

In practice, we find that the fitting of a probability distribution defined as a GP can be computationally demanding and therefore slow, even to the extent that the flow of a face-to-face elicitation meeting would be disrupted. Because of the clear practical advantage in live elicitation of being able quickly to derive densities and distributions from the experts' evidence, in this chapter we examine an approach based on fitting a very simply described probability distribution to what might be a minimal set of knots. To do this we will need to relax our requirement that the density must be infinitely differentiable. 254

§ 9.1 Elicitation plan using minimal knots

To provide a specific example we assume that we are using an elicitation plan to ask the expert for the following five quantities: 255

(i) The minimum the quantity of interest can be, denoted by the scalar a.

(ii) The maximum, denoted by the scalar b.

(iii) The two tertiles, denoted by t_1, t_2.

(iv) The mode, denoted by w.

For simplicity we make it a requirement that all five of these elicited figures are distinct.

256 It would be possible to fit a PDF to these five elicited quantities provided we used a parametric probability distribution with at least five parameters. It would certainly be possible to fit GP-based distributions as described in Chapter 7. Instead, we examine the effect of using a very simply constructed PDF, as described in the following paragraphs.

257 The context in which we developed this minimal knots model was an elicitation in which the expert was confident in specifying the minima and maxima of the quantities of interest. That justified the decision to set the densities corresponding to the quantities a and b both to zero. If it was thought possible that the expert's idea of the minimum and maximum risked being too conservative, one approach might be to set the densities at a and b to small amounts, to permit the resultant prior to contain lower or higher figures than the expert's minimum and maximum.

258 Alternatively, to avoid the need to add two extra knots, the analyst could debate with the expert about decreasing a and increasing b. Another situation in which the maximum or minimum might be attained with non-zero probability is when there is some physical limit that is known to be attainable. For example, a study of water levels in a reservoir might well assign a non-zero probability to the risk that the water might over-top the dam. Another example of an attainable maximum would be a study of wave heights in coastal waters, where, as we have stated in Section 3.3, wave heights cannot exceed the depth of the sea.

259 Let the densities corresponding to t_1, t_2, w be T_1, T_2, W respectively. Consistently with our view that the densities at the end-points a and b are both zero, the assumption then is that the desired PDF to be fitted to the given knots is defined graphically by straight lines joining the points $(a, 0), (b, 0), (t_1, T1), (t_2, T2), (w, W)$ arranged with their abscissae in order of magnitude. The distribution function can be obtained from the

PDF by integration. Because, in this case, the PDF is defined so simply, the integration can be performed analytically if desired, although in practical application numerical integration might well be used.

The quantities T_1, T_2, W do not need to be elicited because they can be calculated. The 260
details of calculating the densities at the knots depend on the location of the mode in relation to the two tertiles, giving rise to three cases, as follows:

(i) Case 1: $w < t_1$

(ii) Case 2: $t_1 < w < t_2$

(iii) Case 3: $t_2 < w$

§ 9.2 Case 1: Mode less than lower tertile

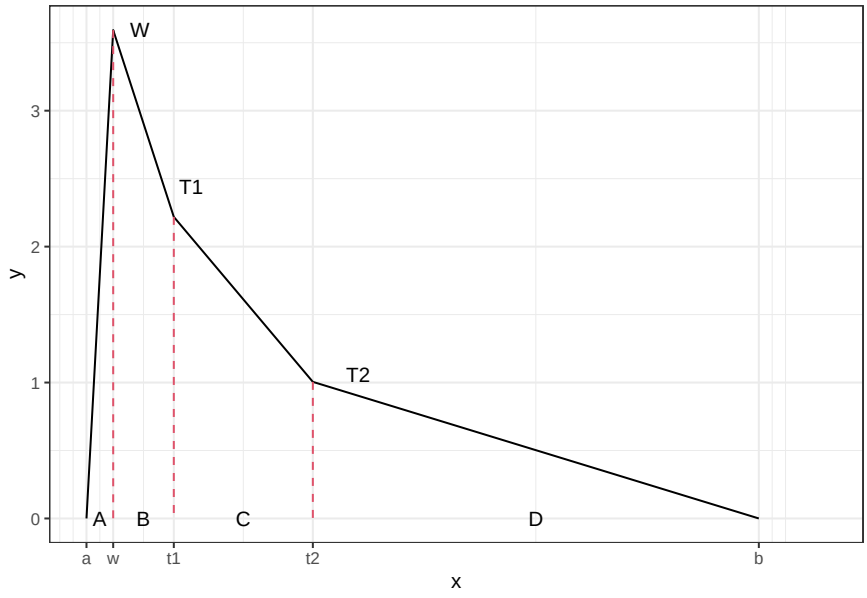


Figure 9.1: Minimal knots Case 1: $w < t_1$

Let 261

$$\begin{aligned} A &= t_1 - a \\ B &= t_1 - w \\ C &= t_2 - t_1 \\ D &= b - t_2 \end{aligned} \tag{9.1}$$

Note that, by definition, the quantities A, B, C, D are all strictly positive. A similar condition holds for the corresponding quantities in the other two cases set out below.

262 The two tertiles define three regions all of area equal to $1/3$. Whence we get three equations:

$$\begin{aligned} \frac{1}{3} &= \frac{1}{2}T_2(b - t_2) = \frac{1}{2}T_2 B \\ \frac{1}{3} &= \frac{1}{2}(T_1 + T_2)(t_2 - t_1) = \frac{1}{2}(T_1 + T_2)C \\ \frac{1}{3} &= \frac{1}{2}W(t_1 - a) + \frac{1}{2}T_1(t_1 - w) = \frac{1}{2}W(A + B) + \frac{1}{2}T_1 B \end{aligned} \tag{9.2}$$

263 Solving these equations produces:

$$\begin{aligned} T_2 &= \frac{2}{3}\,\frac{1}{D} \\ T_1 &= \frac{2}{3}\left(\frac{1}{C} - \frac{1}{D}\right) \\ W &= \frac{2}{3A}\left(1 - \frac{B}{C} + \frac{B}{D}\right) \end{aligned} \tag{9.3}$$

§ 9.3 Case 2: Mode between lower and upper tertiles

264 In this case we define:

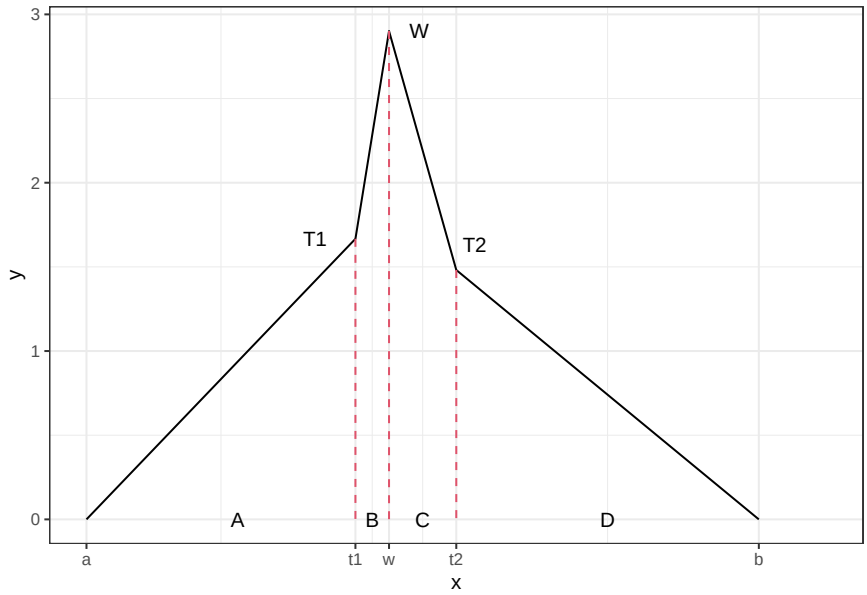


Figure 9.2: Minimal knots Case 2: $t_1 < w < t_2$

$$\begin{aligned} A &= t_1 - a \\ B &= w - t_1 \\ C &= t_2 - w \\ D &= b - t_2 \end{aligned} \tag{9.4}$$

By similarly constructing equations for the area of each of the thirds into which the 265
tertiles divide the PDF, we obtain the following:

$$\begin{aligned} T_1 &= \frac{2}{3}\frac{1}{A} \\ T_2 &= \frac{2}{3}\frac{1}{D} \\ W &= \frac{2}{3}\frac{1}{B+C}\left(1 - \frac{B}{A} - \frac{C}{D}\right) \end{aligned} \tag{9.5}$$

§ 9.4 Case 3: Mode greater than upper tertile

In this case we define: 266

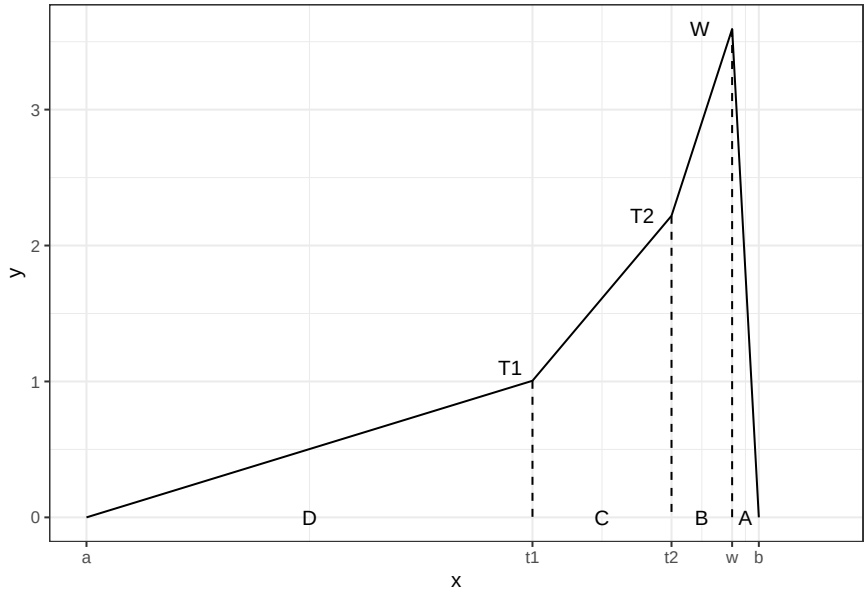


Figure 9.3: Minimal knots Case 3: $t_2 < w$

$$\begin{aligned} A &= b - w \\ B &= w - t_2 \\ C &= t_2 - t_1 \\ D &= t_1 - a \end{aligned} \tag{9.6}$$

267 By an argument on similar lines to the above two cases we then obtain

$$\begin{aligned} T_1 &= \frac{2}{3}\left(\frac{1}{C} - \frac{1}{D}\right) \\ T_2 &= \frac{2}{3}\frac{1}{D} \\ W &= \frac{2}{3}\frac{1}{A+B}\left(1 - \frac{B}{C} + \frac{B}{D}\right) \end{aligned} \tag{9.7}$$

268 There are further conditions in each case that must be enforced because the calculations described above do not take into account the requirements (a) that the PDF be unimodal and (b) that the calculated density, W, at the mode, w, is indeed the maximum of the density across the whole of the support $[a, b]$. The further conditions are as follows:

(i) Case 1: $W > T_1 > T_2$

That is equivalent to:

$$\begin{aligned} (A+2B)\left(\frac{1}{C}-\frac{1}{D}\right) &< 1 \\ D &> 2C \end{aligned} \tag{9.8}$$

(ii) Case 2:

$$\begin{aligned} W &> T_1 \\ W &> T_2 \end{aligned} \tag{9.9}$$

That is equivalent to 269

$$AD - BD - CA > (B+C)\max(A, D) \tag{9.10}$$

(iii) Case 3: $W > T_2 > T_1$

That is equivalent to: 270

$$\begin{aligned} (A+2B)\left(\frac{1}{C}-\frac{1}{D}\right) &< 1 \\ 2D &> C \end{aligned} \tag{9.11}$$

The equations above permit PDF and distribution functions to be calculated very 271
rapidly for review by the expert and possible revision. In a later chapter we will describe an experiment using these minimal knots to fit a PDF to elicited data. The practical question is whether the relatively tight constraints of the minimal set of knots inhibit the expert's choice.

272 Another approach to minimal knots, which we do not pursue here, would be to define a GP CDF F so that the following would hold:

$$0 = F(a) < F(t_1) < F(t_2) < F(b) = 1 \tag{9.12}$$

as well as exactly one of the following:

$$\begin{aligned} F(a) < F(w) < F(t_1) \\ F(t_1) < F(w) < F(t_2) \\ F(t_2) < F(w) < 1. \end{aligned} \tag{9.13}$$

273 Although some kind of probability distribution could be fitted to any choice the expert might wish to make for the two tertiles, the expert might well be unhappy with some distributions even if they do fit the quantiles. We shall see below, for example, (Section 10.3) that as simple a requirement from the expert that the density should approach zero near the elicited maximum and minimum imposes quite strict constraints on the choice of tertiles. And even if the expert is content with the distribution fitted to chosen tertiles, the range of admissible values of the location of the mode may well be limited.

274 It is worth considering the contrast between priors fitted with minimal knots and those fitted to a given parametric probability distribution. Both kinds of prior are very likely to be "wrong" in some sense because real physical phenomena can only rarely be sure to fit any particular distribution. Both are simplifications: minimal knots because of, in particular, the assumption that the PDF is piecewise linear; the parametric because, in particular, it is only one of very many parametric distributions that could be fitted. For many people, their earliest encounters as students with probability theory will have exposed them to textbook examples of smooth continuous PDFs, so it is possible that the spiky picture presented by the minimal knots PDF might be at first off-putting to an expert with that kind of education. But it is true that there is no reason to suppose that probability distributions outside the textbook world should always be smooth.

275 In physical measurements in practice, the smoothness with which PDFs can be estimated is affected by the sensitivity with which observations may be measured. In many cases measurement may be so precise that the estimated PDFs would be effectively smooth. But that is not true in all cases. Although rainfall obviously varies continuously over a range of values, in practice it can only be measured at discrete intervals. In a standard tipping bucket rain-gauge as used by the Meteorological Office, for example, an event is recorded each time a rainfall increment of 0.2 mm is detected.

Chapter 10

Estimating extreme values using elicited priors

In this chapter we present the results of a specific elicitation exercise conducted in two parts held some 20 months apart. The focus of the exercise was the estimation of extreme wave heights at a specific deep-sea location at which a measuring device, mounted on a discus buoy, is located. It records, amongst other data, wave heights at hourly intervals. The analysis of the data was to be based on priors generated from a small number of quantiles given by an expert. Two different methods were used to define the priors. In addition to presenting the results we make some observations on the effectiveness of the elicitation process followed. 276

§ 10.1 The selected location

The location is a 3-meter discus buoy, Station 46085, owned and managed by the United States National Data Buoy Center. It is located in the Gulf of Alaska in the north-eastern Pacific at 55°52′42″N 142°52′32″W where the water depth is 3745 metres. Figure 10.1 shows a similar discus buoy. 277

Station 46085 has been collecting hourly readings of significant wave height, H_s, since 5 May 2007 but with some large gaps, as can be seen in Figure 10.2. By inspection it 278

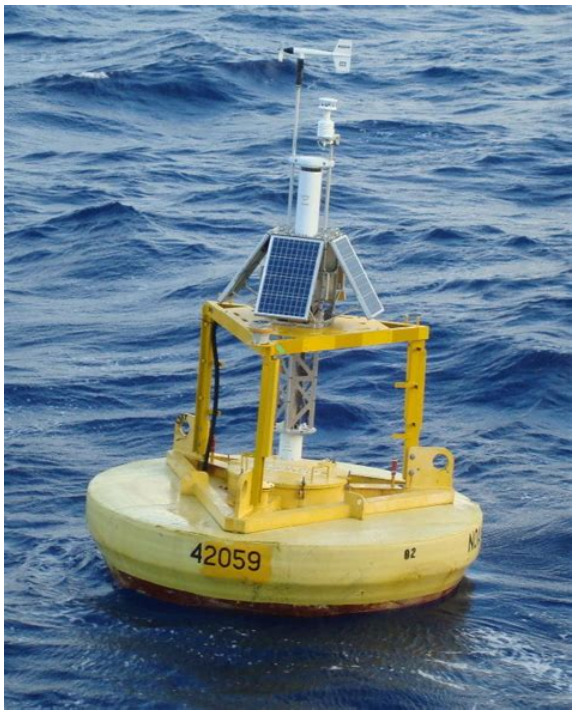


Figure 10.1: A discus buoy similar that at Station 46085

is obvious that the data are subject to annual seasonal variation but there appears to be no obvious trend. For the purposes of this exercise focussed on the defining priors based on expert elicitation we will ignore the seasonal variation and therefore will use the dataset without any further adjustment.

279 We shall use the POT method (see Section 2.4) and define the high threshold u by inspection of the MRL plot, as described in Section 2.7.1. The plot in this case is shown at Figure 10.3. As usual, the interpretation of the plot is debatable but it appears to us to become linear at a little below 8 metres. We set $u = 7.7$ as shown in red in Figure 10.2.

§ 10.2 The First Elicitation Exercise

280 The first elicitation meeting was held on 29 May 2022 at which the expert was Rod Rainey, an eminent offshore engineer and recognised expert on extreme waves, and I acted as facilitator. Rainey was asked to give his opinion about extreme waves at the

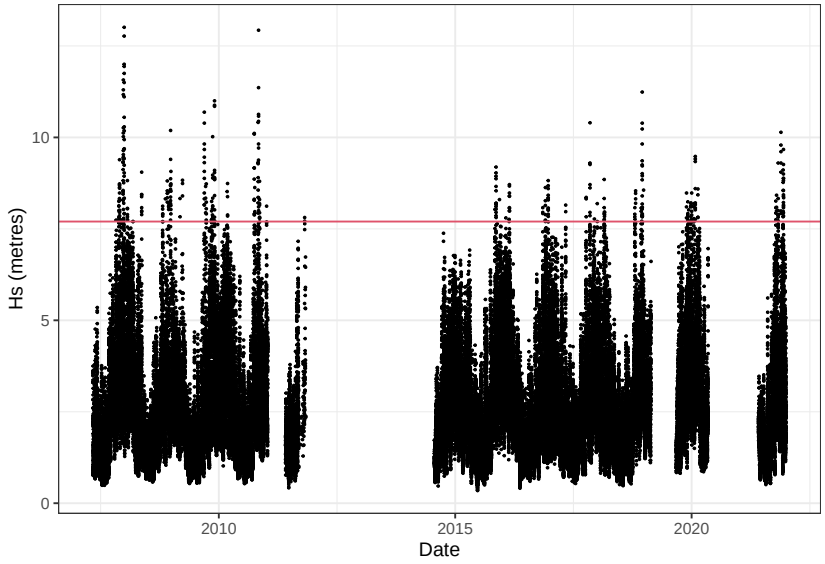


Figure 10.2: Observations of significant wave height at Station 46085

selected location in the North East Pacific Ocean. He has no particular knowledge of the Pacific and none about the specific location.

Rainey was informed of the location and water depth. He considered that a comparable 281
Atlantic Ocean location would be off the coast of Donegal where, in the absence of specific meteorological information, he assumed the winds would be comparable to similar latitudes in the Gulf of Alaska. He was also assuming that for this exercise weather trends due, for example, to global warming, were to be disregarded.

Before proceeding with the elicitation exercise it was necessary to agree how wave heights 282
would be defined. Rainey's view was that without doubt the appropriate measure should be significant wave height (as defined in Section 3.3). The use of that measure was very much a standard in oceanography. If forecasts of individual wave heights were required then the methods mentioned in that Section could be applied to H_s forecasts.

The focus of the elicitation followed that of Coles and Tawn (Chapter 5): the questions 283
aimed to gather data from which probability distributions of three independent variables could be constructed that would be satisfactory to the expert. The three variables were:

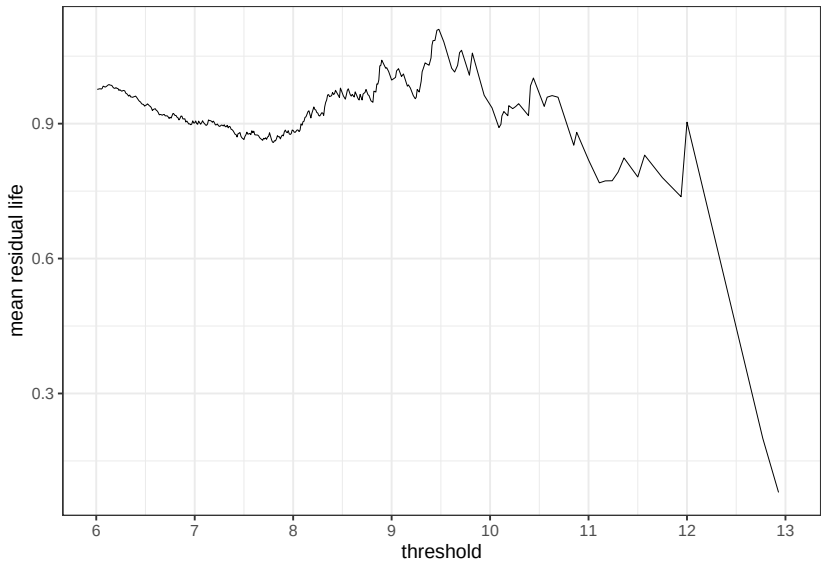


Figure 10.3: Mean residual life plot of H_s at Station 46085

(i) the 10-year wave (that is the significant wave height whose probability of being exceeded in any year would be 0.1 (1/10).

(ii) the difference between the 100-year wave and the 10-year wave.

(iii) the difference between the 1000-year wave and the 10-year wave.

284 The questions asked about each were not the same as used by Coles and Tawn but drew on the guidance in SHELF (Oakley and O'Hagan, 2019). That guidance describes approaches based on eliciting quartiles and on eliciting tertiles but does not recommend one of these approaches over the other. There is very little literature on the use of tertiles, so it was decided to use that method and see if any conclusions could be drawn on its merits. The questions that were to be asked were, accordingly:

(i) what is the minimum figure that the variable could never fall below?

(ii) what is the maximum above which it could not rise?

(iii) what is the lower tertile of the distribution?

(iv) what is the upper tertile?

Rainey quickly answered these questions about the 10-year wave but considerable 285
discussion was required before he could answer those questions relating to the second variable. In the light of that discussion the facilitator rephrased his description of what was required along the following lines:

(i) The terms 10-, 100- and 1000-year wave ("the n-year wave") refer in each case to the a significant wave height, a single number, associated with a single probability. In practice, however, there do not exist observations over sufficiently numerous cycles to reveal the single number.

(ii) At best, a probability distribution of estimates of these values might be based on specific local data. Even then the distribution of the 10-year wave might be possible but specific data for estimating a distribution for 100-year waves are very thin, if they exist at all. And that is even more so in the case of 1000-year waves.

(iii) The aim of using elicitation is to construct priors for use in a Bayesian approach to analysing the (thin) data for the chosen location leading to a probability distribution of the n-year wave from which a central estimate of its size could be derived.

(iv) When we ask an opinion about an n-year wave, it is an opinion about that central estimate, not about individual observations over periods of n years. So, for example, if we ask for an opinion about the minimum of the 10-year wave we mean the lowest possible value the central estimate could take, not the lowest wave height observed during a number, even a large number, of 10-year periods.

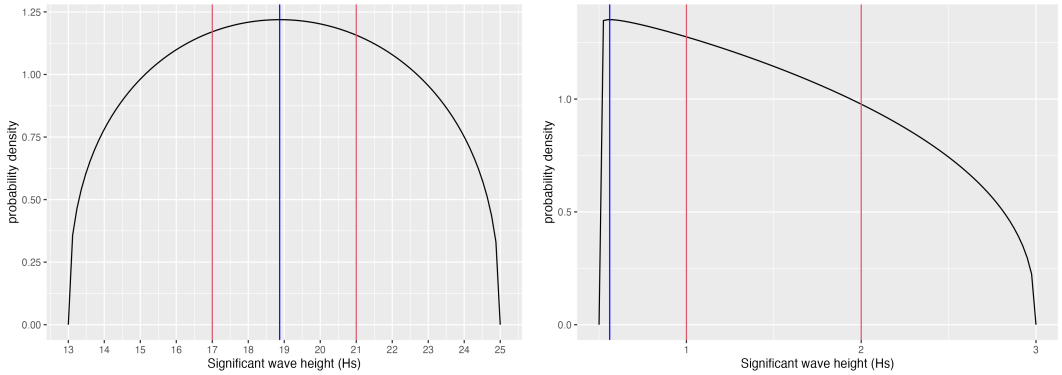
(v) And similarly when we ask about the difference between the 100-year wave, or the 1000-year wave, and the 10-year wave it is the difference in each case between the two (unknown) central estimates that interest us.

The data elicited from Rainey are shown in Table 10.1

p	Minimum	Lower tertile	Upper tertile	Maximum
0.100	13.0	17.0	21.0	25
0.010	0.5	1.0	2.0	3.0
0.001	1.0	2.0	4.0	6.0

Table 10.1: Figures elicited in first elicitation exercise

286 The elicited figures were fitted to beta distributions, as shown in Figure 10.4 (The distributions for $\tilde{q}_2$ and $\tilde{q}_3$ being the same, only one image is shown for those priors). The vertical red lines indicate the elicited tertiles and the vertical blue lines are the calculated modes of the fitted beta distributions.



$\tilde{q}_1 : \alpha = 1.38, \quad \beta = 1.4$ $\tilde{q}_i : i = \{2, 3\}. \quad \alpha = 1.01, \quad \beta = 1.4$

Figure 10.4: First elicitation. Beta distributions for the three $\tilde{q}_i : i = 1, \ldots, 3$

§ 10.3 The Second Elicitation Exercise

287 On reflexion, it was considered a weakness of the chosen elicitation method that the modes were calculated and might therefore not be close to the expert's idea of where the modes should be. Accordingly a second elicitation session was arranged in which for each of the three return levels of interest the following five quantities would be elicited in the order shown:

(i) the minimum, a;

(ii) the maximum, b;

(iii) the lower tertile, t_1;

(iv) the upper tertile, t_2; and

(v) the mode w.

A further change was that the elicited figures would be fitted to minimal knots probability 288
distributions as defined in Chapter 9.

The second elicitation session took place on 30 January 2024. Rainey had not reminded 289
himself of the opinions he had given in the first session some 20 months earlier, and indeed he had not kept any record of what he had said. Before any figures were asked for, Rainey discussed the likely behaviour of waves at the location of the coast of Donegal that he considered comparable to the Gulf of Alaska location. By stating his understanding of the sea conditions there he was setting the scene for him to give his opinions on the figures.

As before, the 10-year return level was covered first. Rainey gave minimum and 290
maximum figures of $a = 13$ and $b = 19$ metres respectively. Then he gave tertiles of $t_1 = 15$ and $t_2 = 17$ metres and a mode of $w = 16$. Quick analysis showed that such values were incompatible with the minimal knots distribution. We know from Equation 9.10 that

$$AD - BD - CA > (B + C)\max(A, D) \tag{10.1}$$

where, in the case of the quantities given by Rainey,

$$A = 2 \quad B = 1 \quad C = 1 \quad D = 2. \tag{10.2}$$

But $AD - BD - CA = 4 - 2 - 2 = 0$ whereas $(B + C)\max(A, D) = 4$, so Equation 9.10 cannot be satisfied using the elicited quantities. Even worse, as Rainey himself worked out, no unimodal density $f(\cdot)$ with its mode located between the tertiles t_1, t_2 could

work, for the following reasons. The tertiles at $t_1 = 15$, $t_2 = 17$ divide the area under the curve into thirds, so we must have:

$$\begin{aligned} \frac{1}{3} &= \int_{t_1}^{t_2} f(x)\,dx \\ \frac{1}{3} &= \int_{a}^{t_1} f(x)\,dx \\ \frac{1}{3} &= \int_{t_2}^{b} f(x)\,dx \end{aligned} \tag{10.3}$$

If $f(x)$ is unimodal with a mode $W = f(w)$ at w between t_1 and t_2, then

$$\frac{1}{3} = \int_{t_1}^{t_2} f(x)\,dx > \frac{1}{2}(T_1 + T_2)W \tag{10.4}$$

where $T_i = f(t_i)$ and the inequality is strict because the inequality $W > \max(T_1, T_2)$ must also be strict. But we have

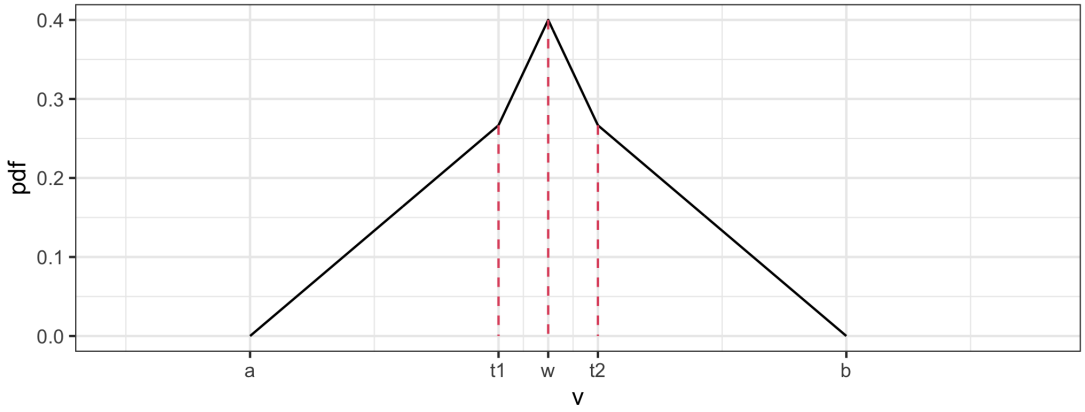
$$\begin{aligned} \frac{1}{3} &= \int_{a}^{t_1} f(x)\,dx < T_1(t_1 - a) = 2f(t_1) \\ \frac{1}{3} &= \int_{t_2}^{b} f(x)\,dx < T_2(b - t_2) = 2f(t_2) \end{aligned} \tag{10.5}$$

where both inequalities are also strict, and hence the contradiction

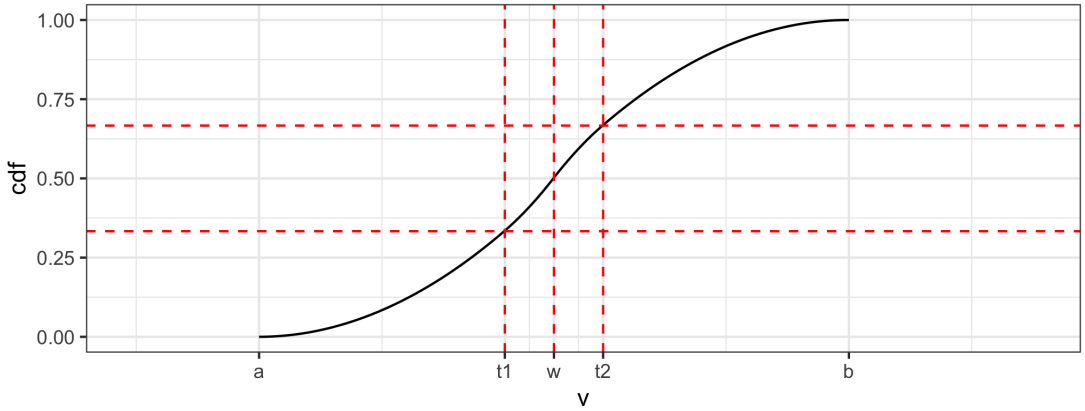
$$\frac{1}{3} < f(t_1) + f(t_2) < f(t_1) + f(t_2) < \int_{t_1}^{t_2} f(x)\,dx = \frac{1}{3}. \tag{10.6}$$

291 Rainey revised his tertiles to $t_1 = 15.5, t_2 = 16.5$, which were now consistent with the other three figures. He stated that the choice of a symmetrical PDF was deliberate. In subsequent discussion, Rainey said that symmetry struck him as a neutral way to deal with the fact that he saw no reason for the mode to be to the left or the right of the distribution. For him to provide an asymmetric PDF would be, he thought, claiming more knowledge than he had. Likewise, he was unwilling to give opinions on a finer scale than the two tertiles. The fitted PDF and CDF are shown in Figures 10.5a and 10.5b.

292 Proceeding to the elicitation of figures for the 100-year difference and the 1000-year difference, after some discussion Rainey concluded that the same distribution would apply to both: $a = 1, b = 3, t_1 = 1\frac{5}{6}, t_2 = 2\frac{1}{6}, w = 2$. The common PDF and CDF are shown in Figures 10.6a and 10.6b



(a)



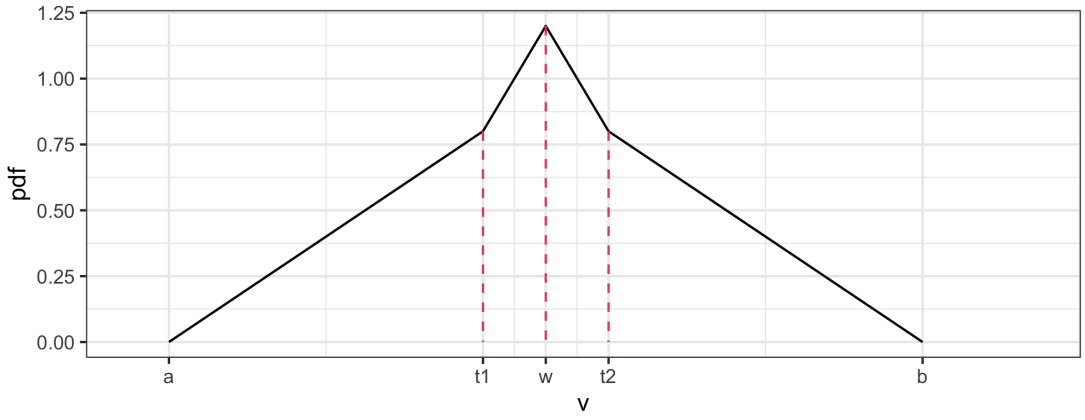
(b)

Figure 10.5: Second elicitation: 10-year return level (a) fitted PDF, (b) fitted CDF, using $a = 13, b = 19, t_1 = 15.5, t_2 = 16.5, w = 16$

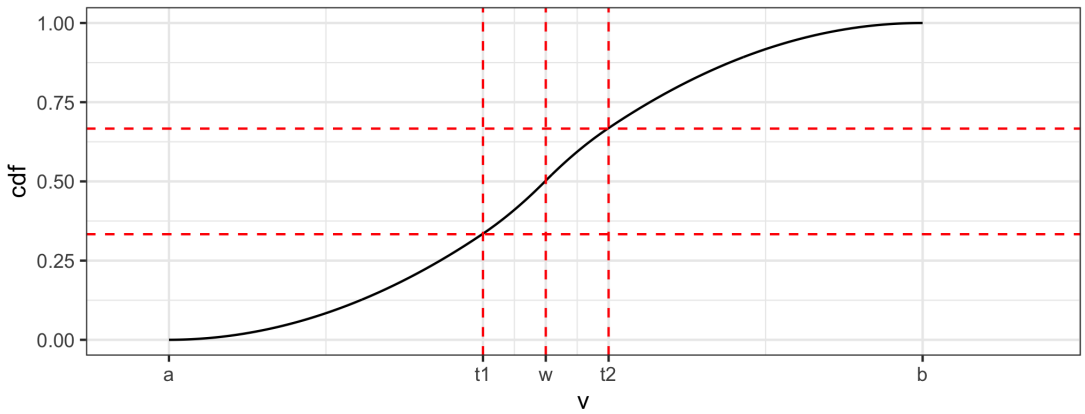
§ 10.4 Reflections on the elicitation exercise

The course of any elicitation exercise will vary depending on the personal characteristics, skills and weaknesses of the expert. Rainey has worked for over 40 years as a professional hydrologist in the context of the oil industry. Somewhat unusually for an engineer his first degree included two years of full-time study of mathematics. This study included probability theory so it was not necessary to explain, for example, what PDF andCDF were, or what is meant by tertiles. 293

Nevertheless, during the elicitation, as noted above, we were able very quickly to detect that Rainey's first ideas about the tertiles were impossible. He commented that his 294



(a)



(b)

Figure 10.6: Second elicitation: differences between 100-year and 1000-year return levels and the 10-year return levels: (a) fitted PDF, (b) fitted CDF, using $a = 1, b = 3, t_1 = 1\frac{5}{6}, t_2 = 2\frac{1}{6}, w = 2$

unfamiliarity with using tertiles rather than quartiles might have led him into that error. It might have been worthwhile spending a little time before the elicitation in talking generally about tertiles, the constraints to which they are subject, and how they are related to the mode. The material for such a conversation was to hand: it is included in Chapter 9 and had been used in writing the code that was being used to process the elicited figures. In addition, a trial run using illustrative figures would have been helpful.

295 The elicitation plan sought to obtain Rainey's figures in a very particular order so as to reduce the risk that his ideas might become unhelpfully anchored too early. In the light of his experience giving impossible tertiles, Rainey in effect refused to comply fully with

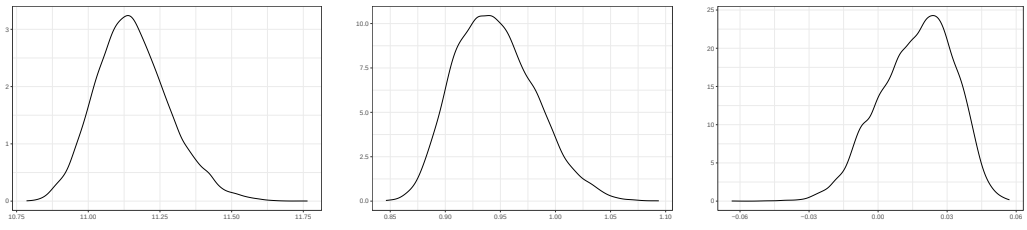
the elicitation plan. He would give the minimum and maximum figures, following the plan, but then insisted on discussing the mode and then the tertiles. He commented that his error had taught him that it was necessary for him to keep in mind the whole picture.

The software being used quickly produced both the PDF and CDF corresponding to 296
the elicited figures. It was obvious that Rainey quickly examined the PDF image to check that it corresponded to what he thought it should be. He looked at the CDF image only cursorily and did not refer to or appear to make any use of it. When asked whether the unusual appearance of the PDF, piecewise linear, with sharp corners, had hindered his interpretation of it, Rainey said that it caused no difficulty at all. He commented that he doubted that any engineer would have difficulty with the appearance of the PDF.

Rainey said that he had found the second elicitation exercise challenging. The questions 297
asked had forced him to probe the edge of his knowledge.

§ 10.5 Inference about extremes

The quantities elicited from Rainey were used to construct the probability distributions 298
and fitted to a PPP in a Stan program. The output from that program in tabular form is shown in Table 10.2 in which each row gives information about the posterior distribution of the individual parameters. The elicited quantities were the $\tilde{q}_i$ and in simulating their posterior distributions the program also calculated the corresponding distributions of the parameters, μ, σ and ξ of the PPP extreme value function. The posteriors of these parameters are shown in graphical form in Figure 10.7. The values of $\hat{R}$ all close to unity and the high values of $n_{\texttt{eff}}$ give some assurance that the simulation has converged.



(a) Location: μ (b) Size: σ (c) Shape: ξ

Figure 10.7: Posterior distributions of parameters of the PPP

Parameter	$\hat{R}$	n_{eff}	mean	sd	2.5%	50%	97.5%
log-posterior	1.004	1931	775.361	1.399	771.655	775.754	776.884
qtilde1	1.002	2540	13.377	0.196	13.070	13.355	13.814
qtilde2	1.002	1830	2.312	0.184	1.962	2.322	2.634
qtilde3	1.002	1849	2.411	0.272	1.898	2.423	2.890
mu	1.003	3524	11.157	0.129	10.929	11.148	11.432
nu	1.003	2102	1.040	0.038	0.963	1.043	1.105
xi	1.003	2117	0.017	0.016	-0.016	0.018	0.044
sigma	1.001	2199	0.945	0.036	0.883	0.943	1.023

Table 10.2: Summary of results of fitting the 2nd set of priors

299 As previously, we can now plot the forecast return levels over a range of periods (Figure 10.8). It is noticeable that the plot of the medians is very nearly a straight line and that the spread between the 2.5% and 97.5% quantiles is remarkably tight. Looking at Table 10.2, it is clear that the forecast of the median of the shape parameter, ξ, is very close to zero: that is the condition for the return level graph to be linear. The tightness of the forecasts appears to be a consequence of the tightness of the priors, and that, as we saw from the contradiction in Equation 10.6, is a consequence of the combination of hard upper and lower limits a and b with the assumption of unimodality. Such tight credibility bands raise the possibility that the expert had an unrealistically optimistic view of his own certainty.

§ 10.6 Parametric priors

300 In this case it so happens that the expert has chosen symmetrical densities around the mode, which is then equal to the mean. That means that we could easily fit the expert's quantiles to a very well-known distribution: the normal distribution. By comparing the results of the two different types of distribution, namely the minimal knots type and the normal, we obtain a cross check on our earlier results. In addition, because the normal densities are non-zero at the minima and maxima, a and b we can test the effect of the adjustment we referred to in Section 9.1 of permitting the density to take small values at those points.

301 By easy trial and error for each of the three $\tilde{q}_i$, we can find the value of σ^2 to give one-third and two-thirds quantiles to a normal distribution $\mathcal{N}(w, \sigma^2)$ at the points t_1 and t_2 whose mode/mean is w. We show in Figure 10.9 an example of such a density. The image shows that the densities at the points a and b are indeed small.

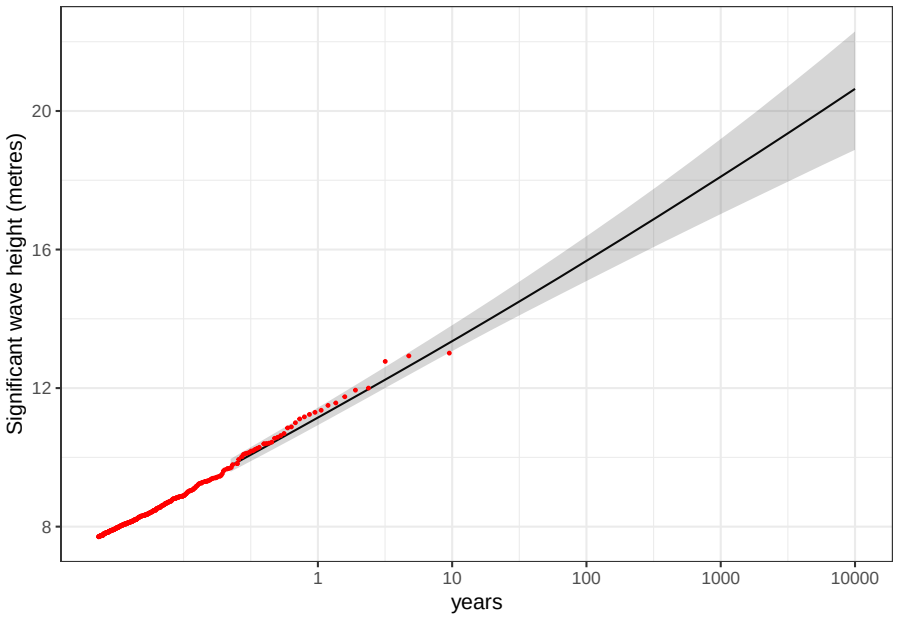


Figure 10.8: Forecast return levels over a range of periods: minimal knots priors
Key: black = means; grey shading = 95% credible range; red: empirical return levels of observations

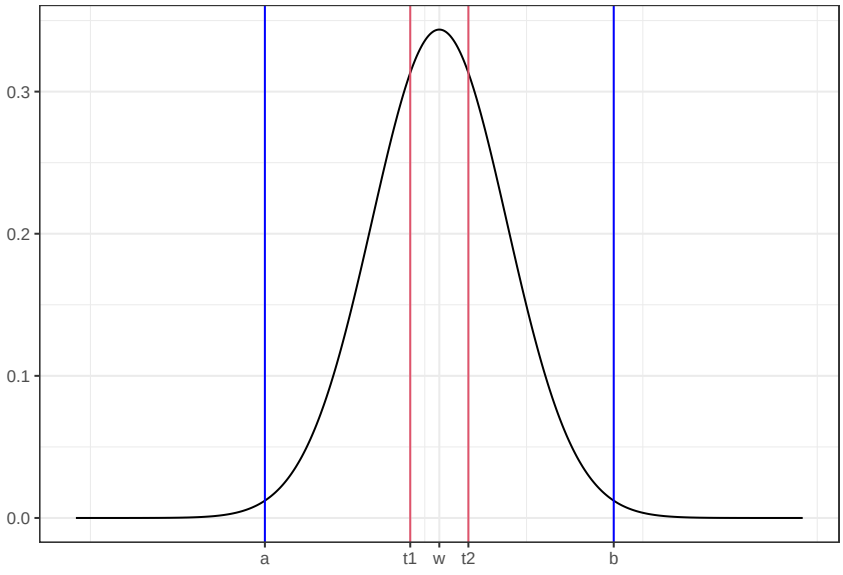


Figure 10.9: Normal densities fitted to elicited quantiles

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
log-posterior	1.0019	2533	782.4688	1.2610	779.2838	782.7840	783.8670
qtilde1	1.0003	2597	13.1875	0.2653	12.6788	13.1814	13.7245
qtilde2	1.0008	2016	2.1875	0.1990	1.8051	2.1849	2.5807
qtilde3	1.0009	2059	2.2530	0.2794	1.7341	2.2472	2.8212
mu	1.0003	3275	11.0607	0.1630	10.7534	11.0584	11.3975
nu	1.0012	2519	1.0270	0.0380	0.9536	1.0272	1.1013
xi	1.0012	2511	0.0113	0.0161	-0.0206	0.0117	0.0419
sigma	1.0003	2315	0.9111	0.0470	0.8219	0.9105	1.0045

Table 10.3: Fit using normal priors

302 Using these normal priors in place of the minimal knots priors used before we can similarly produce forecasts of extremes. The summary of results is shown in Table 10.3. It is obvious that the results are very similar to those obtained before, as can be seen also in the graph of forecast return levels by return period (Figure 10.10).

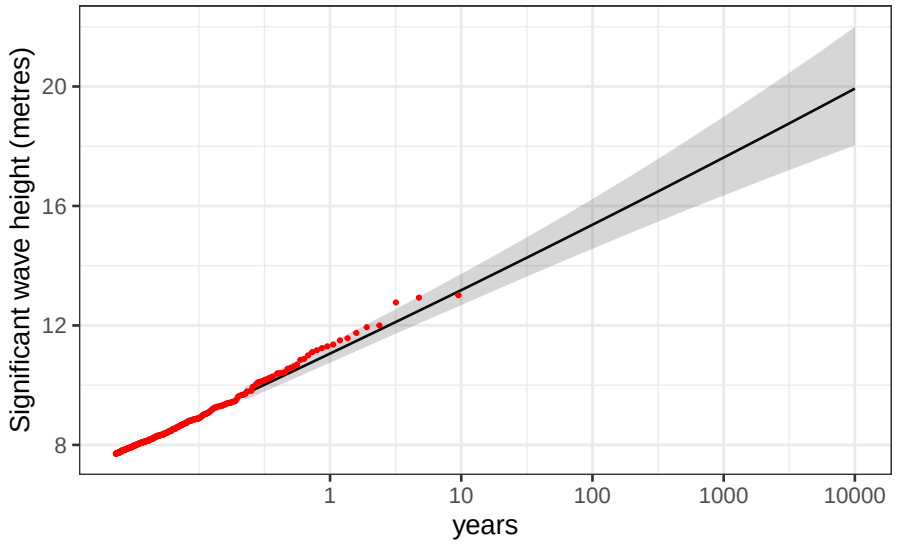


Figure 10.10: Forecast return levels over a range of periods: normal priors
Key: black = means; grey shading = 95% credible range; red: empirical return levels of observations

§ 10.7 Conclusions

303 Using quantities elicited from the expert we have calculated priors based on a minimal set of knots and then simulated forecasts of extreme significant wave heights. The speed of computation was very fast and the elicitation session ran briskly without long waits

for results to be calculated. Because the expert happened to give tertiles symmetrically placed round the mode, we were able to fit normal distributions to the elicited data and using those as priors we could recalculate the forecasts of extreme values. The fact that the results of the two methods were extremely close gave some assurance that the minimal knots approach, though apparently crude, was not necessarily a poorer method than the use of parametric distributions as priors. It also showed that the device of allowing minima below and maxima above the elicited values of a and b might not affect the results in a big way.

Chapter 11

Concluding remarks

We have shown that EVT in its modern developed state provides sophisticated models 304
that can be used to forecast the size and likelihood of extreme events. It can take into account dependency in observations and clustering, as well as trends. The Bayesian paradigm brings clear advantages to the task of modelling extremes particularly if informative priors are derived from expert elicitation.

Where forecasts of extremes are to be used as a basis for important decisions, such as 305
the height of sea defences, it is necessary, however, to pay attention to the uncertainty surrounding the forecasts. A particular source of uncertainty arises from the choice of priors. We have seen that if conducted properly, and rigorously in compliance with established elicitation protocols, elicitation can produce no more than a few numbers relating to each of the quantities of interest. We have shown that those numbers are consistent with countless probability distributions, and we have presented a method by which such distributions can be generated. There appears to be no reason to believe that all of those distributions will lead to identical or even similar posteriors in the Bayesian analysis. It follows that estimating the uncertainty of forecasts of extremes, particularly the most extreme extremes, needs to take account of the uncertainty that exists in the choice of priors.

Given the possibility of generating vast numbers of alternative priors, it would not be 306

straightforward to examine each potential prior and quantify its effects on the forecasts of extremes. We suggest that workers in this field should instead examine the range of broad options, light tail or heavy tail, minimal knots or many knots and test their likely effect on the forecasts.

307 It is striking that although very differently derived priors may indeed give different results, we have been able to show examples where such priors do yield very similar results despite the sparsity of observations. That prompts the thought that, just as a MDA might contain many very different probability distributions (Theorem 2), so a given small set of observations may define a wide class of priors. There is scope for research on ways to classify GP priors.

Acronyms

CDF	Cumulative Distribution Function. xii, 4, 21, 22, 65, 83, 96, 99, 100, 112, 120, 121, 122, 123
EVT	Extreme Value Theory. 2, 3, 4, 5, 7, 8, 10, 13, 15, 25, 30, 129
GEV	Generalised Extreme Value. 9, 16, 45, 48, 51, 53, 54
GP	Gaussian process. 4, 62, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 78, 80, 82, 83, 84, 91, 92, 95, 96, 97, 98, 100, 105, 106, 112, 130
GPD	Generalised Pareto Distribution. 13, 14, 15, 20, 21, 22, 23, 45
HCD	health care decision-making. 34, 35, 36
HMC	Hamiltonian Monte Carlo. 51, 87, 88, 89, 91, 95, 96, 99
iid	independent identically distributed. 8, 13, 18, 19
MCMC	Markov chain Monte Carlo. 22, 23, 31, 43, 51, 53, 68, 87, 95
MDA	maximum domain of attraction. 9, 10, 11, 12, 14, 15, 24, 130
MLE	Maximum likelihood estimate. 48
MRL	mean residual life. 45, 53, 114
PDF	Probability Density Function. xii, 4, 46, 63, 83, 84, 96, 98, 99, 100, 102, 106, 107, 109, 110, 111, 112, 120, 121, 122, 123
POT	Peaks Over Threshold. 13, 16, 17, 18, 114
PPP	Poisson point process. xii, 16, 17, 22, 23, 31, 45, 46, 95, 123
SHELF	Sheffield Elicitation Framework. 36, 39

Definitions

significant wave height	Originally defined as the mean of the highest one-third of the waves, as measured from the trough to the crest of the waves but now formally defined as $Hs = 4\sqrt{m_0}$ where m_0 is the variance of the wave heights observed during a given period . xii, 30, 31, 113, 115, 116, 117, 126
return level	The quantile of a quantity corresponding to a particular probability in a given time period. 1, 46
return period	The reciprocal of the probability that a quantity will exceed a given quantile in a given period.. 1
extreme value distribution	The limit distribution of properly normalised maxima of a sequence of independent and identically distributed random variables. 8, 9, 12, 14, 16, 29
extreme value index	The parameter of the Generalised Extreme Value distribution that determines the shape of the density function. 9, 29, 32
shape parameter	The parameter of the Generalised Extreme Value distribution that determines the shape of the density function. Also known as the extreme value index. 9, 45
Fréchet	An extreme value distribution with a positive shape parameter, consistent with no upper bound to the values that might be attained.. 10

Weibull	An extreme value distribution with a negative shape parameter, possessing therefore an upper bound to the values that might be attained.. 10
Gumbel	An extreme value distribution with a zero shape parameter, consistent with no upper bound to the values that might be attained.. 10
exceedances	The cases in which a scalar random number exceeds some given threshold . 14
mean excess function	The mean of those values of a random variable that exceed a given threshold . 18
Markov chain Monte Carlo	A class of algorithms for sampling from a probability distribution by constructing a Markov chain that has the desired distribution as its equilibrium distribution.. 47
realisation	A single sample drawn from a Gaussian process, comprising a set of values at a finite number of points. . 64, 73, 74, 84
scaled stationary covariance function	A covariance function to represent the expectation that the smaller the values of a function the smaller its variance is likely to be. . 67
emulator	In computing, an emulator is hardware or software that enables one computer system to behave like another computer system.. 73
inequality	The state of being unequal. 73, 74, 75, 86, 90
interpolation	Setting fixed values for a function to take at certain points in its support.. 74, 75, 82, 85, 90
training data	a set of examples used to fit the parameters of a model.. 76

References

Behrens, C. N., Lopes, H. F. and Gamerman, D. (2004) Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling 2004; 4: 227–244*, **4**, 227–244.

Berger, J. O. and Sun, D. (1993) Bayesian analysis for the poly-Weibull distribution. *Journal of the American Statistical Association*, **88**, 1412–1418. URL: `http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476426`.

Berman, S. M. (1964) Limit theorems for the maximum term in stationary sequences. *The Annals of Mathematical Statistics*, **35**, 502–516.

Bojke, L., Soares, M., Claxton, K., Colson, A., Fox, A., Jackson, C., Jankovic, D., Morton, A., Sharples, L. D. and Taylor, A. (2021) Developing a reference protocol for structured expert elicitation in health-care decision-making : a mixed-methods study. Health Technology Assessment (Winchester, England).

Bojke, L., Soares, M. O., Claxton, K., Colson, A., Fox, A., Jackson, C., Jankovic, D., Morton, A., Sharples, L. D. and Taylor, A. (2022) Reference case methods for expert elicitation in health care decision making. *Medical Decision Making*, **42**, 182–193.

Carreau, J. and Bengio, Y. (2009) A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes*, **12**, 53–76. URL: `http://dx.doi.org/10.1007/s10687-008-0068-0`.

Cohen, J. P. (1982) The penultimate form of approximation to normal extremes. *Advances in Applied Probability*, **14**, 324–339. URL: `http://www.jstor.org/stable/1426524`.

Coles, S., Pericchi, L. R. and Sisson, S. (2003) A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, **273**, 35 – 50. URL: `http://www.sciencedirect.com/science/article/pii/S0022169402003530`.

Coles, S. G. and Tawn, J. A. (1996) A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **45**, 463–478. URL: `http://www.jstor.org/stable/2986068`.

Colson, A. R. and Cooke, R. M. (2018) Expert elicitation: using the classical model to validate experts' judgments. *Review of Environmental Economics and Policy*.

Cooke, R. M. (1991) *Experts In Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press. URL: https://doi.org/10.1093/oso/9780195064650.001.0001.

Cooley, D. S. (2005) *Statistical Analysis of Extremes Motivated by Weather and Climate Studies: Applied and Theoretical Advances*. Ph.D. thesis, University of Colorado.

Davison, A. C. and Smith, R. L. (1990) Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, **52**, 393–442. URL: http://www.jstor.org/stable/2345667.

DuMouchel, W. H. (1983) Estimating the stable index alpha in order to measure tail thickness: A critique. *The Annals of Statistics*, **11**, 1019–1031. URL: http://www.jstor.org/stable/2241294.

Egozcue, J. J., Pawlowsky-Glahn, V. and Ortego, M. (2005) Wave-height hazard analysis in Eastern Coast of Spain – Bayesian approach using generalized Pareto distribution. *Advances in Geosciences*, **2**, 25–30.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (2013) *Modelling Extremal Events: for Insurance and Finance*, vol. 33. Springer Science and Business Media.

Engelund, S. and Rackwitz, R. (1992) On predictive distribution functions for the three asymptotic extreme value distributions. *Structural Safety*, **11**, 255 – 258. URL: http://www.sciencedirect.com/science/article/pii/016747309290018I.

European Food Safety Agency (2014) Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*, **12**, 3734.

Fawcett, L. and Walshaw, D. (2006) Markov chain models for extreme wind speeds. *Environmetrics*, **17**, 795–809. URL: http://dx.doi.org/10.1002/env.794.

— (2007) Improved estimation for temporally clustered extremes. *Environmetrics*, **18**, 173–188. URL: http://dx.doi.org/10.1002/env.810.

Fisher, R. A. and Tippett, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

Forristall, G. Z. . (1978) On the statistical distribution of wave heights in a storm. *Journal of Geophysical Research*, **83**, 2353–2358.

Frigessi, A., Haug, O. and Rue, H. (2002) A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, **5**, 219–235. URL: http://dx.doi.org/10.1023/A:1024072610684.

Gnedenko, B. (1943) Sur la distribution limite du terme maximum d'une serie aléatoire. *Annals of Mathematics*, **44**, 423–453. URL: `http://www.jstor.org/stable/1968974`.

Gosling, J. P. (2005) *Elicitation: a non-parametric view*. Ph.D. thesis, University of Sheffield.

Gosling, J. P., Oakley, J. E. and O'Hagan, A. (2007) Nonparametric elicitation for heavy-tailed prior distributions. *Bayesian Anal.*, **2**, 693–718. URL: `http://dx.doi.org/10.1214/07-BA228`.

Gumbel, E. J. (1958) *Statistics of Extremes*. Echo Point Books and Media.

de Haan, L. (1990) Fighting the arch–enemy with mathematics. *Statistica Neerlandica*, **44**, 45–68.

Hall, P. (1979) On the rate of convergence of normal extremes. *Journal of Applied Probability*, **16**, 433–439. URL: `http://www.jstor.org/stable/3212912`.

Helmer, O. (1968) Analysis of the future: The delphi method. *Technological Forecasting for Industry and*.

Holliday, N. P., Yelland, M. J., Pascal, R., Swail, V. R., Taylor, P. K., Griffiths, C. R. and Kent, E. (2006) Were extreme waves in the Rockall Trough the largest ever recorded? *Geophysical Research Letters*, **33**.

International Organization for Standardization (2015) Petroleum and natural gas industries – Specific requirements for offshore structures. *Standard ISO 19901 –1:2015*, International Organization for Standardization, Geneva, CH.

Jaynes, E. T. (2003) *Probability theory : the logic of science*. Cambridge: Cambridge University Press.

Kang, S. and Serfozo, R. F. (1999) Extreme values of phase-type and mixed random variables with parallel-processing examples. *Journal of applied probability*, **36**, 194–210.

Kvingedal, B., Bruserud, K. and Nygaard, E. (2018) Individual wave height and wave crest distributions based on field measurements from the northern North Sea. *Ocean Dynamics*, **68**, 1727–1738. URL: `https://doi.org/10.1007/s10236-018-1216-y`.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983) *Extremes and related properties of random sequences and processes*. Springer Science.

Lindley, D. V., Tversky, A. and Brown, R. V. (1979) On the reconciliation of probability assessments. *Journal of the Royal Statistical Society. Series A (General)*, **142**, 146–180. URL: `http://www.jstor.org/stable/2345078`.

Longuet-Higgins, M. S. (1953) On the statistical distribution of the heights of sea waves. *Journal of Marine Research*, **11**, 245–266.

López-Lopera, A. F., Bachoc, F., Durrande, N. and Roustant, O. (2017) Finite-dimensional Gaussian approximation with linear inequality constraints. *arXiv preprint arXiv:1710.07453*.

Loynes, R. M. (1965) Extreme values in uniformly mixing stationary stochastic processes. *Ann. Math. Statist.*, **36**, 993–999.

Maatouk, H. and Bay, X. (2017) Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, **49**, 557–582.

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**, 2137 – 2157. URL: `http://www.sciencedirect.com/science/article/pii/S0167947311000077`.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

do Nascimento, F. F., Gamerman, D. and Lopes, H. F. (2012) A semiparametric Bayesian approach to extreme value estimation. *Statistics and Computing*, **22**, 661–675. URL: `http://dx.doi.org/10.1007/s11222-011-9270-z`.

Neal, R. M. (2011) MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (eds. S. Brooks, A. Gelman, G. Jones and X.-L. Meng). Chapman and Hall.CRC Boca Raton, FL, USA. URL: `http://arxiv.org/abs/1206.1901v1`.

Nocedal, J. and Wright, S. (1999) *Numerical Optimization*. Springer Science & Business Media.

Northrop, P. J., Attalides, N. and Jonathan, P. (2017) Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **66**, 93–120.

Oakley, J. E. and O'Hagan, A. (2007) Uncertainty in prior elicitations: A nonparametric approach. *Biometrika*, **94**, 427–441. URL: `http://www.jstor.org/stable/20441382`.

— (2019) SHELF: the Sheffield Elicitation Framework (version 4). *Tech. rep.*, School of Mathematics and Statistics, University of Sheffield.

O'Hagan, A. (1988) *Probability: methods and measurement*. Chapman and Hall.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E. and Rakow, T. (2006) *Uncertain judgements: Eliciting Experts' Probabilities*. John Wiley and Sons.

O'Hagan, A. and Forster, J. J. (1994) *Kendall's Advanced Theory of Statistics, Vol. 2b: Bayesian Statistics*. John Wiley and Sons.

Ortego, M., Gibergans-Báguena, J., Tolosana-Delgado, R., Egozcue, J. and Llasat, M. (2010) Bayesian trend analysis for daily rainfall series of Barcelona. *Advances in Geosciences*, **26**, 71–76.

Pakman, A. and Paninski, L. (2014) Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, **23**, 518–542. URL: `https://doi.org/10.1080/10618600.2013.788448`.

Pickands, J. (1975) Statistical inference using extreme order statistics. *The Annals of Statistics*, **3**, 119–131. URL: `http://www.jstor.org/stable/2958083`.

R Core Team (2024) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rainey, R. C. T. and Colman, J. G. (2014) 100-year and 10,000-year extreme significant wave heights – how sure can we be of these figures? URL: `http://www.turing-gateway.cam.ac.uk/presentation/2014-07-30/19399`.

Rasmussen, C. and Williams, C. (2006) *Gaussian Processes for Machine Learning*. MIT Press.

Reed, D. (2023) Rainfall depth frequency curves. Private correspondence with J G Colman.

Renard, B., Garreta, V. and Lang, M. (2006) An application of Bayesian analysis and Markov chain Monte Carlo methods to the estimation of a regional trend in annual maxima. *Water resources research*, **42**.

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

Singpurwalla, N. D. (1988) An interactive PC-based procedure for reliability assessment incorporating expert opinion and survival data. *Journal of the American Statistical Association*, **83**, 43–51.

Singpurwalla, N. D. and Song, M. S. (1989) Reliability analysis using Weibull lifetime data and expert opinion. *Microelectronics Reliability*, **29**, 1098 –. URL: `http://www.sciencedirect.com/science/article/pii/0026271489900863`.

Smith, R. (1990) Regional estimation from spatially dependent data. *Preprint. http://www. stat. unc. edu/postscript/rs/regest. pdf*.

Smith, R. L. (1989) Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science*, **4**, 367–377. URL: `http://www.jstor.org/stable/2245845`.

Smith, R. L. and Naylor, J. C. (1987) A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **36**, 358–369. URL: http://www.jstor.org/stable/2347795.

Stan Development Team (2024) *Stan Modeling Language Users Guide and Reference Manual, Version 2.35.0.* .

Tancredi, A., Anderson, C. and O'Hagan, A. (2006) Accounting for threshold uncertainty in extreme value estimation. *Extremes*, **9**, 87–106. URL: http://dx.doi.org/10.1007/s10687-006-0009-8.

Thom, H. (1954) Frequency of maximum wind speed. *Proceedings of the American Society of Civil Engineers*, **80**, 1–13.

Todorovic, P. and Rousselle, J. (1971) Some problems of flood analysis. *Water Resources Research*, **7**, 1144–1150. URL: http://dx.doi.org/10.1029/WR007i005p01144.

Todorovic, P. and Zelenhasic, E. (1970) A stochastic model for flood analysis. *Water Resources Research*, **6**, 1641–1648. URL: http://dx.doi.org/10.1029/WR006i006p01641.

Winkler, R. L. (1967) The Assessment of Prior Distributions in Bayesian Analysis. *Journal of the American Statistical Association*, **62**, 776–800.

de Zea Bermudez, P., Amaral Turkman, M. and Turkman, K. (2001) A predictive approach to tail probability estimation. *Extremes*, **4**, 295–314. URL: http://dx.doi.org/10.1023/A:1016546027962.

de Zea Bermudez, P. and Turkman, M. A. A. (2003) Bayesian approach to parameter estimation of the generalized Pareto distribution. *Test*, **12**, 259–277. URL: http://dx.doi.org/10.1007/BF02595822.