2009

# Statistical Techniques for Extreme Wave Condition Analysis in Coastal Design

Thompson, Paul


# Statistical Techniques for Extreme Wave Condition Analysis in Coastal Design

By Paul Thompson

A thesis submitted to the University of Plymouth

in partial fulfillment for the degree of

**Doctor of Philosophy**

School of Engineering, Faculty of Technology,

University of Plymouth

April 2009

# Abstract

The study of the behaviour of the extreme values of a variable such as wave height is very important in engineering applications such as flood risk assessment and coastal design. Storm wave modelling usually adopts a univariate extreme value theory approach, essentially identifying the extreme observations of one variable and fitting a standard extreme value distribution to these values. Often it is of interest to understand how extremes of a variable such as wave height depend on a covariate such as wave direction. An important associated concept is that of return level, a value that is expected to be exceeded once in a certain time period.

The main areas of research discussed in this thesis involve making improvements to the way that extreme observations are identified and to the use of quantile regression as an alternative methodology for understanding the dependence of extreme values on a covariate. Both areas of research provide developments to existing return level methodology so enhancing the accuracy of predicted future storm wave events. We illustrate the methodology that we have developed using both coastal and offshore wave data sets.

In particular, we present an automated and computationally inexpensive method to select the threshold used to identify observations for extreme value modelling. Our method is based on the distribution of model parameter estimates across a range of thresholds. We also assess the effect of the uncertainty associated with threshold selection on return level estimation by using a bootstrap procedure. Furthermore, we extend our approach so that the selection of the threshold can also depends on the value of a covariate such as wave direction. As a biproduct of our methodological development we have improved existing techniques for estimating and making inference about the parameters of a standard extreme value distribution.

We also present a new technique that extends existing Bayesian quantile regression methodology by modelling the dependence of a quantile of one variable on the values of

another using a natural cubic spline. Inference is based on the posterior density of the spline and an associated smoothing parameter and is performed by means of a specially tuned Markov chain Monte Carlo algorithm. We show that our nonparametric methodology provides more flexible modelling than the current polynomial based approach for a range of examples.

# Contents

# List of Figures

# Glossary of Terms

This glossary of terms was constructed using definitions from Porkess (2005) and CHL (2008).

**Breaching** Formation of a channel through a barrier spit or island by storm waves, tidal action, or river flow. Usually occurs after a greater than normal flow, such as during a hurricane. Alternatively, Failure of a dike allowing flooding.

**Conditional Probability** The probability of an event occurring, given that another event has already occurred.

**Confidence Interval** A random interval, calculated from a random sample that contains the value of parameter with a predetermined probability, known as the confidence level. For example, a 95% confidence interval for the population mean will contain the mean with probability 0.95. This means that on average 95% of repeatedly sampled intervals will contain the population mean.

**Confidence Level** The probability that a confidence interval includes the true value or accepted reference value.

**Confidence Limits** The maximum and minimum values which define the confidence interval.

**Correlation** An index, taking values between $-1$ and 1, that quantify the linear relationship between a pair of variables. The sign indicates the direction of the relationship and the numerical magnitude its strength.

**Covariance** The mean of the product of the derivations of two random variables from their respective means.

**Crest** Highest point on a beach face, breakwater, seawall, dam, dike, spillway or weir.

**Design Wave** In the design of harbours, harbour works, etc., the type or types of waves selected as having the characteristics against which protection is desired.

**Design Wave Condition** Usually an extreme wave condition with a specified return period used in the design of coastal works.

**Design Storm** A hypothetical extreme storm whose waves coastal protection structures will often be designed to withstand. The severity of the storm (i.e. return period) is chosen in view of the acceptable level of risk of damage or failure. A design storm consists of a design wave condition, a design water level and a duration.

**Empirical Distribution** Given a random sample of size $n$, the value of the Empirical Distribution Function at $x$ is the number of elements in the sample less than or equal to $x$, divided by the sample size.

**Fetch** The area in which seas are generated by a wind having a fairly constant direction and speed. Sometimes used synonymously with Fetch length.

**Fetch Length** The horizontal distance (in the direction of the wind) over which a wind generates seas or creates a wind setup.

**Groundwater** The water contained in interconnected pores located below the water table.

**Interpolation** Estimation of a value of a variable between two known values.

**Inference** Drawing conclusions about a parent population on the basis of evidence obtained from a sample.

**Joint Probability** The probability of two (or more) events occurring together.

**Joint Probability Density** Function specifying the joint distribution of two (or more) variables: $F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y)$.

**Joint Return Period** Average period of time between occurrences of a given joint probability event.

**Least-Squares** The method of minimizing the sum of the squares of the residuals (residual=observed value-fitted value) as a method of fitting models to data.

**Likelihood** The probability mass function or probability density function of a random variable $X$ from a given parametric probability distribution interpreted as a function of the parameters given the value $x$ of $X$, instead of as a function of $x$ given the values of the parameters.

**Marginal Probability** The probability of a single variable in the context of a joint probability analysis.

**Marginal Return Period** The return period of a single variable in the context of a joint probability analysis.

**Multivariate Distribution** A probability distribution involving a number of distinct, but not necessarily independent, variables. If two variables are involved it is called bivariate.

**Overtopping** Passing of water over the top of a structure as a result of wave runup or surge action.

**Piping** Erosion of closed flow channels (tunnels) by the passage of water through soil; flow underneath structures, carrying away particles, may endanger the stability of the structure.

**Population Distribution** The probability distribution of the entire set of values of a variable.

**Posterior Probability** A posterior probability is a measure of belief about a situation after collecting experimental data. The posterior probability density is proportional to the product of the likelihood and the prior probability.

**Prior Probability** A prior probability is a measure of beliefs about a situation prior to doing any experiments at all; this is often based on subjective judgement.

**Probabilistic Model** A mathematical model in which the behavior of one or more of the variables is either completely or partially subject to probability laws.

**Probability** The chance that a prescribed event will occur, represented by a number ($p$) in the range 0 to 1. It can be estimated empirically from the relative frequency (i.e. the number of times the particular event occurs, divided by the total count of all events in the class considered).

**Probability Density** A positive function, the area under which is unity, having the area between $a$ and $b$ is the probability associated random variable takes values between $a$ and $b$.

**Quantile** The probability that a random variable takes values below its $p^{\text{th}}$ quantile is $p$.

**Regression** A functional relationship between two or more variables that is of ten empirically determined from data and is used to predict values of one variable when values of the other variables are known.

**Sampling Distribution** The distribution of a statistic obtained from a sample of a particular size, from a given population.

**Sea wall** A structure separating land and water areas, primarily designed to prevent erosion and other damage due to wave action.

**Seepage** The movement of water through small cracks, pores, interstices, out of a body of surface of subsurface water. The loss of water by infiltration from a canal, reservoir or other body of water or from a field. It is generally expressed as flow volume per unit of time.

**Significant Wave Height** The average height of the one-third highest waves of a given wave group.

**Significant Wave Period** An arbitrary period generally taken as the period of the one-third highest waves within a given group.

**Standard Deviation** A measure of spread of test data about the average, or mean, value; the square root of variance.

**Standard Error** This is the standard deviation of the sampling distribution of a statistic.

**Statistic** A numerical characteristic of a sample.

**Stochastic Process** A series of random variable that develops over a period of time.

**Time Series** A set of values of a variable recorded over a period of time.

**Variance** A measure of spread of test data about the average, or mean value, defined as the average squared difference between each data point and the mean.

**Variance-Covariance Matrix** A symmetric matrix in which the off-diagonal elements are covariates of pairs of variables and the elements on the diagonal are the variances of the variables.

**Wave Steepness** The ratio of wave height to wavelength also known as sea steepness.

# Acknowledgements

I would like to express sincere thanks to my supervisors, Professor Dominic Reeve, Dr Yuzhi Cai and Dr Julian Stander for their excellent supervision throughout this research. Their guidance, patience and support was invaluable and is greatly appreciated.

I would also like to thank Dr Anna Zacharoudaki at University of Plymouth and Dr Peter Hawkes at HR Wallingford for supplying some of the data analysed in this thesis and associated supplementary information. Further thanks are due to Dr Rana Moyeed for his advice and thoughts on Bayesian Quantile regression.

I would like to acknowledge the support of a doctoral scholarship from the University of Plymouth and funding from the EPSRC projects RF-PeBLE (grant No. EP/C005392/1), LEACOAST 2 (grant No: EP/C013085/1) and BVANG (grant No: EP/C002172/1).

Thanks to the Coastal Engineering group at the University of Plymouth for being so welcoming and helpful during my time here, particularly to José, "the donut king", for his good humour and friendship throughout. Finally I would particularly like to thank Miss Lorna Turner, my family and friends for their continual support and patience during my studies.

# Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. This programme of advanced study was financed with the aid of funding from the Engineering and Physical Sciences Research Council, the Flood Risk Management Research Consortium and the Faculty of Technology at the University of Plymouth.

## Publications:

P. Thompson, Y. Cai, R. Moyeed, D. Reeve and J. Stander, "Bayesian nonparametric quantile regression using splines", *Journal of Statistical Computing and Data Analysis*, Submitted September 2008, under revision.

P. Thompson, Y. Cai, R. Moyeed, D. Reeve and J. Stander, "Automated threshold selection methods for extreme wave analysis", *Coastal Engineering*, Submitted September 2008, under review.

P. Thompson, D. Reeve, Y. Cai, J. Stander and R. Moyeed "Bayesian nonparametric quantile regression using splines," Conference proceedings, *Flood Risk 2008 Conference, Keble College, Oxford University, UK*, September 2008.

P. Thompson, D. Reeve, Y. Cai and J. Stander "Automated threshold selection methods for coastal design," Conference proceedings, *2nd IMA Flood Risk Assessment Conference, Plymouth, Devon, UK*, September 2007.

A copy of these papers can be found at the end of the thesis.

## Presentation and Conferences Attended:

P. Thompson, D. Reeve, Y. Cai, J. Stander and R. Moyeed "Bayesian nonparametric quantile regression using splines," Oral and Poster Presentation, *Flood Risk 2008 Conference, Keble College, Oxford University, UK*, September 2008.

P. Thompson, D. Reeve, Y. Cai, and J. Stander "Automated threshold selection methods for coastal design," Oral Presentation, *2nd IMA Flood Risk Assessment Conference, Plymouth, Devon, UK*, September 2007.

P. Thompson, D. Reeve, Y. Cai, and J. Stander "Applied time series methods for coastal development," Poster Presentation, *YCSEC'07: Young Coastal Scientists and Engineers Conference 2007, University of Plymouth, UK*, April 2007.

*YCSEC'06: Young Coastal Scientists and Engineers Conference 2006, Southampton University, UK*, April 2006.

Signed..P Thompson..... Dated..14th June 2009....



Word count of main body of thesis: **33347**

# 1

# Introduction

The purpose of any coastal defence is to prevent flooding or erosion which can endanger local properties and their inhabitants. The design of a coastal defence structure which is both reliable and effective is a complex task. The most effective defence structures can be associated primarily with knowledge of the future conditions which the defence must withstand over its design life. This research aims to improve some of the techniques for producing forecasts of future sea conditions, specifically extreme values of wave height.

Coastal defences are relatively difficult to design due to the complicated nature of the sea's behaviour. The overall engineering aim is to create a design which balances as effectively as possible cost and the level of protection (e.g. protection against an extreme wave event that occurs once in 50 years or once in 100 years, known as 50- or 100-year return levels). The success of any structural design relies on the availability and suitable analysis of data; poor or inaccurately analysed data lead to an unsuccessful design. The aim of this thesis is therefore to present and extend some existing techniques that play an important part in coastal defence design and to make them readily available to the wider engineering community.

## 1.1 Introduction to Flood Risk Assessment

In recent years the problem of coastal flooding and erosion has become more evident through increasing media attention. This has led to a greatly increased requirement for improved strategies for flood prevention and flood risk assessment. Flooding in the UK is incurring annual costs of around £1 billion in damages with the current level of protection (Sayers et al. (2002)).


Figure 1.1: Source-Pathway-Receptor-Consequence flow diagram

Defra studies valued UK assets at risk from coastal flooding to be in the region of £132.2 billion and with another £7.8 billion of assets at risk from coastal erosion. However these valuations are set to rise in line with economic growth and climate change.

The term 'flood risk' in coastal locations can be defined as the probability of occurrence of extreme events (storms, tsunamis) leading to coastal erosion or flooding, multiplied by the (socio-)economic damage caused by the extreme event. The link between hazard and consequence can be simplified concisely into the flow diagram shown in Figure 1.1 (Sayers et al. (2002)).

To assess risk we therefore require a method to estimate the probability of the occurrence of extreme events. These probabilities can be transformed into return level plots from which

future levels of a subject wave condition, for example, wave height, can be forecast. To estimate accurately the required probabilities we also need to be able to identify extreme events so that a statistical model can be fitted to them. Existing techniques for extreme value identification and modelling are discussed in greater detail in Chapter 2.

### 1.1.1 Wave Condition Variables

In this section we introduce some definitions that are relevant to the wave condition variables from the data sets used in this thesis. Each variable describes a measurable element of the coastal wave conditions (see Sorenson (1978) and Reeve et al. (2004)). The variables include:

- **Significant wave height** ($H_s$): This is the mean height of the highest third of waves in a given duration or wave group.

- **Wave direction** ($\theta$): This is direction that a wave is travelling. Current analysis techniques group wave direction into 10 degree sections and generate a model which is based on the elements within each section. The use of wave direction dependent thresholds in extreme value modelling is discussed in Chapter 6. In Chapter 7 we present a new modelling technique that also uses wave direction as a covariate.

- **Wave period** ($T_z$): This is the mean zero upcrossing period.

- **Still water level** ($SWL$): It is a combination of three main components: astronomical tide level, mean sea level and meteorological surge level. Tide and surge components are the main sources of variation as mean sea level is assumed to be relatively constant.

- **Steepness** ($S$): Steepness is different from the other variables as it is calculated from wave period and wave height. Its definition is as follows:

$$S = \frac{2\pi H_s}{gT_z^2} \tag{1.1.1}$$

where $g$ is the gravitational constant. This is a variable that has a theoretical upper boundary due to the restriction that solitary waves with $S$ larger than 1/7 will break (Reeve et al., 2004).

## 1.2 Engineering Considerations for Coastal Defence Works

In this section we focus on the technical performance of the defence design as our overall aim is to develop the statistical methods used to analyze wave conditions that provide insight into future extreme behaviour. By increasing the potential efficiency and sustainability of a defence design through the use of the techniques developed in this thesis, we aim to further improve the design's performance and durability. We now define specific terminology used to describe the performance and durability of coastal defence designs.

### 1.2.1 Design or System Life

The term 'design life' is temporal estimate of an object's capability to perform to a satisfactory level, often used in the context of coastal defence design. Sayers et al. (2002) clarifies 'design life' as one of the following:

**Service Life** The period of time over which the owner expects the structure to perform. This is the 'design life' on which guidance is often given in codes of practice.

**Appraisal Life** The period of time over which the client and respective funders or risk owners expect to see a return on their investment.

**Element Life** The period of time over which a certain element will provide sufficient strength to the structure with or without maintenance.

**Residual Life** The period of time to when the defence is no longer able to achieve minimum acceptable values of defined performance indicators in terms of its serviceability function or structural strength. The residual life is often assessed when an inspection of the defence takes place.

In this thesis we think of the 'design life' of a structure as referring to its 'service life'. Service life is dependent on several different interacting factors. The design limitations are the points at which failure can occur. Reeve et al. (2004) describe some modes of failure of sea defences that include:

- excessive overtopping (see Section 1.3.3) without structural failure;

- failure of surface protection leading to a crest level reduction which in turn leads to increased overtopping, washout and breaching (see Reeve et al. (2004) for definitions);
- geotechnical failure of the structure or its foundation leading to reduction of crest level and breaching;
- seepage or piping and internal erosion leading to breaching (see Reeve et al. (2004) for definitions).

### 1.2.2 Performance Measures

To assess the performance of a coastal defence we require criteria which the defence must satisfy. In commercial defence design and construction these criteria are specified as a basis of the contractual agreement. The primary criterion is cost; the defence must be produced on a realistic budget which is cost effective in relation to the value of the assets which it protects. The second most important criterion is the time allocated to the project including conception, design and construction. The defence should be constructed in a relatively short time frame so that it does not become redundant shortly after implementation, i.e. design life of 10 years with a construction time of 4 years is not satisfactory.

The defence must be constructed to a high standard due to its purpose. Potential risk failures are highly correlated to the quality and standard of the construction. Therefore, the defence should be designed using the most efficient and up-to-date methods which lead to quality defects that are acceptable under the ISO9000 series. Following on from this, health and safety during construction must also be a primary concern. The defence must comply with all statutory health and safety regulations. Regular safety reports and maintenance schedules should be in place from the beginning. Finally, the defence must be environmentally balanced allowing for sustainability and mitigating environmental impact.

## 1.3 Design Criteria of Coastal Defences

In this section we aim to highlight coastal defence design criteria and the uncertainty associated with them; see Sayers et al. (2002). The techniques developed within this thesis are aimed to reduce the uncertainty associated with these design criteria.

### 1.3.1 Event Frequency Considerations

#### Definition of Extremes Values

The correct definition and identification of extreme values such as storm events can lead to improved analysis accuracy as only appropriate data would be used in the modelling process.

#### Joint Probability Method

The joint probability method looks at the effect of two variables such as wave height and wave period having extreme values simultaneously. Failure to take account of such joint probability considerations can amplify the effect of any resultant conditions and so can increase the chance of defence failure due to overtopping. Conversely, incorrect application of joint probability techniques can also have an adverse effect as defence design conditions can be overestimated leading to harmful cost or environmental implications.

### 1.3.2 Modelling Considerations

#### Checking Model Results

When using modelling techniques to determine design conditions it is necessary to validate the results that are obtained, to ensure that the most appropriate model is being used. Modelling must be validated using actual data rather than just simulated data. Also, oversimplification of the process to be modelled can result in its misrepresentation leading to design errors. There is a fine balance in any modelling activity to ensure that the model provides a good representation of the data without losing predictive accuracy due to over specification. In Chapter 6 we explain how to make the extreme modelling process more specific by incorporating a covariate based on wave direction into the extreme wave height model. In Chapter 7 we present a new modelling technique that also uses wave direction as a covariate.

### 1.3.3 Overtopping

An engineering motivation for much of the work in this thesis is overtopping as it is regarded as one of the major sources of flooding and erosion. In Chapter 7 will present new modelling techniques and illustrate their engineering relevance by discussing improvements to overtopping return level estimation.

EurOtop (2007) describe wave overtopping as the mean discharge $Q_m$ per linear metre of width. Wave overtopping is very random in time and volume, and hence there is no constant discharge over the crest of a structure during overtopping.

Extreme waves will discharge a large volume of water over the crest of a coastal defence in a short period of time, often less than a wave period, and so are the main cause of overtopping. Smaller waves will produce little, if any, overtopping.

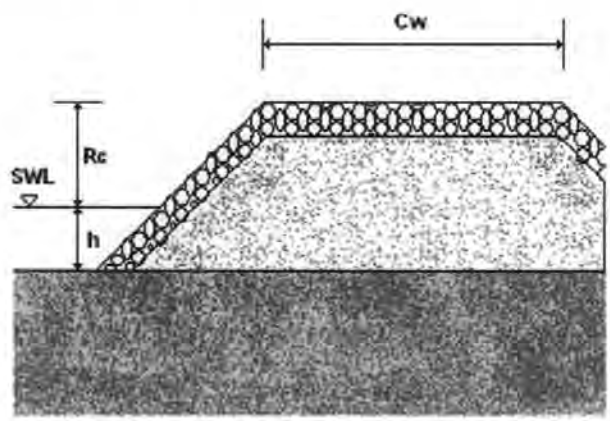



Figure 1.2: Diagram of a typical plane rough-armoured slope defence from Reeve et al. (2004). SWL stands for still water level.

To illustrate the calculations of mean overtopping $Q_m$, we follow the simple calculations from Reeve et al. (2004) applied to a plain rough-armoured slope as shown in Figure 1.2. We present this simple example calculation to motivate the use of techniques developed in this thesis. These calculations were presented in Owen (1980b) which was a technical guideline for overtopping calculations. We also note that a range of variations of this calculation are available for more complex structures, see EurOtop (2007). Firstly, we find the dimensionless

freeboard parameter $R_*$ which is calculated by

$$R_* = \frac{R_c}{T_z(gH_s)^{0.5}}, \tag{1.3.1}$$

where $R_c$ is the height of the crest above the still water level called freeboard, $g$ is the gravitational constant, $T_z$ is wave period and $H_s$ is significant wave height; formal definitions will be given in Section 1.1.1.

Besley (1999) limits the validity of methodology, from Owen (1980b), for predicting mean overtopping discharges to dimensionless freeboard values falling in the range $0.05 < R_* < 0.30$. However, more recent analysis of data by Allsop and Pullen (2003) for 1:2 smooth slopes has shown that this range can be extended to $0.05 < R_* < 0.50$. The next stage is to define a second parameter which we refer to as a dimensionless overtopping rate coefficient as

$$Q_* = A \exp\left(-\frac{BR_*}{r}\right), \tag{1.3.2}$$

where $A$ and $B$ are empirical coefficients dependent on the slope of the structure and $r$ is the roughness coefficient; see Reeve et al. (2004) for tabulated values. The mean overtopping discharge rate $Q_m$ per metre length of structure is then defined as

$$Q_m = Q_* \, T_z \, m \, g \, H_s \tag{1.3.3}$$

Reeve et al. (2004) and EurOtop (2007) go into much more detail and present examples of the calculation of $Q_m$ for more complex structures. In our overtopping calculations of Chapter 7 we continue to make the simplifying assumption of a perpendicular wave approach angle. Taking account of varying wave approach angles is considered to be a topic for further work as discussed in Chapter 8.

## 1.4 Sustainability of Coastal Defences

The sustainability of a coastal defence is another engineering consideration that underpins much of the work in this thesis. The following section, taken from UK (2001), discusses sustainability. The areas defined within this discussion provides some of the key motivations

for our developments of improved techniques for modelling extreme wave conditions. By reducing the uncertainty in extreme value modelling we attempt to optimize the performance considerations discussed in Section 1.2 which dictate the sustainability issues listed below.

**Preserving and Enhancing the Environment**

In conducting any engineering development there is likely to be a potential impact on the local environment. Minimizing the adverse impact of the development on the environment and hence on society can be achieved by ensuring all works must be environmentally neutral or positive. Furthermore, every effort must be made to minimize and otherwise avoid any resultant pollution caused by the works or completed development.

**Using Resources Efficiently**

It is possible for contractors to optimize efficient management of their resources. This can be achieved by aiming to use renewable or recycled construction materials, minimizing the volume of materials used, or even reducing or recycling surplus materials. A less obvious saving can also be made by optimizing energy efficiency in the transportation of materials and the operation of the works or completed development.

**Long-term Viability**

Any defence design aims primarily to provide a good level of protection to local property and its inhabitants over a long term period. It is important to balance the efficiency of the maintenance and operation of the design with the cost of the materials used to create a structure which is designed for long term viability. If possible, the structure should be adaptable during its entire design life to natural processes and climate change. In Section 1.5 we present a case study based on a seaside town near Plymouth that illustrates some of the above concepts.

## 1.5 Case Study: Dawlish, Devon

Located between Torquay and Exeter on the south coast of Devon, Dawlish is a small seaside town situated slightly inland from the coast above high cliffs. Dawlish has a large

area of commercial and private property which is vulnerable to coastal flooding and erosion, including the train line which serves as a main railway route into the Devon and Cornwall region as well as providing local connections. This train line has been the subject of much media interest in recent years as under severe weather it has to be closed due to dangerous conditions or damage to tracks. The track is situated behind a seawall defence which has

Figure 1.3: Image showing waves impacting the Dawlish seawall situated next to the train line. The train shown connects London with the Devon and Cornwall region.

Figure 1.4: Image showing extreme waves impacting and overtopping the Dawlish seawall and flooding the train line.

a beach directly in front of it, as shown in Figure 1.3. Under extreme wave conditions the seawall can be overtopped causing flooding or erosion or both in certain circumstances; an example is shown in Figure 1.4, whilst Figure 1.5 provides a diagrammatic cross-section representation. These effects are costly to repair and have safety and financial implications

for commuters and other train passengers.

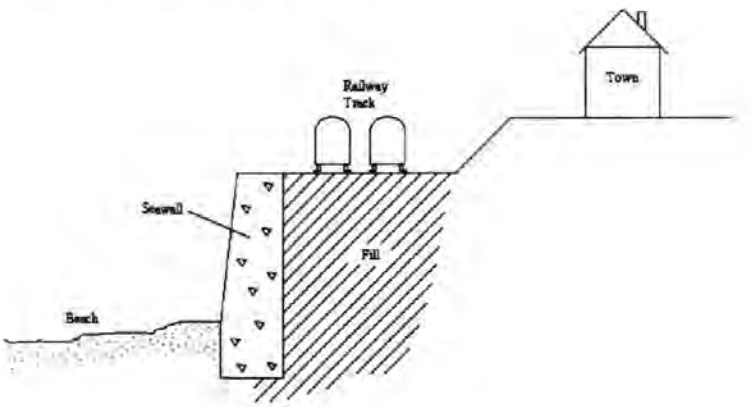



Figure 1.5: A diagrammatic cross-section representation of the Dawlish seawall area from Mockett and Simm (2002)

In 1997 Hyder Consulting were commissioned to undertake a feasibility study of the maintenance or upgrade costs for the existing sea wall at Dawlish; see Mockett and Simm (2002). The aim of the study was to reduce the overall annual maintenance cost of over £1 million for minor repairs and major works by developing a 20–25 strategy. The study highlighted several primary factor influencing the ineffective performance of the defence:

**Reduction in beach levels** It was found that beach levels varied greatly along the stretch of sea wall. Some engineering works already existed to prevent sediment transport and erosion of the beach, but at particular locations these schemes had become ineffective.

**Undermining of the wall** Due to erosion from wave action the toe of the sea wall had become exposed in certain locations causing removal of the infill material behind the structure. This action leaves voids within the defence that can lead to structural failure.

**Voids behind the wall** Voids created by both groundwater and wave action had led to the instability in the defences and structural support of the train line. Some maintenance operations using grouting techniques had provided an adequate repair but not a permanent solution to the voids.

**Overtopping** As a result of low beach levels due to sediment movement and erosion, the sea

wall had become exposed to large waves which are capable of overtopping the defence and causing train line closure due to damage of the track or waves impacting on the trains. Examples of wave overtopping can be seen in Figures 1.4 and 1.6.

Hyder Consulting identified the key aims of the structure and devised a number of potential engineering proposals to reduce the maintenance costs and improve the defence's performance. These proposals were investigated further and some were approved. These decisions depended on the proposal's feasibility, cost and achievable benefit to clients. The proposals were categorized as follows:

*Rejected Proposals*

- The proposed coping for the seawall was considered to be not viable because, using probability analysis of waves and water levels, it was concluded that there was a small chance of green water overtopping within the next 50 years. Therefore, the proposed coping for the seawall was considered to be not viable.
- The proposal to raise beach erosion protectors called groynes, by 500 mm would lessen impact on the beach. However, it was found that the additional forces on the groynes would reduce their design life considerably, and so this proposal was also rejected.
- Offshore structures and beach de-watering systems were found to be expensive and not viable.

*Approved Proposals*

- Detection and repair of voids.
- Construction of a new stepped toe on the seawall.
- New facing works at Dawlish train station.
- Masonry repairs to the face of the wall in sections which were significantly damaged.
- Concrete spraying at the toe as an emergency measure.

In the 2003–04 autumn and winter period bad storm surges caused serious damage to the line causing delays and safety issues. The BBC news web page quoted Professor Laurence Mee,

Figure 1.6: Image showing overtopping of the Dawlish seawall with potential flooding of the train line.

director of the University of Plymouth Marine Institute, who said that "...the more worrying concern is if it [the train line in the Dawlish area] is catastrophically washed away, which would really cut off a major part of the South West and affect the South West economy".

In general it is the balance between associated risk and cost that is the driving factor of any defence scheme. The Dawlish seawall provides adequate performance for protection of the railway structure but allows considerable overtopping in certain circumstances. In an ideal situation a defence would be designed to allow no overtopping, but in real engineering scenarios this is a very impractical and costly operation.

In this thesis we aim to improve some of the probability analysis techniques used to make assessment of potential overtopping in schemes such as the Dawlish sea wall. We achieve this by providing some improvement to the methodology used in wave height modelling to produce return levels. Accurate specification of return levels for wave height provides a great deal of information about the potential overtopping of existing and planned defence schemes and can therefore reduce the risk associated with overtopping of coastal defence schemes. Our methodological improvements were all developed using real data set, which we now discuss.

## 1.6 Introduction to Data Sets

The data used in this thesis were generated using a technique called Hindcasting. The hindcasting technique used by HR Wallingford combines both a wave prediction model (HINDWAVE) and a wave refraction model (TELURAY) to obtain the best possible generated wave condition results (see Wallingford (2005a,b) and Reeve et al. (2004)). Technical explanation of the HR Wallingford Hindcasting model is found in Appendix B. This technique uses wind records to generate wave condition data sets. The main benefit for using this technique is that large wind record data sets are more readily available than large wave condition data sets. The use of generated data rather than directly measured data overcomes the scarcity of good quality, reliable data which span an adequate time scale. All the methods discussed this thesis can, however, be applied to either measured or wind generated data.

### 1.6.1 Data Set Information

The data utilized in this thesis come from two sources. The first hindcast data sets were obtained from Dr Peter Hawkes at HR Wallingford. These hindcast data sets are based on wind condition records from the Met Office for the Selsey Bill area, located on the south coast of the UK, east of Portsmouth and the Isle of Wight. This area has significant existing coast defences to prevent coastal erosion and sediment transport, and has additional defences to prevent flooding of local commercial and private properties.

The data set consists of hourly hindcast measurements of the variables significant wave height, wave period and wave direction over an approximate time span of 27 years. There are seven data sets, all of which are approximately 250,000 observations in length and refer to different tide states; hence a full tide spectrum is given. Measurements of still water level are unavailable. The variable $Y$ of interest will be wave height, while the covariate $t$ will be the cosine of wave direction. In this thesis we use a random sample of 10,000 observations for computational and presentational reasons. The data sets were initially stored in a format that is compatible with the Fortran 90 programs used for the current joint probability software JOINSEA. This software is described in Chapter 2. The data files were reformatted to a

universal `.txt` format which could be viewed using most word processing software. From this format they can be imported simply into the R statistical software used in this project, although R itself can handle a vast range of different formats.

The majority of our model development used parts of the HR Wallingford data sets referring to 'high water level'. A range of samples of different sizes were taken from these data sets to reduce computational time when processing the data.

The second source of data was Dr Anna Zacharioudaki from the School of Engineering at the University of Plymouth. The data are also hindcast generated from wind records. The wind records upon which these data are based were courtesy of the Danish Climate Center and Danish Meterological Institute and were generated from a climate model; see Zacharioudaki (May 2008). These wave records refer to an offshore location in Poole Bay, UK (50.5246 N, −1.6410 E). There are three variables: Wave Height, Wave Period and Wave Direction, each having 86,384 observations at three hourly intervals, so amounting to just over 29 years of data. We include this data when validating the techniques that we have developed as these offshore wave data have a different underlying structure than the HR Wallingford coastal wave data. Wave heights in the offshore data are more uniform across wave direction than in the coastal data.

## 1.7 Research Performed and Structure of the Thesis

As already discussed, one of the main aims of our research is to produce more reliable estimates of the future conditions that a coastal defence structure will need to withstand than are currently routinely available. This is achieved by improving existing techniques for identifying data for extreme value modelling and by investigating alternative methodology for understanding the dependence of extreme values on covariates such as wave direction.

To be more specific, the improvements reported in this thesis include automating the threshold approach used in extreme value modelling and discussed in Chapter 2, and improving existing model parameter estimation methodology. We also extended our automated threshold selection technique to allow the selected threshold to depend on a covariate such as wave direction. In addition, we develop current quantile regression techniques for extreme value modelling that assume a polynomial dependence on the

covariate to allow a more general smooth relationship defined by a spline.

The rest of the thesis is structured as follows. In Chapter 2 we present a literature review of current methodology for univariate and bivariate extreme value theory, discussing various approaches formatting inferences about these models. Chapter 3 reviews the JOINSEA software and methodology developed by HR Wallingford and Lancaster University. Chapter 4 is a literature review of current methodology relevant to our quantile regression work in Chapter 7. In addition to introducing quantile regression, Chapter 4 discusses Bayesian inference and nonparametric modelling. In Chapter 5 we introduce a new, simple technique developed to replace missing observations within our data, so that subsequent analysis can be based on complete data sets.

Chapter 6 presents our new, automated threshold selection technique for both univariate and bivariate extreme value modelling. This chapter also extends this technique to allow the selected threshold to depend on a covariate such as wave direction. Chapter 7 discusses an original technique for Bayesian quantile regression based on splines, as a natural extension to the Bayesian quantile regression methodology found in Yu and Moyeed (2001).

Finally, Chapter 8 draws conclusions about the developments presented in this thesis, highlighting their new contributions to coastal engineering design and statistical methodology. Appendices follow containing full details of mathematical calculations relevant to Chapter 6 and of the HR Wallingford Hindcast wave data generating technique.

# 2

# Literature Review of Extreme Value Theory

In this chapter univariate and bivariate techniques for Extreme Value Theory are introduced. These techniques have been developed to characterize values of an extreme nature within a sample space. This chapter discusses techniques which describe and model extreme data and which can subsequently be used as a basis for the improved methodology for the description and prediction of extreme sea conditions presented in Chapter 6.

## 2.1 Univariate Extreme Value Theory

Extreme value distributions have been used in a range of applications for quite sometime now. It is only during the last twenty years or so, the study of the extreme values of processes (especially natural ones) has increased in popularity with significant improvements to the techniques which describe their behaviour. Extreme analysis techniques were initially proposed as an alternative to the use of a model for the entire range of observations not specifically describing the extreme values. This led to two main approaches for the description of extremes. These are grouped maximum and threshold models, both of which we now present.

### 2.1.1 Group Maximum Methodology and the Generalized Extreme Value Distribution

Classic Extreme Value Theory extracts extreme values by dividing time series data into a set of time intervals or blocks and selecting the maximum from each interval, creating a sample of maxima to which an extreme value model can be fitted (see Finkenstadt and Rootzen, 2004). It is also possible to look at minima as extremes, although this study concentrates on maxima.

Consider a sequence of independent random observations $X_1, \ldots, X_n$ of a random variable $X$ with common distribution function, $F$, where $F(x) = \Pr(X \geq x)$. An example of such data would be daily mean rainfall measurements. When considering the extremes of this sequence we adopt the notation

$$M_n = \max\{X_1, \ldots, X_n\} \tag{2.1.1}$$

where $M_n$ is the maximum of the $n$ observations. The development of a model to describe the statistical behaviour of $M_n$ is the basis of the approach that we will describe. It is possible to obtain the distribution of $M_n$ as follows:

$$\begin{aligned} \Pr\{M_n < z\} &= \Pr\{X_1 \leq z, \ldots, X_n \leq z\} \\ &= \Pr\{X_1 \leq z\} \times \cdots \times \Pr\{X_n \leq z\} \\ &= \{F(z)\}^n \end{aligned} \tag{2.1.2}$$

This requires that the distribution function $F$ be known, which is uncommon in practice. Estimation of $F$ results in small discrepancies which may be transferred to $F^n$ and consequently exaggerated, leading to an unreliable estimate of the distribution function of $M_n$. As an alternative, we investigate models for $F^n$ which are then fitted to the $M_n$ data only.

To understand the methodology used to obtain the form of the distribution $F^n$, it is useful to explain the basis of the method, the Central Limit Theorem. We therefore begin by describing the method used to approximate the sampling distribution of the mean of a

random sample of observations (see Devore and Peck, 1994).

**Definition 2.1.1.1.** ***General results concerning the sampling distribution of*** $\overline{X}$.

*Let* $\overline{X}$ *be the mean of the observations in a particular random sample of size* $n$ *from a population with mean* $\mu$ *and standard deviation* $\sigma$. *Let the corresponding mean and standard deviation for* $\overline{X}$ *be* $\mu_X$ *and* $\sigma_X$ *respectively. The following apply:*

- $\mu_X = \mu$
- $\sigma_X = \frac{\sigma}{\sqrt{n}}$
- *When the population is normal, the sampling distribution of* $\overline{X}$ *is also normal for all sample sizes* $n$.
- *If* $n$ *is sufficiently large, the sampling distribution of* $\overline{X}$ *can be effectively approximated by a normal curve, even if the population is not normal.*

The Central Limit Theorem can be used to standardize a sample mean (linear renormalization), so that it will follow a standard normal distribution, provided that $n$ is large enough. The Central Limit Theorem can be stated as follows:

**Theorem 2.1.1.2.** ***The Central Limit Theorem***

*If* $X_1, \ldots, X_n$ *constitute a random sample from an infinite population with mean* $\mu$, *and finite variance* $\sigma^2$, *then the limiting distribution of*

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}, \tag{2.1.3}$$

*or alternatively*

$$\frac{X_1 + \cdots + X_n - n\,E(X)}{\sqrt{n\ Var(X)}}, \qquad as \quad n \to \infty \tag{2.1.4}$$

*is the standard normal distribution.*

Note that this theorem deals with a random variable of the form, $Y_n=(S_n - b_n)/a_n$ , where $S_n=X_1+\cdots+X_n$. The theorem shows that there exist constants $a_n$ and $b_n$, such that $Y_n$ converges in distribution to a non-degenerate distribution.

The Central Limit Theorem tells us that the normal distribution is found as the limit for the sum of a number of random variables. There are, however, exceptions to this result. If the

underlying distribution $F$ has very heavy tails, then a different distribution, called a stable distribution, will be the limit. A distribution is stable provided the following holds. If we have a number of independent identically distributed random variables which follow a stable distribution, then a linear combination of these variables will have the same distribution, with the exception of different location and scale parameters (also see page 22 for further discussion).

Effectively, the distribution of the mean is altered due to the extremes from the sample, and this affects the asymptotic behaviour. The limit distribution is now the Pareto distribution, which corresponds to a random variable with infinite variance.

An alteration to the Central Limit Theorem is required for the case of maxima rather than averages or sums. Linear renormalization is still required, since for $z < z_+$ ($z_+$ being the upper end point of $F$), $F^n(z) \to 0$ as $n \to \infty$,with the result that the distribution of $M_n$ degenerates to point mass on $z_+$. The linear renormalization of $M_n$ is denoted $M_n^*$ and is defined as:

$$M_n^* = (M_n - b_n)/a_n \tag{2.1.5}$$

where $a_n > 0$ and $b_n$ are sequences of constants. The careful selection of values for $a_n > 0$ and $b_n$ stabilizes the location and scale of $M_n^*$ as $n$ increases. Following this renormalization it is now possible to search for the limiting distributions for $M_n^*$. Hence, we seek a distribution function $G$ and a sequence of constants $a_n > 0$ and $b_n$, such that the limiting distribution of $M_n^* = (M_n - b_n)/a_n$ is $G$ as $n \to \infty$ . This is the initial stage of the Extremal Types Theorem; see Coles (2001) and Beirlant et al. (2004).

**Theorem 2.1.1.3.** ***The Extremal Types Theorem***

*If there exists sequences of constants $a_n > 0$ and $b_n$ such that*

$$Pr\{(M_n - b_n)/a_n \leq z\} \to G(z) \qquad as \quad n \to \infty \tag{2.1.6}$$

*where $G(z)$ is a non-degenerate distribution function, then $G(z)$ belongs to one of the following families:*

***Gumbel***

$$G(z) = \exp\left[-\exp\left\{-\left(\frac{z-b}{a}\right)\right\}\right] \quad -\infty < z < \infty \tag{2.1.7}$$

***Fréchet***

$$G(z) = \begin{cases} 0 & z \leq b \\ \exp\{-(\frac{z-b}{a})^{-\alpha}\} & z > b \end{cases} \tag{2.1.8}$$

***Weibull***

$$G(z) = \begin{cases} \exp\{-[-(\frac{z-b}{a})^{\alpha}]\} & z < b \\ 1 & z \geq b \end{cases} \tag{2.1.9}$$

*where $a > 0$ and $b$ are scale and location parameters respectively, and in the Fréchet and Weibull cases, $\alpha > 0$ is a shape parameter.*

The Extremal Types Theorem states that any suitably normalized variable $M_n^*$ has a limiting distribution which must be one of the Gumbel, Fréchet or Weibull distributions. This choice of limiting distribution is dependent on the underlying population distribution $F$. Unfortunately, $F$ is unknown. We can however resolve this problem by combining the three families of models into the following Generalized Extreme Value Distribution (see Kotz and Nadarajah (2002), Coles (2001) and Beirlant et al. (2004)):

$$G(z) = \exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}^{-1/\xi}\right], \tag{2.1.10}$$

provided $1 + \xi(z - \mu)/\sigma > 0$. This model has three specific parameters, each describing an element of the model: a location parameter, $\mu$; a scale parameter, $\sigma$; and a shape parameter, $\xi$. The Generalized Extreme Value (GEV) Model can be reduced to each of the individual families by the selection of the shape parameter,

- **Gumbel** : $\xi = 0$
- **Fréchet** : $\xi > 0$
- **Weibull** : $\xi < 0$

Data comprising annual or monthly maxima, for example, can be effectively modelled using this method by considering the inference made on the shape parameter, as this determines the tail behaviour of the distribution. The Extremal Types Theorem, Theorem 2.1.1.3 can be rewritten for the GEV (see Coles, 2001) as follows:

**Theorem 2.1.1.4.** ***Extremal Types Theorem: GEV adaption***

*If there exists sequences of constants $a_n > 0$ and $b_n$ such that*

$$Pr\{(M_n - b_n)/a_n \leq z\} \to G(z) \qquad as \quad n \to \infty \tag{2.1.11}$$

*for a non-degenerate distribution function $G(z)$, then $G(z)$ is a member of the GEV family*

$$G(z) = \exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}^{-1/\xi}\right]$$

*which is defined on the set $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where the parameters satisfy $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.*

The application of this theorem is restricted to large values of $n$, i.e. the maxima of large blocks is required. It is possible to avoid problems when normalizing constants are unknown as follows: from Theorem 2.1.1.4 we have

$$\Pr\{(M_n - b_n)/a_n \leq z\} \simeq G(z) \tag{2.1.12}$$

hence,

$$\Pr\{M_n \leq z\} \simeq G\{(z - b_n)/a_n\} = G^*(z) \tag{2.1.13}$$

where $G^*$ can be shown to be another member of GEV family.

Hence, the theorem now also provides a method for estimation of the distribution of $M_n$, using a member of the GEV family. The parameters $\mu$, $\sigma$ and $\xi$ of the GEV can be estimated by maximum likelihood estimation, as explained in Section 2.1.2.

To justify the use of the Extremal Types Theorem (Theorem 2.1.1.4) introduction of an informal proof is necessary. We start by making the following definition from Coles (2001):

**Definition 2.1.1.5.** *A distribution $G$ is said to be max stable if, for every $n = 2, 3, \ldots$, there are constants $\alpha_n > 0$ and $\beta_n$ such that*

$$G^n(\alpha_n z + \beta_n) = G(z)$$

since $G^n$ is the distribution function of $M_n = \max\{X_1, \ldots, X_n\}$, when the $X_i$ are independent variables each with distribution function $G$, the max stability property holds for any distribution for which the process of sampling maxima leads to an identical distribution up to change of scale or location. Using the following key theorem, it is possible to determine the connection to the extreme value limit laws.

**Theorem 2.1.1.6.** *A distribution is max stable if, and only if, it is a generalized extreme value distribution.*

This theorem is then inserted as an integral part of the proof of Theorem 2.1.1.4 the Extremal Types Theorem. Let the maximum random variable in a sequence of $n \times k$ variables for any large $n$ be denoted as $M_{nk}$, and let the limit distribution of $\Pr\{(M_n - b_n)/a_n\}$ be $G$. Hence, for large values of $n$,

$$\Pr\{(M_n - b_n)/a_n \leq z\} \approx G(z)$$

. So it follows that for any integer $k$, since $nk$ is large,

$$\Pr\{(M_{nk} - b_{nk})/a_{nk} \leq z\} \approx G(z) \tag{2.1.14}$$

However, $M_{nk}$ is the maximum of $k$ variables with the same distribution as $M_n$, so

$$\Pr\{(M_n - b_n)/a_n \leq z\} = [\Pr\{(M_{nk} - b_n)/a_n \leq z\}]^k \approx \{G(z)\}^k \tag{2.1.15}$$

Utilizing equations (2.1.14) and (2.1.15) respectively,

$$\Pr\{M_{nk} \leq z\} \approx G\left(\frac{z - b_{nk}}{a_{nk}}\right)$$

and

$$\Pr\{M_{nk} \leq z\} \approx G^k\left(\frac{z - b_n}{a_n}\right)$$

. Hence, we can conclude that $G$ and $G^k$ are identical apart from different location and scale coefficients. So we have shown that the limit distribution $G$ is max stable. Theorem 2.1.1.6 then tells us that $G$ must be a member of the GEV family. A justification of this result can

be found in Coles (2001)

We next present the technique used to estimate the parameters of the Extreme Value Models, known as maximum likelihood estimation (MLE).

### 2.1.2 Maximum Likelihood Estimation for GEV Models

Maximum likelihood estimation is a statistical technique used to make inferences about the parameters of the underlying probability density function of a given data set; see Hogg and Craig (1995), Eliason (1993) and Freund (1992).

More precisely, if the underlying density function $f$ depends on an unknown parameter $\theta$, we can use maximum likelihood estimation to estimate $\theta$. To do this, we consider the probability density function of all the data. If this joint probability density function is thought of as a function of $\theta$, it is known as the likelihood function. The maximum likelihood estimation method finds the value for $\theta$ which maximizes the likelihood function. This is then used as an estimate for $\theta$ and is denoted $\hat{\theta}$. Mathematically, we have $x_1, \ldots, x_n$ which are independent realizations of a random variable with probability density function $f(x;\theta) = \frac{d}{dx}F(x;\theta)$, then the likelihood function is defined as

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{2.1.16}$$

For simplicity in further calculations and numerical stability it is usual to take logs of the likelihood and use the log-likelihood function

$$\log \mathrm{L}(\theta) = \ell(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta) \tag{2.1.17}$$

Because log is a monotonically increasing function, the same value of $\theta$ maximizes $\ell$ and L. Often the likelihood cannot be maximized analytically, so numerical optimization is adopted.

When fitting a GEV model, a blocking method is first applied to split daily data, for example, in to $m$ annual blocks. The maximum from each block is then recorded. These block maxima can be denoted $z_1, \ldots, z_m$ and thought of as a sample from the population of maxima. The parameters of the GEV can then be estimated from this sample by means of the maximum likelihood estimation technique.

*GEV maximum likelihood estimate assuming $\xi \neq 0$ (Fréchet and Weibull)*

If $\xi \neq 0$ (corresponding to the Fréchet and Weibull), the log-likelihood for the GEV parameters will be

$$\ell(\mu, \sigma, \xi) = -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi} \tag{2.1.18}$$

provided that the following conditions are satisfied

$$1 + \xi \left(\frac{z_i - \mu}{\sigma}\right) > 0 \qquad \text{for } i = 1, \ldots, m. \tag{2.1.19}$$

This condition is necessary to prevent data exceeding the end point of the distribution at particular combinations of the parameters, $\mu$, $\sigma$, $\xi$ causing the likelihood to be equal to zero and consequently the log-likelihood to equal $-\infty$.

*GEV maximum likelihood estimate assuming $\xi = 0$ (Gumbel)*

If ($\xi = 0$) (corresponding to the Gumbel)form of the log-likelihood must be adapted

$$\ell(\mu, \sigma) = -m \log \sigma - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp \left[-\left(\frac{z_i - \mu}{\sigma}\right)\right] \tag{2.1.20}$$

Given $\ell(\mu, \sigma, \xi)$ in (2.1.19) and its special case $\ell(\mu, \sigma)$ in (2.1.20), a numerical optimization technique can be employed to find the parameters that maximize the log-likelihood. This is implemented in the function 'gev.fit' of the (R) package (ismev) (Coles and Stephenson (2006)). The above described blocking procedure can be wasteful of data as values which would otherwise be considered as extreme are missed if they fall within a block which has a maxima that is higher than these "extremes". We now present the second technique for modelling extremes, threshold modelling, which has the benefit that all extremes are included provided the threshold has been set properly.

## 2.2 Generalized Pareto Distribution

### 2.2.1 Threshold Modelling

The second method to describe and model extremes is based on setting a threshold and declaring exceedances of this threshold to be extreme values (see Davidson and Smith (1990)). However, the selection of the threshold is not a simple task; current methods are based on diagnostic plots from which the threshold is chosen based on the point where change in behaviour of the plot occurs. The identification of this change is performed purely on an empirical basis by the user. We will discuss this further in Section 2.2.2. In Chapter 6 we present methodology to improve upon existing threshold choice methodology which allows the threshold to be chosen automatically. As mentioned, the value of a variable $X_i$ with distribution function $F$ is considered to be extreme if it exceeds a threshold $u$. The behaviour of extremes can be described by the following conditional probability:

$$\Pr(X > u + y | X > u) = \frac{1 - F(u+y)}{1 - F(u)}, \qquad y > 0. \tag{2.2.1}$$

Just as in the Group Maximum methodology, the population distribution $F$ is unknown, so an alternate means to estimate the distributions of extremes for the threshold exceedance approach is required. For the Group Maximum methodology the GEV distribution model was adopted as the non-degenerate distribution function of suitably scaled and shifted block maximum. In the threshold exceedance approach, the model must be altered to allow for the different form of the extremes. It turns out that the appropriate model is the Generalized Pareto Distribution (GPD) as described in the following theorem (see Coles (2001) and Davidson and Smith (1990)):

**Theorem 2.2.1.1.** *The Generalized Pareto Distribution (GPD)*

*Let $X_1, \ldots, X_n$ be a sequence of independent random variables with common distribution function $F$, and let*

$$M_n = \max\{X_1, \ldots, X_n\} \tag{2.2.2}$$

*Suppose that $F$ satisfies the Extremal Types Theorem, so that for large $n$,*

$$Pr\{M_n \leq z\} \approx G(z) \tag{2.2.3}$$

$$where \quad G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{2.2.4}$$

*for some $\mu$, $\sigma > u$ and $\xi$. Then for large enough $u$, the distribution function of $Y = X - u$, conditional on $X > u$, is approximately*

$$H(y) = 1 - \left[1+\frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi} \tag{2.2.5}$$

$$where \quad \tilde{\sigma} = \sigma + \xi(u-\mu) \tag{2.2.6}$$

*defined on $\{y : y > 0$ and $1 + \xi\, y/\tilde{\sigma}) > 0\}$.*

As mentioned the threshold exceedance approach is usually considered to be much less wasteful of information then the group maximum methodology. Theorem 2.2.1.1 that if block maxima have an approximating GEV distribution, then extremes above a threshold should have an equivalent GPD. Furthermore, the parameters of the GPD are uniquely determined from the threshold $u$ and the GEV parameters. In particular the GPD scale parameter $\tilde{\sigma}$ depends on the GEV parameters $\mu$,$\sigma$ and $\xi$ and upon the threshold $u$ through $\tilde{\sigma} = \sigma + \xi(u-\mu)$. We shall make considerable use of this result later. We find that the shape parameter is common in both distributions. The GEV has two additional parameters which are dependent on size and location, while the equivalent GPD does not have a location parameter. Because of this, the shape parameter $\xi$ becomes the dominating influence on the GPD's behaviour. The upper limits of excesses are as follows:

- If $\xi < 0$, then the distribution of threshold excesses has upper limit $u - \tilde{\sigma}/\xi$.
- If $\xi > 0$ or $\xi = 0$, then the distribution has no upper limit.

Note that when $\xi = 0$, we should take the limit as $\xi \to 0$ of (2.2.5) so obtaining

$$H(y) = 1 - \exp\left(\frac{y}{\tilde{\sigma}}\right), \quad y > 0 \tag{2.2.7}$$

for the distribution function.

### 2.2.2 Threshold Selection

As mentioned above, current threshold selection procedures are based upon interpretation of plots relating to inferences about the model. The first technique that we will discuss is called the Mean Residual Life plot; for an example of such a plot, see Figure 2.1 reproduced from Coles (2001). This graph plots the mean excess values against a range of thresholds values. Let $Y = (X - u_0|X > u_0)$ and assume that a GPD is valid for threshold excesses over threshold $u_0$, where the GPD has parameters $\sigma_{u_0}$ and $\xi$. Note here that we write $\sigma_{u_0}$ for the GPD parameter $\tilde{\sigma} = \sigma + \xi(u_0 - \mu)$. It can be shown that

$$E(Y) = \frac{\sigma_{u_0}}{1-\xi} \quad \text{provided } \xi < 1, \text{ otherwise the mean is infinite.} \tag{2.2.8}$$

This can be generalized to all thresholds $u > u_0$, with the scale parameter being altered accordingly. Therefore,

$$E(X - u|X > u) = \frac{\sigma_{u_0} + \xi(u - u_0)}{1-\xi}, \tag{2.2.9}$$

since $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$ Now, it is clear that this expected value can be estimated as the mean of the excesses over the threshold $u$. The estimates obtained should depend linearly on $u$, up to sampling error, provided that the GPD approximation holds. Hence, this leads us to produce the following Mean Residual Life Plot

$$\{(u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_{(i)} - u)) : u < x_{\max}\} \tag{2.2.10}$$

where

- $x_{(1)}, \ldots, x_{(n_u)}$ are the $n_u$ observations that exceed $u$
- $x_{\max}$ is the largest $X_i$.

As mentioned above, an example of a Mean Residual Life Plot is shown in Figure 2.1. This is based on the daily rainfall data set, discussed in Example 1.6 of Coles (2001). To interpret the Mean Residual Life Plot, 95% confidence intervals are added to the plot. If we consider a threshold $u_0$ to be the optimum threshold choice. If $u_0$ is the lowest threshold

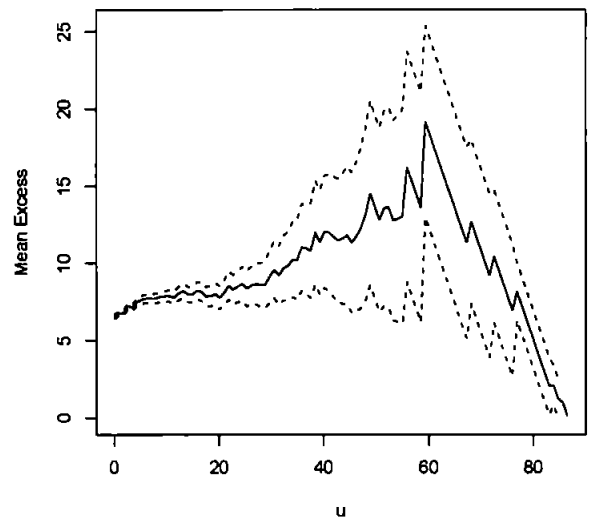



Figure 2.1: The Mean Residual Life plot for the daily rainfall data set of Coles (2001). Confidence intervals are shown as the broken lines.

for which where the GPD provides a valid approximation to the distribution of excesses, then the plot should be approximately linear in $u$ above $u_0$. hence the threshold choice is determined as the point at which linearity in the plot first occurs. For Figure 2.1 this could be $u_0 = 30$, as the plot is linear after 30, up to sampling error, taking account that higher thresholds may have very few values that exceed them. This example makes it clear, however, that threshold choicer based on the Mean Residual Life plot is a very subjective procedure.

The second threshold selection technique provides a much simpler methodology for threshold choice. The process plots the parameter estimates for the GPD model over a range of thresholds. The aim of the plot is to allow the detection of when the model is not sufficient, i.e. when the threshold is set too low or too high.

As before, if it is assumed that the data $X_i$, $i = 1, \ldots, n$ are independent and the GPD is a viable model for the threshold excesses over the lowest possible threshold $u_0$, then excesses over a higher threshold $u > u_0$ will also follow the GPD. As discussed in Section 2.2.1, the shape parameter for the GPD is the same as for the GEV. However, as we have already mentioned, the scale parameter for the GPD is not constant over the thresholds $u$ but takes

the form

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0), \tag{2.2.11}$$

which depends on $u$ unless $\xi = 0$. This dependency is overcome by a simple reparameterization to

$$\sigma^\star = \sigma_u - \xi u \tag{2.2.12}$$

This reparameterized scale parameter is constant with respect to the threshold values $u$. Consequently, both the shape and this new scale parameters will be constant above the threshold $u_0$, up to sampling error, but may not necessarily be constant below $u_0$.

The parameter estimates for $\sigma^\star$ and $\xi$ are then plotted against thresholds $u$; the estimates, $\sigma^\star$ and $\xi$ should remain constant up to sampling error. The threshold $u_0$ is therefore selected as the lowest value for the threshold $u$ for which the estimates are approximately constant. To increase the accuracy of this empirical choice, confidence intervals are added to the parameter estimate plots. These confidence intervals are determined differently for each of the parameters, as follows:

- *Shape parameter* $\xi$:

  The confidence intervals for $\xi$ are determined from variance of the estimate $\hat{\xi}$ found from the numerical optimization procedure. The variance-covariance matrix V:

$$\mathrm{V} = \begin{bmatrix} \mathrm{Var}(\hat{\zeta_u}) & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix} \quad \text{where } \mathrm{Var}(\hat{\zeta_u}) \simeq \hat{\zeta_u}(1 - \hat{\zeta_u})/n \tag{2.2.13}$$

  , in which $\hat{\zeta_u}$ is an estimate of the probability that an individual observation exceeds the threshold $u$ (2.2.6); $v_{i,j}$ is the $(i,j)$ term in the variance covariance matrix $\bar{V}$ of $\hat{\sigma}$ and $\hat{\xi}$, the estimates of the scale and shape parameters respectively. So $\mathrm{Var}(\hat{\xi}) = v_{2,2}$. This variance estimate is used in the standard way to find a confidence for $\xi$.

- *Scale parameter* $\sigma^\star$: the delta method is used, as follows (see Coles, 2001) to find $\mathrm{Var}(\hat{\sigma}^\star)$ from V given in (2.2.13):

$$\text{Var}(\hat{\sigma}^*) \simeq \bigtriangledown \sigma^{*T} \bar{V} \bigtriangledown \sigma^* \qquad \text{where } \bigtriangledown \sigma^{*T} = \left[ \frac{\partial \sigma^*}{\partial \sigma_u}, \frac{\partial \sigma^*}{\partial \xi} \right] \tag{2.2.14}$$

An example of this plot is shown in Figure 2.2. This is based on the daily rainfall data set, discussed in Example 1.6 of Coles (2001).

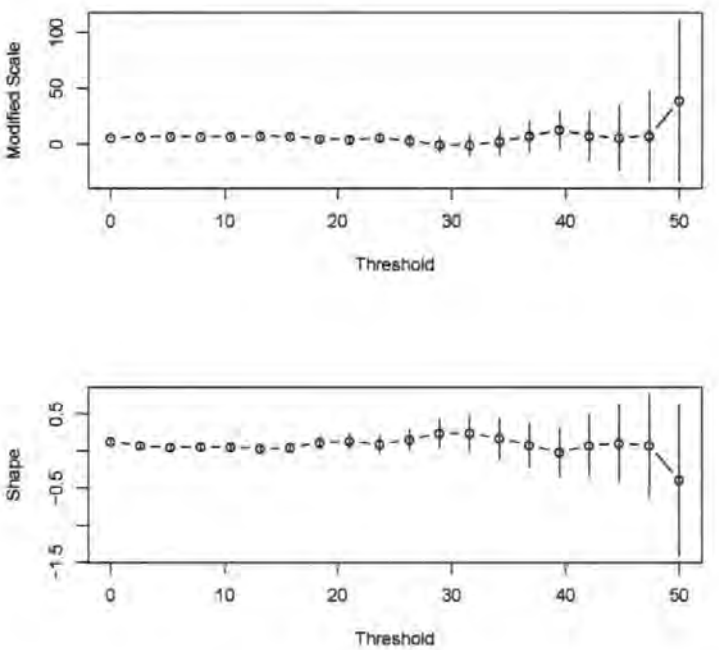



Figure 2.2: Parameter estimate against threshold for the daily rainfall data set of Coles (2001).

### 2.2.3 Parameter Estimation

The estimation of the GPD model parameters is achieved using maximum likelihood estimation (MLE) as discussed for the GEV distribution in Section 2.1.2; see also (see Eliason, 1993). Other parameter estimation techniques are available including Method of moments (MoM) and Probability-Weighted Moments (PWM). The Method of Moments technique works by equating sample moments to the corresponding population moments and solving for the required parameter estimates, i.e. shape and scale parameters, $\xi$ and $\sigma$ respectively.

Hosking and Wallis (1987) introduce the PWM technique and present comparative results between MLE, MoM and PWM, concluding that MoM and PWM are more reliable than

ML when sample size is less than 500. However, it is the case that the techniques presented in Chapter 6 are developed for large data sets due to the nature of the required output, i.e. techniques are developed to produce accurate wave condition forecasts for up to 1000 years ahead. As all the data for this work is collected on an hourly basis as discussed in Chapter 1, 500 observations would amount to just over three weeks of data which would not produce suitably reliable long term forecasts. Hence larger data sets are used and the ML technique should produce sufficiently accurate parameter estimates, see Tawn and Coles (1994).

Initially, one of the threshold selection techniques must be utilized to select an appropriate threshold choice. Then MLE can be applied (see Davidson and Smith (1990)). The excesses of the variable of interest over a threshold $u$ may be denoted $y_1, \ldots, y_k$ if there are $k$ such excesses. In a similar way to the GEV case, the log-likelihood functions for the GPD parameters can be divided into two cases, depending on the value of the shape parameter,

$\xi \neq 0$ :

$$\ell(\sigma, \xi) = -k \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \log \left(1 + \frac{\xi y_i}{\sigma}\right) \tag{2.2.15}$$

provided $1 + \sigma^{-1}\xi y_i > 0$ for $i = 1, \ldots, k$; otherwise $\ell(\sigma, \xi) = -\infty$.

$\xi = 0$ :

$$\ell(\sigma) = -k \log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i \tag{2.2.16}$$

Analytic maximization of the log-likelihood is not possible; hence a numerical optimization algorithm is used. This is implemented in the function 'gpd.fit' of the `R` package `ismev` (see Coles and Stephenson (2006)).

### 2.2.4 Model Fit Assessment

The majority of work conducted within this thesis uses the Threshold Excesses Approach which involves modelling using a GPD, as described in Section 2.2.1 and 2.2.2. Hence only the diagnostic procedures for GPD models are described in this section although similar techniques can be applied to the GEV. When assessing the goodness of the GPD fit, there are several graphical diagnostic methods that can be applied. These include the following:

*Probability Plots:*

The probability plot compares an empirical distribution against the fitted GPD (for more details, see Beirlant et al. (2004) and Finkenstadt and Rootzen (2004)). The more accurate the fit, the nearer to linear the probability plot should be. The points on this plot are defined as

$$\left(\frac{i}{k+1}, \hat{H}(y_i)\right), \qquad i = 1, \ldots, k, \tag{2.2.17}$$

where $i/(k+1)$ is the value of the empirical distribution function corresponding to the $i^{\text{th}}$ point in the ordered data set,

$$\hat{H}(y_i) = 1 - \left(1 + \frac{\hat{\xi} y}{\hat{\sigma}}\right)^{-1/\hat{\xi}} \qquad \text{if } \hat{\xi} \neq 0. \tag{2.2.18}$$

$$\hat{H}(y_i) = 1 - \exp\left(\frac{y}{\hat{\sigma}}\right) \qquad \text{if } \hat{\xi} = 0. \tag{2.2.19}$$

*Quantile Plots*:

The quantile plot also compares an empirical distribution against the fitted GPD (for more details, see Beirlant et al. (2004) and Finkenstadt and Rootzen (2004)). However, this comparison is made on the scale of the quantiles, and not on a probability scale, using the following points

$$\left(\hat{H}^{-1}\left(\frac{i}{k+1}\right), y_i\right), \qquad i = 1, \ldots, k. \tag{2.2.20}$$

where

$$\hat{H}^{-1}(p) = u + \frac{\hat{\sigma}}{\hat{\xi}}\left\{(1-y)^{-\hat{\xi}} - 1\right\} \qquad \text{if } \hat{\xi} \neq 0. \tag{2.2.21}$$

and

$$\hat{H}^{-1}(p) = -\sigma \log(y) \qquad \text{if } \hat{\xi} = 0. \tag{2.2.22}$$

When the Quantile plot is approximately linear this indicates a satisfactory fit has been achieved.

*Density Plots*:

The density plot displays the density function of the fitted GPD overlaid on a histogram of the original values. An example of these plots is shown in Figure 2.5.

### 2.2.5 Return Levels and Periods (GPD only)

The idea of the return period can be best described with an example. In terms of an engineering coastal defence project, for instance a breakwater design, the breakwater will be constructed to particular criteria. More specifically, the coastal defence "fails" if and only if the "worst event" occurs, and so the mean life of the coastal defence is the **return period** of the "worst" event (Castillo et al. (2005) and Coles (2001)). Inferences about return levels can often be more useful than inferences about individual model parameters, particularly in an engineering design scenario. As we concentrate on the GPD in much of this thesis, we shall describe how to calculate return levels from the GPD.

We make the assumption that a GPD with parameters $\sigma$ and $\xi$ is an appropriate model for the excesses of some threshold $u$ by a variable $X$. Hence, for $x > u$, we have

$$\Pr\{X > x | X > u\} = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi} \tag{2.2.23}$$

This implies that

$$\Pr\{X > x\} = \zeta_u \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi} \tag{2.2.24}$$

where $\zeta_u = \Pr\{X > u\}$. It is now possible to define the level $x_m$, known as the $m$-**observation return level**, that is exceeded on average every $m$ observations or, more precisely with probability $1/m$, by rearranging the following equation:

$$\zeta_u \left[1 + \xi\left(\frac{x_m - u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}, \tag{2.2.25}$$

where we are assuming that $m$ is sufficiently large for $x_m$ to be greater than $u$. Solving for $x_m$ we obtain:

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right]. \tag{2.2.26}$$

This is of course only valid for $\xi \neq 0$; when $\xi = 0$, we obtain by a similar argument:

$$x_m = u + \sigma \log(m\zeta_u). \tag{2.2.27}$$

The return level can be shown graphically (Figure 2.5), by plotting $x_m$ against $m$. A

logarithmic scale is used so that linearity is observed when the shape parameter has no effect, i.e. the return level plot is concave and tends to infinity if $\xi > 0$ and convex with a finite asymptote if $\xi < 0$, and so that return levels for small return periods can be seen.

The scale of the return levels used is often important in practical application, for example coastal defence designs often uses return levels on an annual scale as this corresponds to a time scale for this type of design. It is possible to adopt this annual scale, as follows:

The simple transformation requires the number of excesses per year, $n_y$. Therefore, the $N$ year return level, $Z_N$ is

$\xi \neq 0$ :

$$Z_N = u + \frac{\sigma}{\xi}\left[(Nn_y\zeta_u)^{\xi} - 1\right] \qquad \text{setting} \quad m = Nn_y \tag{2.2.28}$$

$\xi = 0$ :

$$Z_N = u + \log(Nn_y\zeta_u). \tag{2.2.29}$$

### 2.2.6 Return Level Estimation

Estimates of return levels are acquired by substituting maximum likelihood estimates of the parameters $\sigma$ and $\xi$ into the above expressions. We also require an estimate of $\zeta_u$, the probability of an individual observation exceeding the threshold $u$. The probability $\zeta_u$ is estimated as

$$\hat{\zeta}_u = \frac{k}{n} \tag{2.2.30}$$

where $k$ is the number of the $n$ events that exceed $u$. In fact $\hat{\zeta}_u$ can easily be seen to be the maximum likelihood estimate of $\zeta_u$, since because of the independence the number of excesses of $u$ follows a binomial distribution $\text{Bin}(n, \zeta_u)$. Furthermore, it is possible to use the properties of the binomial distribution to obtain that $\text{Var}(\hat{\zeta}_u) \simeq \hat{\zeta}_u(1 - \hat{\zeta}_u)/n$ or more precisely that $\hat{\text{Var}}(\hat{\zeta}) = \hat{\zeta}_u(1 - \hat{\zeta}_u)/n$, and so the Variance-Covariance matrix for the parameter estimates can be written as

$$\text{V} = \begin{bmatrix} \text{Var}(\hat{\zeta}_u) & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix}. \tag{2.2.31}$$

Using the delta method (see Freund (1992)), we obtain the variance of the return level as

$$\mathrm{Var}(\hat{x}_m) \simeq \nabla x_m^T V \nabla x_m$$

where

$$\begin{aligned} \nabla x_m^T &= \left[\frac{\partial x_m}{\partial \zeta_u}, \frac{\partial x_m}{\partial \sigma}, \frac{\partial x_m}{\partial \xi}\right] \\ &= \left[\sigma m^{\xi}\zeta_u^{\xi-1}, \xi^{-1}\left\{(m\zeta_u)^{\xi}-1\right\},\right. \\ &\quad \left. -\sigma\xi^{-2}\left\{(m\zeta_u)^{\xi}-1\right\}+\sigma\xi^{-1}(m\zeta_u)^{\xi}\log(m\zeta_u)\right], \end{aligned} \tag{2.2.32}$$

evaluated at $(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$

### 2.2.7 Applied Example of Threshold Exceedance Approach

The example in this section uses the p5data Hindcast data set (which refers to shallow water waves hindcast for a water level corresponding to high water springs) from Dr. Peter Hawkes at HR Wallingford (see Chapter 1). We will concentrate on significant Wave height ($H_s$) as the variable to be modelled.

Using the techniques discussed in Section 2.2 we are able to produce a model for the excesses of an appropriate threshold. To fit the threshold model we must firstly decide on an appropriate threshold choice. The threshold selection techniques based on the Mean Residual Life plot and plots of parameter estimates versus thresholds, as discussed in Section 2.2.2, are now employed as an empirical methods to aid threshold choice.

Figure 2.3 shows the Mean Residual Life plot from which we can see an area of linearity in $u$ between the red lines at $u = 1.65m$ and $u = 2.20m$. This region's lowest value indicates the appropriate threshold choice, i.e. $1.65m$. We also see that we have other regions of linearity but thresholds lower than $1.65m$ are discounted as goodness of fit is lost. This can be confirmed by comparing the widths of confidence intervals on return level plots, or by calculating log-likelihood values. Similarly higher thresholds do not specify enough extremes to justify a satisfactory model representation of the data.

Figure 2.4 shows parameter estimates against a range of corresponding threshold values

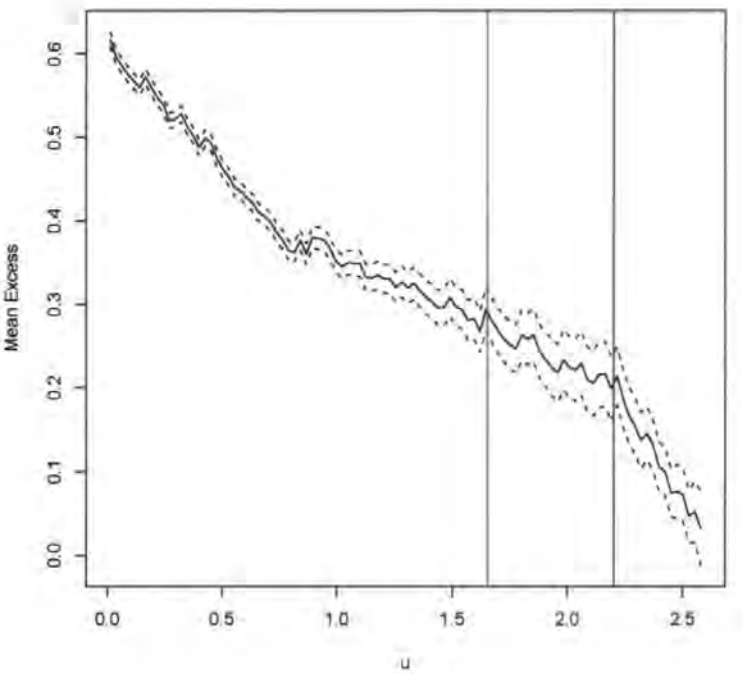



Figure 2.3: The Mean Residual Life plot for the p5data Hindcast data set for the wave height variable. 95% confidence intervals are shown as the broken lines.

and indicates a significant change in behaviour at approximately 1.65m.

As all threshold selection plots are in approximate agreement for an appropriate threshold at $u = 1.65$, we can find the maximum likelihood estimates of the parameters $\sigma$ and $\xi$:

$$(\hat{\sigma}, \hat{\xi}) = (0.392, -0.323)$$

with a corresponding maximized log-likelihood of $-77.38$. The variance covariance matrix takes the following form

$$\bar{\mathrm{V}}(\sigma, \xi) = \begin{bmatrix} 0.00091 & -0.00143 \\ -0.00143 & 0.00292 \end{bmatrix} \qquad (2.2.33)$$

which leads to the corresponding standard errors of 0.0302 and 0.0541 for $\hat{\sigma}$ and $\hat{\xi}$ respectively. The overall number of observations are 10000 of which 299 are deemed as extreme values being in excess of the threshold $u$ set at 1.65m. Using these figures it is possible to complete

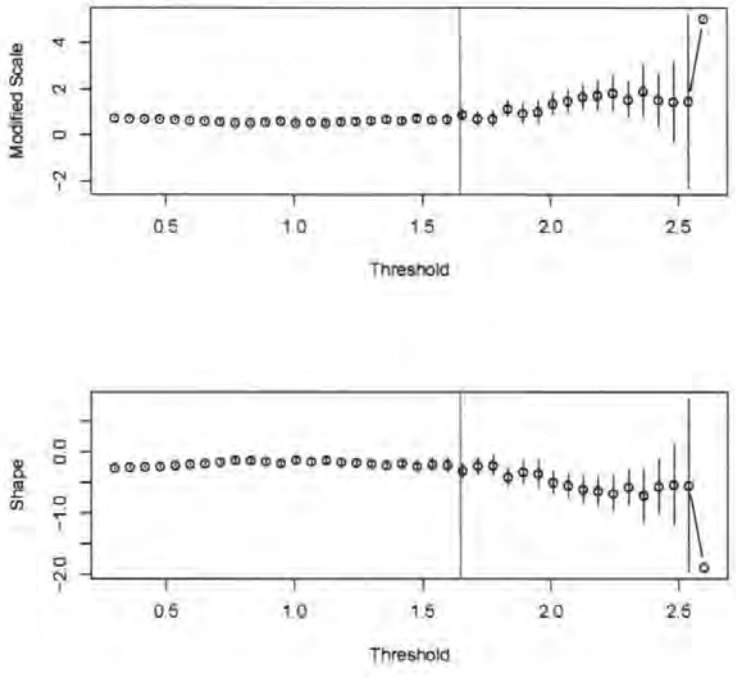



Figure 2.4: Graph showing parameter estimates against a range of thresholds.

the variance covariance matrix for the $(\hat{\zeta}, \hat{\sigma}, \hat{\xi})$. This is achieved by calculation of the maximum likelihood estimate of $\hat{\zeta}_u = 299/10000 = 0.0299$ and its corresponding variance estimate $\hat{\text{Var}}(\hat{\zeta}_u) = \hat{\zeta}_u(1 - \hat{\zeta}_u)/10000 = 0.0000029$. Inserting $\hat{\text{Var}}(\hat{\zeta}_u)$ into the variance-covariance matrix V for $(\hat{\zeta}, \hat{\sigma}, \hat{\xi})$ we obtain from the sub-matrix $\bar{\text{V}}$ in (2.2.33).

$$\text{V} = \begin{bmatrix} 0.0000029 & 0 & 0 \\ 0 & 0.00091 & -0.00143 \\ 0 & -0.00143 & 0.00292 \end{bmatrix} \qquad (2.2.34)$$

The quality of the model can be assessed using the diagnostic plots in Figure 2.5, discussed in Section 2.2.4. The model provides a relatively good fit because the points on the probability plot and the quantile plot lie close to the straight line the points on the return level plot lie within the computed return level confidence envelope, and the fitted GPD density and histogram are similar. In Chapter 6 will present methodology to chase a suitable threshold in an automatic way.

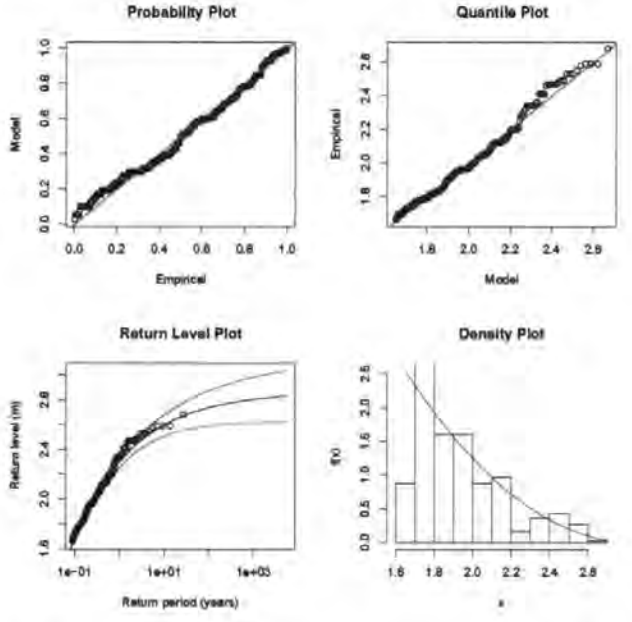



Figure 2.5: Diagnostic Plots for the GPD fit with threshold $1.65m$ for the variable wave height from the p5data.

## 2.3 Bivariate Extreme Value Theory

Let us now assume that we have observations on not one but two random variables $X$ and $Y$. Previously in this chapter our approach would have been to model each variable separately, taking no account of the dependence between them. This is an assumption that is often made but can frequently be incorrect in practical applications. In this section we present characterizations and models for multivariate extremes, with a focus on bivariate extremes. We will revisit the techniques from earlier in this chapter and generalize these to the bivariate case. Initially, we begin this section with some useful definitions to aid explanation of the joint probability techniques.

### 2.3.1 Probability Definitions

The probability of some event $A$, given the occurrence of another event $B$, is known as a conditional probability and can be written $\Pr(A|B)$. The definition of a joint probability would be the probability of two (or more) simultaneously occurring events of interest to produce a particular outcome. A joint probability could be written as $\Pr(A \bigcap B)$ or $\Pr(A, B)$.

Consider our two jointly distributed random variables $X$ and $Y$; the probability distribution of $X$ without consideration of $Y$ is known as the marginal distribution of $X$. This is calculated either by summation (discrete random variables) or integration (continuous random variables) of the joint probability distribution over $Y$. This can be expressed in mathematical notation as follows:

*Discrete: (Probability Mass Function, pmf)*

$$\Pr(X = x) = \sum_y \Pr(X = x, Y = y) = \sum_y \Pr(X = x|Y = y)\Pr(Y = y) \qquad (2.3.1)$$

where $\Pr(X = x, Y = y)$ is the joint distribution of $X$ and $Y$, and $\Pr(X = x|Y = y)$ is the conditional distribution of $X$ given $Y = y$ defined to be $\Pr(X = x, Y = y)/\Pr(Y = y)$.

*Continuous: (Probability Density Function, pdf)*

$$f_X(x) = \int_y f_{X,Y}(x, y)dy = \int_y f_{X|Y}(x|y)f_Y(y)dy \qquad (2.3.2)$$

where $f_{X,Y}(x, y)$ is the joint distribution of $X$ and $Y$, and $f_{X|Y}(x|y)$ is the conditional distribution of $X$ given $Yy$ =defined to be $f_{X,Y}(x, y)/f_Y(y)$. Figure 2.6, taken from Annis (2006), clearly illustrates the relationship between the marginal, conditional and joint probability density functions.

### 2.3.2 The Bivariate Group Maximum Approach

In our situation where information is available for several continuous variables, techniques are required to characterize the behaviour of these variables and their relationship to one another. Let $(X_1, Y_1), (X_2, Y_2)\ldots$ be a sequence of vectors that are independent observations of a random vector having distribution function $F(x, y)$. Recall that classic univariate Extreme Value Theory is based on blocking the data and extracting group maxima; this idea can be transferred to the bivariate case by setting:

$$M_{x,n} = \max_{i=1,\ldots,n}\{X_i\} \quad \text{and} \quad M_{y,n} = \max_{i=1,\ldots,n}\{Y_i\} \qquad (2.3.3)$$

$$\mathbf{M_n} = (M_{x,n}, M_{y,n}) \qquad (2.3.4)$$

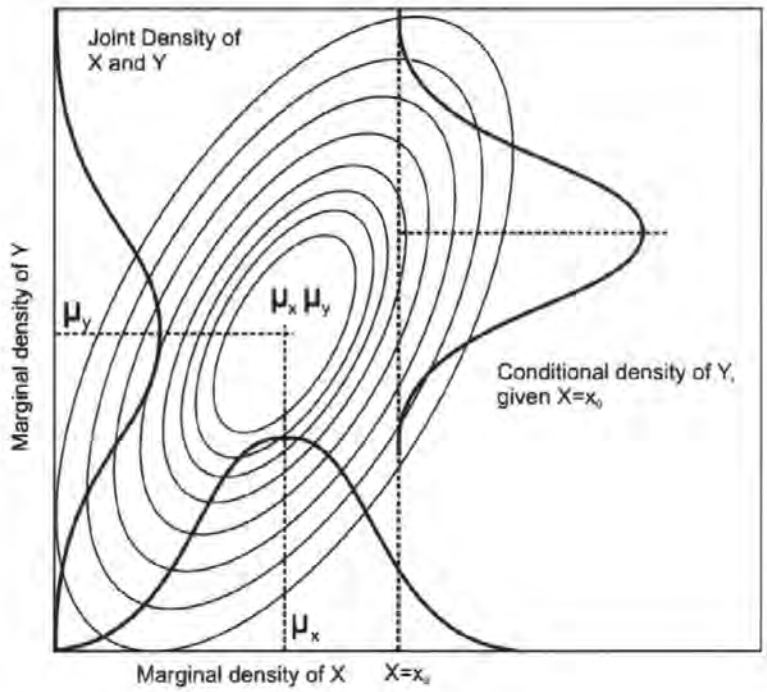



Figure 2.6: Diagram illustrating the relationship between the marginal, conditional and joint probability density functions, taken from Annis (2006). The contours indicate the joint density of continuous random variables $X$ and $Y$, denoted $f_{X,Y}$.

where $\mathbf{M_n}$ is a vector of componentwise maxima and the index $i$ where the maximum of the $X_i$ sequence occurs is not necessarily the same as the index of the maximum of the $Y_i$ sequence. In order to build the multivariate theory, first consider each variable separately. Let us assume that both $X_i$ and $Y_i$ follow the standard Fréchet distribution, with distribution function

$$F(z) = \exp(-1/z), \quad z > 0. \tag{2.3.5}$$

Using **Theorem 2.1.1.3**, this is a special case of the GEV distribution with parameters $\mu = 0$, $\sigma = 1$ and $\xi = 1$, after transformation by addition of 1. Now if $X_1, X_2, \ldots$ is a sequence of independent standard Fréchet variables, and if $a_n = n$ and $b_n = 0$, we have that

$$\begin{aligned} \Pr\{(M_{x,n} - b_n)/a_n \leq z\} &= F^n(nz) \\ &= |\exp\{-1/(nz)\}|^n \\ &= \exp(-1/z) \end{aligned} \tag{2.3.6}$$

for all $n$, for each fixed $z > 0$. A similar result would hold for $Y_1, Y_2, \ldots$ and $M_{y,n}$. Hence the distribution of each suitably scaled maximum of Fréchet random variables. Hence in the multivariate case, if we want to obtain standard univariate results for each margin, the $M_n$ must be re-scaled to

$$\mathbf{M}_\mathbf{n}^\star = (M_{x,n}^\star, M_{y,n}^\star) = \left( \max_{i=1,\ldots,n} \{X_i\}/n, \max_{i=1,\ldots,n} \{Y_i\}/n \right) \tag{2.3.7}$$

The following theorem presents the characterization of the limiting joint distribution of $\mathbf{M}_\mathbf{n}^\star$, as $n \to \infty$, providing a bivariate interpretation of the Extremal Types Theorem (see **Theorem 2.1.1.3** and Coles (2001))

**Theorem 2.3.2.1.** *Let $\mathbf{M}_\mathbf{n}^\star = (M_{x,n}^\star, M_{y,n}^\star)$ be defined by (2.3.7), where the $(X_i, Y_i)$ are independent vectors with standard Fréchet marginal distributions. Then if*

$$Pr\{M_{x,n}^\star \leq x, M_{y,n}^\star \leq y\} \xrightarrow{d} G(x,y), \tag{2.3.8}$$

*where $G$ is a non-degenerate distribution function, then $G$ takes the form*

$$G(x,y) = \exp\{-V(x,y)\}, \quad x > 0,\ y > 0 \tag{2.3.9}$$

*where*

$$V(x,y) = 2\int_0^1 \max\left(\frac{w}{x}, \frac{1-w}{y}\right) dH(w) \tag{2.3.10}$$

*and $H$ is a distribution function on $[0,1]$ satisfying the mean constraint*

$$\int_0^1 w dH(w) = 1/2 \tag{2.3.11}$$

From this theorem, the family of distributions obtained as the limit of equation (2.3.8) is known as the class of bivariate extreme value distributions. This class is in direct correspondence with the set of distribution functions $H$ on $[0,1]$. If $H$ is differentiable with density $h$ satisfying $\int_0^1 wh(w)dw = 1/2$, the integral for $V(x,y)$ becomes

$$V(x,y) = 2\int_0^1 \max\left(\frac{w}{x}, \frac{1-w}{y}\right) h(w)dw.$$

Bivariate extreme value distributions are also created using measures $H$ that are not differentiable. As an example – see also Coles (2001), pages $146 - 147$ – if $H$ is a measure that allocates mass 0.5 on $w = 0$ and $w = 1$, then equation (2.3.11) is satisfied, and it can be shown that

$$V(x, y) = x^{-1} + y^{-1}$$

using equation (2.3.10). The corresponding bivariate extreme value distribution is

$$G(x, y) = \exp\{-(x^{-1} + y^{-1})\}, \quad x > 0, y > 0. \tag{2.3.12}$$

This function can be factorized across $x$ and $y$, and hence corresponds to independent variables. If $H$ is a measure that places all its mass (i.e. unity) on $w = 0.5$, then following the same procedure we obtain a different bivariate extreme value distribution

$$G(x, y) = \exp\{-\max(x^{-1}, y^{-1})\}, \quad x > 0, y > 0. \tag{2.3.13}$$

This is the distribution function of variables which are marginally standard Fréchet, but which are perfectly dependent: $X = Y$ with probability 1. It is possible to obtain the complete class of bivariate limits using the Generalized Extreme Value distribution (GEV). This can be achieved by letting

$$\tilde{x} = \left[1 + \xi_x \left(\frac{x - \mu_x}{\sigma_x}\right)\right]^{1/\xi_x} \quad \text{and} \quad \tilde{y} = \left[1 + \xi_y \left(\frac{y - \mu_y}{\sigma_y}\right)\right]^{1/\xi_y}.$$

Hence, the complete family of bivariate extreme value distributions, with GEV margins with parameters $(\mu_x, \sigma_x, \xi_x)$ and $(\mu_y, \sigma_y, \xi_y)$, has distribution function of the form

$$G(x, y) = \exp\{-V(\tilde{x}, \tilde{y})\}, \tag{2.3.14}$$

provided that $1 + \xi_x(x - \mu_x)/\sigma_x > 0$ and $1 + \xi_y(y - \mu_y)/\sigma_y > 0$, if the function $V$ satisfies Equation (2.3.10) for some $H$ that satisfies the mean constraint (2.3.11).

We now present a sketch justification of Theorem 2.3.2.1. From (2.3.10) we can see that

for any given constant $a > 0$ the following holds:

$$V(a^{-1}x, a^{-1}y) = aV(x, y);$$

we say that $V$ is homogeneous of order $-1$. It then follows from equation (2.3.9) that

$$G^n(x, y) = G(n^{-1}x, n^{-1}y) \qquad \text{for } n = 2, 3, \ldots. \tag{2.3.15}$$

This means that if $(X, Y)$ has $G$ as its distribution function, then after re-scaling by $n^{-1}$, $\mathbf{M_n^*}$ will also have distribution function $G$. Hence, $G$ now has an equivalent multivariate version of the max stability property given in Definition 2.1.1.5. This property forms the proof of Theorem 2.3.2.1, using the argument that limit distributions in equation (2.3.8) must have the property of max stability. Using equation (2.3.15) it is possible to show that distributions of the same type as equation (2.3.9) have the max stability property of max stability and are also the only distributions with this property, conditional on the marginal specification.

From Theorem 2.3.2.1 we have a complete characterization of bivariate limit distributions. However, we have a very wide class of possible limits which are only constrained by equations (2.3.10) and 2.3.11, causing the limit family to have no general finite parameterization. This problem can be overcome using parametric sub-families of distributions for $G$; hence we work with a small subset of the complete class of limit distributions $G$.

Parametric families for $H$ on $[0, 1]$ are required to have mean equal to 0.5 for every value of their defining parameters. When substituted into equations (2.3.10) and (2.3.9) , the corresponding family for $G$ is obtained. There are several possible families for $G$ that can be considered; see Kotz and Nadarajah (2002), Tawn and Coles (1994), and Joe (1997).

### 2.3.3 Bivariate Distribution Functions (bdf)

*Logistic*

In the special case when $H$ is a symmetric logistic distribution we have the following bdf (see Tawn (1988) and Kotz and Nadarajah (2002)):

$$G(x,y) = \exp\left\{-\left(x^{-1/\alpha} + y^{-1/\alpha}\right)^{\alpha}\right\}, \quad x > 0, y > 0, \tag{2.3.16}$$

for a parameter $\alpha \in (0,1)$. This distribution function is derived from (2.3.10) by letting the density function be

$$h(w) = \frac{1}{2}(\alpha^{-1} - 1)\{w(1-w)\}^{-1-1/\alpha}\{w^{-1/\alpha} + (1-w)^{-1/\alpha}\}^{\alpha-2} \tag{2.3.17}$$

on $0 < w < 1$.

*Asymmetric Logistic*

Alternatively, if $H$ is an asymmetric logistic distribution we have the following bdf (see Tawn (1988) and Kotz and Nadarajah (2002)):

$$G(x,y) = \exp\left[-\frac{1-\psi_x}{x} - \frac{1-\psi_y}{y} - \left\{\left(\frac{\psi_x}{x}\right)^{1/\alpha} + \left(\frac{\psi_y}{y}\right)^{1/\alpha}\right\}^{\alpha}\right] \tag{2.3.18}$$

where $0 \leq \psi_x, \psi_y \leq 1$. Similarly to the symmetric Logistic bdf, we require the following density function to be used in (2.3.10) to obtain (2.3.18),

$$h(w) = (\alpha^{-1} - 1)\psi_x^{1/\alpha}\psi_y^{1/\alpha}\{w(1-w)\}^{1/\alpha-2}\{\psi_y w^{1/\alpha} + \psi_x(1-w)^{1/\alpha}\}^{\alpha-2} \tag{2.3.19}$$

*Negative Logistic*

The negative asymmetric logistic distribution has the following bdf (see Kotz and Nadarajah (2002)):

$$G(x,y) = \exp\left[-\frac{1}{x} - \frac{1}{y} + \left\{\left(\frac{\psi_x}{x}\right)^{\alpha} + \left(\frac{\psi_y}{y}\right)^{\alpha}\right\}^{1/\alpha}\right] \tag{2.3.20}$$

where $0 \leq \psi_x, \psi_y \leq 1$. Similarly to the asymmetric logistic bdf, we require the following

density function to be used in (2.3.10) to obtain (2.3.20):

$$h(w) = (1-\alpha^{-1})\psi_x^{1/\alpha}\psi_y^{1/\alpha}\{w(1-w)\}^{1/\alpha-2}\{\psi_y w^{1/\alpha} + \psi_x(1-w)^{1/\alpha}\}^{\alpha-2} \tag{2.3.21}$$

*Bilogistic*

The bilogistic distribution function with parameters $\alpha$ and $\beta$ is (see Coles (2001) and Tawn and Coles (1994))

$$G(x,y) = \exp\left\{-\left(\frac{q}{x}\right)^{1-\alpha} - \left(\frac{1-q}{y}\right)^{1-\beta}\right\} \tag{2.3.22}$$

which is obtained by inserting the following density function into (2.3.10)

$$h(w) = \frac{1}{2}(1-\alpha)(1-w)^{-1}w^{-2}(1-q)q^{1-\alpha}\{\alpha(1-q)+\beta q\}^{-1} \tag{2.3.23}$$

on $0 < w < 1$, where $0 < \alpha < 1$ and $0 < \beta < 1$ and $q = q(w;\alpha,\beta)$ is the root of

$$(1-\alpha)(1-w)(1-q)^{\beta} - (1-\beta)wq^{\alpha} = 0$$

*Negative Bilogistic*

The negative bilogistic distribution function is as follows (see Kotz and Nadarajah (2002) and Tawn and Coles (1994)),

$$G(x,y) = \exp\left\{-\frac{1}{x} - \frac{1}{y} + \left(\frac{q}{x}\right)^{1+\alpha} + \left(\frac{1-q}{y}\right)^{1+\beta}\right\} \tag{2.3.24}$$

This is obtained by inserting the following density function into (2.3.10),

$$h(w) = -\frac{1}{2}(1-\alpha)(1-w)^{-1}w^{-2}(1-q)q^{1-\alpha}\{\alpha(1-q)+\beta q\}^{-1} \tag{2.3.25}$$

on $0 < w < 1$, where $\alpha > 0$ and $\beta > 0$ and $q = q(w;\alpha,\beta)$ is the root of

$$(1+\alpha)(1-w)q^{\alpha} - (1+\beta)w(1-q)^{\beta} = 0$$

*Dirichlet (Coles-Tawn)*

The Dirichlet or Coles-Tawn distribution function takes the following form (see Tawn and Coles (1994) and Coles and Tawn (1991))

$$G(x,y) = \exp\left\{-\frac{1}{x}\left[1 - \mathrm{Be}(q;\alpha+1,\beta)\right] - \frac{1}{y}\mathrm{Be}(q;\alpha,\beta+1)\right\}, \tag{2.3.26}$$

which is obtained by inserting the following density function into (2.3.10)

$$h(w) = \frac{\alpha\beta\Gamma(\alpha+\beta+1)(\alpha w)^{\alpha-1}(\beta(1-w))^{\beta-1}}{2\Gamma(\alpha)\Gamma(\beta)(\alpha w+\beta(1-w))^{\alpha+\beta+1}} \tag{2.3.27}$$

on $0 < w < 1$, where $\alpha > 0$ and $\beta > 0$. Here Be and $\Gamma$ are the usual beta and gamma functions respectively.

### 2.3.4 Modelling the Group Maximum Approach

Coles (2001) describes when modelling data from the Group Maximum approach we adopt Theorem 2.3.2.1 as a basis to work from. Consider the sequence of componentwise block maxima $(z_{1,1}, z_{2,1}), \ldots, (z_{1,m}, z_{2,m})$ created from the original series $(x_1, y_1), \ldots, (x_n, y_n)$ of independent data vectors by blocking into $m$ blocks. Assume the block maxima can be marginally modelled using the GEV distribution; more specifically for each $j$, $z_{i,j}$ is considered an independent realization of a random variable $Z_i$, for $i = 1, 2$, following a GEV:

$$Z_i \sim \mathrm{GEV}(\mu_i, \sigma_i, \xi_i).$$

Now to obtain estimates $(\hat{\mu}_i, \hat{\sigma}_i, \hat{\xi}_i)$ we apply maximum likelihood estimation to the separate series. Hence, we may transform the variables into

$$\tilde{Z}_i = \left[1 + \hat{\xi}_i\left(\frac{Z_i - \hat{\mu}_i}{\hat{\sigma}_i}\right)\right]^{1/\hat{\xi}_i}, \tag{2.3.28}$$

which are approximately distributed according to a standard Fréchet distribution. Substitution of observations $(z_{1,j}, z_{2,j})$ into equation (2.3.28) returns $(\tilde{z}_{1,j}, \tilde{z}_{2,j})$ which is a sequence of independent realizations of a vector with bivariate extreme value distribution $G$ given

by (2.3.9).By differentiation, the probability density function is

$$g(x,y) = \{V_x(x,y)V_y(x,y) - V_{x,y}(x,y)\}\exp\{-V(x,y)\}, \quad x>0, y>0,$$

where $V_x, V_y$ and $V_{x,y}$ denote the partial and mixed derivatives of $V$ respectively. This leads to the likelihood

$$L(\theta) = \prod_{i=1}^{m} g(\tilde{z}_{1,j}, \tilde{z}_{2,j}), \tag{2.3.29}$$

with the corresponding log-likelihood

$$\ell(\theta) = \sum_{i=1}^{m} \log g(\tilde{z}_{1,j}, \tilde{z}_{2,j}). \tag{2.3.30}$$

where $\theta$ represents the parameters of the adopted model for $G$ or $g$, as discussed in Section 2.3.3, and of the marginal GEV parameters in (2.3.4). In this way we combine the above transformation and model fitting steps.

### 2.3.5 Bivariate Threshold Excess Model

When discussing univariate extreme value theory, we highlighted the disadvantage of the group maximum approach in comparison to the 'Threshold Excess Approach'. We now look to extend the univariate 'Threshold Excess Approach' to the bivariate case. The bivariate theory presented in this section can be used in future applied work. The class of approximations to the tail of univariate distribution function $F$ is described by the following family that derives from Theorem 2.2.1.1 via equation (2.3.23)

$$G(x) = 1 - \zeta\left\{1 + \frac{\xi(x-u)}{\sigma}\right\}^{-1/\xi}, \quad x>u.$$

Based on this equation there are parameters $\zeta, \xi$ and $\sigma$ that, for a sufficiently high threshold $u$, imply that $F(x) \approx G(x)$ for $x > u$. Using Equation (2.3.31), it is possible to create a bivariate equivalent, which will give an approximation to the arbitrary joint distribution $F(x,y)$ on the regions $x > u_x, y > u_y$, for large enough thresholds $u_x$ and $u_y$.

Let $(x_1,y_1),\ldots,(x_n,y_n)$ be independent realizations of the random variable $(X,Y)$ with

joint distribution function $F$. From (2.3.23) each random variable has an associated threshold and marginal distribution defined by a set of parameters. For example the random variable $X$ will have threshold $u_x$ and marginal distribution of the form given equation (2.3.31) with parameter set $(\zeta_x, \sigma_x, \xi_x)$.

The following transformations create variables $(\tilde{X}, \tilde{Y})$ which have distribution functions which have standard Fréchet margins for $X > u_x$ and $Y > u_y$:

$$\tilde{X} = -\left(\log\left\{1 - \zeta_x\left[1 + \frac{\xi_x(X - u_x)}{\sigma_x}\right]^{-1/\xi_x}\right\}\right)^{-1} \tag{2.3.31}$$

$$\tilde{Y} = -\left(\log\left\{1 - \zeta_y\left[1 + \frac{\xi_y(Y - u_y)}{\sigma_y}\right]^{-1/\xi_y}\right\}\right)^{-1} \tag{2.3.32}$$

Now using Theorem 2.3.2.1 and assuming $n$ is large, we have

$$\begin{aligned}
\tilde{F}(\tilde{x}, \tilde{y}) &= \left\{\tilde{F}^n(\tilde{x}, \tilde{y})\right\}^{1/n} \\
&\approx [\exp\{-V(\tilde{x}/n, \tilde{y}/n)\}]^{1/n} \\
&= \exp\{-V(\tilde{x}, \tilde{y})\}
\end{aligned}$$

where $\tilde{F}$ is the joint distribution function of $\tilde{X}$ and $\tilde{Y}$; the last equality is due to the homogeneity property of $V$. Therefore, since $F(x, y) = \tilde{F}(\tilde{x}, \tilde{y})$, the following holds

$$F(x, y) \approx G(x, y) = \exp\{-V(\tilde{x}, \tilde{y})\}, \quad x > u_x, y > u_y \tag{2.3.33}$$

so that the bivariate distribution function $G(x, y) = \exp\{-V(\tilde{x}, \tilde{y})\}$ is an approximation to the bivariate tail. Making inference about this model is more complicated than for the group maximum case, as the thresholds define four regions. These regions are the following:

| | |
|---|---|
| **no excess** | $R_{0,0} = (-\infty, u_x) \times (-\infty, u_y)$ |
| **excess in $X$** | $R_{1,0} = [u_x, \infty) \times (-\infty, u_y)$ |
| **excess in $Y$** | $R_{0,1} = (-\infty, u_x) \times [u_y, \infty)$ |
| **excess in both $X$ and $Y$** | $R_{1,1} = [u_x, \infty) \times [u_y, \infty)$ |

If the data point lies in $R_{1,1}$ then the model defined in equation (2.3.33) holds. For all the other regions, $\tilde{F}$ is not defined and so the likelihood must be altered to allow for this.

The final likelihood function is defined as

$$L(\theta;(x_1,y_1),\ldots,(x_n,y_n)) = \prod_{i=1}^{n} \psi(\theta;(x_i,y_i)), \tag{2.3.34}$$

where $\theta$ is the vector of parameters of $F$ and

$$\psi(\theta;(x,y)) = \begin{cases} \frac{\partial^2 F}{\partial x \partial y}\,|_{(x,y)} & \text{if } (x,y) \in R_{1,1} \\ \frac{\partial F}{\partial x}\,|_{(x,u_y)} & \text{if } (x,y) \in R_{1,0} \\ \frac{\partial F}{\partial y}\,|_{(u_x,y)} & \text{if } (x,y) \in R_{0,1} \\ F(u_x,u_y) & \text{if } (x,y) \in R_{0,0} \end{cases}$$

All the terms within the likelihood are derived from the joint tail approximation, given in equation (2.3.33).It is for this reason that $\psi(\theta;(x,y)) = F(u_x,u_y)$ if $(x,y) \in R_{0,0}$ for example: since $f(x,y)$ is not known for $(x,y) \in R_{0,0}$, the contribution to the likelihood is replaced by the probability $\Pr(X < u_x, Y < u_y) = F(u_x,u_y)$, the form of which is known (by continuity of $G$).

The standard dependence modelling techniques, for bivariate extreme values, which have been reviewed in this chapter, so far, have all relied on the assumption of max-stability. This type of modelling can often be insufficient, due to the lack of flexibility in the models,when looking at weaker forms of dependence or near independence. Ledford and Tawn (1996) present an alternate technique to characterize the joint tail region. The development was based on a simple bivariate case and introduced a coefficient of tail dependence parameter. They were able to show that the new coefficient dictated the dependence structure and could be manipulated to encompass a range of dependence structures from bivariate distributions. Ledford and Tawn (1997) then used this coefficient to develop a specific joint probability model structure, establishing the coefficient as a driving factor in dependence modelling.

In an impressive recent paper, Ramos and Ledford (2009) address some of the limitations of previous work by Ledford and Tawn (1996) and Ledford and Tawn (1997). In particular they propose a modelling framework based on a specially developed limit distribution in place of the methodology developed for specific examples in Ledford and Tawn (1997).

### 2.3.6 Applied Example of a Bivariate Threshold Excess Approach

In Section 2.2.7, we saw an example of a univariate threshold excess model. We now present a bivariate equivalent using the same data set and selecting the two variables wave height $H_s$ and wave period $T_z$ to be modelled. The dependence between the variables can be modelled using a range of different dependence models discussed in Section 2.3.2. For this example we will be using the Logistic model, although this may not be the most appropriate choice of dependence model. The process uses (2.3.16) to produce a maximum likelihood estimate of the dependence parameter $\alpha$. When deciding on the most suitable choice of dependence

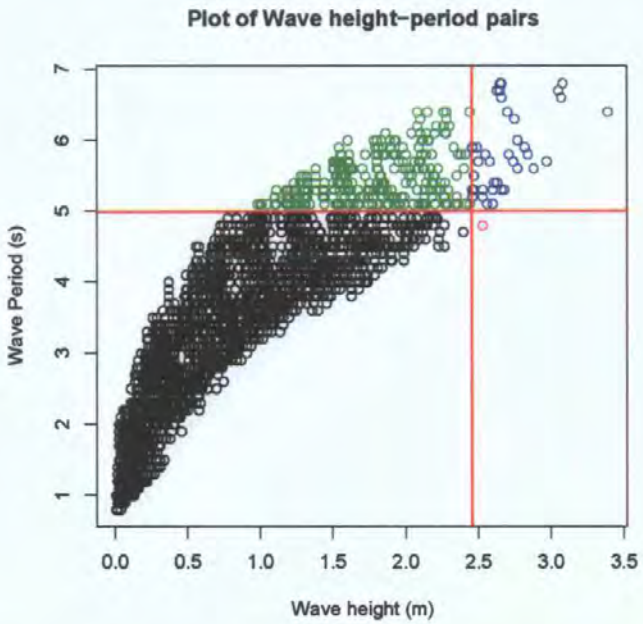



Figure 2.7: Scatter plot of wave period against wave height values. The horizontal and vertical lines indicate the thresholds defining the regions of excess or no excess.

model we have several different methods available to us. The first would be the negative log-likelihood (NLLH) as discussed in Tawn and Coles (1994), which for fixed thresholds chooses the model with the smallest NLLH as the most appropriate; we adopt this approach in this project. An alternate method for dependence model choice would be investigation of AIC (Akaike's Information Criterion) values. AIC compares the goodness-of-fit of several competing models and ranks them according to their AIC, the one having the lowest AIC being the best. The general case of AIC is described as follows. If we have some statistical

model with parameters determined by the method of maximum likelihood, AIC was then defined by

$$AIC = 2k - 2\ln(L) \tag{2.3.35}$$

where $k$ is the number of parameters in our statistical model, and $L$ is the maximized value of the likelihood function for the estimated model (see Akaike (1974) for further details).

We begin by looking at Figure 2.7 which shows wave period plotted against wave height. Sufficiently large marginal thresholds $u_x$ and $u_y$ are also shown. The determination of threshold values in the univariate case was discussed in Section 2.2.2. Using the thresholds $u_x = 2.45$ and $u_y = 5$ for the wave height and wave period marginals respectively, maximization of the likelihood (2.3.34) gave the estimate $\hat{\alpha} = 0.711$ for the model defined in (2.3.16) with a standard error of 0.011. This indicates a model with a reasonably weak level of dependence, but is however significantly different from independence. Further investigations into the effect of using different dependence functions are discussed in Tawn and Coles (1994) and Joe (1997).

# 3

# JOINSEA: The Joint Probability of Waves and Water Levels

In this chapter we review the JOINSEA software, discussing the methodology used and its implementation into the software. We review this software here as in Chapter 6 we introduce a new methodology to improve the modelling of extreme values and provide a comparison to the techniques used in JOINSEA.

## 3.1 What is JOINSEA?

HR Wallingford and Lancaster University were jointly commissioned by D.E.F.R.A (Department for Environment, Food and Rural Affairs, formerly MAFF) to research joint probability techniques for use in coastal defence strategies. The aim of the research was to produce a reliable technique and consequently software to improve the design of coastal defence structures. The resulting reports present a new approach for joint probability modelling of large wave heights and high water levels (see Wallingford, 1998a,b). The methods that were available before JOINSEA's development had limitations which seriously affected the accuracy of their results. These included:

- The empirical estimates for quantifying dependence between variables, in combination with statistical estimates for the distributions of individual variables, were generally considered unreliable at larger values as a means of joint probabilty modeling.

- The assumption that wave period is given exactly by significant wave height and the estimated constant wave steepness was considered invalid.
- The assumption that the estimate of the probability of failure is based on a subset of the true failure region was considered to reduce the accuracy of the method.

The JOINSEA software was designed to overcome these limitations. It was written in the FORTRAN 90 computer programming language which is used in many engineering communities for developing computer code. The software was written as five interlinked programs:

- **BVN**(Bivariate Normal Distribution)
- **MIX**(A mixture of two Bivariate Normal Distributions)
- **SIMBVN**(Simulation of realizations from a Bivariate Normal Distribution)
- **SIMMIX**(Simulation of realizations from a mixture of two Bivariate Normal Distributions)
- **ANALYSIS**

These programs will be discussed in detail in Section 3.3. The program layout is given in Figure 3.1. The programs must be run in sequence as the output from a previous program is the input to the next. The first program is either BVN or MIX with the choice of this being dependent on the data; usually BVN is fitted to data where the extreme wave conditions at the location all come from a single population, and MIX is used when the wave conditions are from two sources (e.g. swell waves as well as wind waves).

The techniques employed in the JOINSEA software were developed based on the assumption that the data would be well estimated by the distributions used in this modelling. However it should be noted that this is a limitation to this methodology, as we believe a large amount of data is required for these assumptions to hold. Furthermore, we note that JOINSEA specifies the thresholds in the marginal extreme model as the 95% quantile and so makes the assumption that the model is relatively insensitive to threshold choice. However, we believe threshold choice can influence model goodness-of-fit and we will discuss this in detail in Chapter 6.

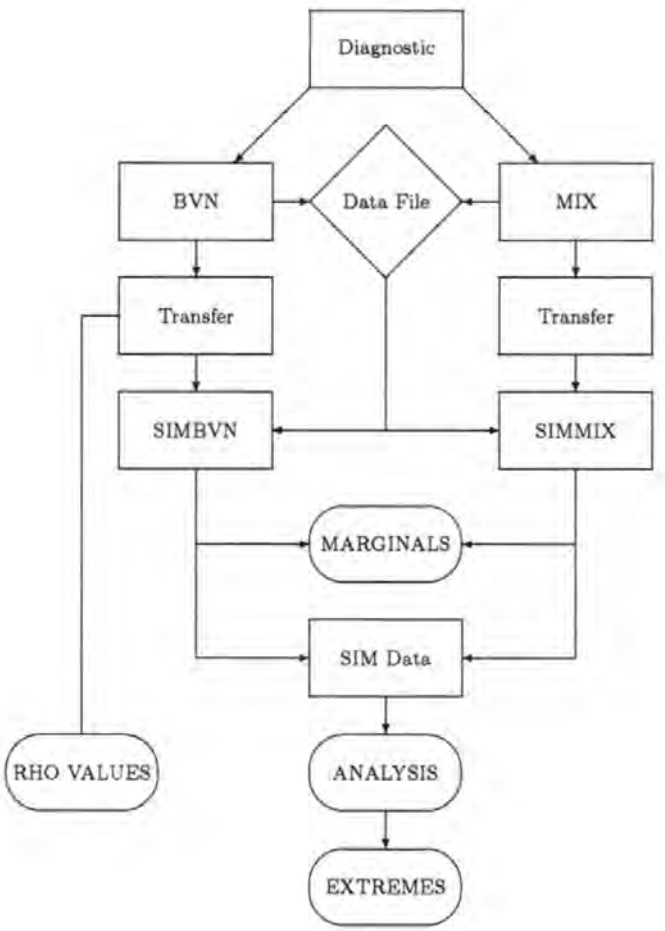



Figure 3.1: Flow diagram showing the JOINSEA program structure, taken from Wallingford (1998a,b).

## 3.2 Joint Probability in JOINSEA

JOINSEA utilizes a joint probability technique to obtain an estimate of the probability that a structure variable would exceed a specified critical level so resulting in the failure of the engineering structure. This is highly dependent on the joint analysis of the bivariate sea condition variables $\boldsymbol{X}$ (wave height, wave period) from which the distribution of the structure variable is established.

In engineering terms, the structure variable is typically a variable which characterizes the behaviour of the structure based on the effect of specific forces. An example of this type of variable would be crest level which is the height of a sea defence. If crest level is poorly specified as a result of underestimation of extreme sea conditions, it is possible that overtopping of coastal defences can occur causing flooding of flood risk areas. Crest level is not, however, the sole factor determining overtopping.

Failure of the coastal defence occurs when the structure variable $\triangle(X)$ is greater than a critical level $u$: $\triangle(X) > u$. The extreme values of $X$ are a set $A_u$, as follows:

$$A_u = \{x : \triangle(x) \geq u\}. \tag{3.2.1}$$

The estimated joint distribution of sea condition variables can be used as a preliminary design tool for coastal defences; it uses the joint probability of specified sea conditions to calculate the probability of failure of a particular design variable.

The key aim is to find an appropriate model that fits the data from the sea condition variable and that can then be used to simulate future conditions, which the defence can then be designed to withstand; in this way the sustainability and accuracy of the design specification is increased.

The joint probability technique adopted in JOINSEA is summarized as follows:

1. An estimate of the joint density of the sea conditions is calculated, using the following approaches:
   - Estimate distributions for each separate variable,
   - Estimate the dependence between the separate variables.

2. An estimate of the probability of failure can be calculated by integration of this estimated joint density over the failure region of the sea conditions given by the set $A_u$.

3. The final stage is a simple conversion of time scales. The time scale is updated from that of the observations to annual time unit.

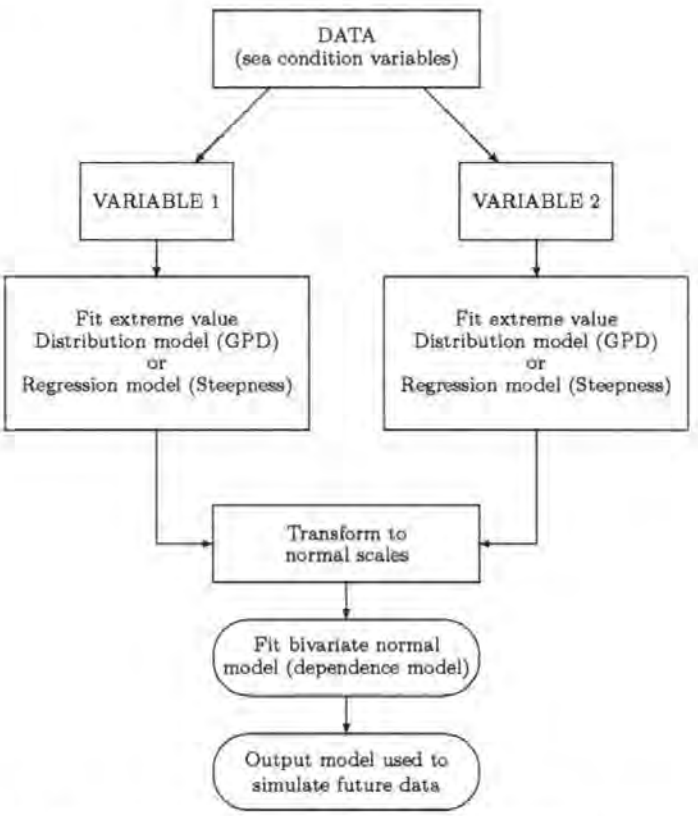



Figure 3.2: Flow diagram showing the BVN program procedure, taken from Wallingford (1998a,b)

## 3.3 Program Structure

We now discuss the individual elements of the JOINSEA program, as shown in Figure 3.2.

### 3.3.1 Bivariate Normal Distribution Program (BVN)

This program's main function is to assess the upper joint tail of the distribution of the variables of interest and then fit to a Bivariate Normal Distribution. This is performed in

two stages: first, separate Generalized Pareto distribution (GPD) models are fitted to the top 5% of each marginal. This allows each tail to be transformed to normality. A bivariate normal model is fitted to the transformed data.

The program selects extreme values by setting a threshold for exceedance, hence determining those values to be modelled. A numerical study using several data sets suggested that extremes predictions were relatively insensitive to the threshold chosen and that the 95% quantile was a reasonable value for the threshold, meaning that the GPD is fitted to the top 5% of observed values of the variable. The choice of threshold for GPD modelling was discussed in greater detail in Chapter 2.

We can think of the above procedure of defining thresholds and then transforming marginal tails to normality so that a bivariate normal distribution can be fitted as a bivariate normal threshold modelling procedure.

To describe the this procedure it is useful to introduce the Multivariate Normal Distribution and build from this idea. Firstly, we denote $\mathbf{X} = (X_1, \ldots, X_k)^T$ as a random variable which follows a multivariate normal distribution with normal marginal distributions $X_i \sim N(\mu_i, \sigma_i^2)$. Let the mean vector $\mu = (\mu_1, \ldots, \mu_k)^T$. Then $\mathbf{X}$ has joint density function

$$f_{\mathbf{x}}(\mathbf{x}) = \phi(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}, \quad \mathbf{x} \in \mathbb{R}^k \qquad (3.3.1)$$

where $|\Sigma|$ is the determinant of the variance covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \ldots & \ldots & \sigma_1\sigma_k\rho_{1k} \\ \ldots & \ldots & \sigma_i\sigma_j\rho_{ij} & \ldots \\ \ldots & \sigma_j\sigma_i\rho_{ji} & \ldots & \ldots \\ \sigma_k\sigma_1\rho_{k1} & \ldots & \ldots & \sigma_k^2 \end{pmatrix} \qquad (3.3.2)$$

in which $\rho_{ij}$ is the correlation of $X_i$ and $X_j$, which is related to the covariance $\sigma_{ij}$ between variables $X_i$ and $X_j$ as follows:

$$\rho_{ij} = \mathrm{Corr}(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{(\mathrm{Var}(X_i)\mathrm{Var}(X_j))^{1/2}} = \frac{\sigma_{ij}}{\sigma_i\sigma_j}. \qquad (3.3.3)$$

It follows from matrix algebra that the joint density for the bivariate case can be written as

follows

$$f_{\mathbf{x}}(x_1, x_2) = \phi(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right] \quad (3.3.4)$$

For the JOINSEA approach it is necessary to transform each GPD marginal to follow a standard normal distribution ($\mu_1 = \mu_2 = 0, \sigma_1 = \sigma_2 = 1$), so that the dependence model fitted may be assumed to be a bivariate normal distribution with standard normal margins. The joint distribution of the bivariate normal random variable is denoted

$$\Phi(x_1^\star, x_2^\star) = \Pr(X_1^\star \leq x_1^\star, X_2^\star \leq x_1^\star) \quad (3.3.5)$$

where $(X_1^\star, X_2^\star)$ denotes the original input variables after transformation to normality. Assuming that the original marginals $X_1$ and $X_2$ are fitted to a GPD above thresholds $u_1$ and $u_2$, it follows that from (2.2.23) that $F_{X_i}(x_i) = 1 - \zeta_{u_i}\{1 + \xi_i(x_i - u_i)/\sigma_i\}_+^{-1/\xi_i}$, for$x_i > u_i$ where $\zeta_{u_i} = \Pr(X_i > u_i)$ and $\sigma_i > 0$. We transform the excesses $x_i > u_i$ to standard normality using the probability integral transform

$$X_i^\star = \Phi^{-1}(F_{X_i}(X_i)), \quad \text{for } i = 1, 2, \quad (3.3.6)$$

with transformed thresholds

$$u_i^\star = \Phi^{-1}(F_{X_i}(u_i)), \quad \text{for } i = 1, 2, \quad (3.3.7)$$

where here $\Phi$ is the cumulative distribution function of the standard normal. Fitting the dependence model is possible once the variables are in the correct form, using maximum likelihood estimation. The likelihood contribution for the observation $(x_1^\star, x_2^\star)$ is

$$\frac{\partial\Phi}{\partial x_1^\star \partial x_2^\star}(x_1^\star, x_2^\star) = \phi(x_1^\star, x_2^\star) \quad (3.3.8)$$

where $\phi(x_1^\star, x_2^\star)$ is the joint density function of the bivariate normal distribution with standard

normal marginals.

As mentioned previously, the GPD is fitted marginally to the data from each variable using maximum likelihood estimation. As this procedure is applied to both variables, two thresholds are required. These define four distinct regions for all the data. The diagram in Figure 7.16 shows these regions for a particular choice of $u_1$ and $u_2$. A data point $(x_1, x_2)$

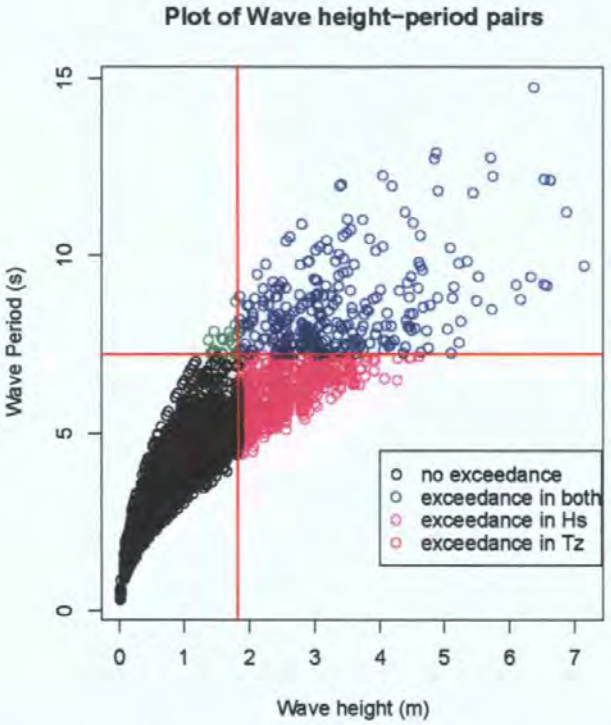



Figure 3.3: Graph shows Wave height against wave period with associated thresholds for each marginal. This defines four distinct regions for all the data.

belongs to one of the four regions. Thus a similar situation to that discussed in Section 2.3.5 results, with the likelihood contribution from point $(x_1, x_2)$ depending on the region to which $(x_1, x_2)$ belongs. The specific likelihood contributions for each region are as follows:

- $x_1 \leq u_1$ and $x_2 \leq u_2$: likelihood contribution:

$$\Phi(u_1^\star, u_2^\star)$$

- $x_1 > u_1$ and $x_2 \leq u_2$: likelihood contribution:

$$\frac{\partial \Phi}{\partial x_1^\star}(x_1^\star, u_2^\star)\frac{dx_1^\star}{dx_1}$$

- $x_1 \leq u_1$ and $x_2 > u_2$: likelihood contribution:

$$\frac{\partial \Phi}{\partial x_2^\star}(u_1^\star, x_2^\star)\frac{dx_2^\star}{dx_2}$$

- $x_1 > u_1$ and $x_2 > u_2$: likelihood contribution:

$$\phi(x_1^\star, x_2^\star)\frac{dx_1^\star}{dx_1}\frac{dx_2^\star}{dx_2}$$

where

$$\frac{dx_i^\star}{dx_i} = \frac{\zeta_{u_i}}{\phi(x_i^\star)\sigma_i}\{1 + \xi_i(x_i - u_i)/\sigma_i\}_+^{-1-1/\xi_i}. \qquad (3.3.9)$$

following from (3.3.6), where here $\phi$ is the standard normal probability density function. The estimates of the model parameters obtained by maximizing the associated likelihood are fed directly into the corresponding simulation program which can be used to generate time series for future sea conditions. These conditions then form the basis of creating an effective and sustainable coastal defence design.

### 3.3.2 Two Bivariate Normal Distributions Program (MIX)

This program follows a similar process to the BVN, with subtle changes to account for a mixture of distributions instead of just one distribution. The main difference is in using a dependence structure based on a mixture of two bivariate normal random variables. This would be appropriate when waves conditions come from two differing populations; for example, swell generated waves entering the system and locally generated wind waves from within the system.

The MIX program allows variation in the dependence above the threshold unlike the single BVN program where dependence is assumed to be constant. This assumption of constant dependence can lead to inaccuracy when the model is used to simulate data through

extrapolation. In the programs the correlation between the variables is checked for constancy. If it is constant, then BVN is sufficient; otherwise the MIX program must be used to account for the addition populations of wave conditions.

Unlike the single BVN, the MIX program assumes standard normal marginal variables rather than those from an extreme value distribution, and also uses a dependence model that utilizes a mixture of bivariate normal random variables.

The MIX program models have several parameters to describe the differing form of dependence, and are again fitted using maximum likelihood. The parameters can be divided into three categories:

1. $p_M$ is a single parameter which describes the proportion of data related to one type of dependence: $p_M = 0$ or $p_M = 1$ would indicate that a single dependence type is present.

2. $\rho_1$ and $\rho_2$ are the correlation parameters which are associated to each dependence type.

3. There are four $\mu$ parameters which indicate the change in mean level among events generated from the differing populations once variables have been transformed to the standard normal marginal scale.

It is possible to obtain the joint distribution function and associated likelihood; see Wallingford (1998a) and Wallingford (1998b) for full details.

### 3.3.3 SIMBVN and SIMMIX Programs

Both programs take their inputs from their respective joint probability prequel programs, BVN and MIX. SIMBVN utilizes the parameter values from the BVN that was fitted to the marginals. These are used to simulate larger data sets for designated return periods, for example 50 years worth of simulated wave heights.

To begin the simulation of realizations, the diagnostic file which gives information on the degree of correlation at different thresholds is assessed. Then, if the correlations are assessed to be constant a threshold is chosen. As the inputs are focused on the use of return levels in years, another input specifying the number of events per year is required.

From resulting simulated data, the extremes can also be calculated and hence a design condition can be extracted, i.e. the worst case scenario can be identified and the design can be made appropriate to this.

### 3.3.4 Analysis Program of Joint Exceedance Extremes and Structural Response Functions

The last sub program in the JOINSEA software is called ANALYSIS. This program is not based on fitting distributions, but by using a count back approach to specify the extremes for use in the design and assessment of sea defences. The structural response functions including overtopping, run up and force, can be calculated using the generated future sea conditions. The structural response variables included in the ANALYSIS program are:

- **Overtopping rate** This is the overtopping rate on a smooth slope calculated using the method of Owen (1980a);
- **Run-up** The runup levels on a smooth slope is calculated using the formulae described in CIRIA/CUR (1991);
- **Wave force on a vertical wall** Methodology in this section of the program is based on Allsop et al. (1996) which calculates the wave forces on vertical walls;
- **Armour size** The rock armour size for a sea wall is calculated using the formulae described in CIRIA/CUR (1991).

Details of all structural response functions can be found in Reeve et al. (2004) and Sorenson (1978). The main outputs of this program are marginal extremes for wave height and water level which are used to return a tabulated summary of return levels, and the joint probability extremes which also give the joint return levels, at return periods specified by the user.

## 3.4 Modern Approaches

In Section 2.3 we discussed in detail existing methodology for modelling bivariate extremes. An impressive very recent paper by Ramos and Ledford (2009) has extended the existing treatment of multivariate extremes by developing an asymptotically motivated representation of extremal dependence that also encompasses asymptotic independence.

Ramos and Ledford (2009) construct parametric models that can accommodate asymptotic dependence, asymptotic independence and asymmetry within a straightforward parsimonious parameterization. They provide a fast simulation algorithm and detail likelihood-based inference including tests for asymptotic dependence and symmetry which are useful for submodel selection. In this way Ramos and Ledford (2009) provide significant extensions of both the theoretical and the practical tools that are available for joint tail modelling. This more recent work offers many advantages over the JOINSEA approach.

# 4

# An Overview of the Bayesian Approach, and Nonparametric and Quantile Regression

In this chapter we provide a review of the Bayesian approach to statistical inference, and of the standard techniques of nonparametric and quantile regression. We include this literature review of statistical techniques as they are the basis of the new methodologies developed in Chapter 7.

We propose to use quantile regression as an improved method of modelling wave condition data, such as the data set shown in Figure 4.2. Quantile regression has been shown to provide significant benefits in modelling data in areas such as finance or medical statistics as quantile regression curves can provide a better inferential picture from the data compared to a standard regression approach. Yu et al. (2003) highlight the potential benefits in modelling extremes values using quantile regression, showing the relation between quantiles and return levels. We will build on their suggestion to illustrate the potential benefit of quantile regression techniques in coastal engineering applications.

We begin by introducing the main concepts of the Bayesian approach and follow this with an outline of the nonparametric spline based techniques of regression modelling. We finally introduce the idea of quantile regression and give a summary of some fundamental concepts of this regression procedure. We finish the chapter by reviewing more recent work showing the developments due to combinations of the concepts presented in the chapter. We

draw particular attention to the paper by Yu and Moyeed (2001) as this provides a basis for techniques developed in Chapter 7.

## 4.1 The Bayesian Approach to Statistical Inference

Both Bayesian and non-Bayesian approaches to statistical inferences draw conclusions about model (or population) parameters from data. Both approaches are based on a similar framework of components:

- A set of data $x$.
- A set of model (or population) parameters $\beta$
- A data model $\pi(x|\beta)$

Bayesian inference differs from non-Bayesian inference as it uses Bayes theorem to obtain $\pi(\beta|x)$, the conditional probability density of the set of parameters given the data. Inference is based on this conditional probability density. In non-Bayesian inference, conclusions are based on $\pi(x|\beta)$ the conditional probability density of the data given the parameter. Bayes Theorem takes the following form:

**Theorem 4.1.0.1.** ***Bayes Theorem*** *If $A$ and $B$ are two events with $P(A)>0$. Then*

$$\begin{aligned} P(B|A) &= \frac{P(B)P(A|B)}{P(A)} \\ &\propto P(B)L(A|B) \end{aligned} \tag{4.1.1}$$

*where*

- *$P(B|A)$ is the conditional probability of $B$ given $A$, also known as the "posterior" probability of $B$ given that the event $A$ has occurred.*
- *$P(B)$ is the probability of $B$ also known as the "prior".*
- *$P(A|B)$ is the conditional probability of $A$ given $B$.*
- *$P(A)$ is the probability of $A$.*

- *$L(A|B)$ is referred to as the "likelihood" when thought of as a function of $B$.*

To transfer Bayes theorem in Bayesian inference requires specification of the "prior" density $\pi(\beta)$ of the set of model parameters $\beta$ now thought of as random variables, and of a likelihood function $\pi(x|\beta)$ for random variables rather than for events. We therefore alter Bayes Theorem to:

**Theorem 4.1.0.2.** ***Bayes Theorem (restated)***

$$\begin{aligned} \pi(\beta|x) &= \frac{\pi(\beta)\pi(x|\beta)}{\pi(x)} \\ &= \frac{\pi(\beta)\pi(x|\beta)}{\int \pi(\beta)\pi(x|\beta)d\beta} \\ &\propto \pi(\beta)\pi(x|\beta). \end{aligned} \tag{4.1.2}$$

Hence, the posterior probability density $\pi(\beta|x)$ is proportional to the prior probability density $\pi(\beta)$ multiplied by the data model $\pi(x|\beta)$, referred to as the likelihood when thought of as a function of $\beta$. Bayesian inference can depend on the choice of prior density for $\beta$ as this represents the prior belief about $\beta$ before the information in the data is introduced. From this posterior density we can also obtain posterior moments, quantiles, etc by expressing them as the posterior expectation of a function $g$ of $\beta$,

$$E[g(\beta)|x] = \frac{\int g(\beta)\pi(\beta)\pi(x|\beta)d\beta}{\int \pi(\beta)\pi(x|\beta)d\beta} \tag{4.1.3}$$

The Bayesian approach to statistical inference can therefore be summarized as the following steps:

- Specification of a data model or likelihood $\pi(x|\beta)$,
- Specification of a prior density $\pi(\beta)$,
- Calculation of the posterior density $\pi(\beta|x)$ using Bayes Theorem;
- Extracting inference about the model parameters $\beta$ from the posterior distribution.

## 4.2 MCMC: Markov chain Monte Carlo

When using Bayesian methodology in practice, it is often the case that the computation of the posterior density $\pi(\beta|x)$ given in (4.1.2) is not simple due to difficulties associated with computing the possibly multidimensional integral $\int \pi(\beta)\pi(x|\beta)d\beta$. To overcome difficulties associated with this integral we can use a numerical simulation based technique called Markov chain Monte Carlo (MCMC). We will now introduce MCMC, discussing its constituent parts separately.

### 4.2.1 Monte Carlo Integration

To simplify notation let us assume that we have a possibly multidimensional random variable $X$ distributed according to probability density function known upto a constant of proportionality. Then

$$E[g(X)] = \frac{\int g(x)\pi(x)dx}{\int \pi(x)dx}, \tag{4.2.1}$$

for some function of interest $g(X)$, in which $\pi$ is proportional to the probability density function of $X$. The purpose of Monte Carlo integration is to use realizations $X_t, t = 1, \ldots, n$, of $X$ and using these to approximate $E[g(X)]$ as

$$E[g(X)] \approx \frac{1}{n}\sum_{t=1}^{n} g(X_t) \tag{4.2.2}$$

Therefore the 'population' mean of $g(X)$ is approximated by the sample mean of $g(X_1), \ldots, g(X_n)$. When the realizations $X_t$ are independent, the accuracy of the approximation to the expectation is proportional to the sample size $n$. However the assumption that independent realizations can be drawn often does not hold when the probability density of $X$ takes a complicated form, as can occur in Bayesian modelling; see Gilks et al. (1996) and Gamerman (1997). In that case the accuracy of the approximation (4.2.2) is reduced. The reason why the realizations $X_t$, $t = 1, \ldots, n$, may not be independent is that they may have to be simulated using the MCMC class of algorithms, which, as we will see, will yield correlated realizations.

### 4.2.2 Markov Chain

Gamerman (1997) describes a Markov chain as a specific stochastic process which characterizes sequences of random variables. This process satisfies the Markov property which means that given the present state, future states are independent of past states. This can be made more precise, as Gilks et al. (1996) for example point out: if we create a sequence of random variables, $X_t, t = 1, 2, \ldots$ a "future" element $X_{t+1}$ is sampled from a density that depends on only the "present" state $X_t$. This means that if we know $X_t$, then $X_{t+1}$ is not dependent on previous elements $X_{t-1}, X_{t-2}, \ldots$ in the chain. The resulting sequence $X_t$, $t = 1, 2, \ldots$, is said to be a Markov chain. The probability density function that determines how the process moves from $X_t$ to $X_{t+1}$ is referred to as a transition kernel. It turns out that under certain regulatory conditions the distribution of these realizations $X_t$ settles down as $t \to \infty$ to what is referred to as a stationary distribution. Moreover it is possible to define the transition kernel of a Markov chain in such a way that the stationary distribution takes a given form. In the case of the Bayesian approach the stationary distribution of the Markov chain has associated probability density function $\pi(\beta|x)$ defined in (4.2.1).

We can understand the posterior density using the realizations from it eventually produced by our Markov chain. We can, for example, approximate expectations using these realizations and expression (4.2.2).

In summary, under specific regularity conditions a suitably generated Markov chain will converge to a unique stationary distribution with probability density function $\pi(\beta|x)$. In other words, if $\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(t)}, \ldots$ are realizations from an appropriate chain, then as $t \to \infty$, $\beta^{(t)}$ will have probability density function $\pi(\beta|x)$. Hence, after a suitably chosen time $B$ say, the realization $\beta^{(t)}$, $t = B+1, B+2, \ldots$ can be thought of as a dependent sample from $\pi(\beta|x)$. The realizations $\beta^{(t)}$, $t = 1, \ldots, B$, up to $B$ are said to come from the burn-in phase and are discarded. Further details of this approach can be found in Gilks et al. (1996) and Gamerman (1997).

### 4.2.3 Metropolis-Hastings Algorithm

There are two main algorithms used to define a Markov Chain with a desired stationary distribution. These are the Metropolis-Hastings (Metropolis et al. (1953) and Hastings (1970)) and Gibbs sampler (Geman and Geman (1984)). The Gibbs sampler is not appropriate for our coastal engineering application as this requires sampling from the full condtional density which in are case is difficult, hence we only discuss the Metropolis-Hastings algorithm. We begin with an initial value $\beta^{(0)}$. The following steps define the transition kernel from $\beta_t$ to $\beta_{t+1}$:

- Sample a candidate point $\beta^*$ from a proposal density $q(\beta|\beta^{(t)})$.
- Accept $\beta^*$ as the next state $\beta^{(t+1)}$ with probability

$$\alpha(\beta^{(t)}, \beta^*) = \min\left\{1, \frac{\pi(\beta^*|x)q(\beta^{(t)}|\beta^*)}{\pi(\beta^{(t)}|x)q(\beta^*|\beta^{(t)})}\right\} \tag{4.2.3}$$

  If $\beta^*$ is accepted , the next state becomes $\beta^{(t+1)} = \beta^*$; if $\beta^*$ is rejected the chain does not move, and $\beta^{(t)} = \beta^{(t-1)}$. As already stated this procedure will yield a sequence of values $\beta^{(0)}, \beta^{(1)}, \beta^{(2)}, \ldots$ such that provided the length $B$ of the burn-in is sufficiently large, we can take $\beta^{(B+1)}, \beta^{(B+2)}, \ldots$ to be a sample from the posterior density $\pi(\beta|x)$. Burn-in length can be determined by examination of trace plots of the Markov chains or via the Gelman-Rubin statistics which checks convergence (see Section 7.4.2 for details). The trace plot of the parameter of interest should be stable after removal of burn-in, hence we remove the initial record of the movement of the chain from its starting value. Figure 4.1 shows an example of burn-in and a typical burn-in allowance. As explained above, since this procedure defines a Markov chain, this is a dependent sample. Nevertheless, it can be used, in conjunction with (4.2.2) to understand $\pi(\beta|x)$ and associated posterior quantiles.

### 4.2.4 Random walk Metropolis-Hastings

The choice of the proposal density $q(\beta|\beta^{(t)})$ is up to the user. This density may or may not depend on $\beta^{(t)}$. For example, if $q$ is taken to be a uniform density over all possible values of

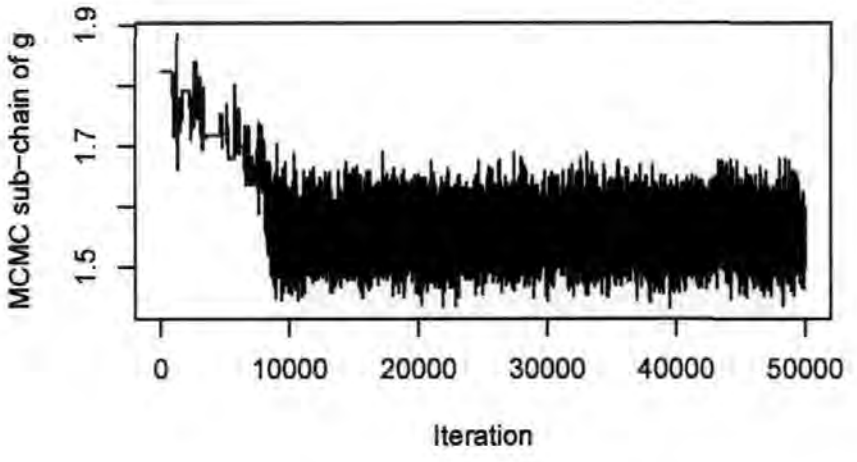



Figure 4.1: Example of a trace plot of some parameter g.

$\beta$, then the candidate $\beta^*$ does not depend on $\beta^{(t)}$. If, on the other hand, $q$ is, for example, taken to be a normal or $t$-density centred on $\beta^{(t)}$, then $\beta^*$ will depend on $\beta^{(t)}$. We shall say that such a choice of $q$ yields a random walk Metropolis-Hastings algorithm as the next position is chosen with reference to the current position. In Chapter 7, we shall implement a random walk Metropolis-Hastings algorithm.

### 4.2.5 Obtaining Posterior Credible Intervals

The realizations $\beta^{(t)}$, $t = B+1, \ldots, T$, produced by the Metropolis-Hastings algorithm can be used to help us understand the posterior $\pi(\beta|x)$. For example, the posterior mean $\mathrm{E}[\beta|x] = \int \beta\pi(\beta|x)d\beta$ can be approximated by the sample mean $\sum_{t=B+1}^{T} \beta^{(t)}/(T-B)$. In a similar way, a 95% posterior credible interval for $\beta$ can be obtained by ordering the $\beta^{(t)}$, $t = B+1, \ldots, T$, and taking the $0.025(T-B)^{\text{th}}$ and $0.975(T-B)^{\text{th}}$ elements of this ordered sample.

## 4.3 Nonparametric Regression Techniques

Nonparametric regression can be thought of as an extension of standard polynomial regression for modelling bivariate data of the form $(t_i, Y_i)$, $i = 1, \ldots, n$, where here $n$ is the number of data points.

First a model of the form $Y = g(t) + \epsilon$, in which $\mathrm{E}[\epsilon] = 0$ for all values of $t$, is postulated. The task is then one of estimating the curve $g$ from the available data $(t_i, Y_i)$, $i = 1, \ldots, n$. Note that since $\mathrm{E}[Y] = g(t)$ we can refer to this model as a mean regression. Nonparametric approaches can offer more flexible estimates of $g$ than standard polynomial regression models, and are not formulated in terms of a parametric model. An example of a nonparametric regression model is the smoothing spline, which will be discussed in detail in Section 4.3.1. Figure 4.2 shows an example of a smoothing spline and a standard polynomial (cubic) regression curve; here we can see the clear difference in flexibility of the smoothing spline. In general, nonparametric regression usually follows one of two approaches, said to be either kernel or spline based. Both methodologies can perform equally well for a range of smoothing problems. However, in this thesis we are mainly interested in spline based approaches. Hence we shall focus on the basic principles of a univariate roughness penalty spline based approach. Details of kernel methods can be found in Gamerman (1997) or Green and Silverman (1994).

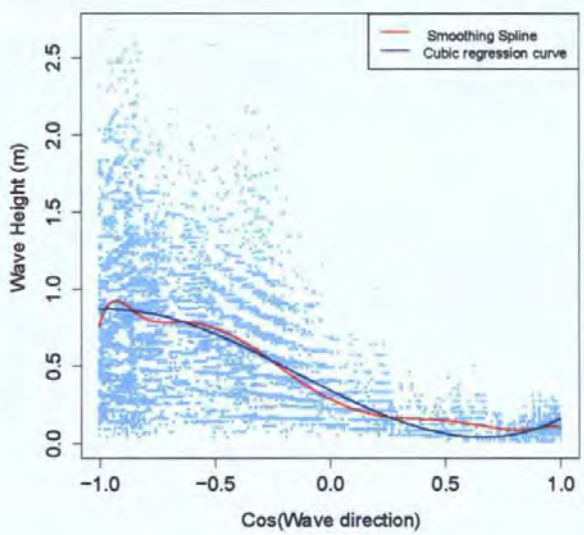



Figure 4.2: Scatter plot of wave height against transformed wave direction with a cubic regression curve and a smoothing spline.

Nonparametric spline based regression curves are in essence a series of polynomials regression curves which have been glued together to create one complete continuous curve; see de Boor (1978). The data is split into sections along the horizontal axis, here denoted $t$, and a curve is fitted in each section rather than across the entire data set. These individual

curves are constrained to fit in a sensitive way. Hence the overall fit is much more accurate as residual error can be greatly reduced due to the more localized fitting without producing a curve that is too rough in the sense that it fluctuates too rapidly. The boundaries between data sections are called knots. The choice of knots is up to the user.

As mentioned, when fitting a curve through a bivariate data set, one important consideration is the roughness of the curve, i.e. how "wiggly" it is. More specifically, we tend to prefer smooth curves that have a reduced amount of rapid fluctuation, hence we wish to study the more slowly moving trend in the data, regarding very rapid variation as 'noise'. It should be noted that this is not always the case, in some situations modelling of the rapid variation is desirable. We are able to quantify the roughness of a curve $g$ with continuous second derivative on the interval $[a, b]$ by means of a roughness penalty which is defined here as the integrated squared second derivative $\int_a^b g''(t)^2 dt$; see Green and Silverman (1994). A standard approach to curve fitting is based on a trade-off between the lack-of-fit of a curve to the data and its roughness, or, equivalently, between goodness-of-fit and smoothness, as discussed in Green and Silverman (1994). These authors also shown how this approach can be formalized within the Bayesian framework (see Gamerman (1997)) by having a prior distribution which quantifies probabilistically the roughness of the fitted curve.

### 4.3.1 Formal Spline Definitions

There are many different types of spline, for example, linear, quadratic, cubic,... These are defined and discussed in detail in de Boor (1978) and Hastie et al. (2001). We shall shortly define a natural cubic spline as this will be the type of spline that we will use later. There are also several different techniques for using splines to make inferences from data, or, in simpler terms for fitting splines to data $(t_i, Y_i)$, $i = 1, \ldots, n$. These techniques include interpolating splines, smoothing splines and quantile regression splines. They will be discussed later in this chapter.

#### The Natural Cubic Spline

A function $g$ is said to be a cubic spline with $N \geq 2$ knots $\tau_1, \ldots, \tau_N$, if $g$ is a cubic polynomial between knots $\tau_{i-1}$ and $\tau_i$, $i = 2, \ldots, N$, and if $g$ has continuous first and second derivatives

at $\tau_i$, $i = 2, \ldots, N-1$. Let $a < \tau_1$ and $b > \tau_N$. The curve $g$ is said to be a natural cubic spline (NCS) on $[a, b]$ if it is linear on the intervals $[a, \tau_1]$ and $[\tau_N, b]$ and if it has continuous first and second derivatives at $\tau_1$ and $\tau_N$; see Green and Silverman (1994), de Boor (1978), Hastie et al. (2001) and Venables and Ripley (2002) for further discussion.

Let $g_i = g(\tau_i)$ and $\gamma_i = g''(\tau_i)$ for $i = 1, \ldots, N$, and let $\boldsymbol{g} = (g_1, \ldots, g_N)^T$ be a column vector of curve values at the knots. Since by definition of a NCS it follows that $g''(\tau_1) = g''(\tau_n) = 0$, we can represent these second derivatives as the vector $\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_{N-1})^T$. Any given vectors $\boldsymbol{g}$ and $\boldsymbol{\gamma}$ are consistent with coming from a NCS provided a certain condition holds. Before stating this condition, we need some further definitions. Let $h_i = \tau_{i+1} - \tau_i$ for $i = 1, \ldots, N-1$. Let the banded matrix $Q$ be the $N \times (N-2)$ matrix with entries $q_{ij}$, for $i = 1, \ldots, N$ and $j = 2, \ldots, N-1$, given by

$$q_{j-1,j} = h_{j-1}^{-1}, \quad q_{jj} = -h_{j-1}^{-1} - h_j^{-1}, \quad q_{j+1,j} = h_j^{-1}$$

and $q_{ij} = 0$ for $|i - j| \geq 2$. Numbering of the elements of $Q$ is based on $\boldsymbol{\gamma}$, and hence the top left element is $q_{12}$. The banded symmetric matrix $R$ of dimension $(N-2) \times (N-2)$ is defined as follows:

$$\begin{aligned} r_{ii} &= \frac{1}{3}(h_{i-1} + h_i) \quad \text{for} \quad i = 2, \ldots, N-1, \\ r_{i,i+1} &= r_{i+1,i} = \frac{1}{6}h_i \quad \text{for} \quad i = 2, \ldots, N-2, \end{aligned}$$

and $r_{ij} = 0$ for $|i - j| \geq 2$. Since $R$ is strictly positive definite, we can define the $N \times N$ matrix $K$ as

$$K = QR^{-1}Q^T. \tag{4.3.1}$$

The symmetric matrix $K$ has rank $N-2$. We can now state the above mentioned condition; the full proof of this theorem can be found in Green and Silverman (1994).

**Theorem 4.3.1.1.** *The vectors $\boldsymbol{g}$ and $\boldsymbol{\gamma}$ specify a NCS $g$ if and only if the condition*

$$Q^T\boldsymbol{g} = R\boldsymbol{\gamma} \tag{4.3.2}$$

*holds. If* $Q^T\boldsymbol{g} = R\boldsymbol{\gamma}$, *then the roughness penalty will satisfy*

$$\int_a^b g''(t)^2 dt = \boldsymbol{\gamma}^T R \boldsymbol{\gamma} = \boldsymbol{g}^T K \boldsymbol{g}. \tag{4.3.3}$$

We shall make use of this expression for the roughness penalty in Section 4.3.2. We now move on to discuss interpolating splines for the points $(\tau_1, g_1), \ldots, (\tau_N, g_N)$.

### Interpolating Splines

Given points $(\tau_1, g_1), \ldots, (\tau_N, g_N)$, an interpolating function $g$ through these points has the property that $g(\tau_i) = g_i$, $i = 1, \ldots, N$. To find such a $g$, we could take $g$ to be piecewise linear or polynomial between the points $(\tau_i, g_i)$. We have already stated that we prefer smooth curves so we immediately disregard the piecewise linear approach: the resulting curve $g$ may have discontinuous derivatives at $\tau_i$, $i = 1, \ldots, N$, and would not appear smooth. If we now consider piecewise polynomials, provided we have chosen the polynomial correctly and are careful to ensure that derivatives are continuous at $\tau_i$, we may produce a curve $g$ that is usually smooth. To further refine this idea to potentially the 'best' or 'smoothest possible' curve, we could use as our interpolating curve $g$, the one with continuous second derivatives that minimizes the roughness penalty $\int g''(t)^2 dt$. Such a curve would be a natural cubic spline with knots at $\tau_1, \ldots, \tau_N$.

Green and Silverman (1994) state the following theorem that asserts the uniqueness of this interpolating natural cubic spline:

**Theorem 4.3.1.2.** *Suppose* $N \geq 2$ *and that* $\tau_1 < \cdots < \tau_N$. *Given any values* $g_1, \ldots, g_N$, *there is a unique natural cubic spline* $g$ *with knots at the points* $\tau_i$ *satisfying*

$$g(\tau_i) = g_i \text{ for } i = 1, \ldots, N. \tag{4.3.4}$$

See Green and Silverman (1994), page 15 for the full proof of this result. We will use this result again in Chapter 7.

### Smoothing Splines

Recall that we have data $(t_i, Y_i)$, $i = 1, \ldots, n$, where $n \geq 3$. Assume that $t_1, \ldots, t_n$ are such that $a < t_1 < \cdots < t_n < b$. Let $\mathcal{S}_2[a, b]$ be the space of functions with continuous second derivatives on $[a, b]$. If we are given any function $g$ in $\mathcal{S}_2[a, b]$, we can define $\mathcal{S}(g)$ to be the penalized sum of squares

$$\sum_{i=1}^{n} \{Y_i - g(t_i)\}^2 + \alpha \int_a^b \{g''(t)\}^2 dt \qquad (4.3.5)$$

where $\alpha$ is a positive smoothing parameter. This penalized sum of squares consists of two main elements, a measure of lack-of-fit to the data $\sum_{i=1}^{n} \{Y_i - g(t_i)\}^2$ and a roughness penalty $\int_a^b \{g''(t)\}^2 dt$. The measure of lack-of-fit of $g$ is the residual sum of squares which represents the discrepancy between our model $g$ and the data. The estimate $\hat{g}$ of the curve is defined to be the minimizer of $\mathcal{S}(g)$ over the class $\mathcal{S}_2[a, b]$. The estimate $\hat{g}$ represents a trade-off between lack-of-fit of the curve to the data and its roughness, a trade-off controlled by the smoothing parameter $\alpha$. The choice of this smoothing parameter (or a parameter directly related to it) is discussed in Section 4.3.3. We shall refer to this $\hat{g}$ as a smoothing spline. Green and Silverman (1994) show that $\hat{g}$ is a natural cubic spline with $n$ knots at $t_1, \ldots, t_n$; note that the linear segments beyond the range of the data do not contribute to the value of the functional $\mathcal{S}(g)$ defined in (4.3.5) since their second derivative is zero.

## 4.3.2 Nonparametric Regression in a Bayesian Framework

We can embed the ideas of the previous section in the Bayesian framework. To do this we adopt a prior density over curves $g \in \mathcal{S}_2[a, b]$ which is proportional to $\exp(-\frac{1}{2}\lambda \int_a^b \{g''(t)\}^2)$, where $\lambda > 0$. Let us further assume that $Y_i = g(t_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, independently. From this prior and data model we can determine the posterior log density of $g$ given the observed data values, $Y_1, \ldots, Y_n$ as

$$\ell_{post}(g) \stackrel{c}{=} -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \{Y_i - g(t_i)\}^2 - \frac{1}{2}\lambda \int_a^b \{g''(t)\}^2 \qquad (4.3.6)$$

Where we use the notation $\stackrel{c}{=}$ to indicate equality up to a constant. Hence, the smoothing spline $\hat{g}$ corresponds to the mode of this posterior. We can now use Theorem 4.3.1.1 to turn what seems to be an infinite dimensional problem into a finite dimensional one; see Green and Silverman (1994), for further details. This considerably simplifies the way in which we may think of this Bayesian approach.

If $g$ is a natural cubic spline taking values $g_1, \ldots, g_n$ at knots $t_1, \ldots, t_n$, then Theorem 4.3.1.1 tells us that the log prior density can be represented as

$$\ell_{prior}(g) \stackrel{c}{=} -\frac{1}{2}\lambda \int_a^b \{g''(t)\}^2 dt = -\frac{1}{2}\lambda \boldsymbol{g}^T K \boldsymbol{g} \qquad (4.3.7)$$

where $\boldsymbol{g} = (g_1, \ldots, g_n)^T$. We see that the higher the value of the roughness $\boldsymbol{g}^T K \boldsymbol{g}$ associated with $\boldsymbol{g}$, the lower the value of the associated prior density, with this effect being controlled by the smoothing parameter $\lambda$; higher values of $\lambda$ result in less prior weight being given to curves with high roughness. The associated log posterior density now takes the form

$$\ell_{post}(g) \stackrel{c}{=} -\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{g})^T(\boldsymbol{Y} - \boldsymbol{g}) - \frac{1}{2}\lambda \boldsymbol{g}^T K \boldsymbol{g}, \qquad (4.3.8)$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$.

### 4.3.3 Choosing the smoothing parameter $\lambda$

For any smoothing problem, the choice of the smoothing parameter is crucial. With our smoothing spline approach, there is a computationally fast 'automated' approach for choosing the smoothing parameter in a way that is informed by the data. This methodology is called cross-validation and is based on the idea of prediction. Green and Silverman (1994) explain that $\hat{g}(t)$ should provide a good prediction of $Y$ at a new value $t$, in the sense that the squared residual $\{Y(t) - \hat{g}(t)\}^2$ should be small. Unfortunately a new observation $(t, Y)$ is not available. To overcome this, the cross-validation procedure generates a 'new' observation by omitting $(t_i, Y_i)$ from the original data. The value of the smoothing spline fitted to the reduced data set at $t_i$ is denoted $\hat{g}^{(-i)}(t_i; \lambda)$.

As the observation that we omitted from the original data was specified in an arbitrary way the overall predictive performance when $\lambda$ is the smoothing parameter can be quantified

by the cross-validation score

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{g}^{(-i)}(t_i;\lambda)\}^2. \tag{4.3.9}$$

Green and Silverman (1994) show that $CV(\lambda)$ can be simplified to

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i - \hat{g}(t_i;\lambda)}{1 - A_{ii}(\lambda)}\right\}^2 \tag{4.3.10}$$

where $A_{ii}(\lambda)$ is the $i^{\text{th}}$ diagonal element of the 'hat' matrix $A(\lambda)$ such that $\hat{\boldsymbol{g}} = A(\lambda)\boldsymbol{Y}$ where $\hat{\boldsymbol{g}} = (\hat{g}(t_1),\ldots,\hat{g}(t_n))^T$. The form (4.3.10) of $CV(\lambda)$ requires the computation of smoothing spline $\hat{g}(t,\lambda)$ for each value of $\lambda$ instead of the $n$ computations $\hat{g}^{(-i)}(t;\lambda)$; $i = 1,\ldots,n$, for each $\lambda$ required in (4.3.9). The form (4.3.10) can be modified to the possibly more stable version

$$GCV(\lambda) = \frac{1}{n\left\{1 - \frac{\text{trace}A(\lambda)}{n}\right\}^2}\sum_{i=1}^{n}\{Y_i - \hat{g}(t_i;\lambda)\}^2, \tag{4.3.11}$$

by replacing $A_{ii}(\lambda)$ by the average value $\text{trace}A(\lambda)/n$. As with $CV(\lambda)$, $GCV(\lambda)$ is minimized over $\lambda$ to yield an estimate of the smoothing parameter. The generalized cross validation estimate of $\lambda$ is usually preferred to the cross validation estimate, although often these estimates of $\lambda$ can be very similar. Further discussion can be found in Green and Silverman (1994).

## 4.4 Quantile regression

In this section we introduce the key elements of quantile regression highlighting the main differences from the standard regression approach.

### 4.4.1 Definitions

We begin by defining the term quantile. The $p$th quantile, $0 \leq p \leq 1$, of a random variable $X$ is a value $q$ such that $\Pr(X \leq q) = p$.

Consider now a regression model with covariate column vector $\boldsymbol{x}^T$ and response variable

$Y$:

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon \tag{4.4.1}$$

where $\boldsymbol{\beta}$ is a vector of parameters and $\mathrm{E}[\epsilon] = 0$ for all covariate values. This formulation effectively models the relationship between $\boldsymbol{x}$ and the conditional mean of $Y$ given $\boldsymbol{X} = \boldsymbol{x}$. Such a model for $E[Y|\boldsymbol{X} = \boldsymbol{x}]$ helps us to understand how the mean of $Y$ depends on $\boldsymbol{x}$, but fails to deliver a complete picture of the behaviour of the distribution of $Y$ as a function of $\boldsymbol{x}$. An alternative method of modelling is therefore required which is not based on the means of $Y$, but which can capture its full distribution. This technique is known as Quantile regression. Quantile regression therefore models the conditional quantiles of $Y$ given $\boldsymbol{X} = \boldsymbol{x}$, denoted $Q_p(Y|\boldsymbol{X} = \boldsymbol{x})$, where $Q_p(Y|\boldsymbol{X} = \boldsymbol{x})$ is such that

$$P(Y \leq Q_p(Y|\boldsymbol{X} = \boldsymbol{x})|\boldsymbol{X} = \boldsymbol{x}) = p \tag{4.4.2}$$

We take the following passage from Koenker and Hallock (2000) as it provides an excellent summary for the reasoning behind the use of quantile regression as an alternate to ordinary least-squares regression:

*"What the [mean] regression curve does is give a grand summary for the averages of the distributions corresponding to the set of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions."*

### Check Function

In standard mean regression, the unknown parameter vector $\boldsymbol{\beta}$ is estimated by minimizing over $\boldsymbol{\beta}$ the residual sum of squares

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2, \tag{4.4.3}$$

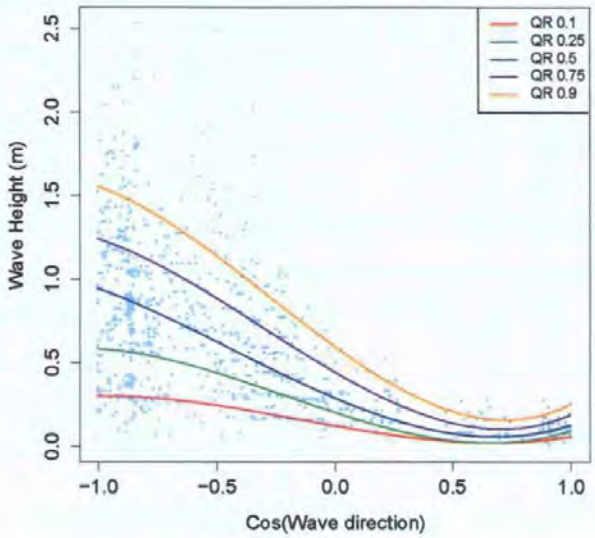



Figure 4.3: Scatter plot of wave height against transformed wave direction with a range of quantile regression curves using cubic polynomials, i.e. setting $\boldsymbol{x}^T = (1, x, x^2, x^3)$, where $x$ is the cosine of wave direction.

where $\boldsymbol{x}_i^T$ is the $i^{\text{th}}$ row of the covariate matrix $X$ over $\boldsymbol{\beta}$. This can be written as $\sum_{i=1}^n r(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$ where $r$ is the quadratic loss function defined as $r(u) = u^2$. In quantile regression the equivalent loss function can be written as $\rho_p(u) = u(p - I(u < 0))$ in which $p$ is the quantile of interest and $I$ is the usual indicator function. The function $\rho_p$ is known as the check function. So just as in mean regression the parameters $\boldsymbol{\beta}$ are estimated by minimizing a sample estimate of $\mathrm{E}[r(Y - \boldsymbol{x}^T\boldsymbol{\beta})]$, so in quantile regression $\boldsymbol{\beta}$ minimizes a sample estimate of $\mathrm{E}[\rho_p(Y - \boldsymbol{x}^T\boldsymbol{\beta})]$. Further discussion can be found in the book by Koenker (2005). The package **quantreg** (Koenker (2008)) that can be run in R can be used to fit quantile regression models by minimizing $\sum_{i=1}^n \rho_p(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$, Figure 4.3 was produced using the function `qr` of this package. As an alternative inferential approach, we now place the above check function based minimization approach to quantile regression in a likelihood framework using the asymmetric Laplace density.

### Asymmetric Laplace Density Approach

We begin by returning to the mean regression model (4.4.1) Let us assume now that the error has a Gaussian distribution $\epsilon \sim N(0, \sigma^2)$, with standard deviation $\sigma$. For our sample,

$\{\boldsymbol{x}_i, Y_i\}_{i=1}^n$, the associated likelihood function for $\boldsymbol{\beta}$ is

$$L(\beta) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right\}. \tag{4.4.4}$$

Least squares estimates can be obtained by maximization $L(\boldsymbol{\beta})$ over $\boldsymbol{\beta}$ and so are equivalent to maximum likelihood estimates. We now move to quantile regression and make the assumption that our model errors now has probability density function

$$f(\epsilon) \propto \exp\left\{-\sum_{i=1}^{n}\rho_p(\epsilon)\right\}, \tag{4.4.5}$$

where $\rho_p$ is the check function. This is known as the asymmetric Laplace density function. The associated likelihood function is $L(\boldsymbol{\beta}) \propto \exp\{-\sum_{i=1}^n \rho_p(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\}$. This has the consequence that an estimate of $\boldsymbol{\beta}$ resulting from the $\sum_{i=1}^n \rho_p(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$ is a maximum likelihood estimate; see Yu et al. (2003).

### 4.4.2 A Nonparametric Approach

Above we discussed the parametric approach to quantile regression based on the model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$. We now focus on the special case where $\boldsymbol{x}^T = (1, x)$ and consider roughness penalty approaches to the quantile regression. Koenker et al. (1994) and Bosch et al. (1995) discuss computational difficulty of estimating what they refer to as a quantile smoothing spline $g$ which minimizes

$$\sum_{i=1}^{n}\rho_p\{Y_i - g(x_i)\} + \lambda\int\{g''(x)\}^2dx, \tag{4.4.6}$$

where the range of integration for the roughness penalty contains $x_1, \ldots, x_n$. As these difficulties are hard to overcome, Koenker et al. (1994) sets up an alternate minimization problem

$$\sum\rho_p\{Y_i - g(x_i)\} + \lambda\left(\int|g''(x)|^q dx\right)^{1/q}, \tag{4.4.7}$$

As we can see the penalty function here is different from the roughness penalty. Koenker et al. (1994) particularly focus on $q = 1$ and $q = \infty$. They found that when $q = \infty$, an upper

bound is imposed on $|g''(x)|$ resulting in a piecewise quadratic estimate which is simple to compute. However, when $q = 1$ the function to be minimized is reduced to

$$\sum_{i=1}^{n} \rho_p\{Y_i - g(x_i)\} + \lambda \int |g''(x)|dx. \tag{4.4.8}$$

This was introduced in earlier work by Koenker (2005), who claimed that the solution to (4.4.8) is a parabolic spline. However, at a later stage this was found to be incorrect. This led Koenker et al. (1994) to reformulate the $q = 1$ penalty term. The paper by Bosch et al. (1995) considers a different approach to estimating the quantile functions that yields solutions that are cubic splines. Koenker (2005) also provides a short summary of these approaches.

### 4.4.3 Bayesian Approach

In this section we discuss the implementation of parametric quantile regression in a Bayesian framework presented by Yu and Moyeed (2001). Sections 4.1 and the above parts of this Section 4.4 provide the necessary foundations for the construction of the Bayesian quantile regression model. We adopt the model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$, where $\epsilon$ follows an asymmetric Laplace distribution with density function given by (4.4.5). This leads to the likelihood function for $\boldsymbol{\beta}$

$$L(\boldsymbol{\beta}) = f(\boldsymbol{Y}|\boldsymbol{\beta}) = p^n(1-p)^n \left\{ -\sum_{i=1}^{n} \rho_p(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}) \right\}. \tag{4.4.9}$$

As we saw in Section 4.1, we now need to specify a prior density $\pi(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$. Any prior could potentially be used in this formulation but without substantial information on which to base this choice, an improper uniform prior distribution was adopted by Yu and Moyeed (2001); see their paper for a complete discussion justifying their prior choice. Now the likelihood and prior can be combined using Bayes theorem to find the posterior for $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto \pi(\beta)L(\boldsymbol{\beta}). \tag{4.4.10}$$

As this posterior density is not available in closed form, inferences about $\boldsymbol{\beta}$ are based on the output of an MCMC algorithm. Yu and Moyeed (2001) take $\boldsymbol{x}^T = (1, x, x^2, x^3)$ and

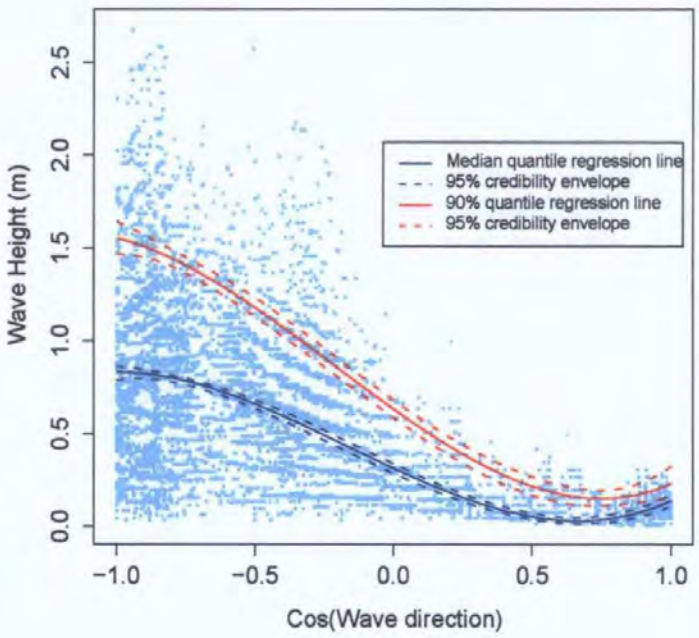



Figure 4.4: Scatter plot of wave height against transformed wave direction with 50% ($p = 0.5$) and 90% ($p = 0.9$) Bayesian quantile regression curves using cubic polynomials. 95% credible intervals are shown for both quantiles.

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$, so that

$$Q_p(Y|x) = \beta_0(p) + \beta_1(p)x + \beta_2(p)x^2 + \beta_3(p)x^3, \qquad (4.4.11)$$

where dependence on $p$ has been made explicit. Hence, Yu and Moyeed (2001) are performing inference on cubic quantile functions. We apply the approach of Yu and Moyeed (2001) to the data set introduced in Section 1.6. Posterior mean quantile regression curves for quantiles at $p = 0.5$ (median) and $p = 0.9$ are shown in Figure 4.4. A 95% credible interval is shown for both quantiles. This was calculated using the approach of Section 4.2.5.

# 5

# Dealing with Missing Data

## 5.1 Introduction

In this chapter we describe a method to impute missing information between existing time series observations. Our technique was primarily developed for data used within this thesis as the other methodology that we have presented depends on complete time series information.

With any data set there is a possibility of recording errors including completely missing values. These may be due to equipment failure, human error or unforeseen circumstances. They can be particularly prevalent in time series data. For our purposes it is important when analyzing time series that the data are complete since working with an incomplete data set can lead to biased inferences; for instance, extracting the maximum value within monthly time intervals may return a value below the actual maximum for the interval if the actual maximum is not present due to a recording error.

Since the Hindcast data that we analyse (see 1.6 for a detailed discussion) has been simulated from wind records spanning many years, it is highly likely that some errors are present. Due to the volume of observations, it was particularly difficult to visually identify errors or missing readings directly from the data. Initial time series plots were produced to try to aid identification, but due to scaling issues, errors were extremely difficult to identify.

## 5.2 Replacing Missing Data using LOESS

The remedy to the problem of identifying missing data automatically was found by considering the format of the data; as the data formed a time series, the temporal increments between observations were known. Because of this it was possible to create a temporal template to which the data should correspond. An R (R Development Core Team (2008)) function was created to generate such a template and to merge it with the existing time series. The resultant output inserted an "NA" (or Not Available) value at the times when no data value was recorded, revealing the missing values or gaps in the data.

We needed to replace missing information that our template method identified with generated values that would follow the time series pattern, so future modelling of the data would only be affected in a limited way by missing values. Accordingly, we extended our R function to search for "NA"s and then to identify the pattern of readings in a designated time period before and after each "NA". This created a window of information upon which to base the estimation of each missing value. By using a sufficiently large, but localized window around each problem area to provide sufficient information either side of the void, we were able to replace the missing values between the known blocks of information.

To achieve this replacement, a locally weighted least squares regression or *loess* model was fitted to the window of observations; see Harrell. Jr (2001) and Venables and Ripley (2002) for details. Fitted values could be extracted from the model to replace the missing values. We now briefly outline the loess technique. If we have bivariate data $(X, Y)$, then to obtain a smoothed value of $Y$ at $X = x$, we set a window around $x$ that contains a fixed number of data points. We fit a weighted linear regression to these points rather than the full data set. The predicted value from this weighted regression at $X = x$ is now our smoothed value of $Y$ at $X = x$. As loess involves weighted least squares regression, the weights must be chosen appropriately: points closest to $x$ are given the largest weighting and as the distance from the point $x$ increases the weighting reduces. Hence data points which lie near the window boundaries are given a much small weighting than points near $x$. There are two parameters that the user can choose in the loess approach. These are the span of the window, that is the proportion of the full data set used in each window, and the degree of the polynomial fitted.

This can be 0 for locally constant, 1 for locally linear and 2 for locally quadratic regression loess.

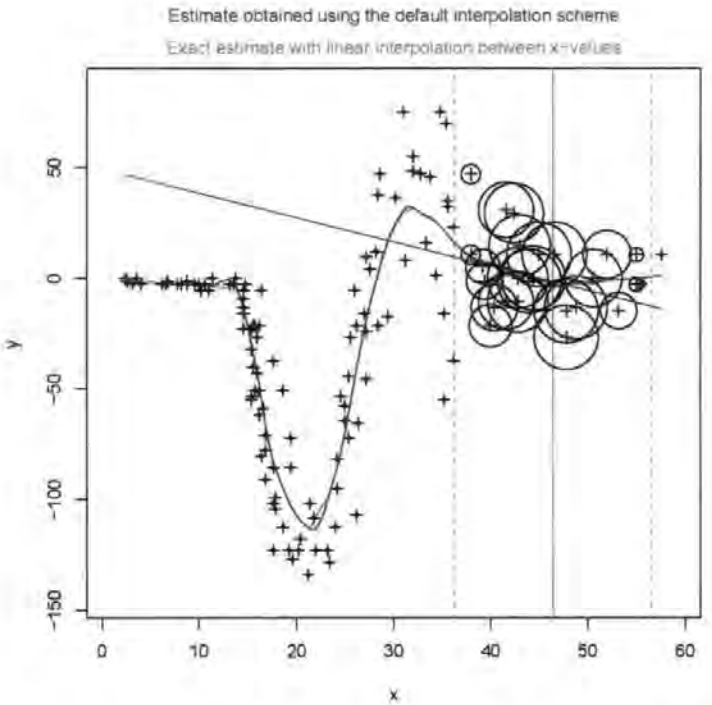



Figure 5.1: Illustration of the loess technique, showing the smooth curve, the window (dashed green vertical lines) at a particular $x$ value (unbroken green vertical line), the weights (circles) applied to each data point and the weighted linear regression fit (purple line). This plot was produced by the function `loess.demo` of the `TeachingDemos` package by Snow (2008).

We have seen that loess effectively uses a moving constant, linear or polynomial regression approach. This ensures good smoothing behaviour throughout the range of $x$. Another benefit provided by the loess procedure is robustness. After making initial local estimates of trend, the loess procedure will identify outliers from this trend; these are then reduced in weight and the trend is recalculated. This process is repeated up to three times to provide a trend approximation that is robust to outlying data values.

Figure 5.1, produced by the function `loess.demo` of the `TeachingDemos` package by Snow (2008), illustrates the loess technique applied to the famous simulated motorcycle accident data of Silverman (1985). The point $x$ is shown by the unbroken vertical line, while the window (here containing the nearest fifth of the data) is illustrated by the dashed vertical lines. The weight applied to each data point in the weighted linear regression within the

window is proportional to the area of the circle at that point. The value of the smooth curve at $x$ is the value at that point of the fitted weighted linear regression linear which is also shown.

The missing values in the original time series data were then replaced with the estimates from the loess procedure to produce a complete data set. We adapted the standard loess procedure in two cases. If there were a lot of missing values towards the beginning or end of our time series, or if there were large runs of missing values, then the window used for the local regression fitting would be expanded to allow a sufficient amount of information to inform our imputations. Figure 5.2 shows a time series plot of a section of hindcast wave period data, from the HR Wallingford data set, to which our loess method has been applied to replace missing information. The results of doing this are shown in red. A simulation study was performed to check that the choices of the models parameters were appropriate. The simulation study for a section of wave period data from the HR Wallingford data set is now described in the following section.

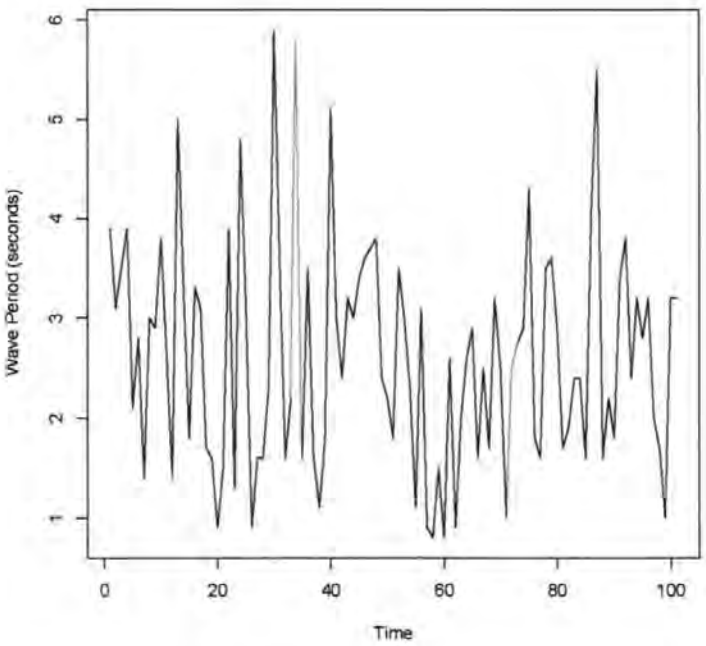



Figure 5.2: Time series plot of a section of the Hindcast Wave Period data from the HR Wallingford date set with missing observations. The gaps in the data have been imputed using the loess filling routine, the results of which are shown in red.

## 5.3 Loess Model Simulation Study

Table 5.1 summarizes the study performed to calibrate our loess based method for replacing missing values. As mentioned above, there are two parameters that the user can choose: span which controls the proportion of data in each window, and degree which determines the type of local regression model. We used the following procedure to recommend good choices of these parameter for our data set. We took a section of data containing two missing values. These two missing values are referred to in Table 5.1 as $NA1$ and $NA2$ and were imputed using our loess technique. The "Empirical Model Fit Quality" provides us with a visual assessment of fit quality. While this is a subjective method of assessment, the results are so clear cut, as we will discuss, that we did not pursue our study further. A more objective approach would have been to omit some known values and to choose the parameters by minimizing over span and degree a badness-of-fit criterion such as root mean square error (RMSE)

$$\sqrt{\frac{1}{J}\sum_{i=1}^{J}(\text{true}_i - \text{imputed}_i)^2} \qquad (5.3.1)$$

where the sum is over the $J$ omitted values, and $\text{true}_i$ and $\text{imputed}_i$ are true and imputed values of the $i^{\text{th}}$ missing value. We can conclude from the results of the visual assessment of fit quality presented in Table 5.1, that degree has a large effect on the fit of the model.

Table 5.1: The empirical model fit quality for 30 combinations of the span and degree parameters in our loess imputation procedure

| Model No. | Parameters | | NA values | | Empirical Model |
|---|---|---|---|---|---|
| | span | degree | $NA1$ | $NA2$ | Fit Quality (%) |
| 1 | 0.2 | 0 | 5.5 | 2.5 | 80 |
| 2 | 0.4 | 0 | 5.701 | 2.901 | 60 |
| 3 | 0.6 | 0 | 5.740 | 3.114 | 40 |
| 4 | 0.8 | 0 | 5.773 | 3.281 | 30 |
| 5 | 1.0 | 0 | 5.882 | 3.706 | 20 |

| | | | | | |
|---|---|---|---|---|---|
| 6 | 1.2 | 0 | 5.878 | 3.800 | 20 |
| 7 | 1.4 | 0 | 5.864 | 3.884 | 10 |
| 8 | 1.6 | 0 | 5.900 | 3.923 | 10 |
| 9 | 1.8 | 0 | 5.877 | 3.987 | 10 |
| 10 | 2.0 | 0 | 5.857 | 4.040 | 10 |
| 11 | 0.2 | 1 | *NA* | *NA* | 0 |
| 12 | 0.4 | 1 | 5.250 | 2.000 | 70 |
| 13 | 0.6 | 1 | 5.594 | 2.678 | 50 |
| 14 | 0.8 | 1 | 5.703 | 3.005 | 50 |
| 15 | 1.0 | 1 | 5.792 | 3.605 | 30 |
| 16 | 1.2 | 1 | 5.790 | 3.751 | 20 |
| 17 | 1.4 | 1 | 5.779 | 3.876 | 10 |
| 18 | 1.6 | 1 | 5.771 | 3.975 | 10 |
| 19 | 1.8 | 1 | 5.760 | 4.048 | 10 |
| 20 | 2.0 | 1 | 5.751 | 4.103 | 0 |
| 21 | 0.2 | 2 | *NA* | *NA* | 0 |
| 22 | 0.4 | 2 | *NA* | *NA* | 0 |
| 23 | 0.6 | 2 | 5.521 | 2.317 | 80 |
| 24 | 0.8 | 2 | 5.598 | 2.551 | 80 |
| 25 | 1.0 | 2 | 5.729 | 2.544 | 80 |
| 26 | 1.2 | 2 | 5.778 | 2.543 | 80 |
| 27 | 1.4 | 2 | 5.820 | 2.550 | 80 |
| 28 | 1.6 | 2 | 5.843 | 2.528 | 90 |
| 29 | 1.8 | 2 | 5.867 | 2.537 | 90 |
| 30 | 2.0 | 2 | 5.884 | 2.543 | 90 |

## 5.4 Performance Assessment

Assessment of the accuracy of the values replaced by our loess methodology is a difficult problem as in practice 'true' values on which we would assess the accuracy are not available. As already mentioned, we can, however, replace some known data values by missing values and try to recover the known data. In this way we have known values on which to base our assessment of accuracy. We judge the accuracy of our replaced values by calculating the RMSE as defined in (5.3.1) for our loess imputation method and a linear interpolation method. Figure 5.3 shows a section of the HR Wallingford data with artificial missing values replaced by imputed values from our loess filling routine and by linear interpolated values shown as the blue and green lines respectively. The RMSE values for our loess filling routine

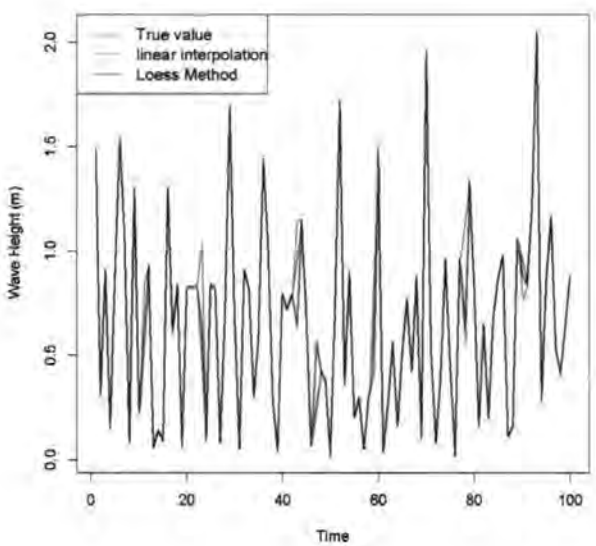



Figure 5.3: Time series plot of a section of known data from the HR Wallingford data set. The gaps in the data have been added at locations where known values are available, then have been imputed using the loess filling routine and linear interpolation methods.

and linear interpolation are 0.080 and 0.095 respectively. Hence we see that both methods recover missing values relatively well, but our loess performs better than linear interpolation.

## 5.5 Summary

We developed a technique to replace observations missing from time series data with imputed values from a loess model. The loess approach was used as it was felt that this was the most locally sensitive and hence appropriate for this application. Loess stands for local weighted polynomial regression, as the model is fitted over a small window at each point using weighted least squares. More weight is given to data near the point at which the response is being imputed and less weight to data further away. We presented a study that informed our choice of the loess parameters. We quantified the performance of our loess based methodology and showed that it performed well, and better than linear interpolation. We applied our technique to all our data with success, so providing us with competes data sets to which to apply the techniques of Chapters 6 and 7.

# 6

# Automated Threshold Selection Methods in Extreme Value Theory

## 6.1 Introduction

In Chapter 2 we discussed methodology based on the Generalized Pareto Distribution (GPD) used to provide inference about extreme values. We described the threshold selection methods employed in this methodology to define sufficiently large values for GPD fitting and illustrating them in Section 2.2.7. Examples of these threshold selection techniques can be found in Coles and Tawn (1991), Tawn and Coles (1994), and Tawn and Bruun (1998). The specification of an accurate threshold plays a major part in the quality of inference obtained. Coles (2001) and other authors explain that poor estimation or specification of the threshold can greatly affect the accuracy and utility of GPD models and their predictions.

Threshold selection has received additional, recent attention in the literature. Dupuis (1999) presents a guide to threshold selection based on robustness considerations, while Tancredi et al. (2006) adopt a Bayesian approach and discuss how to take account of threshold uncertainty; see Section 6.5.1 for further discussion of Tancredi et al. (2006) and Guillou and Hall (2001) for related methodology. In this chapter we present a new threshold selection technique which improves on the performance of existing methods. We illustrate our technique using the sea conditions data discussed in Chapter 1. Our automated threshold selection method requires no external input other than the variable of interest,

and is considerably simpler and easier to implement than the computationally expensive approaches proposed in recent papers.

We have also extended our threshold selection method to allow threshold choice to depend on a covariate such as the cosine of wave direction (we use the cosine transformation to reduce some of the problems associated with directional data), where our specific aim is to account for the directional effect when modelling wave height or wave period using GPDs. The practical advantage of our extended procedure is that it automatically identifies the wave directions associated with the highest waves and consequently can provide better estimation of wave height return levels.

We also present adaptations to the parameter estimation methodology used to fit the GPD. These adaptations were established as a result of problems that arose when implementing our new threshold selection technique. Finally we show some of the software developed during this work by focusing on a Graphical User Interface that we have produced. This software was developed as a potential design tool to facilitate the inclusion of extremes analysis in the coastal design process. Examples are included throughout the chapter to highlight the applications of the techniques developed and to provide comparison to existing methods such as JOINSEA, as discussed in Chapter 3.

## 6.2 Automated Constant Threshold Selection technique

Selection of an appropriate threshold $u$ is routinely performed on a visual basis using plots such as those shown in Section 2.2.7 and so can have a range of associated errors. These visual procedures require prior knowledge and experience of the accurate interpretation of these threshold choice plots to achieve a satisfactory model fit; see Chapter 2 and Davidson and Smith (1990), Walshaw and Coles (1994) and Coles (2001) for example. We now introduce the theoretical basis for our threshold selection methodology.

The form of the GPD is given and discussed in Section 2.2.1. We reproduce it here for convenience. Let $X$ be a random variable (such as Wave Height) and let $u$ be a suitably large threshold. Then, under the condition of Theorem 2.2.1.1, the distribution function of

the exceedance $Y = X - u$, conditional on $Y > 0$, is approximately

$$H(y) = 1 - \left[1 + \frac{\xi y}{\sigma_u}\right]^{-1/\xi}, \tag{6.2.1}$$

where $\xi$ is referred to as the shape parameter and where the scale parameter $\sigma_u > 0$ depends on the threshold $u$; see equations (2.2.27) and (2.2.11) and the discussion in Section 2.2.1.

When fitting the GPD to data, $\sigma_u$ and $\xi$ can be estimated using maximum likelihood estimation as discussed in Sections 2.1.2 and 2.2.3. To achieve a good model fit, we need to choose a suitable value of the threshold $u$. Figure 2.4 illustrates a routinely used threshold selection technique based on a plot of parameter estimates of GPDs fitted using a range of thresholds against the threshold, and is the basis for our automated threshold selection methodology. We now outline our methodology.

Let $u_1, \ldots, u_n$ be $n$ equally spaced increasing candidate thresholds. Let $\hat{\sigma}_{u_j}$ and $\hat{\xi}_{u_j}$ be maximum likelihood estimators of the scale and shape parameter based on data above the threshold $u_j$, $j = 1, \ldots, n$. Finally, let $u$ be a suitable threshold, that is one for which values of $y > u$ can be modelled using the GPD. It follows from equation (2.2.11) that, provided $u \leq u_{j-1} < u_j$,

$$\sigma_{u_{j-1}} = \sigma_u + \xi(u_{j-1} - u) \text{ and } \sigma_{u_j} = \sigma_u + \xi(u_j - u); \tag{6.2.2}$$

see also Coles (2001), page 83. Hence,

$$\sigma_{u_j} - \sigma_{u_{j-1}} = \xi(u_j - u_{j-1}). \tag{6.2.3}$$

Furthermore, standard maximum likelihood theory, as discussed in Coles (2001), tells us that $E[\hat{\sigma}_{u_j}] \approx \sigma_{u_j}$ and $E[\hat{\xi}_{u_j}] \approx \xi$, for any $j$ such that $u_j > u$. Let

$$\tau_{u_j} = \hat{\sigma}_{u_j} - \hat{\xi}_{u_j} u_j, \quad j = 1, \ldots, n, \tag{6.2.4}$$

and consider the differences

$$\tau_{u_j} - \tau_{u_{j-1}}, \; j = 2, \ldots, n; \tag{6.2.5}$$

it follows from (6.2.3) that $E[\tau_{u_j} - \tau_{u_{j-1}}] \approx 0$. Moreover, we can appeal to the same theory to conclude that the $\tau_{u_j} - \tau_{u_{j-1}}$ approximately follow a normal distribution. This result leads us to the following procedure for finding a suitable threshold $u$:

(1) Identify suitable values of equally spaced candidate thresholds $u_1 < u_2 < \cdots < u_n$. We found that setting $n = 100$ gives good results. We take $u_1$ to be the median and $u_n$ to be the 98% quantile of the data, unless fewer than 100 values exceed this value, in which case $u_n$ is set to the $100^{\text{th}}$ data value in descending order. Our procedure performs well in such circumstances. Less reliable results were obtained from smaller data sets.

(2) If $u$ is a suitable threshold, then all differences $\tau_{u_j} - \tau_{u_{j-1}}$ have an approximate normal distribution with mean 0 provided $u \leq u_{j-1} < u_j$. If $u$ is unsuitable, then these differences may not follow a normal distribution. This suggests that a suitably applied test for normality is an effective method to determine $u$.

The Pearson's Chi-square Test is used as a test of goodness-of-fit to establish whether or not the observed differences are consistent with a normal distribution with mean 0; see Greenwood and Nikulin (1996). Initially, we consider $u = u_1$ and perform the Pearson normality test based on all the differences $\tau_{u_2} - \tau_{u_1}, \tau_{u_3} - \tau_{u_2}, \ldots, \tau_{u_n} - \tau_{u_{n-1}}$. If the null hypothesis of normality is not rejected, $u$ is taken to be a suitable threshold. If the null hypothesis is rejected, then we consider $u = u_2$, remove $\tau_{u_2} - \tau_{u_1}$ from the set of differences considered, and repeat the above procedure. We have found that a size 0.2 test generally performs well. Reducing the size of the test has the effect of lowering the chosen threshold.

(3) Step 2 is iterated until the Pearson's Chi-square test indicates that the differences are consistent with a normal distribution with mean 0. If this does not happen, $u_n$ is returned with a warning. Our experience is that this latter situation occurs rarely.

The above steps can be performed quickly, so yielding a procedure that is computationally inexpensive. We implemented our method in the freely available, open source statistical environment `R` (R Development Core Team (2008)). Before presenting examples of the application of our methodology in Section 6.4, we now discuss some adaptation of the

maximum likelihood methodology used to estimate the parameters $\xi$ and $\sigma_u$. We made these adaptations in the light of estimation difficulties that we encountered while developing our automated threshold selection technique.

## 6.3 Adaptation to the Parameter Estimation Methodology

### 6.3.1 Current Parameter Estimation Technique

Throughout this thesis we use maximum likelihood estimation (MLE) to estimate the GPD model parameters $\xi$ and $\sigma$ (we now drop the subscript $u$ for notational convenience). Further details about MLE can be found in Chapter 2; see also Eliason (1993) and Davidson and Smith (1990). Let the excesses of a threshold $u$ be denoted $y_1, \ldots, y_k$ if there are $k$ excesses. As we saw in Section 2.2.3 (equations (2.2.15) and (2.2.16)), the log-likelihood can be divided into two cases, depending on the value of the shape parameter $\xi$:

$\xi \neq 0$ :

$$\ell(\sigma, \xi) = -k \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \log\left(1 + \frac{\xi y_i}{\sigma}\right) \tag{6.3.1}$$

provided $1 + \frac{\xi y_i}{\sigma} > 0$ for $i = 1, \ldots, k$;

$\xi = 0$ :

$$\ell(\sigma) = -k \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{k} y_i; \tag{6.3.2}$$

we note that $\ell(\sigma) = \lim_{\xi \to 0} \ell(\sigma, \xi)$, by Taylor expanding $\log(1 + \xi y_i/\sigma)$. The following facts about the parameter estimates come from Smith (1985) and Coles (2001):

$\xi > -0.5$ Maximum likelihood estimators have their usual asymptotic properties.

$-1 < \xi < -0.5$ Maximum likelihood estimators are obtainable, but do not have usual asymptotic properties.

$\xi < -1$ Maximum likelihood estimators are unlikely to be obtainable.

Analytic maximization of the log-likelihood is not possible; hence a numerical optimization algorithm is used. When obtaining parameter estimates, we also aim to find the standard errors and correlations of these estimates. These can be obtained from the asymptotic

variance-covariance (VC) matrix which can be calculated from knowledge of the Hessian matrix.

**Definition 6.3.1.1.** ***Hessian Matrix*** *The analogue of the second derivative for functions of several variables is called the Hessian Matrix (see Lang (1987), Dineen (1998) and Freund (1992)). If $f$ is a function of $X = (x_1, \ldots, x_n)$, then its Hessian $H_f(X)$ is the matrix*

$$H_f(X) = \left( \frac{\partial^2 f}{\partial x_i \partial x_j} \right). \tag{6.3.3}$$

In the present case our function of several variables is the log-likelihood $\ell$ of the GPD model given in (6.3.1) and (6.3.2). In the $\xi > 0$ case, the symmetric Hessian matrix $H$ takes the general form

$$H = \begin{pmatrix} \frac{\partial^2 \ell(\sigma,\xi)}{\partial \sigma^2} & \frac{\partial^2 \ell(\sigma,\xi)}{\partial \sigma \partial \xi} \\ \frac{\partial^2 \ell(\sigma,\xi)}{\partial \xi \partial \sigma} & \frac{\partial^2 \ell(\sigma,\xi)}{\partial \xi^2} \end{pmatrix}. \tag{6.3.4}$$

An approximate VC matrix can be obtained by inverting the negative Hessian and evaluating the result at the maximum likelihood estimates $\hat{\sigma}$ and $\hat{\xi}$ provided by the numerical optimization routine applied to (6.3.1) and (6.3.2). In practice the Hessian is usually estimated numerically as part of the optimization procedure; see Coles (2001). Development of our automated threshold selection method highlighted problems within the current approach of obtaining parameter estimates and their approximate VC matrix by means of the numerical estimation of the Hessian. The range of problems and their solutions are discussed in the following sections.

### 6.3.2 Analytic Hessian Calculation

The need for an analytic form of the Hessian matrix arose when 'singularities' occurred due to its numerical estimation. The calculation of the approximate VC matrix of the parameter estimates requires the inversion of the negative Hessian Matrix. When determining the standard errors of the parameter estimates from the VC matrix, it is necessary to take the square root of each of the leading diagonal elements; if any of the elements to be square rooted is negative (corresponding to a negative variance estimate), then a complex or undefined

value occurs causing the 'singularity'. The solution to the problem was to discover why a negative value was occurring in the approximate VC matrix.

In general, if the true value of a positive quantity lies very near to zero then sometime numerical techniques can return a negative value. This can occur in the numerical evaluation of the Hessian. Adopting an analytic solution avoided this problem and provided a much tidier general solution, even though the numerical approach is often adequate. We now present analytic expressions for the elements of the Hessian matrix:

$$\frac{\partial^2 \ell(\sigma,\xi)}{\partial \sigma^2} = \frac{1}{\sigma^2}\left\{k - \xi \sum_{i=1}^{k} \frac{y_i(2\sigma + \xi\, y_i)}{(\sigma + \xi\, y_i)^2} - \sum_{i=1}^{k} \frac{y_i(2\sigma + \xi\, y_i)}{(\sigma + \xi\, y_i)^2}\right\} \tag{6.3.5}$$

$$\begin{aligned}\frac{\partial^2 \ell(\sigma,\xi)}{\partial \sigma \partial \xi} &= \frac{\partial^2 \ell(\sigma,\xi)}{\partial \xi \partial \sigma} = \frac{1}{\sigma\xi}\left\{-\sum_{i=1}^{k} \frac{y_i}{(\sigma + \xi\, y_i)}\right. \\ &\quad \left. + \sigma\xi \sum_{i=1}^{k} \frac{y_i}{(\sigma + \xi\, y_i)^2} + \sigma \sum_{i=1}^{k} \frac{y_i}{(\sigma + \xi\, y_i)^2}\right\}\end{aligned} \tag{6.3.6}$$

$$\begin{aligned}\frac{\partial^2 \ell(\sigma,\xi)}{\partial \xi^2} &= \frac{1}{\xi^3}\left\{-2\sum_{i=1}^{k} \log\left(\frac{\sigma + \xi\, y_i}{\sigma}\right) + 2\xi \sum_{i=1}^{k} \frac{y_i}{(\sigma + \xi\, y_i)}\right. \\ &\quad \left. + \xi^3 \sum_{i=1}^{k} \frac{y_i^2}{(\sigma + \xi\, y_i)^2} + \xi^2 \sum_{i=1}^{k} \frac{y_i^2}{(\sigma + \xi\, y_i)^2}\right\}\end{aligned} \tag{6.3.7}$$

The proof of these results is given in Appendix A. As mentioned, this analytical form of the Hessian matrix eliminated 'singularity' problems due to the numerical evaluation of the Hessian.

### 6.3.3 Adapted Log-likelihood Function and Hessian Matrix when $\xi = 0$

In Section 6.3.1, we saw that we needed to take care when defining the likelihood $\ell(\sigma,\xi)$ when $\xi = 0$, and so we defined $\ell(\sigma)$ in (6.3.2), where $\ell(\sigma) = \lim_{\xi \to 0} \ell(\sigma,\xi)$, by using Taylor's Theorem:

$$\ell(\sigma) = -k \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{k} y_i. \tag{6.3.8}$$

It is also possible to use Taylor's Theorem to obtain expressions for the second derivatives of $\ell$ as $\xi \to 0$ and hence for the limiting Hessian. As an example we have

$$\lim_{\xi \to 0} \frac{\partial^2 \ell(\sigma, \xi)}{\partial \xi^2} = \frac{1}{3\sigma^3} \sum_{i=1}^{k} y_i^2 (3\sigma - 6\xi y_i - 2y_i). \tag{6.3.9}$$

We used these limiting version of the log-likelihood function and associated Hessian matrix in our numerical optimization algorithm.

### 6.3.4 New Boundary Conditions on Second Derivatives of the GPD Likelihood Function

During our investigation and implementation of the maximum likelihood method for estimation of the parameters $\xi$ and $\sigma$, it was found that numerical optimization of the likelihood suffered problems due to the boundary conditions associated with these parameters:

$$1 + \frac{\xi y_i}{\sigma} > 0, \; i = 1, \ldots, k_0 \tag{6.3.10}$$

If the maximum likelihood estimate were near the boundary defined by (6.3.10) it was sometimes impossible to evaluate a positive definite approximate VC matrix. Figures 6.1 shows the location of the maximum likelihood estimate for a case when it is situated well inside the region defined by (6.3.10). Figure 6.2 on the other hand, shows a maximum likelihood estimate just inside the region defined by (6.3.10). In this case the numerical form of the approximate VC matrix was not positive definite. As an additional problem, if $\xi < -0.5$ we found that our procedure was unable to establish the approximate VC matrix correctly, as explained in Section 6.3.1.

In order to overcome all these problems we imposed different constraints on the values of the parameters $\sigma$ and $\xi$ when optimizing $\ell(\sigma, \xi)$. These were driven by the need for an approximate VC matrix that was positive definite. We restricted our numerical optimization routine to values of $\sigma$ and $\xi$ for which the Hessian was negative definite, or, equivalently, the approximate VC matrix was positive definite. This eliminated the problems that we encountered up to this point. The benefit of this approach is that we no longer need to consider the categorization over values of $\xi$ given in Section 6.3.1, as our new constraint

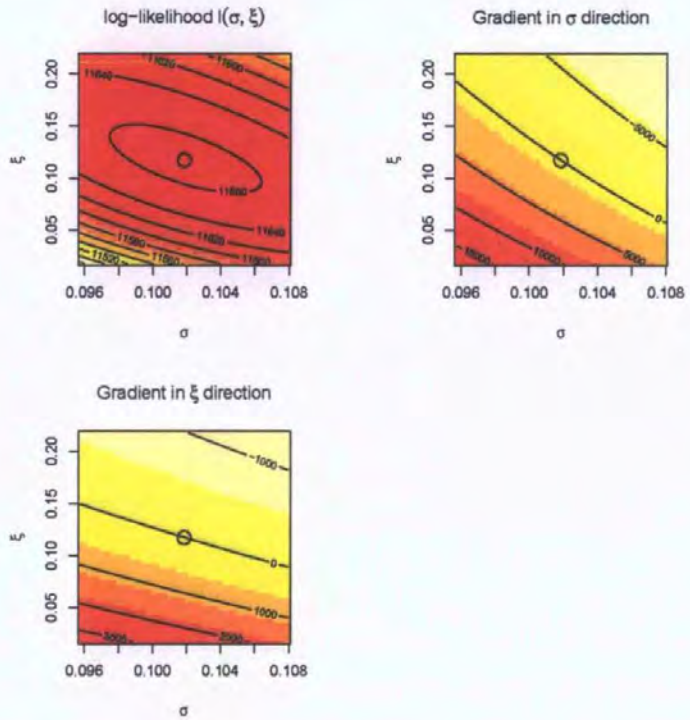



Figure 6.1: Plots corresponding to the case where the maximum likelihood estimates are within the boundary conditions given by (6.3.10). The log-likelihood $\ell(\sigma, \xi)$ together with its gradients $\frac{\partial}{\partial\sigma}\ell(\sigma, \xi)$ and $\frac{\partial}{\partial\xi}\ell(\sigma, \xi)$ in the $\sigma$ and $\xi$ directions are shown. The maximum likelihood estimate is indicated by the circle. In this case the numerical form of the approximate VC matrix was positive definite.

ensures that asymptotic properties are maintained.

## 6.4 Applied Examples

In this section we will apply our automated threshold selection methodology to the HR Wallingford Coastal Wave data set introduced in Chapter 1 and to the Offshore Wave data set from Zacharioudaki (May 2008) to illustrate the successful application of our technique to both types of data.

### 6.4.1 Application to Univariate Coastal Wave Data

We now apply the method presented in Section 6.2 to a real data set. The data used in this example relate to conditions near the Selsey Bill area (Hawkes, personal communication), as discussed in Chapter 1. These data were generated using the hindcast technique (see Reeve

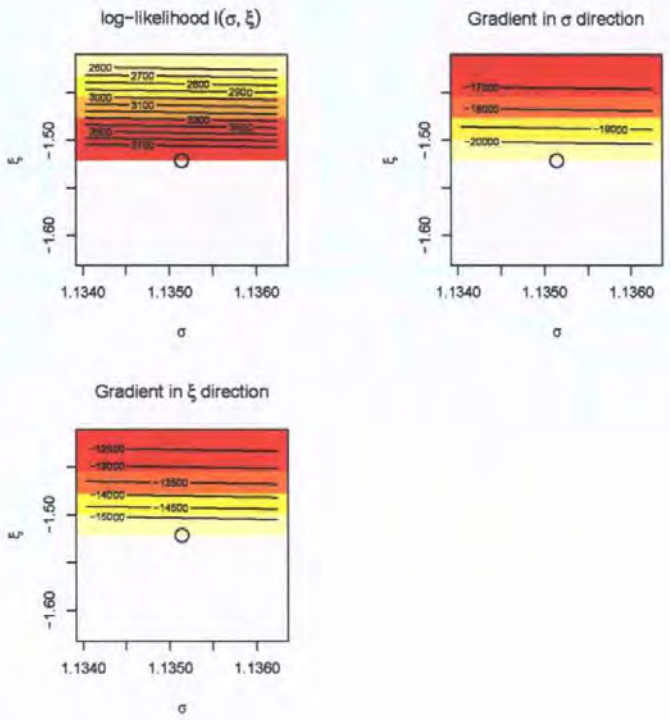



Figure 6.2: The same functions as in Figure 6.1. The maximum likelihood estimate is again indicated by a circle and lies very near the boundary defined by (6.3.10). In such cases the numerical form of the approximate VC matrix may not be positive definitive.

et al. (2004)) based on wind records. The data set consists of hourly hindcast measurements of the variables significant wave height, wave period and wave direction over a time span of 27 years. Wave hindcasting attempts to create the wind-wave conditions, and cannot account for the swell component. In this example we take a random sample of 10,000 observations from the data set. This random sample has the same structure as the full data set; we take a sample to reduce computational processing time significantly. The resulting values are typical of data that are collected in similar studies and can be thought of as satifying the independence assumption that underlie maximum likelihood theory, as discussed in Section 2.1.2. A plot of wave height (in metres) against the cosine of wave direction is shown in Figure 6.3. Our automated threshold selection technique was applied to these wave height observations and indicated 0.487 m as a suitable threshold. This threshold is also shown in Figure 6.3. The values of the cosine of wave direction were not used in finding this threshold. Figure 6.4 plots differences $\tau_{u_j} - \tau_{u_{j-1}}$ against threshold $u_{j-1}$, and as described in Section 6.2.1 is the basis of our threshold selection procedure. Figure 6.5

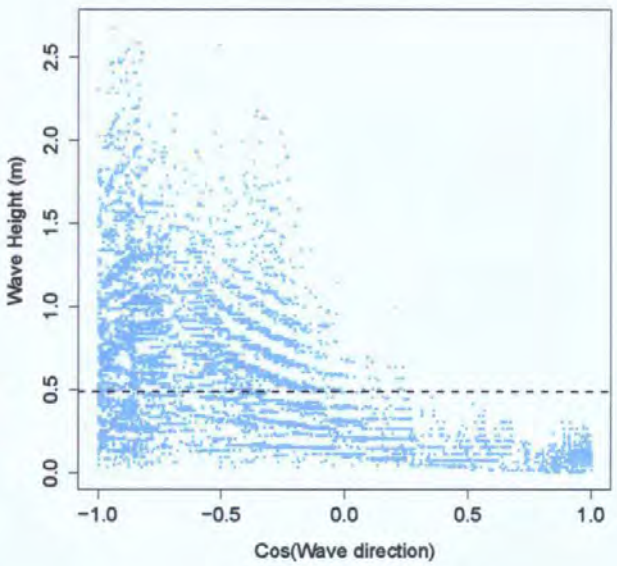



Figure 6.3: Scatter plot of wave height against the cosine of wave direction for 10,000 values from the Selsey Bill Coastal Wave data set. The horizontal line was obtained by applying our automated threshold selection procedure to the wave height observation, taking no account of the cosine of wave direction.

shows diagnostic plots, as discussed by Coles (2001) and produced by the `ismev` package of Coles and Stephenson (2006) run in `R` (R Development Core Team (2008)). These diagnostic plots indicate that the fitted GPD model is satisfactory. Both the probability and quantile plots show that there is little difference between empirical and fitted values from the model, indicating a good fit. Similarly, there is reasonable agreement between the data and the estimated return levels and associated 95% confidence envelope, and between the histogram of the data values above the chosen threshold and the fitted GPD density. This example shows that our proposed methodology can provide an automated, simple and computationally inexpensive threshold selection method that avoids the need for subjective interpretation of threshold choice plots with all their associated errors.

### 6.4.2 Application to Bivariate Coastal Wave Data

The application of the automated threshold selection techique to a bivariate data set is straightforward. The univariate automated threshold selection procedure is applied to each margin separately. The chosen thresholds are then fed into a bivariate model as discussed

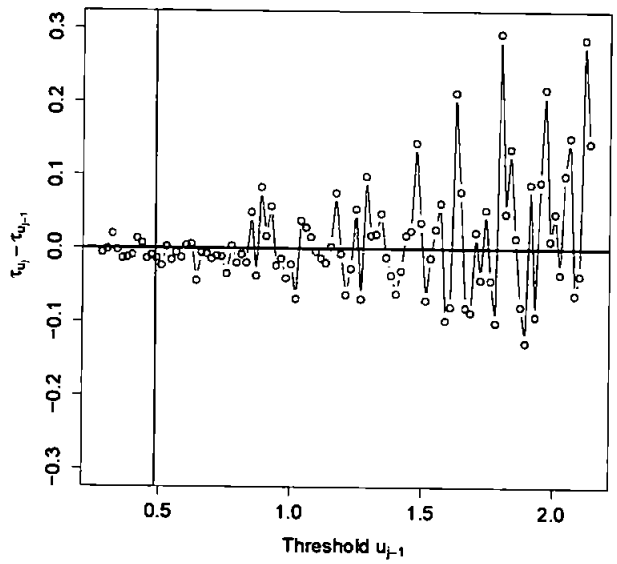



Figure 6.4: Graph of the differences $\tau_{u_j} - \tau_{u_{j-1}}$ against threshold $u_{j-1}$ for the wave height data. The vertical line indicates the automated threshold selection choice.

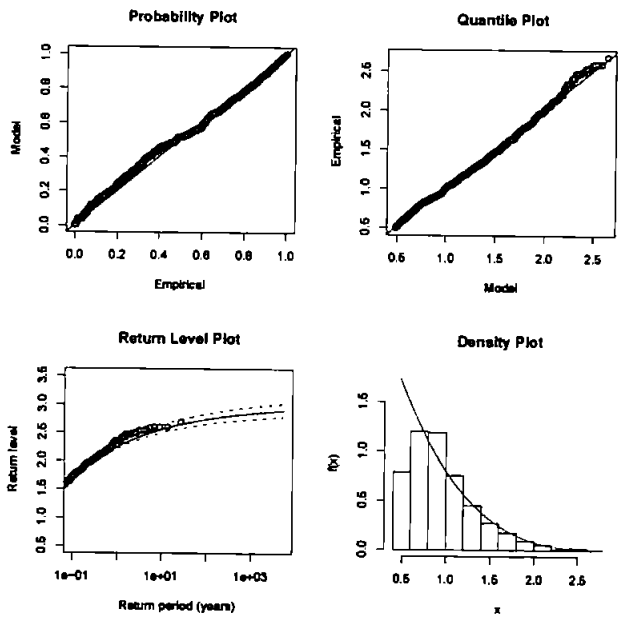



Figure 6.5: Diagnostic plots for the GPD fit when the threshold is chosen using our automated threshold selection approach applied to the wave height data.

in Section 2.3. Essentially the benefits from the univariate procedure are transferred to the bivariate case, improving the accuracy of the definition of exceedances and reducing the dependency on interpretation of diagnostic plots for threshold choice. Figure 7.16 presents

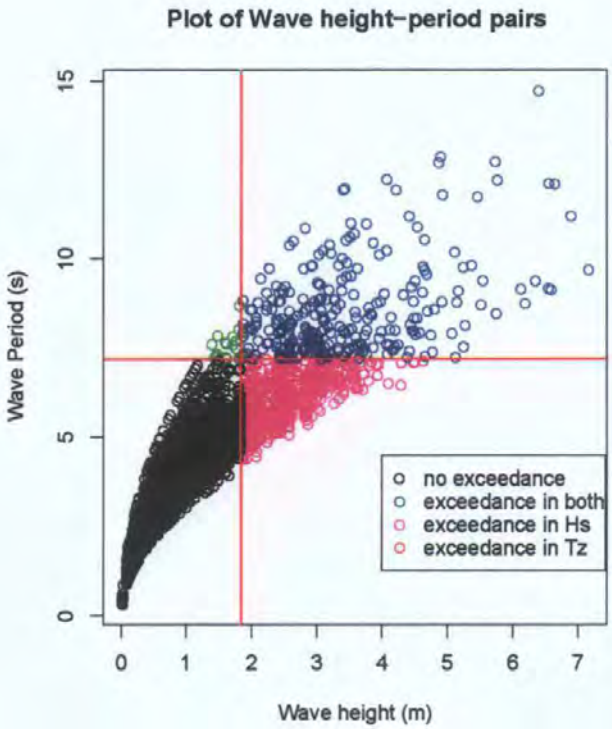



Figure 6.6: Wave period against wave height with associated marginal thresholds. This defines four distinct regions for all the data.

an example of the automated threshold selection technique applied to bivariate Coastal Wave condition data. The four regions identified corresponding to those of Figure 2.7; see Section 2.3 for discussion of the associated modelling uncertainty.

## 6.5 Performance and Uncertainty Assessments

### 6.5.1 Using Bootstrap Percentile Intervals to Assess Return Level Uncertainty

Uncertainty associated with inferences from the GPD model can depend on two sources: firstly, the uncertainty associated with estimating the scale and shape parameters from the available exceedances; secondly, the uncertainty associated with the selection of the threshold that defines these exceedances. Uncertainty in parameter estimation can be relatively small in comparison to the uncertainty in the choice of threshold. It is therefore important when

discussing inferential process to include the effect of the uncertainty associated with threshold choice.

As we saw in Chapter 2, return levels play a vital role in coastal engineering; see Section 2.2 and page 82 of Coles (2001) for a detailed discussion about the estimation of return levels and approximate confidence intervals from GPD fits. Standard software programmes, such as the `ismev` package (Coles and Stephenson (2006)), estimate return levels and approximate confidence intervals, as shown in Figure 6.5, but do not take into account uncertainty due to threshold selection.

Tancredi et al. (2006) present a review of existing model based methodology to account for threshold uncertainty in GPD models, and then introduce their own technique. In contrast to conventional fixed threshold methods, Tancredi et al. (2006) work in the Bayesian framework and assume that the threshold is one of the parameters about which to make inference. To overcome the lack of a natural model below the threshold and to avoid over-restrictive parametric assumptions, they propose a flexible mixture of an unknown number of uniform distributions with unknown range for below-threshold data; we will adopt a somewhat similar approach for our simulation study in Section 6.5.2. They consider it reasonable to expect different estimates of return levels and precision of estimates for different thresholds. This essentially leads to a Bayesian mixing of all reasonable threshold values and parameter estimates to determine an overall estimate of return levels and their uncertainty. Their approach is, however, highly computationally intensive, requiring the use of a reversible jump Markov chain Monte Carlo algorithm to cope with the unknown number of uniform distributions used for below-threshold modelling; see Green (1995). It also requires a number of prior assumptions to be made, although Tancredi et al. (2006) argue that return level estimation is more robust to these assumptions than to threshold choice in a fixed approach. Because of these drawbacks, we take a different approach to assess return level uncertainty based on the bootstrap procedure. Mooney and Duval (1993) and Efron and Tibshirani (1993) provide a basic summary of this procedure as follows:

1. Set $b = 1$.

2. Draw a simple random sample of size $m$ from the original data set $y_1, \ldots, y_m$ with replacement. We call this a bootstrap sample.

3. For the bootstrap sample, calculate the quantity of interest, here a specific return level, and call it $\hat{\theta}_b^\star$. We calculate the return level by first estimating the threshold using the methodology in Section 6.2. We then make use of this threshold when estimating the GPD model. Finally, we use the GPD parameter estimates to calculate a return level estimate.

4. Increase $b$ by 1 and repeat steps 2 and 3 a total of $B$ times, where $B$ is a large number. We set $B = 1000$.

5. Construct a probability distribution by attaching a $1/B$ probability to each point, $\hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_B^\star$.

Uncertainty in the quantity of interest – in this case a specific return level – can be quantified by summarizing this probability distribution using a confidence interval. More precisely, we will use a bootstrap percentile interval. To obtain an $(1-\alpha)$-level interval we sort the $B$ values $\hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_B^\star$ in ascending order and select the $(\frac{\alpha}{2}B)^{\text{th}}$ and $(1-\frac{\alpha}{2})B^{\text{th}}$ values as our confidence interval using the integer below and the integer above if these values are not themselves integers. We set $\alpha = 0.05$, yielding 95% confidence intervals.

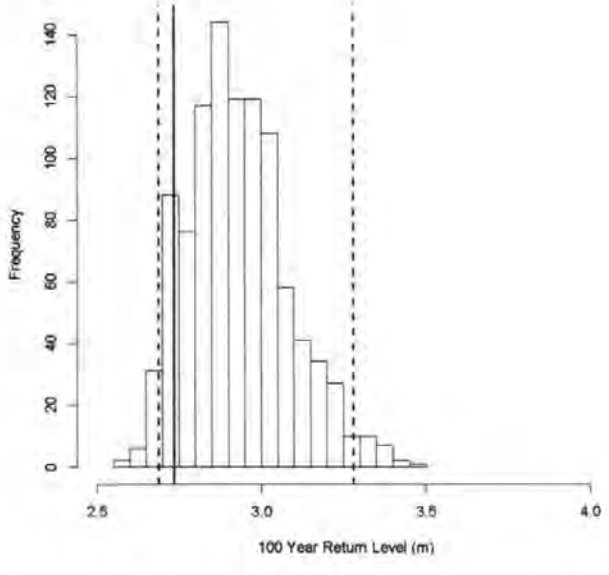



Figure 6.7: Histogram of the bootstrapped 100 year return levels and associated 95% bootstrap percentile interval ($B = 1000$ bootstrap iterations). The dashed lines are the percentile interval and the solid line is the return level based on the original data.

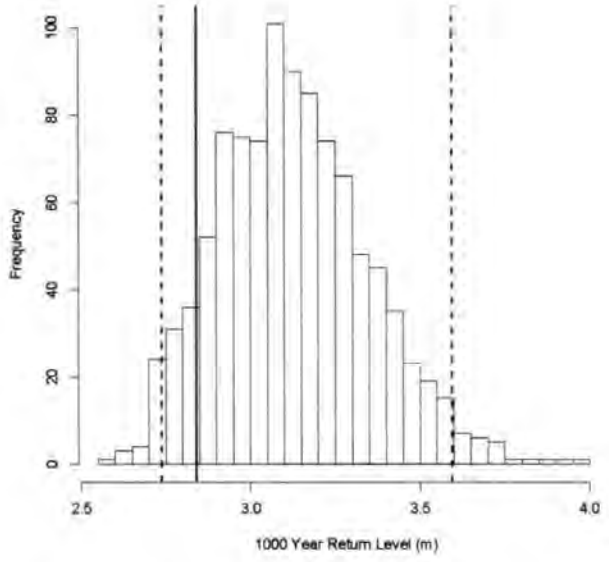



Figure 6.8: Histogram of the bootstrapped 1000 year return levels and associated 95% bootstrap percentile intervals. The dashed lines are the percentile interval and the solid line is the return level based on the original data.

We now present the result of applying the above bootstrap methodology to our Coastal Wave data set. Figure 6.7 shows a histogram of the bootstrapped 100 year return levels $\hat{\theta}_1^\star, \ldots, \hat{\theta}_B^\star$ and the associated bootstrap percentile interval. Figure 6.8 is an analogous plot for the 1000 year return level. These percentile intervals enable us to quantify the uncertainty in return level estimation in an accurate way, without ignoring threshold choice uncertainty and relying on the standard asymptotic theory outlined in Section 2.2.6 and on page 82 of Coles (2001). Figures 6.7 and 6.8 show that the bootstrap percentile interval widths are approximately 0.6 m for the 100 year wave height return level and 0.8 m for the 1000 year return level, indicating that uncertainty about these estimates is not particularly large from an engineering point of view.

### 6.5.2 Simulation Study to Assess the Performance of our Automated Threshold Selection Method

In this section we investigate the performance of our automated threshold selection method by means of a simulation study. Figure 6.9 shows a histogram of a data set comprising 10,000

simulated values of a random variable $X$ with distribution function given by

$$F(x) = \{(1-\beta)G_1(x) + \beta\}I[x > u] + G_2(x)I[x \leq u], \quad x > 0, \tag{6.5.1}$$

where $I$ is the usual indicator function and $\beta = \mathrm{P}(X \leq u)$. $G_1(x)$ is a GPD function with associated density function

$$g_1(x) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-(1/\xi+1)}, \quad x > u, \quad 1 + \frac{\xi(x-u)}{\sigma} > 0; \tag{6.5.2}$$

$G_2(x)$ is a truncated normal distribution function with associated density function

$$g_2(x) = \frac{\frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(x-\gamma)^2}{2\alpha^2}\right)}{\int_0^\infty \frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(x-\gamma)^2}{2\alpha^2}\right)dx}, \quad x > 0. \tag{6.5.3}$$

With this $F$, the distribution of the random variable $X$ can be thought of as a mixture of a normal distribution truncated on $(0, u]$ and a GPD on $(u, \infty)$ with weights $\beta$ and $1-\beta$, with non-extreme values coming from the truncated normal and extreme values from the GPD; this is somewhat similar to the model assumed by Tancredi et al. (2006). Given $\beta$ and the parameters $\gamma$ and $\alpha$ of $g_2$, we can find $u$ from the condition

$$\begin{aligned} \beta &= \Pr(X \leq u) = G_2(u) = \int_0^u g_2(x)dx \\ &= \frac{\int_0^u \frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(y-\gamma)^2}{2\alpha^2}\right)dy}{\int_0^\infty \frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(y-\gamma)^2}{2\alpha^2}\right)dy}. \end{aligned} \tag{6.5.4}$$

For the simulated data set shown in Figure 6.9 we set $\beta = 0.9$, $\gamma = 2$ and $\alpha = 0.7$, and solved for $u$ to obtain $u = 2.90$. We choose the parameter $\sigma$ of the GPD so that there was no discontinuity at $u$ in the probability density function of $X$. To do this we require

$$\beta g_2(u) = (1-\beta)g_1(u) = \frac{1-\beta}{\sigma}. \tag{6.5.5}$$

With $u = 2.90$, this equation can easily be solved to yield $\sigma = 0.40$. We set the shape parameter $\xi$ of the GPD to be 0.2. The resulting probability density function of $X$ is shown

in Figure 6.9, together with the threshold $u = 2.90$ (dotted line).

A random sample $x_1, \ldots, x_N$ can be simulated from $F$ as follows:

- Set $i = 1$. Simulate $y \sim N(\gamma = 2, \alpha^2 = 0.7^2)$;
- If $y < 0$ reject it;
- else if $0 < y < u$, set $x_i = y$ and increase $i$ by 1;
- else if $y > u$ simulate $x \sim \text{GPD}(u = 2.90, \sigma = 0.4, \xi = 0.2)$, set $x_i = x$ and increase $i$ by 1.
- Stop when $i = N + 1$.

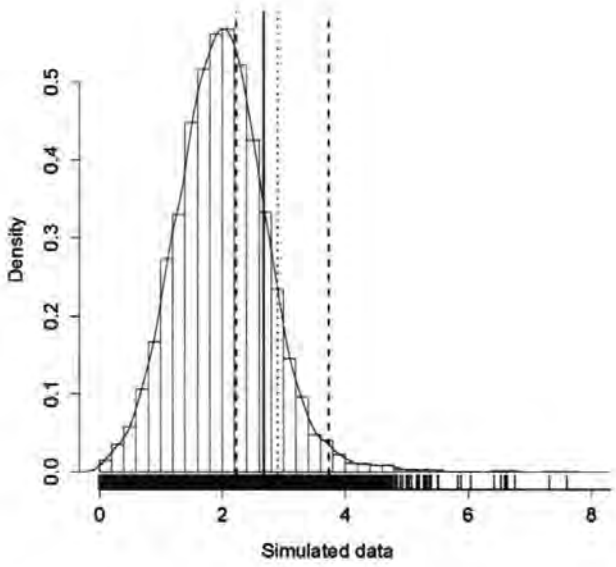



Figure 6.9: Histogram of a data set of 10,000 simulated values of a random variable $X$ with distribution function $F$. The associated probability density function is also shown. The individual values are indicated by a rug of dashes. Our automated threshold choice is indicated by a solid line, with the true threshold $u = 2.90$ being shown by a dotted line. The 95% bootstrap percentile intervals is also presented using dashed lines.

We applied our automated threshold selection method to the simulated data set of size $N = 10,000$ shown in Figure 6.9. The selected threshold took the value 2.678 and can be seen to be close to the true value of $u = 2.90$. We next used the above simulation procedure to generate 1000 random samples of size $N = 10,000$ from $F$. We applied our

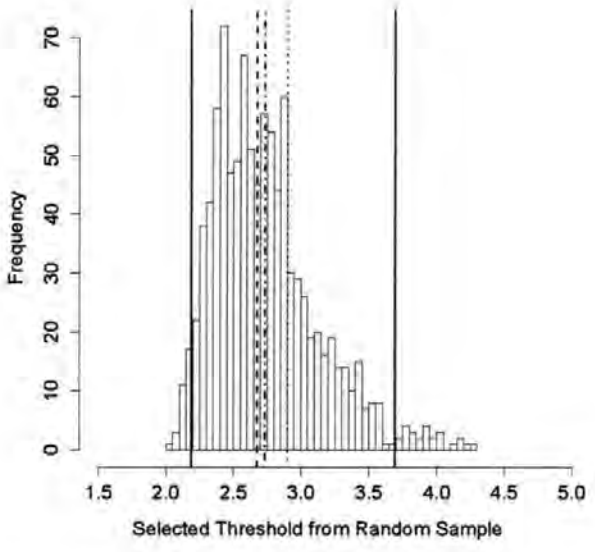



Figure 6.10: Histogram of thresholds selected from 1000 random samples of size $N = 10,000$ from $F$. The mean and median of the automated threshold choices for the simulated data sets are shown by dot-dashed and dashed lines respectively; while the true threshold $u = 2.90$ is shown by a dotted vertical line. The 2.5% and 97.5% quantiles are shown as the outer solid lines.

threshold selection technique to each random sample; a histogram of these 1000 thresholds, together with 2.5% and 97.5% quantiles (2.189,3.694), the true threshold $u = 2.90$ and mean $u_{mean} = 2.73$ and median $u_{med} = 2.67$ values of the distribution of estimated thresholds are shown in Figure 6.10. The selected thresholds seem to be evenly and not very widely spread around the true threshold, suggesting that our method can recover a known threshold to a good degree of accuracy. Our method performed similarly well when applied to data sets simulated using different values of $\beta$, $\gamma$, $\alpha$ and $\xi$.

We now focus on the simulated data set shown in Figure 6.9 and repeat the bootstrap analysis discussed in Section 6.5.1, except that our bootstrap quantity of interest $\hat{\theta}_b^*$ now becomes selected threshold instead of a specific return level. Figure 6.11 shows a histogram of the bootstrap threshold choices together with the 95% bootstrap percentile interval (2.225,3.732), our automated threshold choice of 2.678 for the original simulated data set, mean $u_{mean} = 2.75$ and median $u_{med} = 2.68$ values of the distribution of estimated thresholds and the true threshold $u = 2.90$. The 2.5% and 97.5% quantiles found above have also been added. The 95% bootstrap interval is also shown in Figure 6.9. We can see from these

plots that the 95% bootstrap percentile interval is not very wide and contains the true and selected thresholds. The actual interval values of $(2.225, 3.732)$ compare well with the 2.5% and 97.5% quantiles $(2.189, 3.694)$ indicating that the bootstrap assesses well the uncertainty associated with our threshold choice procedure.

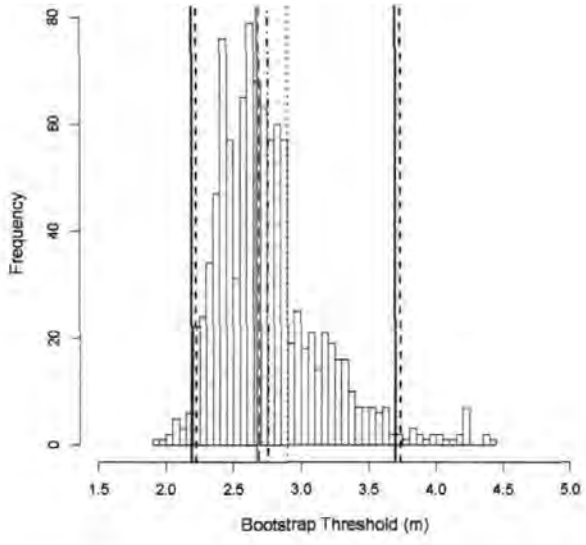



Figure 6.11: Histogram of the bootstrap threshold choices. The automated threshold choice of 2.678 for the original simulated data set is shown as the solid red line. The mean and median of the automated threshold choices for the simulated data sets are shown by dot-dashed and dashed lines respectively; while the true threshold $u = 2.90$ is the dotted line. The 95% bootstrap percentile interval is shown as the dashed lines, with the 2.5% and 97.5% quantiles from Figure 6.10 being given using the outer solid lines.

The conclusion of this simulation study is that our automated and computationally inexpensive procedure can recover a theoretical threshold from simulated data to a good degree of accuracy and that the bootstrap can be successfully used to assess the uncertainty associated with this procedure. In the next section we give a further example of the application of our procedure by comparing it to an existing technique utilized in the JOINSEA software.

Table 6.1: The chosen threshold, number of exceedances, GPD parameter estimates and standard errors from our new automated threshold selection method and the approach adopted in the JOINSEA software.

| | New Technique | JOINSEA |
|---|---|---|
| Threshold Value | 0.487 | 1.480 |
| Number of Exceedances | 5372 | 497 |
| Maximum Likelihood Estimate, $\xi$ | $-0.230$ | $-0.271$ |
| Maximum Likelihood Estimate, $\sigma$ | 0.576 | 0.405 |
| Standard Error, $\xi$ | 0.00952 | 0.04094 |
| Standard Error, $\sigma$ | 0.00940 | 0.02409 |

### 6.5.3 Comparison of our Automated Threshold Selection Techniques with the Approach Used in the JOINSEA Software

We present a review of the JOINSEA software in Chapter 3; see also Wallingford (1998b). In this section we compare our automatic threshold selection technique with an existing approach used in the JOINSEA software. The JOINSEA approach for choosing an appropriate threshold assumes that exceedances can be identified for GPD modelling as values greater than the 95% quantile. We now use the Selsey Bill Coastal Wave data set to compare our choice of threshold and fitted GPD with those obtained from the approach adopted in JOINSEA. Table 6.1 gives the results from the two approaches.

Figure 6.12 shows again a scatter plot of wave height against the cosine of wave direction for the Selsey Bill Coastal Wave data set, together with the two thresholds. The dashed line was obtained using our new threshold technique, while the solid line is the JOINSEA threshold. If we fit two GPD models to the wave height exceedances defined by each threshold then we obtain the results given in Table 6.1.

We see from Table 6.1 and Figure 6.12 that the threshold values are very different, with the automated threshold being almost 1m below the JOINSEA threshold. Figures 6.13 and 6.14 show comparisons of inferences (return levels, confidence intervals and fitted densities) from the fitted models based on each threshold. We can see that the resulting models are actually very similar indicating that our automated threshold selection technique is comparable to that of JOINSEA. The JOINSEA threshold yields fewer exceedances, which is the cause of the increased return level confidence interval widths in Figure 6.13. The

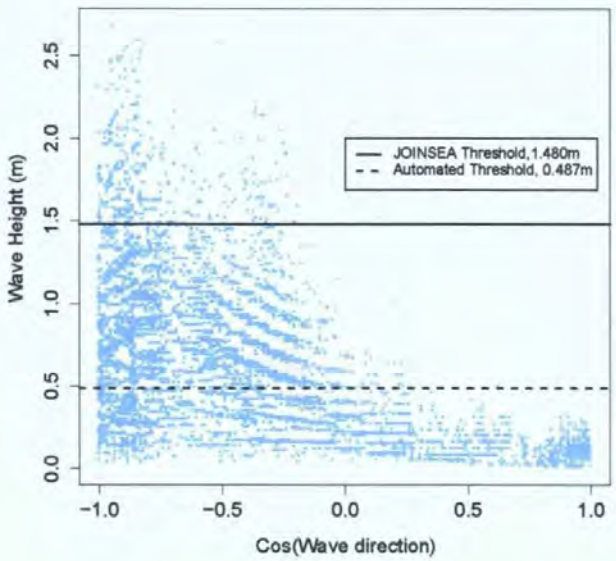



Figure 6.12: Scatter plot of wave height against the cosine of wave direction for 10,000 values from the Selsey Bill Coastal Wave data set. Our automated threshold choice is shown using the dashed line, while the solid line shows the threshold chosen by the JOINSEA software. Both threshold choices take no account of the cosine of wave direction.

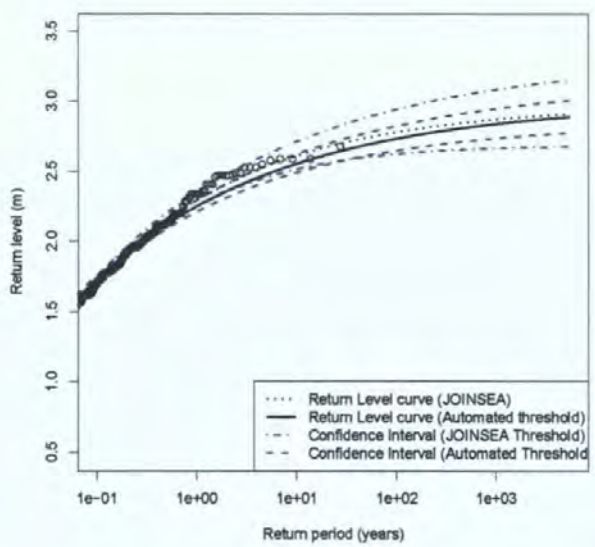



Figure 6.13: Returns level curves and confidence envelopes from both automated and JOINSEA threshold model fits to the Coastal Wave data.

narrower confidence intervals yielded by our threshold selection technique, together with the fact that it is more model based, lead us to prefer our methodology over the JOINSEA

approach. We also note that for data sets such as those simulated in Section 6.5.2 with $\beta > 0.95$ the JOINSEA approach is guaranteed to lead to non-extremes being included in future GPD analyses.

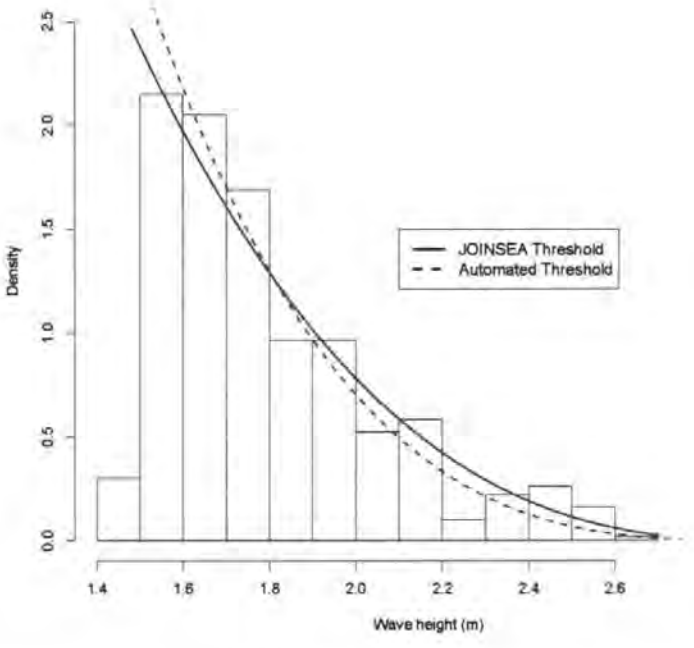



Figure 6.14: Histogram of the exceedances of the Coastal Wave data from the JOINSEA threshold choice, together with the GPD fit (solid line). The GPD fit based on our threshold procedure is also shown (dotted line). This GPD fit has been scaled so that the area under it above the JOINSEA threshold is one.

We applied our automated threshold selection technique to different data sets which varied in size and data collection location, and found it performed consistently well in terms of model goodness-of-fit.

We felt that in the case of the Selsey Bill Coastal Wave data our automated approach chose a relatively low threshold as a type of "average" threshold across the range of direction covariate values. This observation led us to extend our automated technique to allow the chosen threshold to vary with covariate value. We discuss our direction varying threshold methodology in detail in Section 6.6.

Table 6.2: The chosen threshold, number of exceedances, GPD parameter estimates and standard errors for our new automated threshold selection method and the approach adopted in the JOINSEA software.

| | **New Technique** | **JOINSEA** |
|---|---|---|
| Threshold Value | 1.746 | 1.97 |
| No. of Exceedances | 7635 | 4311 |
| Maximum Likelihood Estimate $\xi$ | 0.4101 | 0.4011 |
| Maximum Likelihood Estimate $\sigma$ | -0.076 | -0.085 |
| Standard Error $\xi$ | 0.0065 | 0.0085 |
| Standard Error $\sigma$ | 0.011 | 0.015 |

### Application to Univariate Offshore Wave data

We now a provide comparison between JOINSEA and our threshold selection technique based on Offshore wave data; for details on this data see Zacharioudaki (May 2008).

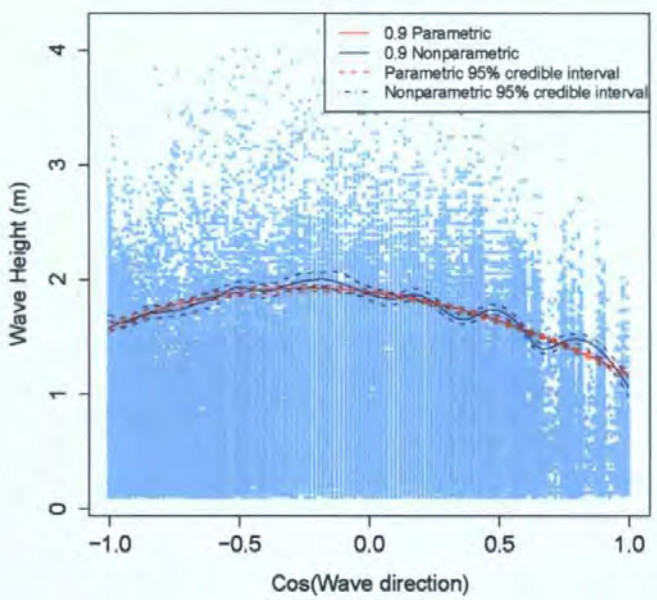



Figure 6.15: Scatter plot of Hs. Our automated threshold choice is shown using the dashed line, while the solid line shows the threshold chosen by the JOINSEA software. Both threshold choices take no account of the cosine of wave direction.

Figure 6.15 shows a scatter plot of wave height against the cosine of wave direction for the Offshore Wave data set, together with the two thresholds. The dashed line was obtained using our new threshold technique, while the solid line is the JOINSEA threshold. Fitting

the two GPD models to the wave height exceedances defined by each threshold yielded the results given in Table 6.2. Figures 6.16 and 6.17 show comparisons of inferences

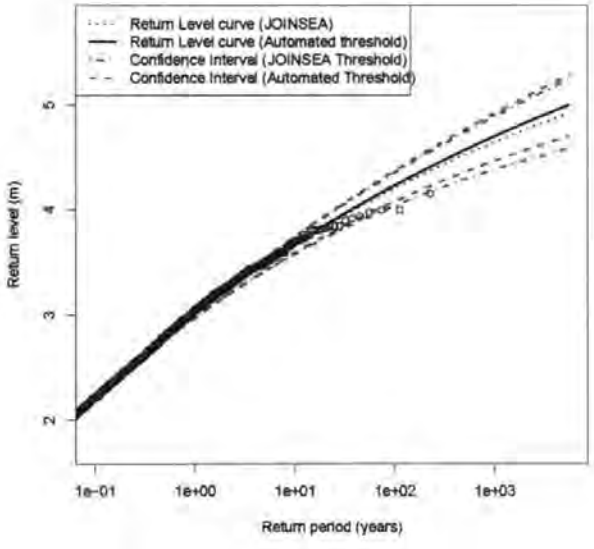



Figure 6.16: Returns level curves and confidence envelopes from both automated and JOINSEA threshold model fits to the Offshore Wave data.

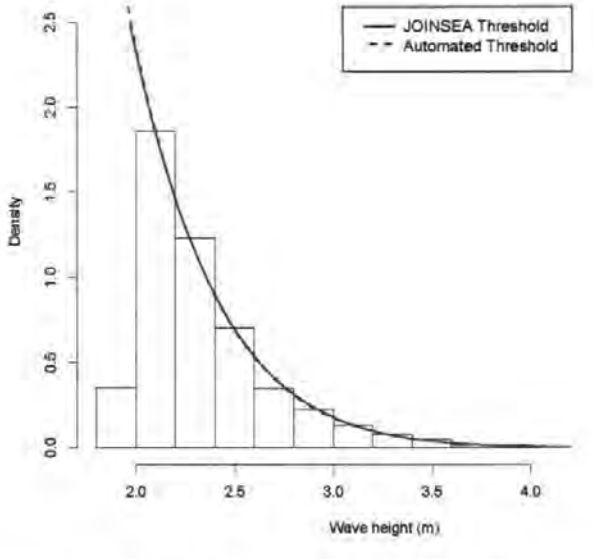



Figure 6.17: Histogram of the exceedances of the Offshore Wave data from the JOINSEA threshold choice, together with the GPD fit (solid line). The GPD fit based on our threshold procedure is also shown (dotted line). This GPD fit has been scaled so that the area under it above the JOINSEA threshold is one.

(return levels, confidence intervals and fitted densities) from the fitted models based on each threshold. Figure 6.16 shows that our automated threshold has a smaller return level confidence interval width than the JOINSEA threshold due to the JOINSEA approach again yielding fewer exceedances. Despite this minor difference we can see from Figure 6.17 that the resulting models perform almost identically indicating that our automated threshold selection technique is comparable to that of JOINSEA.

## 6.6 Extended Automated Threshold Selection Technique

We have seen that the Selsey Bill Coastal Wave data set comprises information about wave direction as well as wave height. So far we have worked only with wave height. It is clear from Figure 6.12 that the behaviour of wave height varies with wave direction. It therefore makes sense to include the directional effect in our automated threshold selection procedure, rather than to have a threshold that is constant over wave direction.

In extreme wave analysis directional effects are usually dealt with using one of two methods: either the data are split according to different directions with each separate data set being modelled independently, or the wave direction is included as a covariate as in Ewans and Jonathan (2006) and Jonathan and Ewans (2007), for example. In this section we propose a new approach to blocking the data.

Our approach is based on the automated threshold selection procedure that we have already presented and is as follows:

(1) First the data set is blocked according to the cosine of wave direction. The number of blocks is initially defined by the user; see Figure 6.18 for example where the covariate axis is split into 40 equal width blocks. Each block is then altered iteratively to its optimum size as described in (2).

(2) The constant automated threshold selection procedure is applied to the data in each block. The block size can then be altered in order to achieve a satisfactory GPD fit in each block. If there is not a sufficient number of observations within the block or if the block's optimal threshold choice does not define enough exceedances to achieve a good GPD fit, then the block is merged with the next consecutive block and the

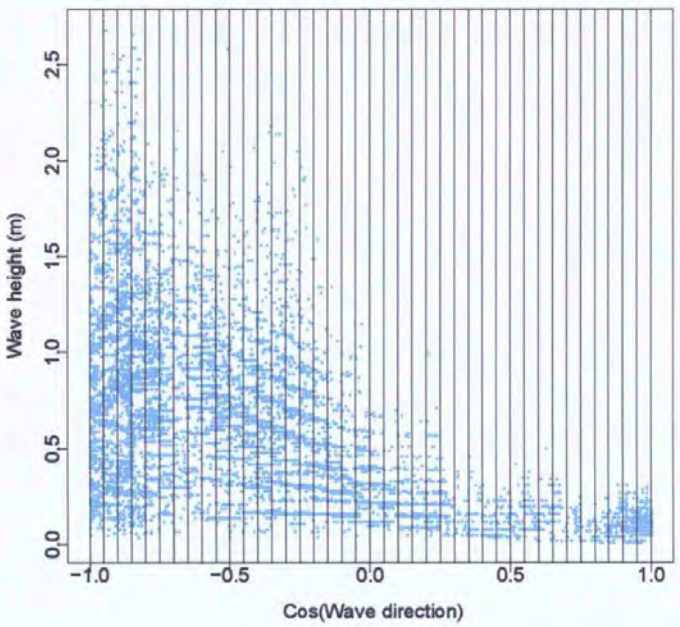



Figure 6.18: Scatter plot of wave height against the cosine of wave direction. The data has been split into 40 sections equally spaced along the covariate axis.

process is repeated. The merging of consecutive blocks is continued until the required minimum values for the number of observations and the number of exceedances for the merged block allows satisfactory fits to be reached. Our optimal blocks are shown in Figure 6.19. The minimum values that we used were determined through a simulation study by fitting a number of GPD models to different data sets and assessing the dependence of model fit quality on these values.

(3) Each block now has a constant optimal threshold associated with it. If these individual block thresholds are considered together a piecewise constant threshold function is defined. A threshold that is continuous in the cosine of wave direction covariate can be obtained by applying a smoothing spline, for example. We did this using the `smooth.spline` function of the R statistical programming language; see Green and Silverman (1994) and R Development Core Team (2008). The resulting smoothed direction varying threshold function is shown in Figure 6.20. The more appropriate thresholds that this extended automated threshold selection technique provides can yield more accurate direction specific return level estimates. These in turn can lead

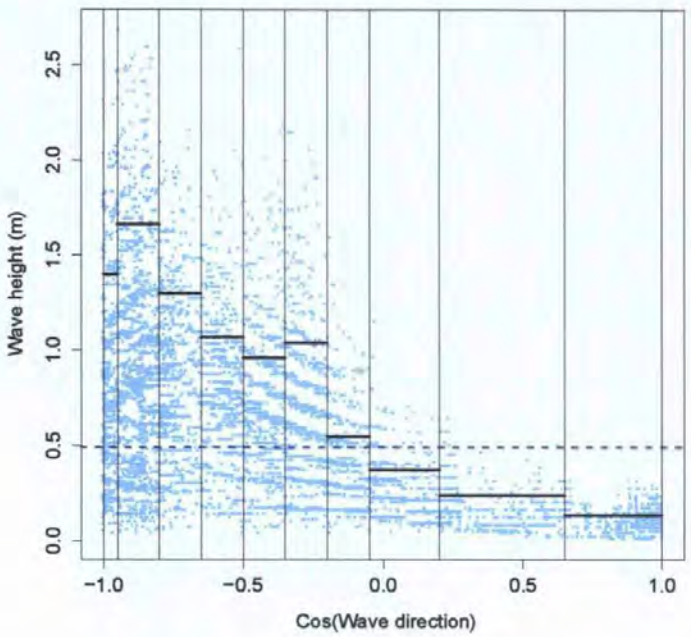



Figure 6.19: Scatter plot of wave height against the cosine of wave direction for the Coastal Wave data. The data has now been split into optimal blocks along the covariate axis. Individual automated thresholds have been chosen for each block and are shown by the solid horizontal lines. The dotted line shows the threshold chosen without reference to cosine of wave direction.

to improved coastal defence designs that account for directional variations in extreme wave heights.

In order to justify further the choice of these direction varying thresholds we show in Figure 6.21 probability density contours for a bivariate kernel density estimate (calculated using the `kde2d` function of the `MASS` library; see Venables and Ripley (2002)) based on wave height and the cosine of wave direction. We see that the chosen thresholds aline well with the tail of this probability density function across the range of cosine wave direction, supporting our direction varying threshold choice procedure.

## 6.7 Developed Software including Graphical User Interface

During the course of our work, we have made considerate use of the JOINSEA software, as discussed in Chapter 3. Although this software provides excellent routines for univariate and bivariate extreme value modelling, we feel that its use may present some difficulties

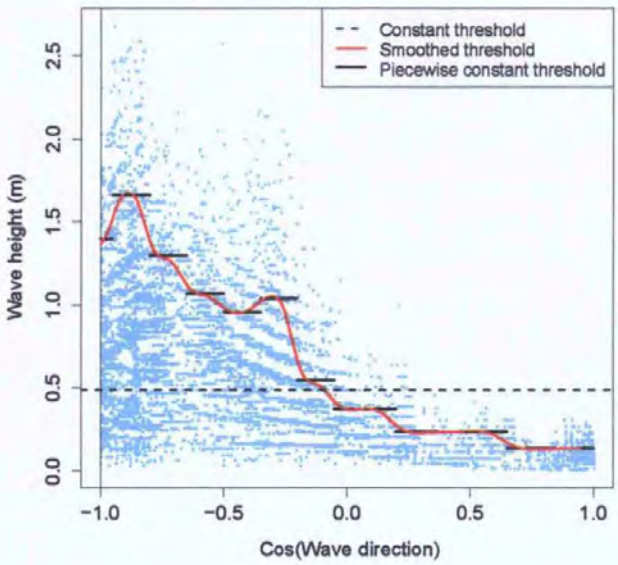



Figure 6.20: The bivariate Coastal Wave data with piecewise constant and smoothed covariate varying thresholds.

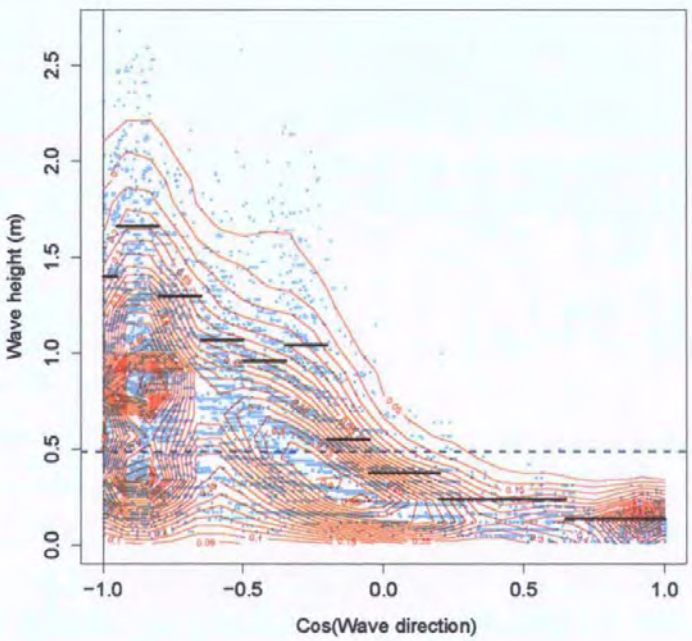



Figure 6.21: Probability density estimate contours overlaid on the scatter plot of wave height against cosine of wave direction. The thresholds selected by the extended automated threshold selection technique are shown using the solid lines.

for practitioners unfamiliar with joint probability techniques or FORTRAN programming. JOINSEA requires the user to input information in several stages to achieve an accurate joint probability analysis. This degree of control on the analysis is useful to the experienced/advanced users, but may limit the uptake of the software by beginner/intermediate users as their knowledge of the procedures will be limited.

These considerations motivated our investigation into automating some of the techniques used to produce a joint probability analysis, such as the new threshold selection techniques introduced in this chapter. A natural development of our methodological work was to implement these techniques into a user friendly interface which required minimum technical knowledge to operate and which produced both univariate and joint extremes analysis.

The software was written in R and utilized the TclTk and TclTk2 packages to create the Graphical User Interface (GUI). We now present and discuss some screen shots from the software windows.

Upon starting the software, the "Xsea" introductory screen appears, shown in the upper image of Figure 6.22. This screen gives version and creator information. Pressing the "Continue..." button takes the user automatically to the data input screen shown in the lower image of Figure 6.22.

The data input screen allows the user to load a text or Excel data file. The format of the required data is simple: for example, three columns of data each with a variable name at the top of each column. This simple format reduces the amount of user formatting to a minimum. Once the data is loaded, the user is able to produce time series plots of each marginal data by following the "Which margin do you want?" prompt. Also on the data input screen the user is asked to decide between "Univariate" or "Bivariate" analysis. The user us now taken to the next screen that depends on the choice of analysis made.

If the user chose univariate analysis, the program moves to the screen shown in the upper image of Figure 6.23. This screen asks the user to input the subject variable name, e.g. Hs, Wave Height or Wave Direction. The Wave Direction variable name is used if the user selected the use of a direction varying threshold. Otherwise, a constant threshold is selected which only requires the subject variable. Once this information has been inserted, the "start" button can be pressed and the software will begin its calculations.

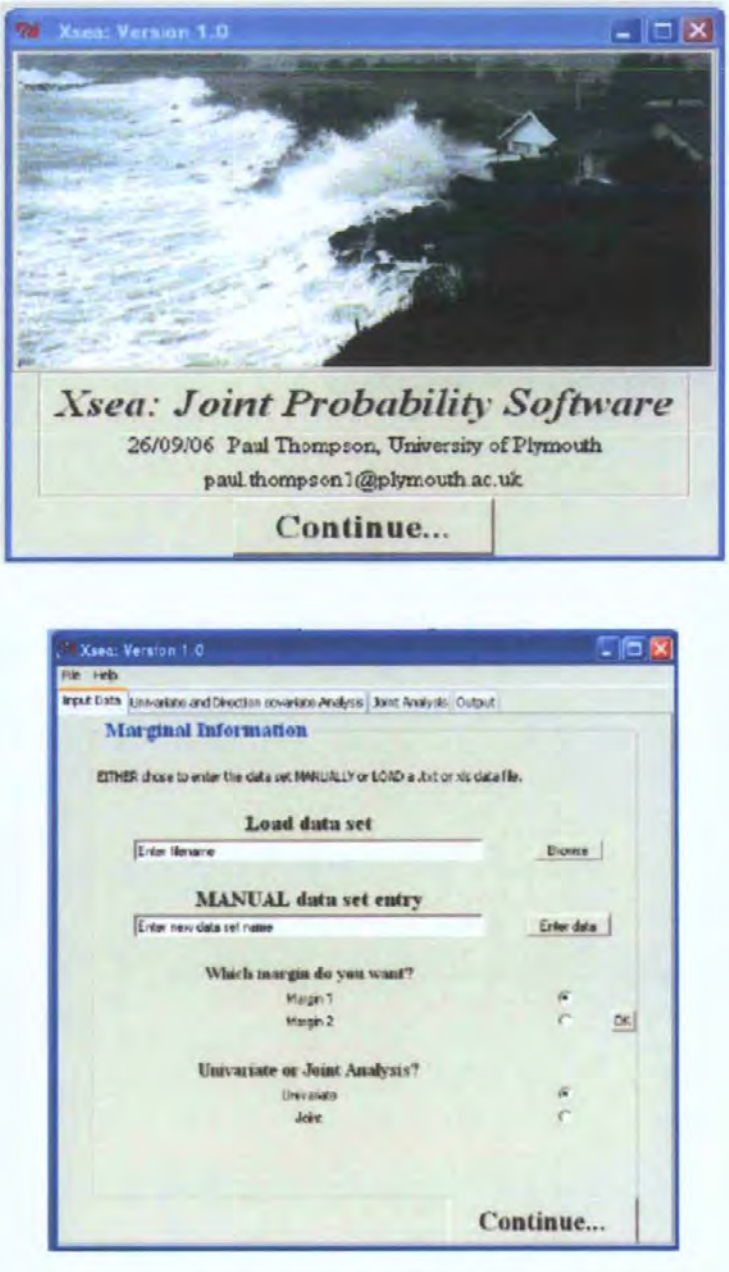



Figure 6.22: Screen shots from the Xsea GUI. The top image shows the introductory screen, while the bottom image shows the data input screen.

Alternatively, the user may chose to perform a bivariate analysis which will lead to the screen shown as the lower image of Figure 6.23. The user is given several options for data presentation and diagnostic plots. The user is also provided with the option to manually chose the dependence function as discussed in Section 2.3.3; if the automatic choice is selected, then the software will fit all the dependence functions and select the most appropriate based on log-likelihood values, as suggested by Tawn and Coles (1994). As


Figure 6.23: Screen shots from the Xsea GUI. The top image shows the univariate, while the bottom image shows the bivariate analysis screen.

before, after making these choices, the "start" button can be pressed and the software will begin its calculations.

The output graphs include probability, quantile and return level plots, histogram (with fitted model curve), plot of parameter estimate differences versus threshold, and displays of joint probability regions. Examples of this output would be figures seen in Section 6.4.

## 6.8 Summary

We have presented a new automated method for selecting the threshold for the GPD in extreme value modelling. We have shown the practical applicability of our method by presenting relevant examples for coastal and offshore wave data. Our method uses a series of normality tests to find an appropriate threshold choice for a given data set. We have carried out a simulation study to check the performance of our approach and have assessed the effect of the uncertainty associated with our method on return level estimation using the bootstrap procedure. The simulation study has shown that our automated technique can recover a known threshold from a simulated data set to a good degree of accuracy, and the bootstrap enables us to obtain bootstrap percentile intervals to assess the accuracy of our automated technique. We have also provided comparisons of our new approach with an existing technique implemented in the JOINSEA software, pointing out why we prefer our method.

We have extended our methodology to incorporate a direction covariate dependant threshold. This extension uses our automated threshold selection technique to segregates the data into optimal blocks based on goodness-of-fit and sample size requirements. Our methodology can lead to more accurate return level estimates, with their uncertainty properly quantified, which can inform and enhance the coastal design process. We have also made adaptations to the standard maximum likelihood based parameter estimation techniques that overcome some of the numerical difficulties that we encountered whilst developing our automated threshold selection methodology.

We have implemented much of this methodology in a friendly GUI. This allows a range of users to perform both univariate and bivariate extreme value modelling without knowledge of FORTRAN or the JOINSEA program.

# 7

# Bayesian Nonparametric Quantile Regression Using Splines

## 7.1 Introduction

In Chapter 4 we discussed Quantile Regression, mentioning both the nonparametric and Bayesian approaches. There we saw that the overall aim was to estimate the conditional $p^n$ quantiles of a variable $Y$ given the value of a covariate $\boldsymbol{X}$: $Q_p(Y|\boldsymbol{X} = \boldsymbol{x})$. Furthermore we reviewed combinations of these approaches which give rise to more complex techniques, such as Bayesian Quantile Regression as presented in Yu and Moyeed (2001). In this chapter we present a new method that we call Bayesian non-parametric quantile regression using splines. The technique links elements of existing Bayesian quantile regression (Yu and Moyeed (2001)) and nonparametric regression using splines.

The Bayesian quantile regression (BQR) methodology developed in Yu and Moyeed (2001) adopts a parametric approach based on polynomial quantile functions; see section 4.4.3 for an example. Although Yu and Moyeed (2001) present excellent results, there are certain drawbacks associated with using polynomials. These include the influence of outliers and the need to choose the degree of the polynomial, possibly for each quantile considered. Also, the data may have a limited local effect on the shape of a polynomial regression curve especially when modelling extreme quantiles. In this chapter we present a nonparametric alternative to the parametric approach of Yu and Moyeed (2001) based on

using natural cubic splines (NCS) as defined in Section 4.3.1 rather than polynomials. Our approach provides a more versatile and flexible method of fitting a quantile regression curve. As our technique utilizes much of the theory described in Chapter 4, hence we will frequently refer to particular sections from that chapter.

## 7.2 Bayesian Modelling and Inference

In this section we present a framework for Bayesian nonparametric quantile regression using splines rather than polynomials as in Yu and Moyeed (2001). In our approach we model a quantile function of a covariate $t$ using a NCS with $N$ knots at points $\tau_1, \ldots, \tau_N$ along the range of the covariate as defined in Section 4.3.1. The NCS is uniquely determined by its values $\boldsymbol{g} = (g_1, \ldots, g_N)^T$ at these knots, since, by Theorem 4.3.1.2, there is a unique NCS that can be drawn through the points $(\tau_i, g_i)$, $i = 1, \ldots, N$. As our approach is Bayesian, we begin by defining the prior density for $\boldsymbol{g}$ as multivariate normal; see Green and Silverman (1994), page 51 for a discussion about the use of the multivariate normal density as a prior in this context.

Our prior for $\boldsymbol{g}$ is defined by means of the multivariate normal density

$$\pi(\boldsymbol{g}|\lambda) = \frac{\lambda^{(N-2)/2}}{(2\pi)^{(N-2)/2}(\mu_1 \cdots \mu_{N-2})^{1/2}} \exp\left(-\frac{1}{2}\lambda\, \boldsymbol{g}^T K \boldsymbol{g}\right), \qquad (7.2.1)$$

in which $\mu_1, \ldots, \mu_{N-2}$ are the inverses of the $N-2$ non-zero eigenvalues of $K$, as defined in equation (4.3.2), and $\lambda > 0$ is an unknown parameter. More details about this multivariate normal distribution can be found in Rao (1973), page 528. Note that (7.2.1) depends on the roughness $\int_a^b g''(t)^2 dt = \boldsymbol{g}^T K \boldsymbol{g}$ of the NCS $g$ uniquely defined by $\boldsymbol{g}$; see Theorem 4.3.1.1. As larger values of $\lambda$ result in more probability density being given to less rough curves $g$, we will refer to $\lambda$ as a smoothing parameter.

We next require a prior on the smoothing parameter $\lambda$ which is constrained by a lower limit of zero. Hence, we follow standard practice by using the gamma distribution as our prior for $\lambda$ which takes the form

$$\pi(\lambda) = \frac{\lambda^{\alpha-1}\exp(-\lambda/\beta)}{\Gamma(\alpha)\beta^{\alpha}}, \quad \lambda > 0, \qquad (7.2.2)$$

in which $\Gamma$ is the usual gamma function. The user is able to specify the hyperparameters $\alpha$ and $\beta$. Under this prior $\mathrm{E}[\lambda] = \alpha\beta$ and $\mathrm{Var}[\lambda] = \alpha\beta^2$, results that can be used to guide hyperparameter choice.

The final step in our Bayesian approach is to define the likelihood of the data $(t_i, y_i)$, $i = 1, \ldots, n$, given $\boldsymbol{g}$. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. We proceed in accordance with the BQR approach of Yu and Moyeed (2001) by substituting our NCS $g$ for their polynomial:

$$L(\boldsymbol{y}|\boldsymbol{g}) = p^n(1-p)^n \exp\left\{-\sum_{i=1}^{n} \rho_p(y_i - g(t_i))\right\} \tag{7.2.3}$$

where $p$ is the probability corresponding to the quantile of interest, $0 < p < 1$, and $\rho_p$ is the standard loss function

$$\rho_p(u) = u(p - I(u < 0)) \tag{7.2.4}$$

in which $I$ is the usual indicator function. The values of $g(t_i)$, $i = 1, \ldots, n$, in (7.2.3) are uniquely determined by $\boldsymbol{g}$. We note that the likelihood is not dependent on $\lambda$. Combining $\pi(\lambda)$, $\pi(\boldsymbol{g}|\lambda)$ and $L(\boldsymbol{y}|\boldsymbol{g})$, we can write the posterior density function of $\boldsymbol{g}$ and $\lambda$ as

$$\pi(\boldsymbol{g}, \lambda|\boldsymbol{y}) \propto L(\boldsymbol{y}|\boldsymbol{g})\pi(\boldsymbol{g}|\lambda)\pi(\lambda). \tag{7.2.5}$$

by means of Bayes Theorem as discussed in Section 4.1. We now simulate realizations of $\boldsymbol{g}$ and $\lambda$ from this posterior density using an MCMC approach implemented via the Metropolis-Hastings algorithm; see Gamerman (1997) and Section 4.2. Our inferences will be based on these posterior realizations. In particular, we shall use equation (4.2.2) to approximate the posterior mean of $\boldsymbol{g}$ as $(\overline{g_1}, \ldots, \overline{g_N})$ yielding an estimate of $q_p(t)$, the $p^n$ quantile at $t$; again see Section 4.2 for a full discussion. Our algorithm can be summarized as follows:

(i) Assign initial values $\boldsymbol{g}^{(0)}$ and $\lambda^{(0)}$ to $\boldsymbol{g}$ and $\lambda$. We set $\boldsymbol{g}^{(0)}$to be the values at $\tau_1, \ldots, \tau_N$ of the posterior mean cubic quantile regression curve obtained using the methodology of Yu and Moyeed (2001); see Section 4.4.3. The cubic quantile regression curve was chosen as this is also an example of a cubic spline, although a very constrained one. We obtain the value of $\lambda^{(0)}$ by applying generalized cross validation (GCV) to the usual mean smoothing spline; see Green and Silverman (1994) and Section 4.4.3. We chose

this value, which we shall refer to as GCV(mean spline), because it can be found easily and quickly using R's (R Development Core Team (2008)) `smooth.spline` function (see Venables and Ripley (2002), for example). We set iteration number $j = 1$.

(ii) We generate a candidate vector $\boldsymbol{g}^*$ from the multivariate normal distribution

$$\boldsymbol{g}^*|\boldsymbol{g}^{(j-1)} \sim MVN(\boldsymbol{g}^{(j-1)}, \Sigma) \tag{7.2.6}$$

with mean $\boldsymbol{g}^{(j-1)}$ and variance-covariance matrix $\Sigma = \sigma^2 K^-/\lambda$, where $K^-$ is the generalized inverse of $K$. The constant $\sigma^2$ is specified by the user; see Section 7.4.

(iii) We then calculate the acceptance probability of a move from $\boldsymbol{g}^{(j-1)}$ to $\boldsymbol{g}^*$ which takes the form:

$$\begin{aligned} \alpha(\boldsymbol{g}^{(j-1)}, \boldsymbol{g}^*) &= \min\left\{1, \frac{\pi(\boldsymbol{g}^*, \lambda^{(j-1)}|\boldsymbol{y})q(\boldsymbol{g}^{(j-1)}|\boldsymbol{g}^*)}{\pi(\boldsymbol{g}^{(j-1)}, \lambda^{(j-1)}|\boldsymbol{y})q(\boldsymbol{g}^*|\boldsymbol{g}^{(j-1)})}\right\} \\ &= \min\left\{1, \frac{L(\boldsymbol{y}|\boldsymbol{g}^*)\pi(\boldsymbol{g}^*|\lambda^{(j-1)})q(\boldsymbol{g}^{(j-1)}|\boldsymbol{g}^*)}{L(\boldsymbol{y}|\boldsymbol{g}^{(j-1)})\pi(\boldsymbol{g}^{(j-1)}|\lambda^{(j-1)})q(\boldsymbol{g}^*|\boldsymbol{g}^{(j-1)})}\right\} \end{aligned} \tag{7.2.7}$$

where the proposal density $q(\boldsymbol{g}^*|\boldsymbol{g}^{(j-1)})$ is the probability density function of the multivariate normal specified in (7.2.6). In fact, because $q$ is symmetric in its arguments, it cancels out of (7.2.7).

(iv) A random variable $u$ is simulated from a uniform distribution $U(0,1)$. If $u \leq \alpha(\boldsymbol{g}^{(j-1)}, \boldsymbol{g}^*)$, then $\boldsymbol{g}^*$ is accepted by setting $\boldsymbol{g}^{(j)} = \boldsymbol{g}^*$, otherwise the chain does not move and $\boldsymbol{g}^{(j)} = \boldsymbol{g}^{(j-1)}$.

(v) We now generate a candidate $\lambda^*$ from the log-normal distribution as follows:

$$\mu^* \sim N(\log(\lambda^{j-1}), \sigma_\lambda^2) \tag{7.2.8}$$

$$\lambda^* = \exp(\mu^*), \tag{7.2.9}$$

where the normal distribution (7.2.8) has mean $\log(\lambda^{(j-1)})$ and variance $\sigma_\lambda^2$. The variance $\sigma_\lambda^2$ can be specified by the user; again see Section 7.4.

(vi) We then calculate the acceptance probability of a move from $\lambda^{(j-1)}$ to $\lambda^*$ which takes the form:

$$\begin{aligned}\alpha(\lambda^{(j-1)}, \lambda^*) &= \min\left\{1, \frac{\pi(\boldsymbol{g}^{(j)}, \lambda^*|\boldsymbol{y})q(\lambda^{(j-1)}|\lambda^*)}{\pi(\boldsymbol{g}^{(j)}, \lambda^{(j-1)}|\boldsymbol{y})q(\lambda^*|\lambda^{(j-1)})}\right\} \\ &= \min\left\{1, \frac{\pi(\boldsymbol{g}^{(j)}|\lambda^*)\pi(\lambda^*)q(\lambda^{(j-1)}|\lambda^*)}{\pi(\boldsymbol{g}^{(j)}|\lambda^{(j-1)})\pi(\lambda^{(j-1)})q(\lambda^*|\lambda^{(j-1)})}\right\} \qquad (7.2.10)\end{aligned}$$

where $q$ is the log-normal probability density function specified through (7.2.8) and (7.2.9). In this case cancellation of the $q$ terms in (7.2.10) is not possible as $q$ is not symmetric in its arguments.

(vii) A random variable $u$ is simulated from a uniform distribution $U(0,1)$. If $u \leq \alpha(\lambda^{(j-1)}, \lambda^*)$, then $\lambda^*$ is accepted by setting $\lambda^{(j)} = \lambda^*$, otherwise the chain does not move and $\lambda^{(j)} = \lambda^{(j-1)}$.

(viii) We now increment $j$ by 1, and repeat steps (ii)-(viii) for a total of $d$ iterations.

Whilst the methodology of Yu and Moyeed (2001) updates the parameters of a fixed degree regression polynomial at each iteration of the Metropolis-Hastings algorithm, our methodology updates both the entire vector of values $\boldsymbol{g}$ at the knots of the NCS and the smoothing parameter $\lambda$. We set the number of iterations $d$ to 500,000. We allow a burn-in of 50,000 iterations. Inference is based on thinned values of $\boldsymbol{g}$ and $\lambda$ produced by the Metropolis-Hastings algorithm after burn-in. We thin by taking every tenth value, partly because of storage consideration; see Section 7.4.2 for further discussion. Convergence issues are discussed in detail in Section 7.4. All code was written in R R Development Core Team (2008), using R's random number generating functions.

A considerable advantage of the Bayesian approach is that we can calculate associated credible intervals to provide an idea of the associated posterior uncertainty. These credible intervals are obtained using the methodology of Section 4.2.5 by ordering the thinned $g^{(j)}(\tau_i)$ sequence over $j > 50,000$ and extracting the values which correspond to, for example, the 2.5% and 97.5% quantiles. A 95% posterior credible interval for $\lambda$ can be obtained in a similar way. In the next section we present some examples of applications of this methodology. We finish this section by remarking that another approach to quantile regression is based on the

minimization over curves $g$ of

$$\sum_{i=1}^{n} \rho_p(y_i - g(t_i)) \tag{7.2.11}$$

Often $g$ is taken to be a B-spline (Hastie et al. (2001)) or a NCS with pre-specified knots and hence smoothness. The minimizing $g$ can be found using the **quantreg** package (Koenker (2008)) running under **R** R Development Core Team (2008); see Koenker (2005) for an example. Some other authors have considered the problem of minimizing over curves $g$ belonging to a suitable space a version of (7.2.11) penalized for roughness such as

$$\sum_{i=1}^{n} \rho_p(y_i - g(t_i)) + \lambda \int_a^b g''(t)^2 dt; \tag{7.2.12}$$

see Bosch et al. (1995) and reference therein, and Koenker et al. (1994) for further discussion. Koenker et al. (1994) also describe a similar minimization approach based on a total variation roughness penalty; software for this is again available in Koenker (2008). As far as we know, none of these approaches routinely yield confidence envelopes for the estimated curve, or choose the amount of smoothing in an automated way once the model has been specified.

## 7.3 Applied Examples

In this section we will apply our Bayesian nonparametric quantile regression to the HR Wallingford Coastal Wave data and Offshore Wave data kindly provided by Dr Anna Zacharioudaki (Zacharioudaki (May 2008)), both introduced in Chapter 1, Section 1.6. We will also apply the developed methodology to the Immunoglobulin-G data set from Yu and Moyeed (2001) to provide a direct comparison to their Bayesian polynomial based quantile regression technique. We initially study the HR Wallingford data and apply the model for $p = 0.9$ and ($p = 0.5$) median quantiles.

### 7.3.1 Application to Coastal Wave Data

The data used in this example relate to conditions near the Selsey Bill area and were generated using a hindcasting technique (see Reeve et al. (2004)) using wind records. The data set consists of hourly hindcast measurements of the variables significant wave height,

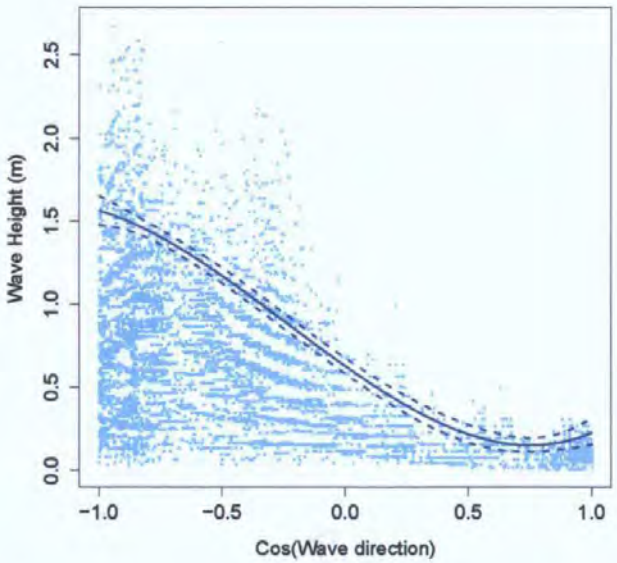



Figure 7.1: Scatter plot of the Coastal Wave data showing the $p = 0.9$ Bayesian quantile regression curve using a cubic polynomial. A 95% credible envelope is also presented.

wave period and wave direction over an approximate time span of 27 years. A good understanding of this type of data is important for the coastal design process, as illustrated by Thompson et al. (2008). Here, our variable $Y$ of interest will be wave height, while the covariate $t$ will be the cosine of wave direction. In this example we take a random sample of 10,000 observations for computational and presentational reasons. The resulting data set is shown in Figure 7.1 and is denoted $(t_1, y_1), \ldots, (t_n, y_n)$, where sample size $n = 10,000$.

This plot also shows the parametric Bayesian quantile cubic regression curve of Yu and Moyeed (2001) for $p = 0.9$ together with a 95% credible envelope. For our spline based approach we set $N = 30$ and used a grid $\tau_1 < \cdots < \tau_{30}$ of equally spaced knots over the range of covariate values $t_1, \ldots, t_n$. We found that such a grid of knots allows flexible modelling without imposing a very high computational burden. We set the hyperparameters $\beta = 10/\text{GCV(mean spline)} = 10^7$ and $\alpha = \text{GCV(mean spline)}/\beta = 10^{-13}$ in which $\text{GCV(mean spline)} = 10^{-6}$. With these hyperparameters the prior mean and variance of $\lambda$ are $\text{E}[\lambda] = 10^{-6}$ and $\text{Var}[\lambda] = 10$, representing a large amount of prior uncertainty about $\lambda$.

Figure 7.2 presents the resulting Bayesian nonparametric quantile regression curve and

95% credible envelope. To obtain the regression curve shown in Figure 7.2, we drew the unique NCS through the points $(\tau_i, \overline{g_i}), i = 1, \ldots, N$. Similarly, we produce our 95% credible envelope by drawing NCSs through the 2.5% and 97.5% posterior quantiles found in Section 7. The more local nature of the fitting procedure is easily seen from Figure 7.2. In order to

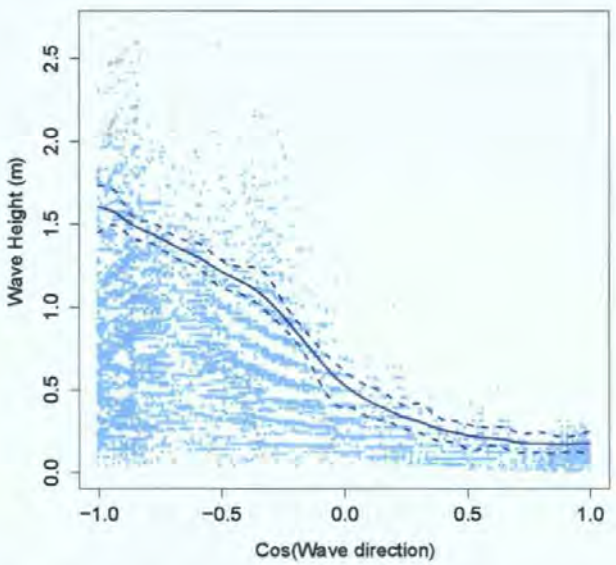



Figure 7.2: Scatter plot of the wave data showing the $p = 0.9$ Bayesian nonparametric quantile regression curve using splines. A 95% credible envelope is also presented.

judge the goodness-of-fit of both approaches we found empirical and fitted quantiles on a grid of size 100 along the covariate. We calculated the 'residual' in each piece of the grid as:

$$\text{residual} = \text{empirical quantile} - \text{fitted quantile}. \qquad (7.3.1)$$

in which the empirical quantile is the $p^{th}$ quantile of the data values in the piece of the grid and the fitted quantile is the value produced by our model at the centre of the piece. As usual, smaller residuals in absolute value are associated with better fits. Figure 7.3 shows the absolute value of the residuals from both the cubic polynomial quantile regression curve shown in Figure 7.1 and the spline based curve shown in Figure 7.2 against the cosine of wave direction. A robust locally linear smoother provided by R's R Development Core Team (2008) `loess` function (see Venables and Ripley (2002), for example) was added through each set of (covariate, |residual|) points. These curves indicate that the spline based quantile curve

gives a better quality of fit through almost the full covariate range than the cubic polynomial quantile curve. This is due to the more local nature of the spline based fitting procedure. We also calculated the mean square error based on the residuals for each model as a further method of assessing goodness-of-fit. We obtained mean square error values of 0.010 and 0.016 for the spline and polynomial based approach respectively. This is a further indication of the improvement that the nonparametric approach provides over its parametric counterpart.

We do remark, however, that we did not build our model with only goodness-of-fit in mind as we have introduced the roughness penalty to relax the fitting to obtain a smoother curve.

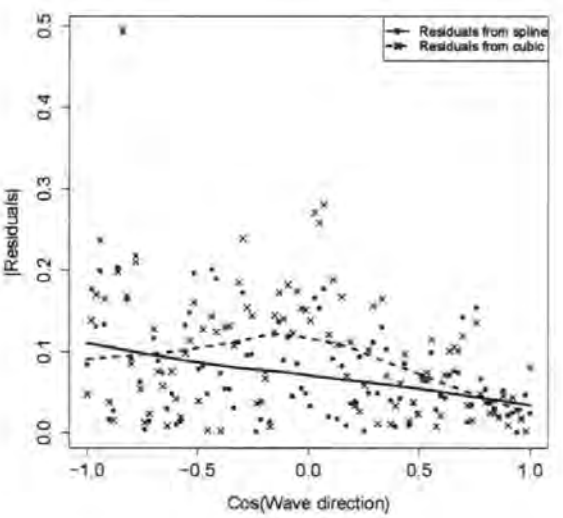



Figure 7.3: The absolute values of the residuals against the cosine of wave direction with associated loess smoother from both the spline (dots, unbroken line) and the cubic (crosses, dashed line) quantile regressions. A grid of size 100 along the covariate was used in the calculation of the residuals.

Finally, we calculated a 95% credible interval for the smoothing parameter $\lambda$, which for this example is (0.0027, 2.746). This wide interval indicates that there is considerable posterior uncertainty associated with $\lambda$. This may be a reflection of the variation in the nature of the data over the direction covariate.

We have chosen to show higher quantiles in this applied example as in Section 7.6 we present a novel application of this methodology for producing covariate dependent return level plots. However, as will be shown in this section we can fit the Bayesian nonparametric

regression at any quantile value. Figure 7.4 shows the Bayesian nonparametric median ($p = 0.5$) regression curve for the HR Wallingford Coastal Wave data.

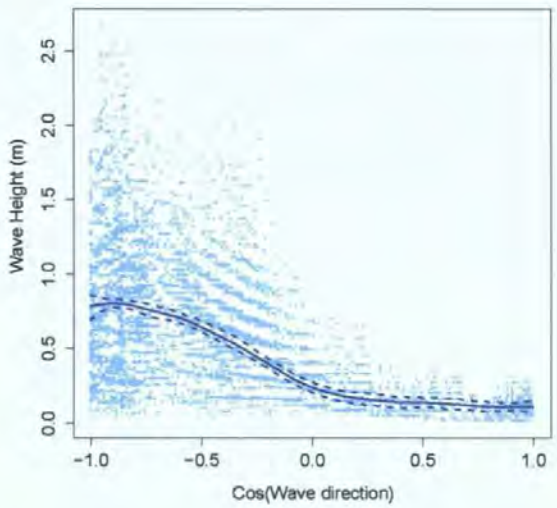



Figure 7.4: Scatter plot of the wave data showing the median Bayesian nonparametric quantile regression curve. A 95% credible envelope is also presented.

### 7.3.2 Application to Immunoglobulin-G Data

Yu and Moyeed (2001) present an application of their Bayesian quantile regression technique using a data set containing the serum concentration (grams per litre) of immunoglobulin-G (IgG) in 298 children aged from 6 months to 6 years; see Yu and Moyeed (2001) and references therein for further information on this data set. We will show that our nonparametric technique performs in a comparable to the parametric case. Figure 7.5 shows a scatter plot of Immunoglobulin-G data with the $p = 0.9$ Bayesian nonparametric quantile regression curve using splines and the $p = 0.9$ parametric Bayesian quantile regression curve. We also include 95% credible envelopes for both Bayesian quantile regression curves.

We can see from the Figure 7.5 that for the immunoglobulin-G data set both approaches produce a similar fit for the $p = 0.9$ quantile. We can also see that the average width of confidence interval in the nonparametric model is less than that of its parametric counterpart, this is emphasized near the covariate end regions. As in the previous applied example we judge the goodness-of-fit by finding empirical and fitted quantiles on a grid of size 30 along

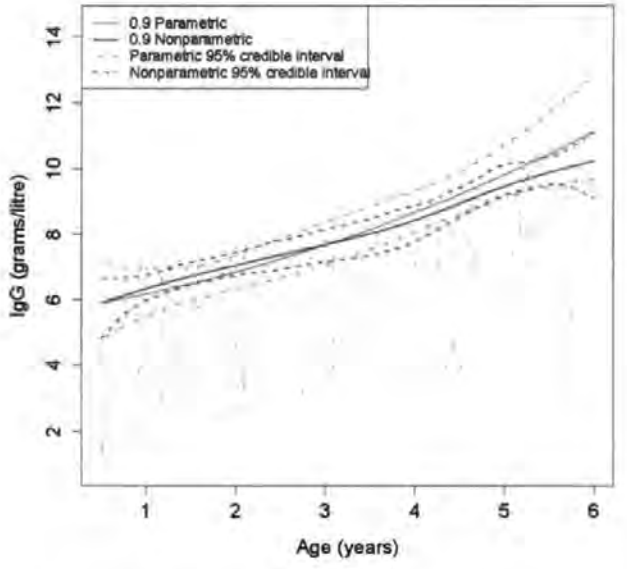



Figure 7.5: Scatter plot of the Immunoglobulin-G data showing the $p = 0.9$ Bayesian nonparametric quantile regression curve using splines and $p = 0.9$ parametric Bayesian quantile regression curve. 95% credible envelopes are also presented.

the covariate. 'Residuals' were again calculated using equation (7.3.1) where smaller residuals in absolute values indicated a better fit. Figure 7.6 shows the absolute values of the residuals plotted against age with associated loess smoother for both the spline and the cubic quantile regressions. We can see from this plot that both approaches provide a comparable result which reflects the behaviour in Figure 7.5. We do however see from Figure 7.6 that the residual seem rather large. This is due to there being fewer data causing the empirical quantile to vary much more than in the previous example, as Figure 7.7 shows. We find that we do not have an improved result as in the previous example. This may be due either to the smaller data set which effects the degree of local variation in the data, or, more likely, to the data being more uniform over the covariate, meaning that a cubic polynomial will provide an adequate fit. However in more complex data we find the spline based approach provides improved accuracy as highlighted in the previous example.

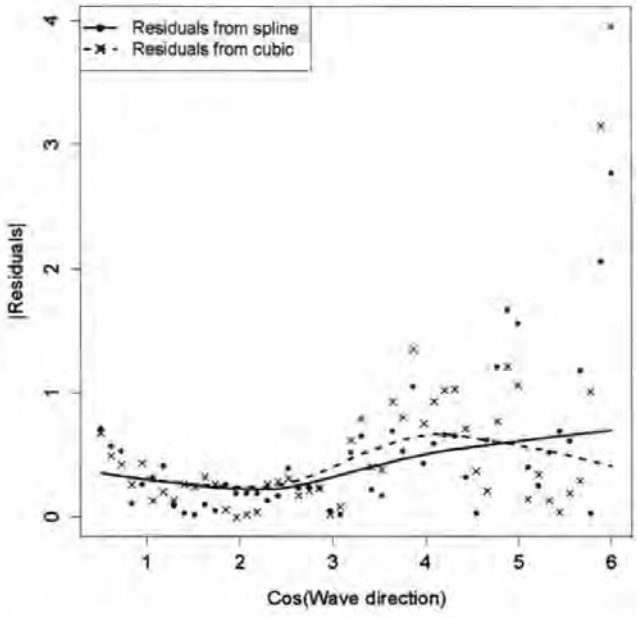



Figure 7.6: The absolute values of the residuals against age with associated loess smoother from both the spline (dots, unbroken line) and the cubic (crosses, dashed line) quantile regressions. A grid of size 30 along the covariate was used in the calculation of the residuals.

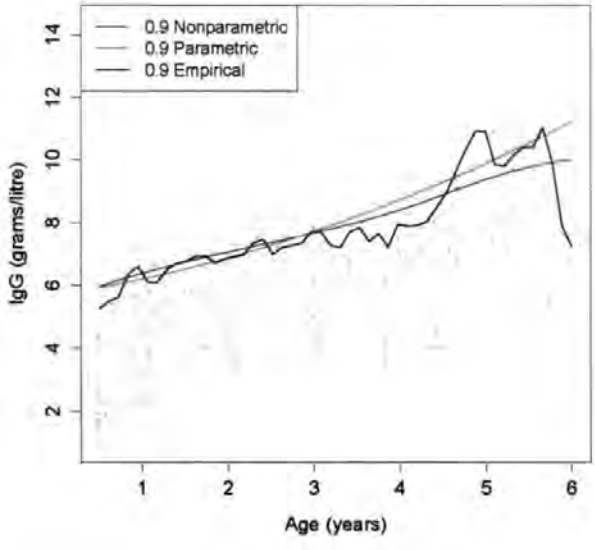



Figure 7.7: Scatter plot of the Immunoglobulin-G data showing the $p = 0.9$ Bayesian nonparametric quantile regression curve using splines and the $p = 0.9$ parametric Bayesian quantile regression curve. The empirical $p = 0.9$ quantile is also shown and can be seen to be highly variable, leading to the relatively large residuals seen in Figure 7.6.

### 7.3.3 Application to Offshore Wave Data

The offshore wave data used in this section was kindly provided by Dr. Anna Zacharioudaki from the School of Engineering, University of Plymouth (Zacharioudaki (May 2008)). As we mention in Chapter 1, these wave records refer to an offshore location in Poole Bay, UK. There are three variables: Wave Height, Wave Period and Wave Direction, each having 86,384 observations at 3 hourly intervals, which amounts to just over 29 years of data. We include this example to further illustrate and validate our approach as this data set had a different underlying structure from the HR Wallingford Coastal Wave data as there is less variation in the magnitude of values (including extremes) over the direction covariate. The data are shown in Figure 7.8, together with the $p = 0.9$ Bayesian quantile regression curves and associated credible intervals. Our nonparametric quantile regression curve using splines may provide us with a better understanding of the fine features of the $p = 0.9$ quantile than the cubic quantile regression curve. This may be particularly helpful with data sets of this size and visual complexity.

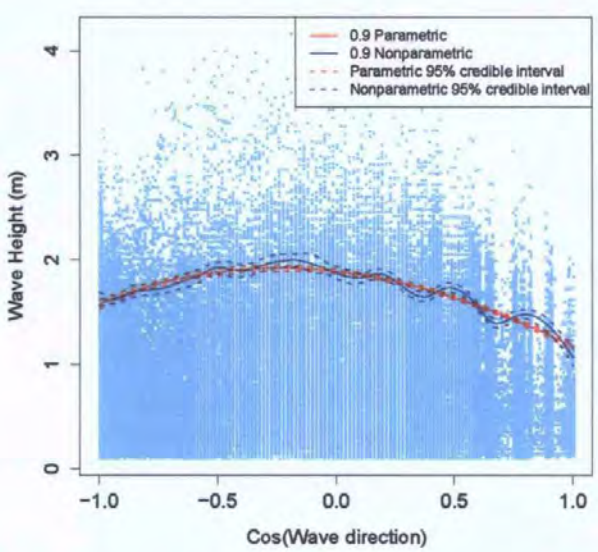



Figure 7.8: Scatter plot of the Offshore Wave data showing the $p = 0.9$ Bayesian nonparametric quantile regression curve using splines and $p = 0.9$ parametric Bayesian quantile regression curve. 95% credible envelopes are also presented.

Figure 7.9 shows the absolute value of the residuals from both the cubic polynomial

quantile regression curve and the spline based curve against the cosine of wave direction. We can clearly see that our spline based approach again provides a better quality of fit through the full covariate range than the cubic polynomial quantile curve. Again this is as a result of the more local nature of the spline based fitting procedure. This is more apparent in this example as we have a greater amount of data points to work with meaning, local variation can be better identified than in smaller data sets.

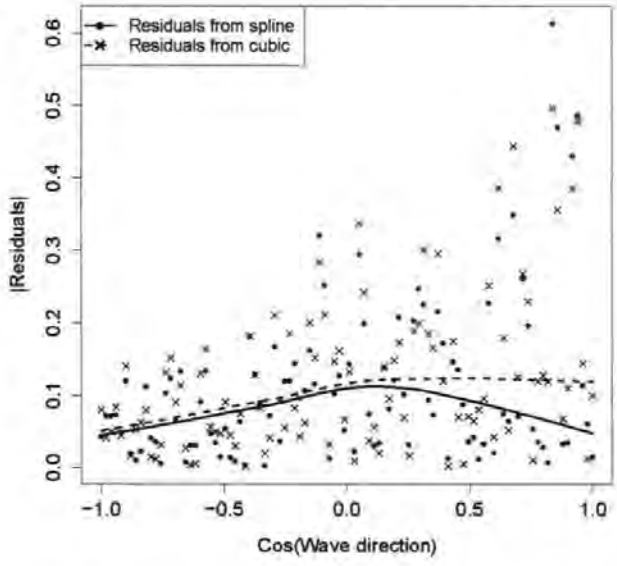



Figure 7.9: The absolute values of the residuals against the cosine of wave direction with associated loess smoother from both the spline (dots, unbroken line) and the cubic (crosses, dashed line) quantile regressions. A grid of size 100 along the covariate was used in the calculation of the residuals.

## 7.4 Markov Chain Monte Carlo Performance

### 7.4.1 Choosing the Proposal Density and Acceptance Rate

In step (ii) of the Metropolis-Hastings algorithm presented in Section 7.2 the candidate vector $\boldsymbol{g}^*$ was drawn from a multivariate normal distribution with variance-covariance matrix $\Sigma = \sigma^2 K^-/\lambda$. In this way a candidate $\boldsymbol{g}^*$ has similar structure to a $\boldsymbol{g}$ from the prior term $\pi(\boldsymbol{g}|\lambda)$ given in equation 7.2.1. We also considered generating $\boldsymbol{g}^*$ from a multivariate normal distribution with $\Sigma = \sigma^2 I_N$ where $I_N$ is the $N \times N$ identity matrix. As a third possibility we

updated a random subset of $g_1, \ldots, g_N$ again using independent normal distributions with variance $\sigma^2$. All three possibilities of generating $\boldsymbol{g}^\star$ performed similarly, with the choice of $\sigma^2$ having the greatest effect on the convergence of the Metropolis-Hastings algorithm.

Bédard (2006a) introduced a technique that can be applied here to optimally choose the parameter $\sigma^2$ that controls the variance $\Sigma = \sigma^2 K^- / \lambda$ of the proposal density $q$ for $\boldsymbol{g}$ in the Metropolis-Hastings algorithms. The technique plots an efficiency criterion against acceptance rates from the Metropolis-Hastings algorithm or against $\sigma^2$. The acceptance rate or value of $\sigma^2$ that corresponds to the maximum efficiency can then be chosen. The key

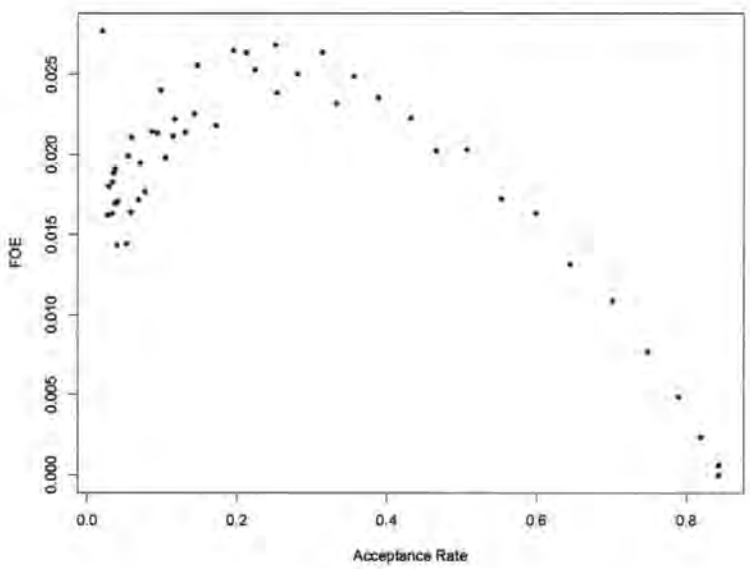



Figure 7.10: Efficiency against acceptance rate when updating $\boldsymbol{g}$ in the Metropolis-Hastings algorithm.

to this procedure is the use of the first order efficiency criterion which measures the average squared jumping distance for each parameter from one iteration to the next. In the case of the polynomial model of Yu and Moyeed (2001) in which the parameters $\beta_0, \beta_1, \beta_2$ and $\beta_3$ are updated individually, Bédard (2006a) would define the first order efficiency criterion (FOE) for the $i$th parameter as

$$\mathrm{FOE}_i = E\left[\left(\beta_i^{(j+1)} - \beta_i^{(j)}\right)^2\right], \qquad (7.4.1)$$

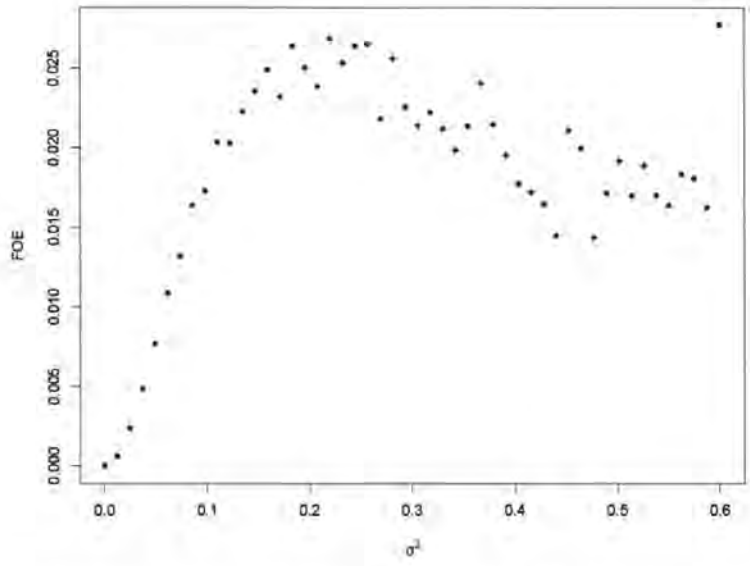



Figure 7.11: Efficiency against $\sigma^2$ for updating $\boldsymbol{g}$ in the Metropolis-Hastings algorithm.

where the expectation is over iterations $j$. The definition can be easily extended to the case of the spline, in which all the parameters $\boldsymbol{g} = (g_1, \ldots, g_N)^T$ are updated simultaneously, by using squared Euclidean distance as follows:

$$\text{FOE} = E\left[\sum_{i=1}^{N}\left(g_i^{(j+1)} - g_i^{(j)}\right)^2\right], \tag{7.4.2}$$

where again the expectation is over iterations $j$. Figures 7.10 and 7.11 show plots of FOE against acceptance rate and against $\sigma^2$ for updating $\boldsymbol{g}$. These plots allow the user to choose the acceptance rate or $\sigma^2$ corresponding to the highest value of FOE. From Figure 7.10 it can be seen that an acceptance rate of about 0.24 is most appropriate. This may seem rather low, but is due to the fact that we are updating a whole vector of parameters $\boldsymbol{g}$ and not just an individual parameter. It is also in agreement with some of the literature about optimal acceptance rates; see Bédard (2006a), Bédard (2006b) and references therein for example. A relatively low acceptance rate corresponds to a relatively high proposal variance which itself allows larger possible jumps for the vector of parameters $\boldsymbol{g}$. A similar approach can be used to choose the value of $\sigma_\lambda^2$ for updating the smoothing parameter $\lambda$ in step (v) of the Metropolis-Hastings algorithm presented in Section 7.2. In our application we fixed a value

for $\sigma_\lambda^2$ and tuned $\sigma^2$. We then fixed our chosen $\sigma^2$ and tuned $\sigma_\lambda^2$. Finally, we fixed our chosen $\sigma_\lambda^2$ and re-tuned $\sigma^2$. We found that we were able to achieve good convergence for both $\boldsymbol{g}$ and $\lambda$ with these tuned values of $\sigma^2$ and $\sigma_\lambda^2$, as we will discuss in Section 7.4.2. We also found that this approach yielded a value of $\sigma_\lambda^2$ that was relatively insensitive to the value of $\sigma^2$.

### 7.4.2 Assessing Markov Chain Monte Carlo Convergence

Visual assessment of the convergence of the Metropolis-Hastings algorithm was found to be difficult as the simulated elements included $N = 30$ points along the spline rather than just a few model parameters. We found that the combination of a large number of sub-chains and an acceptance step based on a vector of points rather than an individual parameter could cause some convergence issues, although these could be overcome with good choices of $\sigma^2$ and $\sigma_\lambda^2$ as discussed in Section 7.4.1. Convergence is generally slower in comparison with more usual parametric models. However this computational sacrifice is balanced by the improved localized fitting of the model which was seen in Section 7.3. The visual assessment of

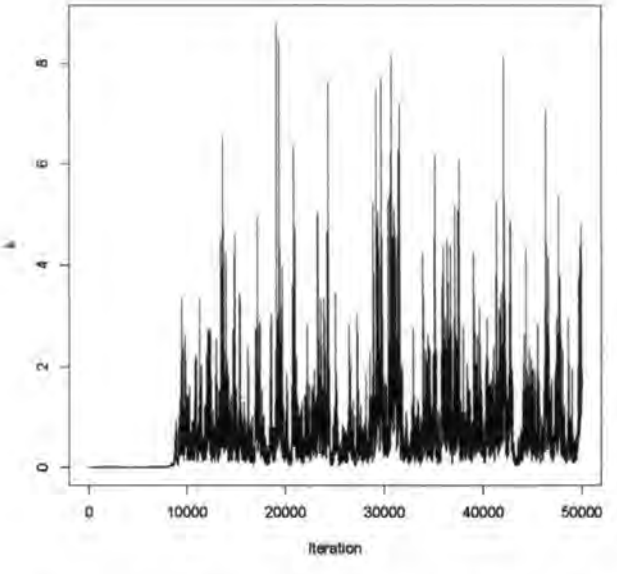



Figure 7.12: Thinned time series plot for $\lambda$.

convergence of $\lambda$ was also difficult as the parameter took a wide range of values as highlighted in Figure 7.12, where we can see that the time series converges around a lower value, with a tendency to jump to higher values (indicating smoother curves). We see that the time series

has moved away from the low initial value of $\lambda^{(0)} = 10^{-6}$ and from the prior mean set to the same value. In fact, values of $\lambda$ as low as $10^{-6}$ produce curves (not shown) that are visually far too rough.

After initially examining time series plots of individual chains, we used the more formal Gelman-Rubin statistic, discussed in Gelman and Rubin (1992), Gelman (1996) and Brooks and Gelman (1998), to assess convergence of $\boldsymbol{g}$ and of $\lambda$. The Gelman-Rubin procedure compares the variances between and within chains to monitor convergence and is based on the 'estimated potential scale reduction factor' $\hat{R}^{1/2}$,see Gelman and Rubin (1992) for details, which represents the estimated factor by which a credible interval for a parameter of interest may shrink if further simulation is carried out. Good performance is indicated by values of $\hat{R}^{1/2}$ close to 1. The value of $\hat{R}^{1/2}$ should certainly not exceed 1.2 as suggested in Kass et al. (1998). We calculated $\hat{R}^{1/2}$ for each sub-chain $g_i$, $i = 1, \ldots, N$, and for $\lambda$ and found that $\hat{R}^{1/2}$ took values between 1.0006 and 1.0152. Thinning was applied by taking every tenth value as particular sub-chains showed strong autocorrelations. As already mentioned, thinning also reduced storage requirements. Our examination of time series plots together with satisfactory values of the Gelman-Rubin statistic gave us confidence that the Metropolis-Hastings algorithm was producing realizations approximately from the posterior distribution $\pi(\boldsymbol{g}, \lambda|\boldsymbol{y})$.

## 7.5 Alternate Techniques for Performing Inference about the Smoothing Parameter $\lambda$

Up to now we have performed inference about the smoothing parameter $\lambda$ introduced in Section 7.2 in the Bayesian framework. However we have also explored an alternative method for estimating $\lambda$ which we will discuss in Section 7.5.1. In Section 7.5.2 we mention methodology employed in de Pasquale et al. (2004) that can be used when performing inference about $\lambda$ in the Bayesian framework if the normalizing constant in (7.2.1) for example were unavailable.

### 7.5.1 Investigating a Range of Smoothing Parameters

In this section we show the results of substituting a range of values for $\lambda$ into our Bayesian nonparametric quantile regression using splines methodology. We base our values of $\lambda$ on the automated approach for the choice of the smoothing parameter in the case of the mean regression problem, with estimate provided by the smoothing spline as discussed in Section 4.3. The actual methodology used was generalized cross-validation which is discussed in detail in Section 4.3.3 and in Green and Silverman (1994).

We first consider the mean regression problem and estimate the parameter $\lambda$ of the associated smoothing spline by generalized cross-validation. We then found the associated spline based $p = 0.9$ quantile regression curves for this value of $\lambda$ and for $c\lambda$ for the following values of $c$ : 0.0001, 0.1, 2, 10, 10000. All these curves are shown in Figure 7.13. We can see from Figure 7.13 that the quantile regression curves differ considerably across these values of $\lambda$, with some curves appearing very rough. We found it hard to select a suitable smoothing parameter using this approach. It is for that reason that we included $\lambda$ in our Bayesian approach effectively making it fully Bayesian.

### 7.5.2 Applying Fully Bayesian Methodology in the Absence of Normalization Constants

In Section 7.2 we defined our prior for $\boldsymbol{g}$, $\pi(\boldsymbol{g}|\lambda)$, through (7.2.1). When we first considered this multivariate normal prior, we did not know its normalization constant. In other words we wrote

$$\pi(\boldsymbol{g}|\lambda) = \frac{1}{C(\lambda)} \exp\left(-\frac{1}{2}\lambda\, \boldsymbol{g}^T K \boldsymbol{g}\right), \qquad (7.5.1)$$

where the function $C(\lambda)$ was unknown. A consequence of not knowing $C(\lambda)$ is that it is impossible to compute the acceptance probability $\alpha(\lambda^{(j-1)}, \lambda^\star)$ given in (7.2.10). The methodology that we now present allows us to apply our fully Bayesian methodology in the absence of knowledge of $C(\lambda)$. We present it as useful general methodology. As we subsequently found a closed form expression for $C(\lambda)$, this methodology became redundant

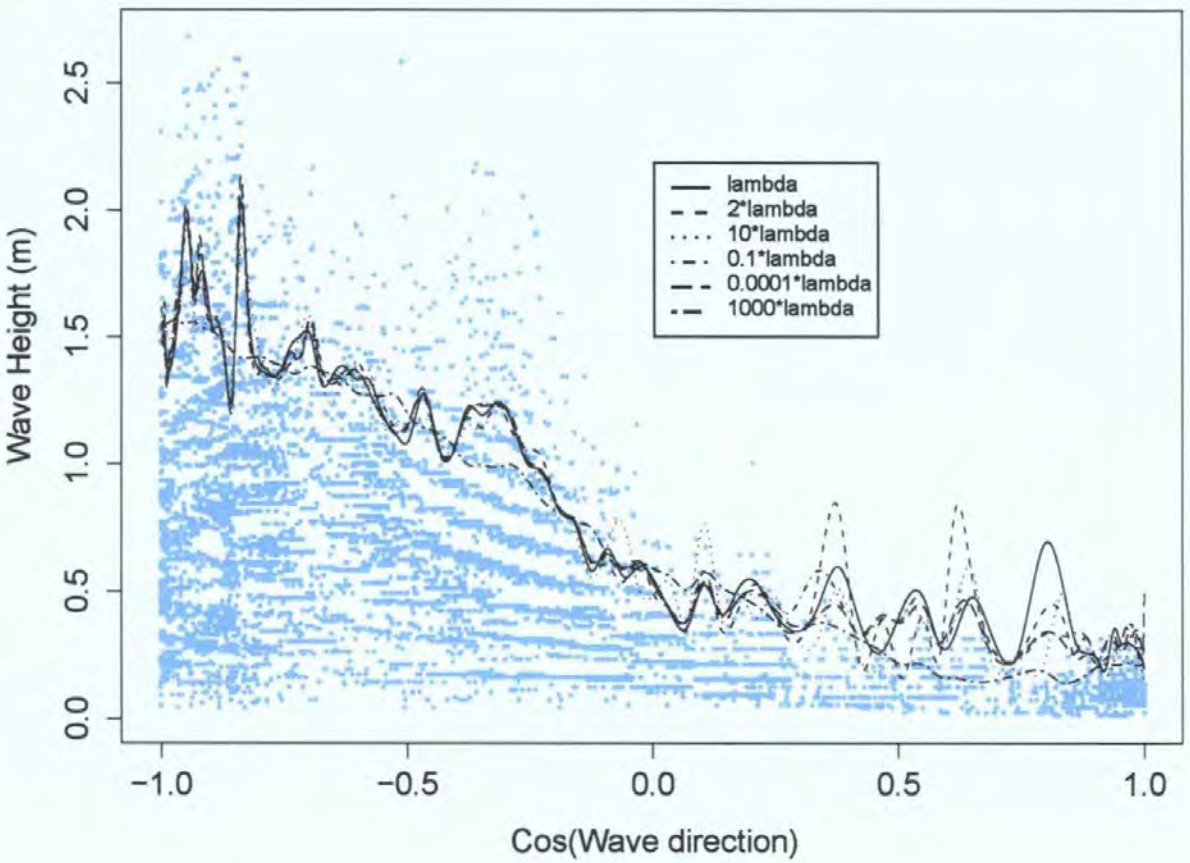



Figure 7.13: Scatter plot of the HR Wallingford Coastal Wave data showing 90% ($p = 0.9$) Bayesian quantile regression curves for a range of smoothing parameter values $c\lambda$, where $\lambda$ is obtained by generalized cross-validation and $c = 0.0001,\ 0.1,\ 2,\ 10,\ 10000$.

in the present case. From (7.5.1) we see that

$$C(\lambda) = \int_{\boldsymbol{g} \in \Re^N} \exp\left(-\frac{1}{2}\lambda\, \boldsymbol{g}^T K \boldsymbol{g}\right) d\boldsymbol{g} \qquad \boxed{7.5.2}$$

and from (7.2.10) we see that we need to be able to evaluate ratios such as $C(\lambda^{(j-1)})/C(\lambda^\star)$. We now explain how this can be done; de Pasquale et al. (2004) use similar methodology for

a discrete random variable. We begin as follows:

$$\begin{aligned}
\frac{d}{d\lambda}\log C(\lambda) &= \frac{1}{C(\lambda)}\frac{d}{d\lambda}C(\lambda) \\
&= \frac{1}{C(\lambda)}\int\limits_{\boldsymbol{g}\in\mathbb{R}^N}\left(-\frac{1}{2}\boldsymbol{g}^T K\boldsymbol{g}\right)\exp\left\{-\frac{1}{2}\lambda\boldsymbol{g}^T K\boldsymbol{g}\right\}d\boldsymbol{g} \\
&= \int\limits_{\boldsymbol{g}\in\mathbb{R}^N}\left(-\frac{1}{2}\boldsymbol{g}^T K\boldsymbol{g}\right)\frac{\exp(-\frac{1}{2}\lambda\boldsymbol{g}^T K\boldsymbol{g})}{C(\lambda)}d\boldsymbol{g} \\
&= \int\limits_{\boldsymbol{g}\in\mathbb{R}^N}\left(-\frac{1}{2}\boldsymbol{g}^T K\boldsymbol{g}\right)\pi(\boldsymbol{g}|\lambda)d\boldsymbol{g} \\
&= \underset{\boldsymbol{g}\sim\pi(\boldsymbol{g}|\lambda)}{E}\left[-\frac{1}{2}\boldsymbol{g}^T K\boldsymbol{g}\right] \qquad (7.5.3)
\end{aligned}$$

This expectation can be calculated using the MCMC output through equation (4.2.2). By integrating (7.5.3)over $\lambda$ we are now able to find an expression for $\log\frac{C(\lambda)}{C(\lambda_0)}$, for a fixed minimum value of $\lambda_0$, as follows:

$$\log C(\lambda) - \log C(\lambda_0) = \int_{\lambda_0}^{\lambda}\underset{\boldsymbol{g}\sim\pi(\boldsymbol{g}|\lambda')}{E}\left[-\frac{1}{2}\boldsymbol{g}^T K\boldsymbol{g}\right]d\lambda'. \qquad (7.5.4)$$

Hence

$$\log\frac{C(\lambda)}{C(\lambda_0)} = \int_{\lambda_0}^{\lambda}\underset{\boldsymbol{g}\sim\pi(\boldsymbol{g}|\lambda')}{E}\left[-\frac{1}{2}\boldsymbol{g}^T K\boldsymbol{g}\right]d\lambda', \qquad (7.5.5)$$

an integration that can be performed numerically using Simpson's rule. If we can find $\log\frac{C(\lambda)}{C(\lambda_0)}$ for any value of $\lambda$, then we can compute

$$\begin{aligned}
\log\frac{C(\lambda)}{C(\lambda^*)} &= \log\frac{C(\lambda)/C(\lambda_0)}{C(\lambda^*)/C(\lambda_0)} \\
&= \log\frac{C(\lambda)}{C(\lambda_0)} - \log\frac{C(\lambda^*)}{C(\lambda_0)}. \qquad (7.5.6)
\end{aligned}$$

In practice we compute $\log\frac{C(\lambda)}{C(\lambda_0)}$ on a grid of $\lambda$'s, and then the value of $C(\lambda^{(j-1)})/C(\lambda^*)$ for all possible $\lambda^{(j-1)}$ and $\lambda^*$ for all possible $\lambda^{(j-1)}$ and $\lambda^*$ by interpolation. All these calculations can be performed once and for all before the main Metropolis-Hastings algorithm is run.

## 7.6 Covariate specific return level plots

In Section 7.2 we introduced our Bayesian quantile regression methodology. We illustrated examples of the methodology in Section 7.3. We now use our method to create return level plots incorporating a directional covariate for use in coastal defence design.

### 7.6.1 Covariate dependent return level plots

A return level plot shows the relationship between return period and return level. These concepts were discussed in detail in Section 2.2. Figure 2.5 shows an example of a return level plot. Reeve et al. (2004) define the return period as a measure of the rarity of an event. For example, if we have a return period of $R$ years and $n_y$ events in a year, then the $R$-year event would be the one with probability $1/(n_y R)$ of being exceeded. The return level is the associated magnitude of the subject variable ( wave height) corresponding to a given return period. Traditionally, engineers need to consider extremes of the sea condition that their coastal defence design must withstand. A return level plot can provide a good indication of the potential extreme conditions of a subject variable such as wave height. We now show how to generate a return level plot using our Bayesian quantile spline functions.

In Section 7.3.1 we modelled the variable wave height with the cosine of wave direction as covariate. The covariate can be incorporated into the return level plot so that the return levels can now be specific to cosine wave direction. This is of particular use where extremes from a specific directions are of interest, for example in the case of defences sheltered by natural geographic features. Let $q_p(t)$ is the value of $p^{th}$ quantile when the covariate (e.g. the cosine of wave direction) takes the value $t$. Then $\Pr(Y > q_p(t)) = 1 - p$, where the variable $Y$ is wave height, with the consequence that the return period in years associated with the return level $q_p(t)$ is $\frac{1}{n_y(1-p)}$, in which $n_y = 10,000/27$, since we are working with a data set of 10,000 observations observed over 27 years; see Section 1.6. Figure 7.14 shows a set of quantile curves for different return periods. Figure 7.15 is similar to Figure 7.14 except that it presents the traditional return level plots for specific directions rather than a curve at a specific return level over all directions. This plot is more useful in the design process than a non-covariate specific return level plot as the latter is effectively an average over the entire

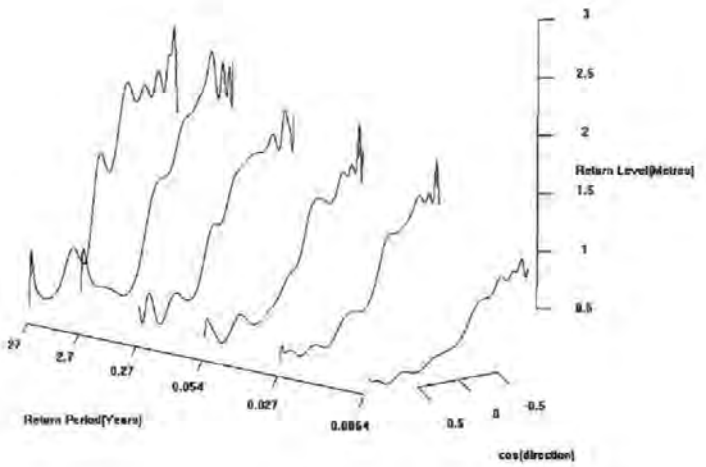



Figure 7.14: Quantile regression splines for the variable wave height with cosine wave direction as a covariate.

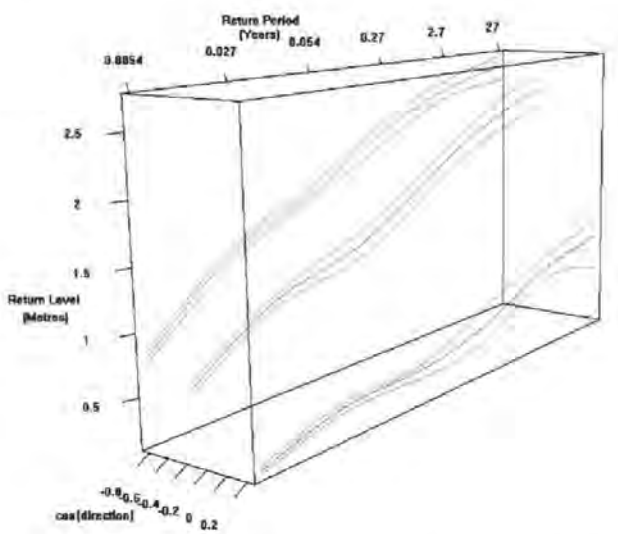



Figure 7.15: Quantile regression return level curves for wave height for three specific directions. Also shown are 95% credibility envelopes.

range of wave direction. We believe therefore that our methodology can improve defence design estimates as it allows direction to be properly included in the design process.

### 7.6.2 Overtopping return level plots

Overtopping occurs when some amount of sea water discharges over the crest (or highest point) of a sea defence such as the one shown in Figure 7.17 from Reeve et al. (2004). The amount of data about overtopping over a period of time depends on the severity of the wave

conditions and the sufficiency of the sea defence to prevent overtopping. Collecting large quantities of overtopping data may be problematic as in calm to moderate conditions no overtopping may be recorded. Effectively we are, in some sense, only dealing with extremes so the methodology would be restricted to defences which were designed to allow a certain degree of overtopping or existing failing defences. It is possible to construct an overtopping

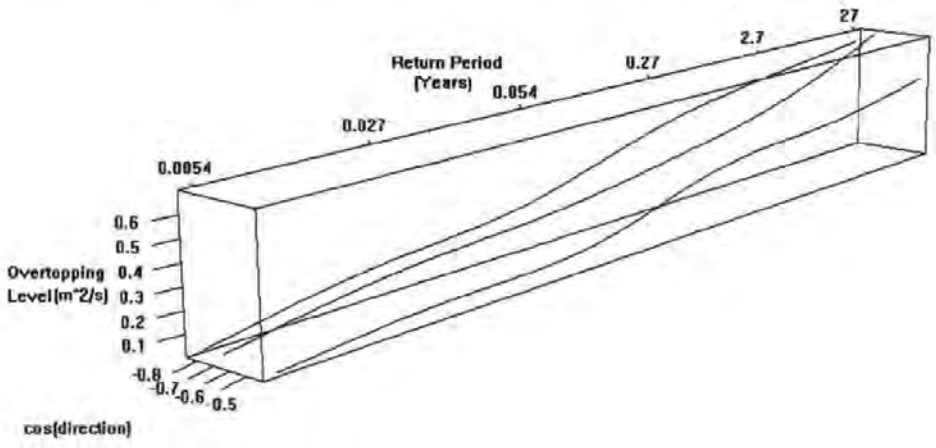



Figure 7.16: Quantile regression based return level curves for overtopping level for three specific directions.

return level plot based on the physical dimension and location of the defence, return level estimates of wave height and data about wave period. Once the dimension and location of the defence are defined, we require data at this location for the variables wave height $H_s$, wave period $T_z$ and wave direction. In this methodology we assume the defence orientation is always perpendicular to the wave approach angle. A plot like Figure 7.17 can be very useful to engineers when assessing the effectiveness of their sea defence designs. Further work could produce overtopping return level plots to allow for different orientations of the defence. This work would incorporate different approach angles as the degree of wave overtopping can be significantly altered when considering angles which are not perpendicular to the defence.

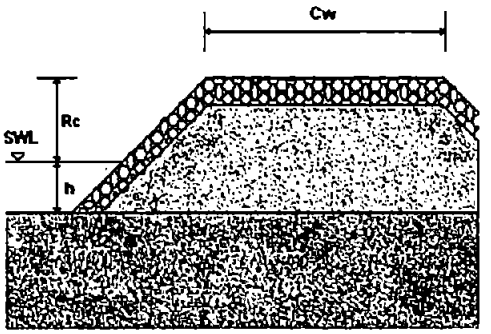



Figure 7.17: Diagram of a typical rough plane slope defence. SWL stands for still water level.

## 7.7 Summary

In this chapter we have developed methodology to extend fixed degree polynomial based quantile regression to nonparametric quantile regression within a Bayesian framework. We achieved this by using a spline based approach. We defined the posterior density of a NCS and an associated smoothing parameter. We sampled from this posterior by means of the Metropolis-Hastings algorithm and used our sample to make inferences that include the quantification of uncertainty by means of credible intervals.

We have also presented applications of our Bayesian nonparametric quantile regression methodology to three examples providing comparisons with the existing parametric method of Yu and Moyeed (2001) that show improvement due to the greater flexibility of our model especially for large data sets. We have made suggestions for increasing the efficiency of our methodology by making a good choice of the proposal density in the Metropolis-Hastings algorithm. We have also created wave height and overtopping return level plots using our Bayesian quantile spline functions that incorporate a direction based covariate. These plots are of considerable use in the coastal design process.

# 8

# Discussion of Results and Future Work

## 8.1 Discussion of Results

In this thesis we extended and improved the modelling techniques used in the analysis of extreme wave conditions. Improved modelling of these extremes can have important consequences for the cost and effectiveness of an engineering structure. The techniques considered in this thesis can be categorized into two general areas: Extreme Value Theory and Quantile Regression.

All our work has been motivated by and tested on real data sets. These data sets can be divided into two broad categories: Coastal Wave data and Offshore Wave data. Some observations were missing from our data due to either breaks in recording or other errors. To ensure that we were working with complete data sets, we developed a technique to replace missing observations with predicted values from a loess model. The loess technique is based on locally weighted polynomial regressions, fitted using weighted least squares, with more weight being given to observations near the point at which the estimation is being performed. We adapted the loess technique to deal with missing values close to the start or end of the data set and with large gaps of missing values. This methodology is discussed in Chapter 5.

In Chapter 2 we introduced extreme value theory, and discussed the Generalized Pareto Distribution (GPD) which models data defined as excesses over a user defined threshold. The GPD has two parameters, called the shape and scale parameters, that are estimated from the data, usually by maximum likelihood estimation. Existing methods for defining the

threshold position rely heavily on prior knowledge of the interpretation of threshold selection plots or on making an assumption that suitable values for GPD modelling are located above a particular quantile. There can be a range of errors associated with these approaches including incorrect assessment of the plots leading to considerable under- or over-estimation of the threshold, or selection of the wrong quantile value.

Hence, in Chapter 6 we developed technique that clarified and automated threshold selection for the GPD model. This method was based on the distribution of parameter estimates across a range of possible threshold values and required no external input other than the data set itself. When compared with the quantile based threshold selection technique used in the JOINSEA software, reviewed in Chapter 3, it was found that our automated approach yielded more favourable results in the majority of cases. Automation of threshold selection also opens up extreme value analysis to a wider range of users as the amount of prior expertise required is reduced. We quantified the effect of uncertainty associated with threshold selection on return level estimation using the bootstrap procedure and in particular bootstrap percentile intervals. We used a simulation study to show that our automated technique can recover a known threshold to a good degree of accuracy.

Development of our automated threshold selection technique required us to make three refinements to existing GPD model parameter estimation methodology. The first refinement was the calculation and use of an analytic Hessian matrix to obtain estimates of the variances of the maximum likelihood parameter estimates. The current numerical approximation of the Hessian matrix sometimes leads to negative variances and hence undefined standard errors. Our analytical version removed this problem. Our second refinement involved tightening the constraints on the parameters of the GPD to ensure that asymptotic properties always held and that the negative Hessian was positive definite. Our third refinement concerned the calculation of the Hessian matrix when the shape parameter was near zero. We used the Taylor expansion of the log function to ensure that the Hessian was correctly computed when the shape parameter was near zero.

We have extended our automated threshold selection methodology to incorporate a covariate dependent threshold. This extension uses our automated threshold selection technique to segregate the data into optimal blocks based on goodness-of-fit and sample size

requirements. Our methodology can lead to more accurate return level estimates, with their uncertainty properly quantified, which can inform and enhance the coastal design process. Towards the end of Chapter 6 we also describe a Graphical User Interface (GUI) that we have produced to allow engineering practitioners easy access to a range of techniques for extreme value modelling.

In Chapter 7 we focused on the use of quantile regression as a modelling technique to understand the behaviour of extreme values as a function of a covariate. We reviewed quantile regression itself and associated methodology in Chapter 4. Our work builds on the methodology for Bayesian quantile regression presented by Yu and Moyeed (2001). They adopt a parametric polynomial quantile regression model and perform inference from the posterior distribution of the parameters by means of a Markov chain Monte Carlo (MCMC) algorithm, also outlined in Chapter 4.

We extended their polynomial based quantile regression methodology to nonparametric quantile regression within the Bayesian framework. We achieved this by using a spline based approach. We explained how to define the posterior density of a natural cubic spline and an associated smoothing parameter. We then sampled from this posterior density by means of a particular MCMC algorithm known as the Metropolis-Hastings algorithm. We used our sample to make inferences that include the quantification of uncertainty. We applied our Bayesian nonparametric quantile regression methodology to our wave data using a covariate based on wave direction and provided comparisons with the polynomial based methodology of Yu and Moyeed (2001) that show improvements due to the greater flexibility of our model. We presented suggestions for increasing the efficiency of our methodology by making a good choice of the proposal density in the Metropolis-Hastings algorithm. We achieved this by modifying a technique presented in Bédard (2006b) for choosing the proposal density in an efficient way.

To our knowledge the quantile regression approach has not been applied to wave condition data before, and hence it provides a different perspective to the analysis of extreme wave conditions. We used our Bayesian nonparametric quantile regression technique to create some engineering design aids including a directional dependent wave height return level plot. From this plot we produced a directional dependent overtopping return level plot. The

advantages of these plots is that we can now understand how these extreme values depend on wave direction. This additional information may be considerably advantageous from a coastal engineering point of view especially when considering the location of a proposed coastal defence.

## 8.2 Future Work

In this section we discuss several areas of future work arising from the methodology presented in this thesis. We provide a brief description of how each area of future development could be undertaken.

**Alternate covariates:** In Chapter 6 we presented an automated threshold selection technique and extended it to depend on a directional covariate so creating a direction varying threshold. Wave direction is not, however, the only potential covariate. We could take other variables, such as fetch length, as a covariate to obtain a threshold which would be dependent on those covariates. We could extend our methodology to allow the threshold to depend on two or more covariates. To do this we would have to extend our method for defining blocks to two or more dimensions.

**Adaptive density gridding:** It may also be possible to improve our blocking technique in the direction varying threshold method by using an adaptive density gridding which is used in numerical methods for parameter estimation as a block scaling method.

**Further applications of Bayesian nonparametric quantile regression using splines:** This methodology that we presented in Chapter 7 is not restricted to the application of extreme wave analysis, and can be used in a range of situations. We have already considered a medical application based on the Immunoglobulin-G data from Yu and Moyeed (2001). Our methodology is particularly appropriate for situations in which there is a clear non-polynomial process underlying the data; the famous but small motor-cycle data set analysed by Silverman (1985) provides an example of the type of data for which the advantage of the spline based approach could be very clearly apparent.

**Increasing the number of knots used in the definition of the natural cubic spline:** In Chapter 7 we adopted a natural cubic spline with $N$ knots and set $N = 30$. We investigated only thirty knots because of computational considerations. We will investigate more sophisticated computational algorithms so that we can use more knots, so potentially increasing the flexibility of the model. It would be interesting to see the effect of having a knot at each unique covariate value.

**Further development of the Graphical User Interface:** In Chapter 6 we presented a GUI that we produced to allow engineering practitioners easy access to a range of useful techniques. We believe that the interface's structure and functionality could be increased further. This could be achieved by creating a "stand-alone" piece of software written in a faster programming language such as `C` or `C++`. We also plan to incorporate our computationally demanding Bayesian nonparametric quantile regression modelling into our GUI.

# A

# Hessian Calculations

The following sections present full details of the calculations of the elements of the Hessian matrix:

$$H = \begin{pmatrix} \frac{\partial^2 \ell(\sigma,\xi)}{\partial \sigma^2} & \frac{\partial^2 \ell(\sigma,\xi)}{\partial \sigma \partial \xi} \\ \frac{\partial^2 \ell(\sigma,\xi)}{\partial \xi \partial \sigma} & \frac{\partial^2 \ell(\sigma,\xi)}{\partial \xi^2} \end{pmatrix} \tag{A.0.1}$$

where the associated log-likelihood is

$$\ell(\sigma,\xi) = -k\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right) \tag{A.0.2}$$

Let $H$ be broken down as

$$H = \begin{pmatrix} \text{Element 1} & \text{Element 2} \\ \text{Element 3} & \text{Element 4} \end{pmatrix} \tag{A.0.3}$$

We will consider each element in order. For all calculations in this appendix some basic differentiation rules are required:

**Addition Rule**

$$\frac{d}{dx}[f(x)+g(x)] = f'(x)+g'(x), \tag{A.0.4}$$

where $'$ denotes differentiation

**Product Rule**

$$\frac{d}{dx}[f(x)g(x)] = f(x)g'(x)+g(x)f'(x) \tag{A.0.5}$$

**Quotient Rule**

$$\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2} \tag{A.0.6}$$

**Chain Rule**

$$\frac{d}{dx}[f(g(x))] = f'(g(x))g'(x) \tag{A.0.7}$$

## A.1 Element 1

Consider,

$$\frac{d}{d\sigma}\left[\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right)\right]=-\frac{\xi}{\sigma^2}\sum_{i=1}^{k}\frac{y_i}{\left(1+\frac{\xi y_i}{\sigma}\right)} \tag{A.1.1}$$

using the Chain Rule (A.0.7) and the Addition rule (A.0.4) with

$$f(u)=\log(u) \qquad f'(u)=\frac{1}{u}$$

$$g(\sigma)=1+\frac{\xi y_i}{\sigma} \qquad g'(\sigma)=-\frac{\xi y_i}{\sigma^2}$$

Now consider

$$\frac{d}{d\sigma}\left[-k\log\sigma\right]=-\frac{k}{\sigma}; \tag{A.1.2}$$

hence, the first derivative of the likelihood function with respect to $\sigma$ is

$$\frac{\partial\ell}{\partial\sigma}=-\frac{k}{\sigma}+\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\frac{\xi y_i}{\sigma^2\left(1+\frac{\xi y_i}{\sigma}\right)}. \tag{A.1.3}$$

The second stage is to calculate the second derivative, so consider the following

$$\frac{d}{d\sigma}\left[\xi\sum_{i=1}^{k}y_i\left(\frac{-1}{\sigma^2+\sigma\xi y_i}\right)\right]=-\xi\sum_{i=1}^{k}y_i\frac{2\sigma+\xi y_i}{(\sigma^2+\sigma\xi y_i)^2} \tag{A.1.4}$$

using both Chain Rule (A.0.7) and Product Rule (A.0.5) with

$$f(u)=\frac{1}{u} \qquad f'(u)=-\frac{1}{u^2}$$

$$g(\sigma)=\sigma^2+\sigma\xi y_i \qquad g'(\sigma)=2\sigma+\xi y_i,$$

giving

$$\frac{d}{d\sigma}[f(g(\sigma))]=\frac{-1}{(\sigma^2+\sigma\xi y_i)^2}(2\sigma+\xi y_i)=\frac{-2\sigma+\xi y_i}{(\sigma^2+\sigma\xi y_i)^2}$$

Therefore, since $\frac{d}{d\sigma}\left[-\frac{k}{\sigma}\right] = \frac{k}{\sigma^2}$, we have

$$\frac{\partial \ell^2(\sigma,\xi)}{\partial \sigma^2} = \frac{k}{\sigma^2} - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\frac{\xi y_i(2\sigma+\xi y_i)}{\sigma^2(\sigma+\xi y_i)^2} \quad \text{(A.1.5)}$$

$$= \frac{1}{\sigma^2}\left\{k - \sum_{i=1}^{k}\frac{y_i(2\sigma+\xi y_i)}{(\sigma+\xi y_i)^2}\xi - \sum_{i=1}^{k}\frac{y_i(2\sigma+\xi y_i)}{(\sigma+\xi y_i)^2}\right\}. \quad \text{(A.1.6)}$$

## A.2 Element 4

First not that

$$\frac{d}{d\xi}[-k\log\sigma] = 0 \quad \text{(A.2.1)}$$

Now consider

$$\frac{d}{d\xi}\left[\underbrace{-\left(1+\frac{1}{\xi}\right)}_{a_1}\underbrace{\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right)}_{a_2}\right]. \quad \text{(A.2.2)}$$

Expression (A.2.2) can be evaluated using the Chain (A.0.7), Product (A.0.5) and Quotient (A.0.6) rules by first applying the Product rule where $f(x) = a_1$ and $g(x) = a_2$, then applying the Quotient rule to $a_2$ and finally the Chain rule to $a_1$.

*Applying the Quotient Rule to $a_2$:*

$$\frac{d}{d\xi}\left[1+\frac{1}{\xi}\right] = -\frac{1}{\xi^2}$$

*Applying the Chain Rule to $a_1$:*

$$\frac{d}{d\xi}\left[\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right)\right] = \frac{y_i}{\sigma+\xi y_i}$$

where the elements of the Chain Rule are

$$f(u) = \log(u) \qquad f'(u) = \frac{1}{u}$$

$$g(\xi) = \sigma + \xi y_i \qquad g'(\xi) = \frac{y_i}{\sigma}$$

Putting the results from $a_1$ and $a_2$ into the Product Rule (A.0.5), we obtain the first derivative for $\frac{\partial \ell(\sigma,\xi)}{\partial \xi}$.

$$\begin{aligned}\frac{\partial \ell(\sigma,\xi)}{\partial \xi} &= -\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}+\frac{1}{\xi^2}\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right)\\ &= \frac{1}{\xi^2}\left\{-\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right)+\xi^2\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}+\right.\\ &\quad \left.\xi\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}\right\} \qquad \text{(A.2.3)}\end{aligned}$$

From the first derivative we can now calculate the second as follows. Consider

$$\frac{\partial^2 \ell}{\partial \xi^2}=\frac{d}{d\xi}\left[\underbrace{-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}}_{b_1}+\underbrace{\frac{1}{\xi^2}\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right)}_{b_2}\right]$$

. To evaluate the derivative of $b_1$ we use the Product Rule (A.0.5). Since $f$ is

$$f(\xi)=-\left(1+\frac{1}{\xi}\right)$$

so the derivative $f'(\xi)$ is

$$\frac{d}{d\xi}\left[-\left(1+\frac{1}{\xi}\right)\right]=\frac{1}{\xi^2};$$

similarly the $g$ element is

$$g(\xi)=\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}$$

, the derivative $g'(\xi)$ is

$$\frac{d}{d\xi}\left[\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}\right]=-\frac{y_i^2}{(\sigma+\xi y_i)^2}$$

. Now applying the Product Rule we obtain

$$-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\frac{y_i^2}{(\sigma+\xi y_i)^2}+\frac{1}{\xi^2}\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}$$

. Again, we need to apply the Product rule (A.0.5) to $b_2$. Since $f$ is

$$f(\xi) = \frac{1}{\xi^2}$$

, the derivative $f'(\xi)$ is

$$\frac{d}{d\xi}\left[\frac{1}{\xi^2}\right] = -\frac{2}{\xi^3}$$

; similarly the $g$ element is

$$g(\xi) = \sum_{i=1}^{k} \log\left(1 + \frac{\xi y_i}{\sigma}\right)$$

, so the derivative $g'(\xi)$ is

$$\frac{d}{d\xi}\left[\sum_{i=1}^{k} \log\left(1 + \frac{\xi y_i}{\sigma}\right)\right] = \sum_{i=1}^{k} \frac{y_i}{(\sigma + \xi y_i)}$$

. Inserting these into the Product rule (A.0.5) applied to $b_2$, we obtain

$$\frac{1}{\xi^2}\sum_{i=1}^{k} \frac{y_i}{(\sigma + \xi y_i)} - \frac{2}{\xi^3}\sum_{i=1}^{k} \log\left(1 + \frac{\xi y_i}{\sigma}\right)$$

The Addition rule (A.0.4) is used to complete the differentiation, giving the second derivative

$$\begin{aligned}\frac{\partial^2 \ell(\sigma, \xi)}{\partial \xi^2} &= -\frac{2}{\xi^3}\sum_{i=1}^{k} \log\left(\frac{\sigma + \xi\, y_i}{\sigma}\right) + \frac{2}{\xi^2}\sum_{i=1}^{k} \frac{y_i}{(\sigma + \xi\, y_i)} \\ &\quad + \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^{k} \frac{y_i^2}{(\sigma + \xi\, y_i)^2},\end{aligned} \tag{A.2.4}$$

which can be simplified to the following

$$\begin{aligned}\frac{\partial^2 \ell(\sigma, \xi)}{\partial \xi^2} &= \frac{1}{\xi^3}\left\{-2\sum_{i=1}^{k} \log\left(\frac{\sigma + \xi\, y_i}{\sigma}\right) + \xi\sum_{i=1}^{k} \frac{y_i}{(\sigma + \xi\, y_i)}\right. \\ &\quad \left. + \xi^3\sum_{i=1}^{k} \frac{y_i^2}{(\sigma + \xi\, y_i)^2} + \xi^2\sum_{i=1}^{k} \frac{y_i^2}{(\sigma + \xi\, y_i)^2}\right\}\end{aligned} \tag{A.2.5}$$

## A.3 Elements 2 & 3

Both these elements should be equal, and we confirm this by presenting both calculations. We initially calculate the second derivative by differentiating for $\xi$ then $\sigma$. From Section **A.2**, we can use the calculation of the first derivative of $\ell$ with respect to $\xi$.

$$\frac{\partial \ell(\sigma,\xi)}{\partial \xi} = \underbrace{-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}}_{c_1} + \underbrace{\frac{1}{\xi^2}\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right)}_{c_2}. \tag{A.3.1}$$

If we evaluate the derivative of (A.3.1) with respect to $\sigma$ using the Addition Rule (A.0.4), it is possible to consider $c_1$ and $c_2$ separately. We now consider $c_1$ when differentiated with respect to $\sigma$. This is found by using the Chain Rule (A.0.7):

$$-\left(1+\frac{1}{\xi}\right)\frac{d}{d\sigma}\left[\sum_{i=1}^{k}\frac{y_i}{\sigma+\xi y_i}\right] = \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)^2} \tag{A.3.2}$$

using

$$f(u)=\frac{1}{u} \qquad f'(u)=-\frac{1}{u^2}$$

$$g(\sigma)=\sigma+\xi y_i \qquad g'(\sigma)=1$$

Now differentiate $c_2$ using the Chain Rule (A.0.7)

$$\left(\frac{1}{\xi^2}\right)\frac{d}{d\sigma}\left[\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right)\right] = -\frac{1}{\xi^2}\sum_{i=1}^{k}\frac{\xi y_i}{\sigma^2\left(1+\frac{\xi y_i}{\sigma}\right)} \tag{A.3.3}$$

using

$$f(u)=\log(u) \qquad f'(u)=\frac{1}{u}$$

$$g(\sigma)=1+\frac{\xi y_i}{\sigma} \qquad g'(\sigma)=-\frac{\xi y_i}{\sigma^2}$$

Therefore, by expanding $c_1$, collecting the terms in $c_1$ and $c_2$ together, and taking out

the common factor, we obtain

$$\frac{\partial^2 \ell(\sigma,\xi)}{\partial\sigma\partial\xi} = \frac{1}{\xi\sigma}\left\{-\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)} + \xi\sigma\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)^2} + \sigma\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)^2}\right\} \tag{A.3.4}$$

To check for consistency we now present the calculation of the second derivative of the log-likelihood function when differentiating with respect to $\sigma$ then $\xi$. From **A.1**, we can use the calculation of the first derivative of $\ell$ with respect to $\sigma$.

$$\frac{\partial \ell(\sigma,\xi)}{\partial\sigma} = \underbrace{-\frac{k}{\sigma}}_{d_1} + \underbrace{\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\frac{\xi y_i}{\sigma^2\left(1+\frac{\xi y_i}{\sigma}\right)}}_{d_2} \tag{A.3.5}$$

Consider $d_1$ and note that $\frac{\partial}{\partial\xi}\left[-\frac{k}{\sigma}\right] = 0$. Now consider splitting $d_2$ into the following components and using Product Rule (A.0.5),

$$\frac{\partial}{\partial\sigma}\left[\underbrace{\overbrace{\left(1+\frac{1}{\xi}\right)}^{e_1}\overbrace{\sum_{i=1}^{k}\frac{\xi y_i}{\sigma^2\left(1+\frac{\xi y_i}{\sigma}\right)}}^{e_2}}_{d_2}\right] \tag{A.3.6}$$

Next evaluate $\frac{\partial}{\partial\xi}[e_1]$:

$$\frac{\partial}{\partial\xi}\left[1+\frac{1}{\xi}\right] = -\frac{1}{\xi^2}; \tag{A.3.7}$$

now evaluate $\frac{\partial}{\partial\xi}[e_2]$:

$$\frac{\partial}{\partial\xi}\left[\sum_{i=1}^{k}\frac{\xi y_i}{\sigma^2\left(1+\frac{\xi y_i}{\sigma}\right)}\right] = \frac{1}{\sigma}\frac{\partial}{\partial\xi}\left[\sum_{i=1}^{k}\frac{\xi y_i}{(\sigma+\xi y_i)}\right]. \tag{A.3.8}$$

We evaluate this using the Quotient Rule (A.0.6)

$$f(\xi) = \xi y_i \qquad f'(\xi) = y_i$$

$$g(\xi) = \sigma + \xi y_i \qquad g'(\xi) = y_i$$

. Substituting back, we obtain

$$\frac{1}{\sigma}\frac{\partial}{\partial\xi}\left[\sum_{i=1}^{k}\frac{\xi y_i}{(\sigma+\xi y_i)}\right]=\frac{1}{\sigma}\sum_{i=1}^{k}\frac{y_i\sigma}{(\sigma+\xi y_i)^2}=\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)^2}. \tag{A.3.9}$$

Now that parts $e_1$ and $e_2$ have been differentiated we can insert the derivative into the Product Rule (A.0.5) to obtain the derivative of $d_2$ which yields

$$\frac{\partial^2\ell(\sigma,\xi)}{\partial\sigma\partial\xi}=\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)^2}-\frac{1}{\sigma\xi}\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}. \tag{A.3.10}$$

We now simplify this to the same form as $\frac{\partial^2\ell(\sigma,\xi)}{\partial\sigma\partial\xi}$. Firstly expand (A.3.10) together,

$$\frac{\partial^2\ell(\sigma,\xi)}{\partial\xi\partial\sigma}=\frac{1}{\xi}\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)^2}+\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)^2}-\frac{1}{\sigma\xi}\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)} \tag{A.3.11}$$

Now we can extract the common term $\frac{1}{\sigma\xi}$ and simplify,

$$\frac{\partial^2\ell(\sigma,\xi)}{\partial\xi\partial\sigma}=\frac{1}{\sigma\xi}\left\{\sigma\xi\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)^2}+\sigma\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)^2}-\sum_{i=1}^{k}\frac{y_i}{(\sigma+\xi y_i)}\right\} \tag{A.3.12}$$

Hence,

$$\frac{\partial^2\ell(\sigma,\xi)}{\partial\xi\partial\sigma}=\frac{\partial^2\ell(\sigma,\xi)}{\partial\sigma\partial\xi}$$

# B

# HR Wallingford Hindcast General Methodology

## B.1 HINDWAVE Wave Generation Model

The purpose of Hindcasting is to provide large quantities of reliable wave data for sites for which otherwise little measured data is available. The process needs to be cost effective, reliable and also relatively fast due to the demand for speed within the engineering industry.

The HINDWAVE wave generation model has two sources of input: the first represents the changing wind velocity at the location of interest and the other refers to the shape of the wave generation area. The program can function with just these inputs to produce directionally dependent wave distributions, or can be used with additional inputs such as measured wave data. Fetch length measured around the prediction point at 10 degree intervals, can be used as another input to the HINDWAVE model; it is primarily used in the sub-model for wave generation called JONSEY. The process which the model follows can be divided into two stages. The initial stage generates several hundred possible wave conditions based on the input wind information. The second stage uses the generated possible wave conditions as a reference and matches the most appropriate wave condition to a set of corresponding wind speed and direction data obtained from the Met Office. The outcome is a time series data set of wave conditions for a particular location.

The model pairs the duration and dominant set of wind conditions at the specified

location with reference wave conditions by looking at each hourly or half hourly time segment. By vectorially averaging the wind velocities over the specified durations preceding them, an average speed and direction is found for each record. The largest of these is then chosen from the corresponding wave height values, along with the corresponding wave period and wave direction generated data to form the wave data sets. The model for particular locations may be based on data collected at a coastal location. If the prediction point is an offshore location then the model has to be calibrated using a speed up function.

## B.2 The TELURAY Refraction Model

Any wave on the sea surface is subject to a number of external and internal forces acting upon it, both simultaneously and independently, which continuously alter its behaviour. The effect of these forces depends to a large extent on the depth of water in comparison to the wave length. When the depth of water is large in comparison to wave length the predominant forces acting are the stresses as a result of wind action and internal viscosity. Alternatively, if the depth of water is comparatively shallow then the effects of the sea bed becomes a predominant factor.

In particular two cases are considered (see Wallingford, 2005b): shoaling and depth refraction. The shoaling effect is due to changes in wave height as a result of the waves slowing down as they travel through water of decreasing depth. Depth refraction generally occurs due to the waves travelling towards the coast undergoing a gradual change in direction as a consequence of a change in depth. This means the wave crests will have a tendency to align with the seabed contours.

Another strong effect is current refraction which is predominant in areas with strong tidal currents that can influence waves. This influence is dependent on the spatial change of the current strength, and its direction relative to wave direction. If we think of waves entering a region of opposing currents of either increasing or decreasing strength then the waves will consequently be either steepened or stretched respectively. This in practice changes the ratio of wave height to wave length, i.e. stretched waves correspond to smaller wave height and longer wave length. However, in the majority of cases the course of the waves will be altered according to the current direction.

Table B.1: Table showing the significance of current effects on waves depends on the peak tidal velocity

| Peak Tidal Velocity | Ranked Significance |
|---|---|
| $< 1$ ms$^{-1}$ | not significant |
| $1 - 2$ ms$^{-1}$ | may be important |
| $> 2$ ms$^{-1}$ | likely to be important |

Further information on this model, please contact Peter Hawkes at HR Wallingford.

# Bibliography


H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 6(19):716–723, 1974.

N. W. H Allsop, D. Vincinanza, M. Calabrese, and L. Centurioni. Breaking wave impact loads on vertical faces. *ISOPE Conference*, 1996.

W. Allsop and T. Pullen. *Wave Overtopping of Simple Embankments: Improved Methods*, 2003. Coastal Flooding Hazard by Wave Overtopping SHADOW Phase 1, R & D Interim guidance note: FD2410/GN2, Environment Agency.

C. Annis. Joint, marginal and conditional distributions, January 2006. URL : `http://www.statisticalengineering.com/joint_marginal_conditional.html`.

M. Bédard. Efficient sampling using algorithms: Applications of optmial scaling results, July 2006a. URL : `http://probability.ca/jeff/ftpdir/mylene3.pdf`.

M. Bédard. Optimal acceptance rates for Metropolis Hastings algorithms: Moving beyond 0.234. *Submitted for publication in the Annals of Statistics*, July 2006b.

J. Beirlant, Y. Goegobeur, J. Segers, and J. Teugels. *Statistics of Extremes*. Wiley, West Sussex, 2004.

P. Besley. *Overtopping of seawalls design and assessment manual*, 1999. R & D Technical Report W 178, ISBN 1 85705 069 X, Environment Agency.

R. Bosch, Y. Ye, and G.G. Woodworth. A convenient algorithm for quantile regression with smoothing splines. *Computational Statistics and Data Analysis*, 19:613–630, 1995.

S.P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

E. Castillo, S. Hadi, N. Balakrishnan, and J. Sarabia. *Extreme Value and Related Models with Applications in Engineering and Science.* Wiley, Hoboken, N.J, 2005.

CHL. Engineering glossary by CHL (Coastal and Hydraulics Laboratory, U.S. Army Corps of Engineers), 2008. URL : `http://chl.erdc.usace.army.mil/`.

CIRIA/CUR. *Manual on the use of rock in coastal and shoreline engineering*, 1991. CIRIA special publication 83/CUR Report 154.

S. Coles. *An Introduction to Statistical modelling of Extreme Values.* Springer, London, 2001.

S. Coles and A. Stephenson. *ismev: An Introduction to Statistical Modeling of Extreme Values*, 2006. URL : `http://www.maths.lancs.ac.uk/stephena/`. R package version 1.2.

S.G. Coles and J.A. Tawn. Modelling extreme multivariate events. *Journal of the Royal Statistical Society: Applied Statistics*, 53(2):377–392, 1991.

A.C. Davidson and R.L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Applied Statistics*, 52(3):393–442, 1990.

C. de Boor. *A Practical Guide to Splines.* Springer-Verlag, New York, 1978.

F. de Pasquale, P. Barone, G. Sebastiani, and J. Stander. Bayesian analysis of dynamic magnetic resonance breast images. *J. R. Stat. Soc: Applied Statistics*, 53(3):475–493, 2004.

J. Devore and R. Peck. *Introductory Statistics.* West Publishing Company, Minneapolis/St.Paul, 2nd edition, 1994.

S. Dineen. *Multivariate Calculus and Geometry.* Springer, London, 1998.

D.J. Dupuis. Exceeedances over high thresholds: A guide to threshold selection. *Extremes*, 1(3):251–261, 1999.

B. Efron and R.J. Tibshirani. *Introduction to the Bootstrap*. Chapman and Hall, London, 1993.

S.R. Eliason. *Maximum Likelihood Estimation: Logic and Practice*. Sage University Paper, Newbury Park, California., 1993.

EurOtop. *EurOtop Manual (collaboration of Environment Agency (UK) and Expertise Netwerk Waterkeren (NL) and Kuratorium fr Forschung im Ksteningenieurwesen (DE).)*, 2007. URL : `http://www.overtopping-manual.com`.

K. Ewans and P. Jonathan. Estimating extreme wave design criteria incorporating directionality. In *9th International Workshop on Wave Hindcasting and Forecasting*, Victoria, Canada, September 2006.

B. Finkenstadt and H. Rootzen. *Extreme Values in Finance, Telecommunications, and the Environment*. Chapman and Hall, Boca Raton, Florida, 2004.

J.E. Freund. *Mathematical Statistics*. Prentice Hall, New Jersey, 5th edition, 1992.

D. Gamerman. *Markov Chain Monte Carlo: Stocastic simulation for Bayesian inference*. Chapman and Hall, London, 1997.

A. Gelman. *Inference and monitoring convergence*. Chapman and Hall, 1996. Featured in: W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in Practice*.

A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7:457–511, 1992.

D. Geman and S. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans.Pattn Anal.Mach.Intell.*, (6):721–741, 1984.

W.R. Gilks, S. Richardson, and D.J. Spielgelhalter. *Markov Chain Monte Carlo In Practice.* Chapman and Hall, Suffolk, 1996.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

P.J. Green and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall/CRC, Cambridge, 1994.

P.E. Greenwood and M.S. Nikulin. *A Guide to Chi-Squared Testing.* Probability and Statistics. John Wiley and Sons, New York, 1996.

A. Guillou and P. Hall. A diagnostic for selecting the threshold in extreme value analysis. *J. R. Stat. Soc. B*, 63(2):293–305, 2001.

F.E. Harrell. Jr. *Regression Modelling Strategies.* Springer, New York, 2001.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York, 2001.

W.K. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57:97–109, 1970.

R.V. Hogg and A.T. Craig. *Introduction to Mathematical Statistics.* Prentice Hall, New Jersey, 5th edition, 1995.

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

H. Joe. *Multivariate Models and Dependence Concepts.* Chapman and Hall, New York, 1997.

P. Jonathan and K. Ewans. The effect of directionality on extreme wave design criteria. *Ocean Engineering*, 34:1977–1994, 2007.

R.E. Kass, B.P. Calin, A. Gelman, and R.M. Neal. MCMC in practice: A roundtable discussion. *The American Statistician:Statistical Practice*, 52(2):93–100, 1998.

R. Koenker. *Quantile Regression*. Cambridge University Press, New York, 2005.

R. Koenker and K. Hallock. Quantile regression: An introduction, 2000. URL : `http://www.econ.uiuc.edu/ roger/research/home.html`.

R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

Roger Koenker. ***quantreg**: Quantile Regression*, 2008. URL `http://www.r-project.org`. R package version 4.24.

S. Kotz and S. Nadarajah. *Extreme Value Distributions*. Imperial College Press, London, 2002.

S. Lang. *Calculus of Several Variables*. Springer-Verlag, New York, 3rd edition, 1987.

A.W. Ledford and J.A. Tawn. Modelling dependence within joint tail regions. *J. R. Stat. Soc. B*, 59(2):475–499, 1997.

A.W. Ledford and J.A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187, 1996.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 6(21): 1087–1092, 1953.

I.D. Mockett and J.D. Simm. *Risk Levels in Coastal Engineering Works*. Thomas Telford, London, 2002.

C.Z. Mooney and R.D. Duval. *Bootstrapping: A Nonparametric Approach to Statistical Inference.* Sage University Paper, London, 1993.

M. W. Owen. *Design of seawalls allowing for wave overtopping*, 1980a. HR Wallingford Report (EX 924).

M.W. Owen. *Design of sea walls allowing for wave overtopping*, 1980b. Report EX 924, Hydraulics Research, Wallingford.

R. Porkess. *Collins Dictionary of Statistics.* Collins, London, 2005.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008. URL : `http://www.R-project.org`. ISBN 3-900051-07-0.

A. Ramos and A. Ledford. A new class of models for bivariate joint tails. *J. R. Stat. Soc. B*, 1(71):219–241, 2009.

C.R. Rao. *Linear Statistical Inference and its Applications (2nd Ed.).* John Wiley & Sons, New York, 1973.

D. Reeve, A. Chadwick, and C. Fleming. *Coastal Engineering: Processes, Theory and Design Practice.* SPON, London, 2004.

P.B. Sayers, B.P. Gouldby, J.D. Simm, I. Meadowcroft, and J. Hall. Defra & environmental agency R&D programme: Risk, Performance and Uncertainty in Flood and Coastal Defence – A review, 2002.

B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric curve fitting. *J. R. Stat. Soc: Statistical Methodology*, 1(47):1–52, 1985.

R. Smith. Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72 (1):67–90, 1985.

G. Snow. `TeachingDemos`: *Demonstrations for teaching and learning*, 2008. R package version 2.3.

R.M. Sorenson. *Basic Coastal Engineering*. Wiley, London, 1978.

A. Tancredi, C. Anderson, and A. O'Hagan. Accounting for threshold uncertainty in extreme value estimation. *Extremes*, 9:86–106, 2006.

J.A. Tawn. Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3): 397–415, 1988.

J.A. Tawn and J.T. Bruun. Comparison of approaches for estimating the probability of coastal flooding. *Journal of the Royal Statistical Society: Applied Statistics*, 47(3):405–423, 1998.

J.A. Tawn and S.G. Coles. Statistical methods for multivariate extremes: an application to structural design. *Journal of the Royal Statistical Society: Applied Statistics*, 43(1):1–48, 1994.

P. Thompson, D. Reeve, Y. Cai, J. Stander, and R. Moyeed. Bayesian non-parametric quantile regression using splines for modelling wave heights. *FloodRisk 2008 Conference Proceedings*, 2008.

Defra UK. Flood and Coastal Defence Project Appraisal Guidance Overview (including general guidance), 2001. URL : `http://www.defra.gov.uk/environ/fcd/pubs/pagn/fcdpag1.pdf`.

W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S (4th Ed)*. Springer-Verlag, New York, 2002.

HR Wallingford. *The HINDWAVE Wave Prediction Model*. HR Wallingford, Howbery Park, Wallingford, Oxon, OX10 8BA, 2005a.

HR Wallingford. *The TELURAY Wave Refraction Model.* HR Wallingford, Howbery Park, Wallingford, Oxon, OX10 8BA, 2005b.

HR Wallingford. *The Joint Probability of Waves and Water Levels: JOINSEA.* HR Wallingford, Howbery Park, Wallingford, Oxon, OX10 8BA, report: tr71 edition, November 1998a.

HR Wallingford. *The Joint Probability of Waves and Water Levels: JOINSEA.* HR Wallingford, Howbery Park, Wallingford, Oxon, OX10 8BA, report: sr537 edition, November 1998b.

D. Walshaw and S.G. Coles. Directional modelling of extreme wind speeds. *Journal of the Royal Statistical Society: Applied Statistics*, 43(1):139–157, 1994.

K. Yu and R. Moyeed. Bayesian quantile regression. *Statistics and Probability Letters*, 54: 437–447, 2001.

K. Yu, Z. Lu, and J. Stander. Quantile regression: application and current research areas. *Journal of the Royal Statistical Society: The Statistician*, 52(3):331–350, 2003.

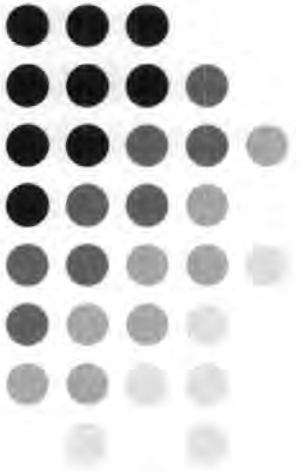
A. Zacharioudaki. *PhD Thesis: Mathematical Modelling of Shoreline Evolution under Climate Change.* University of Plymouth, School of Engineering, May 2008.

# *Statistical Methods for Extreme Wave Condition Analysis in Coastal Design*

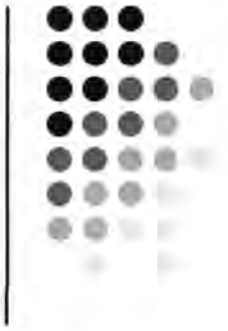
Paul Thompson

PhD presentation

27/02/09



# *Aims*

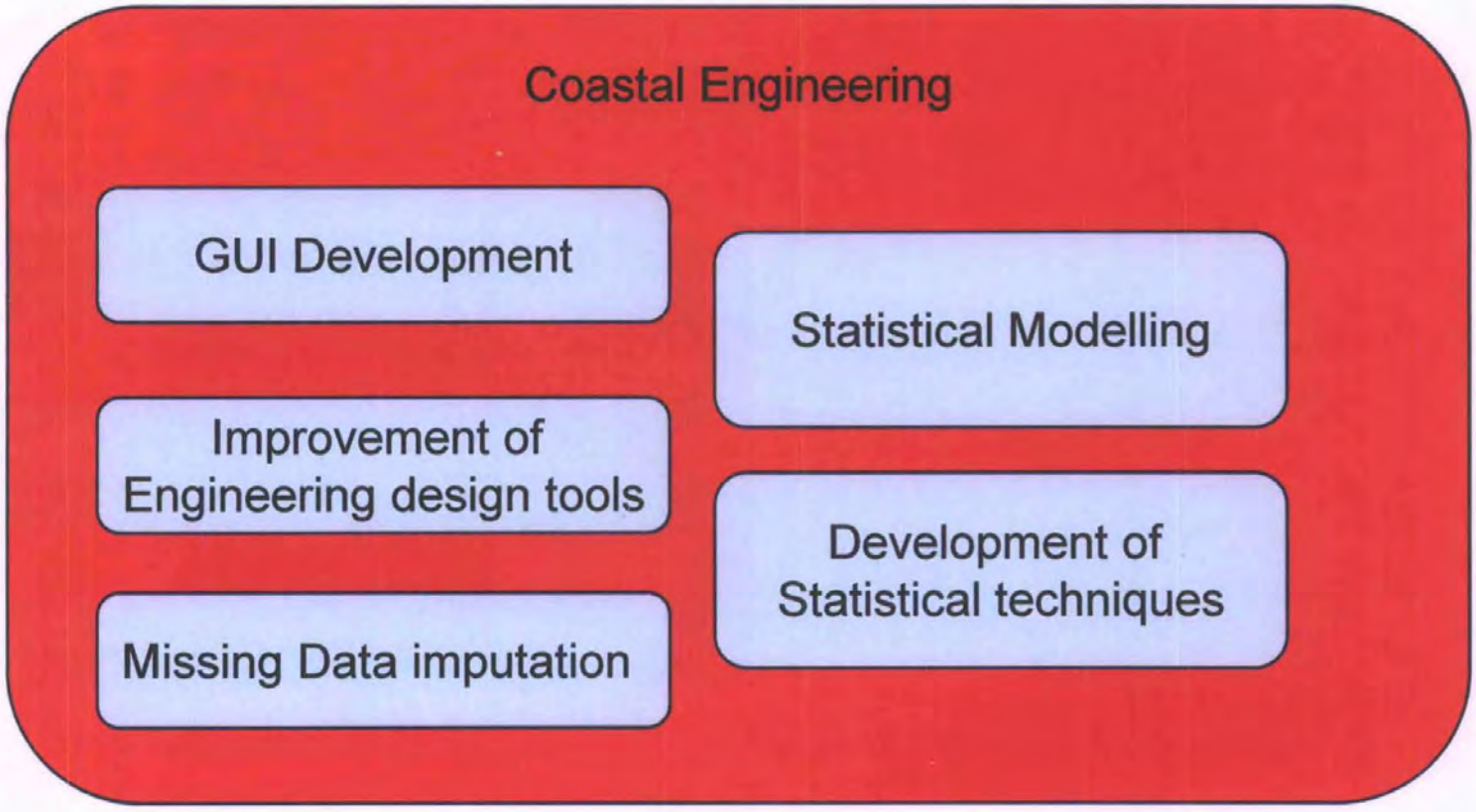


- To produce more reliable estimates of future conditions that a coastal defence structure will need to withstand.

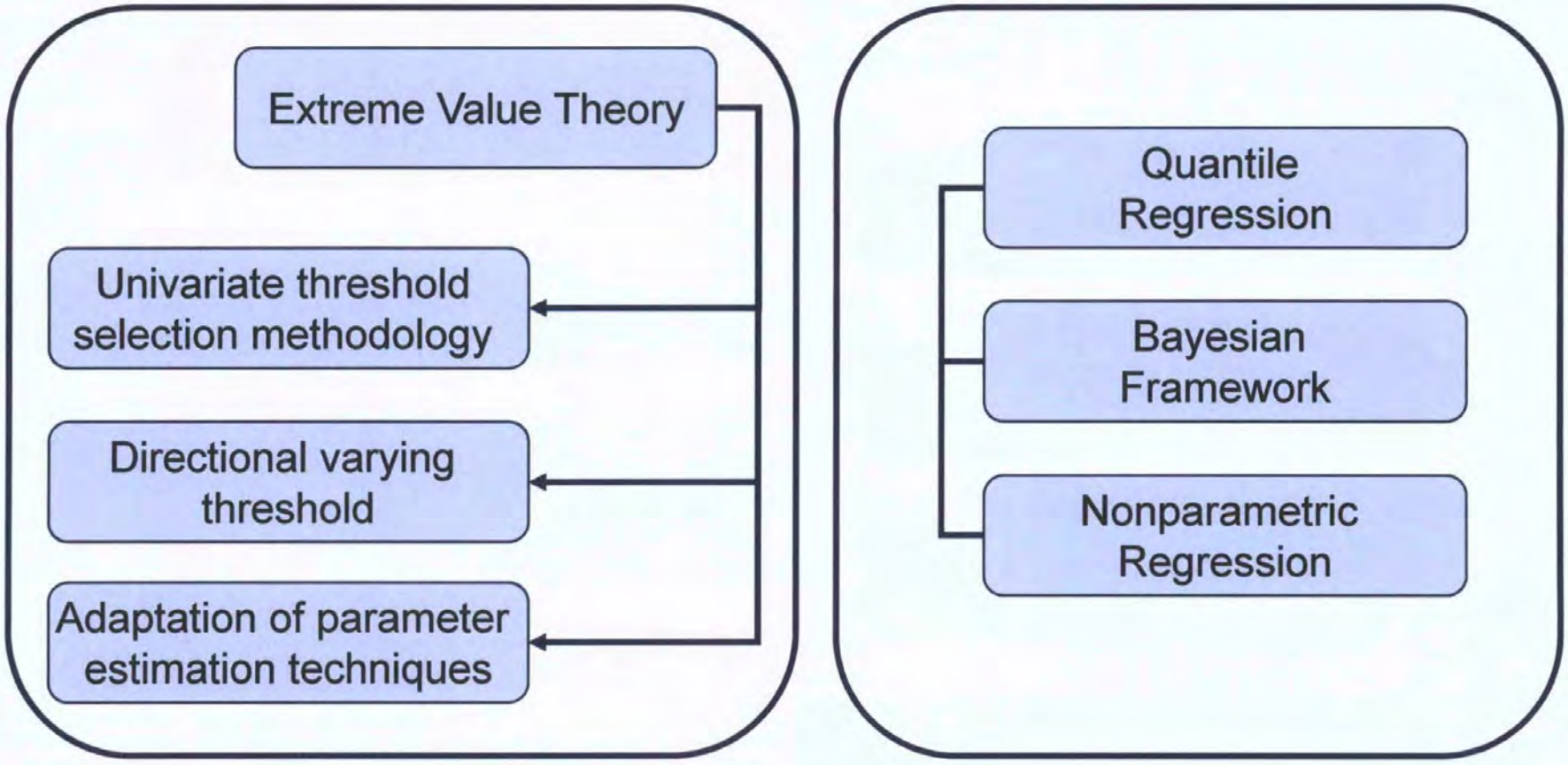
- Improvement of existing techniques for identifying data for extreme value modelling.

- Investigation of alternate methodology for understanding dependence of extreme values on covariates such as wave direction.

# PhD Structure



# *Statistical modelling & development*




PART 1

PART 2

# *Additional development resulting form model improvement*

- Return level plots derived via the quantile regression methodology.
- Directional dependent return level plots.
- Graphical user interface prototype to aid implementation of developed techniques to the wider engineering community.

# *Chapter summary*

- **Chapter 1**: Introduction and broader motivations for the project.
- **Chapter 2**: Literature review of Extreme Value Theory techniques.
- **Chapter 3**: Review of the JOINSEA joint probability software.
- **Chapter 4**: Literature review of Bayesian approach, and nonparametric and quantile regression.
- **Chapter 5**: Technique for replacing missing information in the data sets.
- **Chapter 6**: Presentation of developed techniques for use in Extreme Value Theory.
- **Chapter 7**: Presentation of developed techniques using Bayesian methods, and nonparametric and quantile regression.
- **Chapter 8**: Discussions and Future work.

# Automated Threshold Selection Methods for Extreme Wave Analysis

Paul Thompson [a,*], Yuzhi Cai [b], Dominic Reeve [a], Julian Stander [b]

[a] *C-CoDE, School of Engineering, University of Plymouth, Devon, PL4 8AA, UK*

[b] *School of Mathematics & Statistics, University of Plymouth, Devon, PL4 8AA, UK*

**Abstract**

The study of the extreme values of a variable such as wave height is very important in flood risk assessment and coastal design. Often values above a sufficiently large threshold can be modelled using the Generalized Pareto Distribution, the parameters of which are estimated using maximum likelihood. There are several popular empirical techniques for choosing a suitable threshold, but these require the subjective interpretation of plots by the user.

In this paper we present a pragmatic automated, simple and computationally inexpensive threshold selection method based on the distribution of the difference of parameter estimates when the threshold is changed, and apply it to a published rainfall and a new wave height data set. We assess the effect of the uncertainty associated with our threshold selection technique on return level estimation by using the bootstrap procedure. We illustrate the effectiveness of our methodology by a simulation study and compare it with the approach used in the JOINSEA software. In addition, we present an extension that allow the threshold selected to depend on the value of a covariate such as the cosine of wave direction.

*Key words:* Bootstrap, Covariate dependent thresholds, Distribution with Generalized Pareto tail, Generalized Pareto distribution, GPD, JOINSEA, Return level confidence intervals.

* Corresponding author.
*Email address:* p1thompson@plymouth.ac.uk (Paul Thompson).

## 1 Introduction

The successful design of a reliable and effective coastal defence structure can be associated primarily with knowledge of future extreme conditions which the defence must withstand. Typically, coastal defences are designed to provide sufficient protection against flooding or erosion to a desired return level associated with a particular return period, e.g. 100 years. The estimation of return levels and their uncertainty therefore has considerable engineering importance, especially in the area of coastal defence design. Statistical methodology for such estimation tasks requires as its input data about the extreme values of the conditions of interest.

There are two main methods for defining extremes. The first is based on dividing the time period over which the data are collected into blocks, with the most extreme value in each block being used for future analysis (e.g. daily or monthly maxima). The second method is based on exceedances over a specified threshold. In this paper we concentrate on the excesses over a threshold and provide an automated and computationally inexpensive threshold specification technique. Before presenting our technique, it is necessary to discuss how excesses over a suitable threshold can be modelled and analysed statistically.

Let $y$ be a value taken by the variable of interest, for example wave height, and let $u$ be a threshold. Provided $u$ is sufficiently large, values of $y$ greater than $u$ can be modelled using the generalized Pareto Distribution (GPD); see Coles [1], for example. The cumulative distribution function $H$ of the GPD takes the form:

$$H(y) = 1 - \left[1 + \frac{\xi(y-u)}{\sigma_u}\right]^{-1/\xi}, \tag{1}$$

where $y > u$ and $1 + \xi(y-u)/\sigma_u > 0$. The parameters $\sigma_u$ and $\xi$ control the scale and shape of the distribution. Here we use the notation $\sigma_u$ to emphasize that the scale parameter changes with the threshold $u$, although we will drop the subscript $u$ when this emphasis is no longer needed; the shape parameter $\xi$ does not change with $u$. The parameters $\sigma_u$ and $\xi$ need to be estimated from available data, and this can be done using maximum likelihood estimation, as discussed in detail in Coles [1] and Smith [18]. Usually, selection of an appropriate threshold $u$ is performed on a visual basis and so can have a range of associated errors. These visual procedures require prior knowledge of the accurate interpretation of threshold choice plots, such as the Mean Residual Life plot, to achieve a satisfactory model fit; again see Coles [1] for examples. We illustrate the difficulties associated with the interpretation of the Mean Residual Life plot in Section 2.

Threshold selection has received some additional attention in the literature, for example, Dupuis [4] presents a guide to threshold selection based on robust-

ness considerations, while Tancredi et al. [19] adopt a Bayesian approach and discuss how to take account of threshold uncertainty. The methods presented in these papers are complicated to implement and can be computationally demanding; see Section 2.3 for further discussion of Tancredi et al. [19] and Guillou & Hall [10] for related methodology. The automated threshold selection method that we will present requires little external input other than the variable of interest, and is considerably simpler and easier to implement than the approaches proposed in these papers.

We have also extended our threshold selection method to allow threshold choice to depend on a covariate such as the cosine of wave direction, where our specific aim is to account for the directional effect when modelling wave height or wave period using GPDs. The practical advantage of our extended procedure is that it automatically identifies the wave directions associated with the highest waves and consequently can provide better estimation of wave height return levels.

The rest of this paper is organized as follows. In Section 2 we present our automated threshold selection technique and compare it with one of the currently available subjective approaches. We also describe a bootstrap procedure for assessing the effect of uncertainty on return level estimation. In Section 3 we describe a simulation study aimed at quantifying the effectiveness of our method. In Section 4 we compare our approach with the existing methodology used in the JOINSEA software (see [12] and [13]). In Section 5 we extend our method to allow threshold choice to depend on a covariate. Finally, in Section 6 we present some concluding comments.

## 2 Automated Threshold Selection Technique

### *2.1 Theoretical Basis*

When fitting the GPD to data, the scale and shape parameters $\sigma_u$ and $\xi$ can be estimated using maximum likelihood estimation. To achieve a good model fit, we need to choose a suitable value of the threshold $u$. Commonly used techniques involve visual assessment of threshold choice plots and rely upon prior experience of their interpretation; see Tawn & Coles [20] and Davidson & Smith [3]. Such plots are found in Coles [1] for GPDs fitted to rainfall data. We shall discuss one of these plots, the Mean Residual Life plot, in Section 2.2.2 below. Another of these techniques plots parameter estimates of GPDs fitted using a range of thresholds against the threshold, and is the basis for our automated threshold selection methodology. We now outline our automated method for threshold selection.

Let $u_1, \ldots, u_n$ be $n$ equally spaced increasing candidate thresholds. Let $\hat{\sigma}_{u_j}$ and $\hat{\xi}_{u_j}$ be maximum likelihood estimators of the scale and shape parameter based on data above the threshold $u_j$, $j = 1, \ldots, n$. Finally, let $u$ be a suitable threshold, that is one for which values of $y > u$ can be modelled using the GPD. It follows from Coles [1], page 83 that, provided $u \leq u_{j-1} < u_j$,

$$\sigma_{u_{j-1}} = \sigma_u + \xi(u_{j-1} - u) \text{ and } \sigma_{u_j} = \sigma_u + \xi(u_j - u). \tag{2}$$

Hence,

$$\sigma_{u_j} - \sigma_{u_{j-1}} = \xi(u_j - u_{j-1}). \tag{3}$$

Furthermore, standard maximum likelihood theory, as discussed in Coles [1], tells us that $E[\hat{\sigma}_{u_j}] \approx \sigma_{u_j}$ and $E[\hat{\xi}_{u_j}] \approx \xi$, for any $j$ such that $u_j > u$. Let

$$\tau_{u_j} = \hat{\sigma}_{u_j} - \hat{\xi}_{u_j} u_j, \quad j = 1, \ldots, n, \tag{4}$$

and consider the differences

$$\tau_{u_j} - \tau_{u_{j-1}}, \; j = 2, \ldots, n; \tag{5}$$

it follows from the above results about the expected values of maximum likelihood estimators and from (3) that $E[\tau_{u_j} - \tau_{u_{j-1}}] \approx 0$. Moreover, we can appeal to the same theory to conclude that $\tau_{u_j} - \tau_{u_{j-1}}$ approximately follow a normal distribution. The variability of this difference does not itself measure the variability associated with our threshold selection procedure. This distributional result suggests the following procedure for finding a suitable threshold $u$:

(1) Identify suitable values of equally spaced candidate thresholds $u_1 < u_2 < \cdots < u_n$. We found that setting $n = 100$ gives good results. We take $u_1$ to be the median and $u_n$ to be the 98% quantile of the data, unless fewer than 100 values exceed this value, in which case $u_n$ is set to the $100^{\text{th}}$ data value in descending order. Our procedure performs well in such circumstances. Less reliable results were obtained from smaller data sets.

(2) If $u$ is a suitable threshold, then all differences $\tau_{u_j} - \tau_{u_{j-1}}$ have an approximate normal distribution with mean 0 provided $u \leq u_{j-1} < u_j$. If $u$ is unsuitable, then these differences may not follow a normal distribution. This suggests that a suitably applied test for normality is an effective method to determine $u$.

The Pearson's Chi-square Test is used as a test of goodness of fit to establish whether or not the observed differences are consistent with a normal distribution with mean 0; see Greenwood & Nikulin [9]. Initially,

we consider $u = u_1$ and perform the Pearson normality test based on all the differences $\tau_{u_2} - \tau_{u_1}, \tau_{u_3} - \tau_{u_2}, \ldots, \tau_{u_n} - \tau_{u_{n-1}}$. If the null hypothesis of normality is not rejected, $u$ is taken to be a suitable threshold. If the null hypothesis is rejected, then we consider $u = u_2$, remove $\tau_{u_2} - \tau_{u_1}$ from the set of differences considered, and repeat the above procedure. We have found from a simulation study that a size 0.2 Pearson normality test generally performs most consistently over a range of normality tests and sizes. Reducing the size of the test has the effect of lowering the chosen threshold.

(3) Step 2 is repeated until the Pearson's normality test indicates that the differences are consistent with a normal distribution with mean 0. If this does not happen, $u_n$ is returned with a warning. Our experience is that this latter situation occurs rarely.

The above steps can be performed quickly, so yielding a procedure that is computationally inexpensive. We implemented our method in the freely available, open source statistical environment `R` [16], which is becoming more widely used in engineering and related areas.

### *2.2 Practical Examples*

#### *2.2.1 Coastal Wave Data*

We now apply the method presented in Section 2.1 to a real data set. The data used in this example relate to conditions near the Selsey Bill area (Hawkes, personal communication). They were generated using the hindcast technique (see Reeve et al. [17], for example) based on wind records. The data set consists of hourly hindcast measurements of the variables significant wave height, wave period and wave direction over a time span of 27 years. Wave hindcasting attempts to create the wind-wave conditions, and cannot account for the swell component. In this example we take a random sample of 10,000 observations from the data set. The resulting values are typical of data that are collected in similar studies and satisfy the independence assumption that underlie maximum likelihood theory. A plot of wave height against the cosine of wave direction is shown in Figure 1.

Our automated threshold selection technique was applied to these wave height observations and indicated 0.487 m as a suitable threshold. This threshold is also shown in Figure 1. The values of the cosine of wave direction were not used in finding this threshold. Figure 2 plots differences $\tau_{u_j} - \tau_{u_{j-1}}$ against threshold $u_{j-1}$, and as described in Section 2.1 is the basis of our threshold selection procedure. Figure 3 shows diagnostic plots, as discussed by Coles [1] and produced by the freely available `ismev` package of Coles & Stephenson [2]

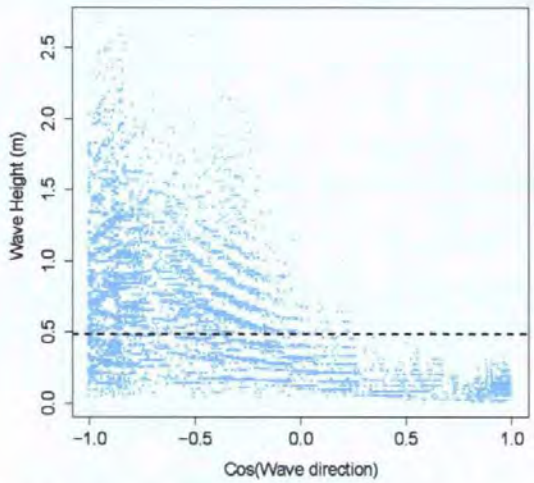



Figure 1: Scatter plot of wave height against the cosine of wave direction for 10,000 values from the Selsey Bill data set. The horizontal line was produced by applying our automated threshold selection procedure to the wave height observation, taking no account of the cosine of wave direction.

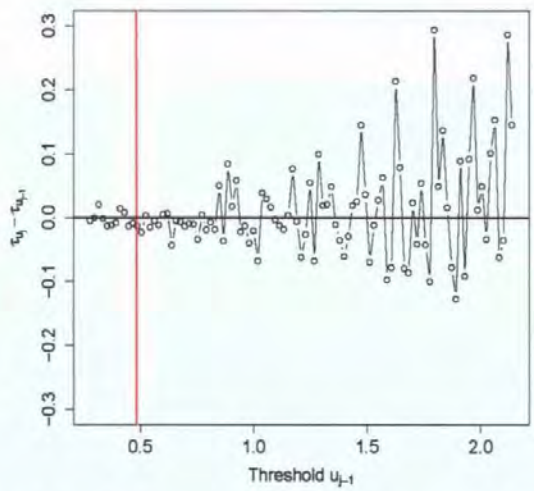



Figure 2: Graph of the differences $\tau_{u_j} - \tau_{u_{j-1}}$ against threshold $u_{j-1}$ for the wave height data. The vertical line indicates the automated threshold selection choice.

run in R [16]. Such plots are now used routinely, and so have not been edited here; detailed explanation is provided in the caption. These diagnostic plots indicate that the fitted GPD model is satisfactory. Both the probability and quantile plots show that there is little difference between empirical and fitted values from the model, indicating a good fit. Similarly, there is reasonable agreement between the data and the estimated return levels and associated

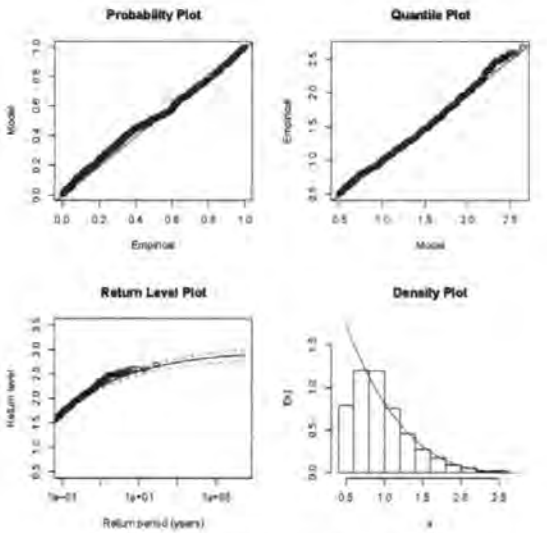



Figure 3: Diagnostic plots for the GPD fit when the threshold is chosen using our automated threshold selection approach applied to the wave height data. This plot was generated using the `ismev` package [2]. In the third plot Return level refers to wave height (m). In the fourth plot $x$ refers to the wave height (m), and $f(x)$ to its probability density. See text for discussion of the individual plots.

95% confidence envelope, and between the histogram of the data values above the chosen threshold and the fitted generalized Pareto density. This example shows that our proposed methodology can provide an automated, simple and computationally inexpensive threshold selection method that avoids the need for subjective interpretation of threshold choice plots with all their possible errors.

### *2.2.2 Daily Rainfall Data*

We now compare the automated threshold selection method presented in Section 2.1 with a currently available subjective method by applying them to a data set considered by Coles [1]. The data comprise daily rainfall accumulations at a location in south west England recorded over the period 1914–1962. Coles [1] presents this example to illustrate the currently available threshold selection techniques. Figure 4 shows a plot of the data together with the threshold of 30 mm as recommended by Coles [1] and our own automated choice of 20 mm. Figure 5 shows the Mean Residual Life plot (see Coles [1] for details) upon which Coles bases his choice. A threshold is usually identified as a value beyond which the plot is linear (up to sampling error). The behaviour of the plot is linear (up to sampling error) beyond 60 mm, but few data points lie above this value. Linearity also occurs between 30 and 60 mm, and so Coles [1] recommends a value of 30 mm. A similar argument could also

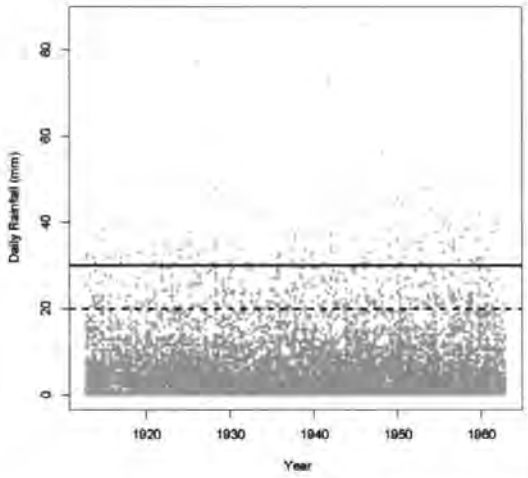



Figure 4: Scatter plot of daily rainfall data against time. The dashed line shows our automated threshold choice, while the unbroken line is the threshold value recommended by Coles [1].

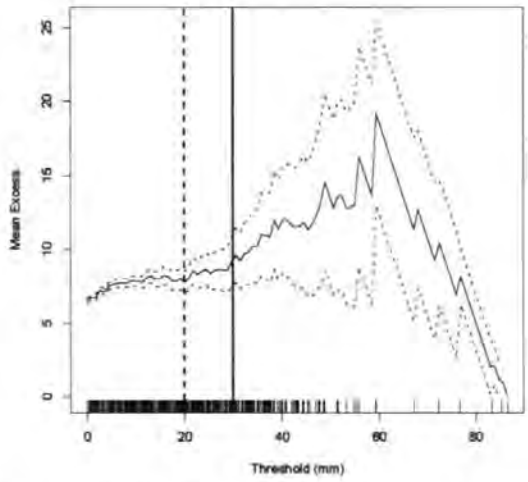
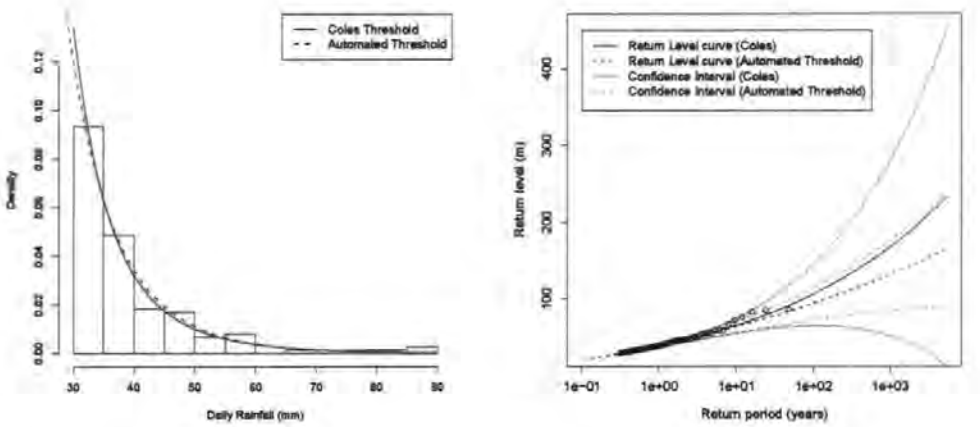



Figure 5: Mean Residual Life plot for the daily rainfall data. The dashed line was produced by our automated threshold selection procedure. We have also added the threshold value recommended by Coles [1] as the unbroken line. The individual values are indicated by a rug of dashes.

be used to justify our automated threshold choice of 20 mm. The subjective nature of and difficulties associated with the interpretation of the Mean Residual Life plot are well illustrated by this example. Figures 6 show comparisons of inferences (fitted densities, return levels and confidence intervals) from the fitted models based on each threshold. We can see that the fitted models are relatively similar indicating that our automated threshold selection technique




(a) (b)

Figure 6:
(a): Histogram of the exceedances from threshold choice of 30 mm recommended by Coles [1], together with the GPD fit (solid line). The GPD fit based on our threshold of 20 mm is also shown (dotted line). This GPD fit has been scaled so that the area under it above 30 mm is one.
(b): Returns level curves and confidence envelopes based on Coles's threshold [1] (unbroken) and our threshold (dashed).

compares well to the subjective procedure. Coles's [1] threshold does yield fewer exceedances, which is the cause of the increased return level confidence interval widths in Figure 6 (b).

### *2.3 Using Bootstrap Percentile Intervals to Assess Return Level Uncertainty*

Uncertainty associated with inferences from the GPD model can depend on two sources: firstly, the uncertainty associated with estimating the scale and shape parameters from the available exceedances; secondly, the uncertainty associated with the selection of the threshold that defines these exceedances. Uncertainty in parameter estimation can be relatively small in comparison to the uncertainty in the choice of threshold. It is therefore important when discussing our technique to include the effect of the uncertainty associated with threshold choice in the inferential procedure.

As we saw in Section 1, return levels play a vital role in coastal engineering; see page 82 of Coles [1] for a detailed discussion about the estimation of return levels and approximate confidence intervals from GPD fits. Standard software programmes, such as the `ismev` package estimate return levels and approximate confidence intervals, as shown in Figure 3, but do not take into account uncertainty due to threshold selection.

In an important and innovative paper Tancredi et al. [19] present a review of existing model based methodology to account for threshold uncertainty in GPD models, and then introduce their own technique. In contrast to conventional fixed threshold methods, Tancredi et al. [19] work in the Bayesian framework and assume that the threshold is one of the parameters about which to make inference. To overcome the lack of a natural model below the threshold and to avoid over-restrictive parametric assumptions, they propose a flexible mixture of an unknown number of uniform distributions with unknown range for below-threshold data. They consider it reasonable to expect different estimates of return levels and precision of estimates for different thresholds. This essentially leads to a Bayesian mixing of all reasonable threshold values and parameter estimates to determine an overall estimate of return levels and their uncertainty. Their approach is, however, highly computationally intensive, requiring the use of a reversible jump Markov chain Monte Carlo algorithm to cope with the unknown number of uniform distributions used for below-threshold modelling; see Green [7]. It also requires a number of prior assumptions to be made, although Tancredi et al. [19] argue that return level estimation is more robust to these assumptions than to threshold choice in a fixed approach. Because of these drawbacks, we take a different approach to assess return level uncertainty based on the bootstrap procedure. Mooney & Duval [15] and Efron & Tibshirani [5] provide a basic summary of this procedure as follows:

(1) Set $b = 1$.
(2) Draw a simple random sample of size $m$ from the original data set $y_1, \ldots, y_m$ with replacement. We call this a bootstrap sample.
(3) For the bootstrap sample, calculate the quantity of interest, for example a specific return level, and call it $\hat{\theta}_b^\star$. We calculate the return level by first estimating the threshold using the methodology in Section 2.1. We then make use of this threshold when estimating the GPD model. Finally, we use the GPD parameter estimates to calculate the return level estimate.
(4) Increase $b$ by 1 and repeat steps (2) and (3) a total of $B$ times, where $B$ is a large number. We present results for $B = 1000$. Other values of $B$, ranging from 250 to 3000, yielded similar results.
(5) Construct a probability distribution by attaching a $1/B$ probability to each point, $\hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_B^\star$.

Uncertainty in the quantity of interest – for example a specific return level – can be quantified by summarizing this probability distribution using a confidence interval. More precisely, we will use a bootstrap percentile interval. To obtain an $(1-\alpha)$-level interval we sort the $B$ values $\hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_B^\star$ in ascending order and select the $(\frac{\alpha}{2}B)^{\text{th}}$ and $(1-\frac{\alpha}{2})B^{\text{th}}$ values as our confidence interval using the integer below and the integer above if these values are not themselves integers. We set $\alpha = 0.05$, yielding 95% confidence intervals. We now present the result of applying the above bootstrap methodology to our data set. Fig-

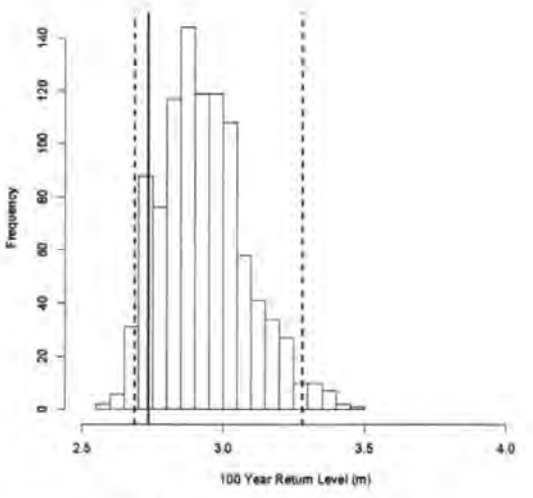



Figure 7: Histogram of the bootstrapped 100 year return levels and associated 95% bootstrap percentile interval ($B = 1000$ bootstrap iterations). The dashed lines are the percentile interval and the solid line is the return level based on the original data.

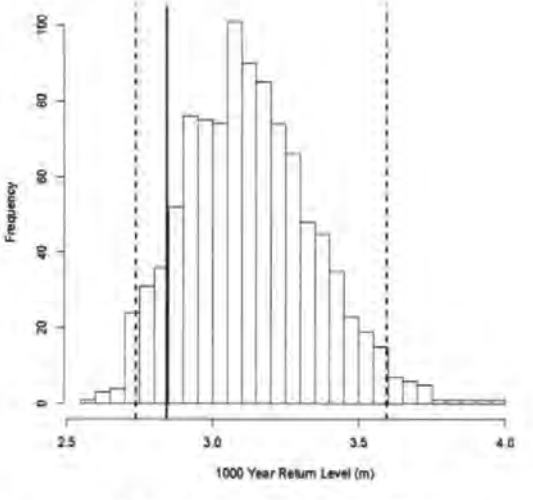



Figure 8: Histogram of the bootstrapped 1000 year return levels and associated 95% bootstrap percentile intervals. The dashed lines are the percentile interval and the solid line is the return level based on the original data.

ure 7 shows a histogram of the bootstrapped 100 year return levels $\hat{\theta}_1^\star, \ldots, \hat{\theta}_B^\star$ and the associated bootstrap percentile interval. Figure 8 is an analogous plot for the 1000 year return level. These percentile intervals enable us to quantify the uncertainty in return level estimation in an accurate way, without ignoring threshold choice uncertainty and relying on the standard asymptotic theory

outlined on page 82 of Coles [1]. Figures 7 and 8 show that the bootstrap percentile interval widths are approximately 0.6 m for the 100 year return level and 0.8 m for the 1000 year return level, indicating that uncertainty about these estimates is not particularly large from an engineering point of view.

We remark that there are more refined methods for obtaining bootstrap confidence intervals. Venables & Ripley [21] discuss 'normal', 'basic', $BC_a$ and studentized confidence intervals, in addition to percentile confidence intervals in their Section 5.7; see Davison & Hinkley [11] and Efron & Tibshirani [5] for excellent and extensive further discussion. We chose to use percentile confidence intervals because they are simple to understand and implement.

## 3 Simulation Study

In this section we investigate the performance of our automated threshold selection method by means of a simulation study. Figure 9 shows a histogram of a data set comprising 10,000 simulated values of a random variable $X$ with distribution function given by

$$F(x) = \{(1-\beta)G_1(x) + \beta\}I[x > u] + G_2(x)I[x \leq u], \quad x > 0, \tag{6}$$

where $I$ is the usual indicator function and $\beta = \mathrm{P}(X \leq u)$. $G_1(x)$ is a GPD function with associated density function

$$g_1(x) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-(1/\xi+1)}, \quad x > u, \quad 1 + \frac{\xi(x-u)}{\sigma} > 0; \tag{7}$$

$G_2(x)$ is a truncated normal distribution function with associated density function

$$g_2(x) = \frac{\frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(x-\gamma)^2}{2\alpha^2}\right)}{\int_0^\infty \frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(x-\gamma)^2}{2\alpha^2}\right)dx}, \quad x > 0. \tag{8}$$

With this $F$, the distribution of the random variable $X$ can be thought of as a mixture of a normal distribution truncated on $(0, u]$ and a GPD on $(u, \infty)$ with weights $\beta$ and $1-\beta$, with non-extreme values coming from the truncated normal and extreme values from the GPD. Given $\beta$ and the parameters $\gamma$ and $\alpha$ of $g_2$, we can find $u$ from the condition

$$\begin{aligned}\beta = \Pr(X \leq u) = G_2(u) &= \int_0^u g_2(x)dx \\ &= \frac{\int_0^u \frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(y-\gamma)^2}{2\alpha^2}\right)dy}{\int_0^\infty \frac{1}{\alpha\sqrt{2\pi}}\exp\left(-\frac{(y-\gamma)^2}{2\alpha^2}\right)dy}.\end{aligned} \tag{9}$$

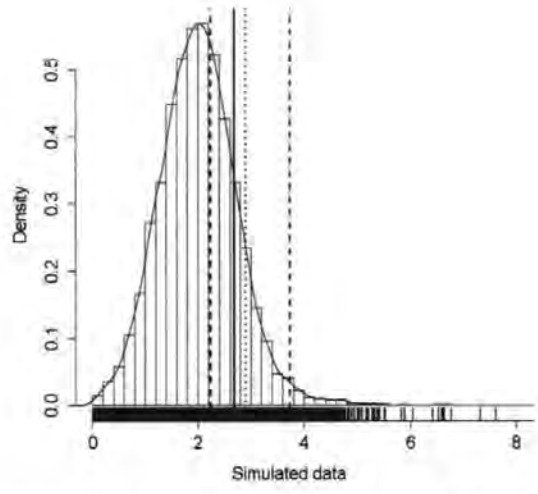



Figure 9: Histogram of a data set of 10,000 simulated values of a random variable $X$ with distribution function $F$. The associated probability density function is also shown. The individual values are indicated by a rug of dashes. Our automated threshold choice is indicated by a solid line, with the true threshold $u = 2.90$ being shown by a dotted line. The 95% bootstrap percentile intervals is also presented using dashed lines.

For the simulated data set shown in Figure 9 we set $\beta = 0.9$, $\gamma = 2$ and $\alpha = 0.7$, and solved for $u$ to obtain $u = 2.90$. We choose the parameter $\sigma$ of the GPD so that there was no discontinuity at $u$ in the probability density function of $X$. To do this we require

$$g_2(u) = (1 - \beta)g_1(u) = \frac{1 - \beta}{\sigma}. \tag{10}$$

With $u = 2.90$, this equation can easily be solved to yield $\sigma = 0.40$. We set the shape parameter $\xi$ of the GPD to be 0.2. The resulting probability density function of $X$ is shown in Figure 9, together with the threshold $u = 2.90$ (dotted line).

A random sample $x_1, \ldots, x_N$ can be simulated from $F$ as follows:

- Set $i = 1$. Simulate $y \sim N(\gamma = 2, \alpha^2 = 0.7^2)$;
- If $y < 0$ reject it;
- else if $0 < y < u$, set $x_i = y$ and increase $i$ by 1;
- else if $y > u$ simulate $x \sim \text{GPD}(u = 2.90, \sigma = 0.4, \xi = 0.2)$, set $x_i = x$ and increase $i$ by 1.
- Stop when $i = N + 1$.

We applied our automated threshold selection method to the simulated data set of size $N = 10,000$ shown in Figure 9. The selected threshold took the

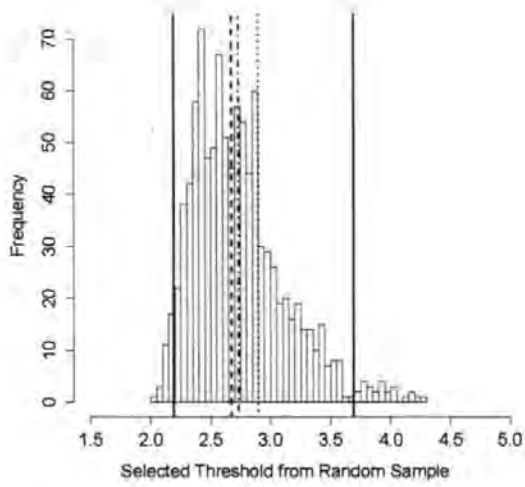



Figure 10: Histogram of thresholds selected from 1000 random samples of size $N = 10,000$ from $F$. The mean and median of the automated threshold choices for the simulated data sets are shown by dot-dash and dashed lines respectively; while the true threshold $u = 2.90$ is shown by a dotted line. The 2.5% and 97.5% quantiles are shown as the outer solid lines.

value 2.678 m and can be seen to be close to the true value of $u = 2.90$. We next used the above simulation procedure to generate 1000 random samples of size $N = 10,000$ from $F$. We applied our threshold selection technique to each random sample; a histogram of these 1000 thresholds, together with 2.5% and 97.5% quantiles (2.189,3.694), the true threshold $u = 2.90$ and mean $u_{mean} = 2.73$ and median $u_{med} = 2.67$ values of the distribution of estimated thresholds are shown in Figure 10. The selected thresholds seem to be evenly and not very widely spread around the true threshold, suggesting that our method can recover a known threshold to a good degree of accuracy. Our method performed similarly well when applied to data sets simulated using different values of $\beta$, $\gamma$, $\alpha$ and $\xi$.

We now focus on the simulated data set shown in Figure 9 and apply the bootstrap analysis discussed in Section 2.3, except that our bootstrap quantity of interest $\hat{\theta}_b^\star$ now becomes selected threshold instead of a specific return level. Figure 11 shows a histogram of the bootstrap threshold choices together with the 95% bootstrap percentile interval (2.225,3.732), our automated threshold choice for the original simulated data set and the true threshold $u = 2.90$. The 2.5% and 97.5% quantiles found above have also been added. The 95% bootstrap interval is also shown in Figures 9 and 11. We can see from these plots that the 95% bootstrap percentile interval is not very wide and contains the true and selected thresholds. The actual interval values of $(2.225, 3.732)$ compare well with the 2.5% and 97.5% quantiles $(2.189, 3.694)$ indicating that the bootstrap assesses well the uncertainty associated with our threshold choice

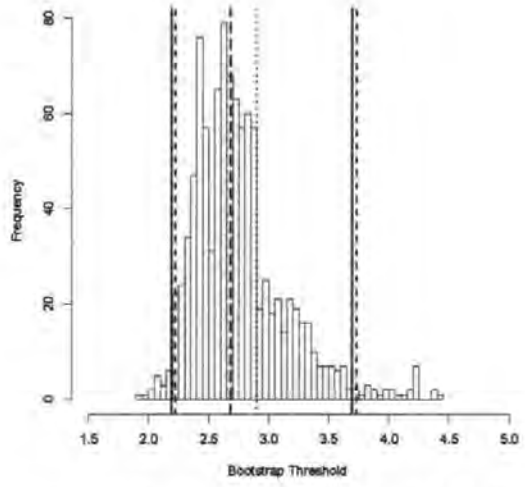



Figure 11: Histogram of the bootstrap threshold choices. The automated threshold choice for the original simulated data set is shown as the large-dash line, while the true threshold $u = 2.90$ is the dotted line. The 95% bootstrap percentile interval is shown as the dashed lines, with the 2.5% and 97.5% quantiles from Figure 10 being given using the outer solid lines.

procedure.

In order to validate our bootstrap procedure further we performed an extensive study based on data sets simulated from distribution (6) to check the coverage of our bootstrap confidence intervals. Good results were obtained. We found that for the 1000 year return level, for example, the true coverage was 94%, very close to its 95% nominal level. The conclusion of all our simulation work is that our automated and computationally inexpensive procedure can recover a theoretical threshold from simulated data to a good degree of accuracy and that the bootstrap can be successfully used to assess associated uncertainties.

In the next section we give a further example of the application of our procedure by comparing it to an existing technique utilized in the JOINSEA software.

## 4 Comparison to the JOINSEA Software

In this section we compare our new method with an existing technique used in the JOINSEA software (see [12] and [13]). The JOINSEA approach for choosing an appropriate threshold assumes that extremes can be identified as exceedances over a 95% quantile. We now use the Selsey Bill data set introduced in Section 2.2 to compare our choice of threshold and fitted GPD

with those obtained from the approach adopted in JOINSEA. Table 1 gives the results from the two approaches.

| | **New Technique** | **JOINSEA** |
|---|---|---|
| Threshold Value | 0.487 | 1.480 |
| Number of Exceedances | 5372 | 497 |
| Maximum Likelihood Estimate, $\xi$ | $-0.230$ | $-0.271$ |
| Maximum Likelihood Estimate, $\sigma$ | 0.576 | 0.405 |
| Standard Error, $\xi$ | 0.00952 | 0.04094 |
| Standard Error, $\sigma$ | 0.00940 | 0.02409 |

Table 1: The chosen threshold, number of exceedances, GPD parameter estimates and standard errors for our new automated threshold selection method and the approach adopted in the JOINSEA software.

Figure 12 shows again a scatter plot of wave height against the cosine of wave direction for the Selsey Bill data set, together with the two thresholds. The dashed line was obtained using our new threshold technique, while the solid line is the JOINSEA threshold. We see from Table 1 and Figure 12 that the threshold values are very different, with the automated threshold being almost 1 m below the JOINSEA threshold. Figures 13 (a) and 13 (b) show comparisons of inferences (return levels, confidence intervals and fitted densities) from the fitted models based on each threshold. We can see that the resulting models are actually very similar indicating that our automated threshold selection technique is comparable to that of JOINSEA. The JOINSEA threshold yields fewer exceedances, which is the cause of the increased return level confidence interval widths in Figure 13 (a). The narrower confidence intervals yielded by our threshold selection technique, together with the fact that it is more model based, lead us to prefer our methodology over the JOINSEA approach. We also note that for data sets such as those simulated in Section 3 with $\beta > 0.95$ the JOINSEA approach is guaranteed to lead to non-extremes being included in future GPD analyses.

We applied our automated threshold selection technique to different data sets which varied in size and data collection location, and found it performed consistently well in terms of model goodness of fit. We felt that in the case of the Selsey Bill data our automated approach chose a relatively low threshold as a type of "average" threshold across the range of direction covariate values. This observation led us to extended our automated technique to allow the chosen threshold to vary with covariate value. We discuss our direction varying threshold methodology in detail in Section 5.

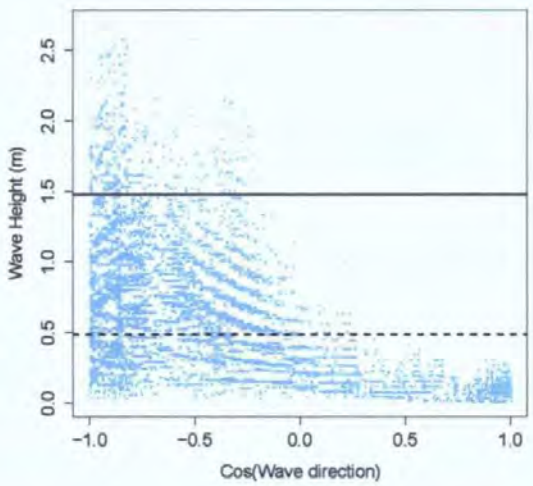
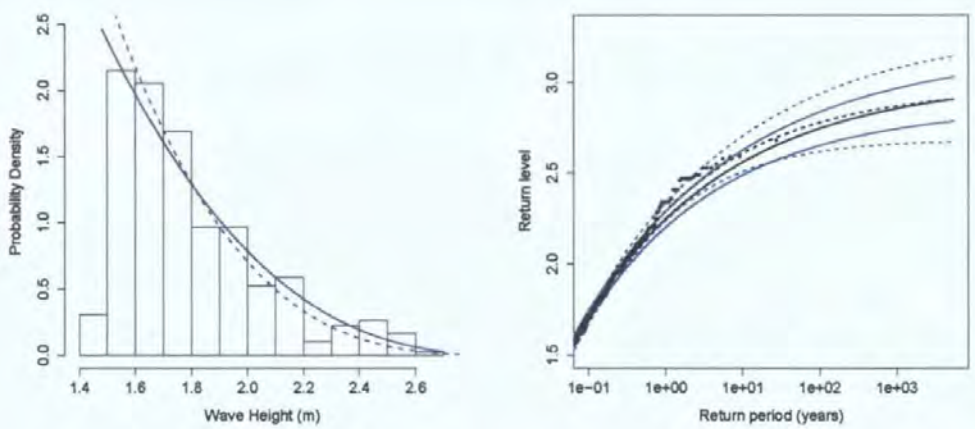
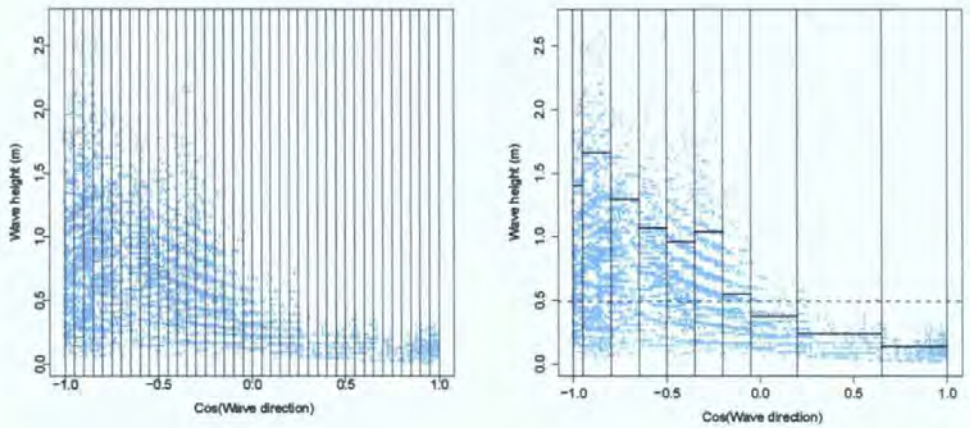



Figure 12: Scatter plot of wave height against the cosine of wave direction for 10,000 values from the Selsey Bill data set. Our automated threshold choice is shown using the dashed line, while the solid line shows the threshold chosen by the JOINSEA software. Both threshold choices take no account of the cosine of wave direction.




(a) (b)

Figure 13:
(a): Histogram of the exceedances from the JOINSEA threshold choice, together with the GPD fit (dashed line). The GPD fit based on our threshold procedure is also shown (unbroken line). This GPD fit has been scaled so that the area under it above the JOINSEA threshold is one.
(b): Return level curves and confidence envelopes from both automated (unbroken)and JOINSEA (dashed) threshold model fits.




(a) (b)

Figure 14:
(a): Scatter plot of wave height against the cosine of wave direction. The data has been split into 40 sections equally spaced along the covariate axis.
(b): Scatter plot of wave height against the cosine of wave direction. The data has now been split into optimal blocks along the covariate axis. Individual automated thresholds have been chosen for each block and are shown by the solid horizontal lines. The dotted line shows the threshold chosen without reference to cosine of wave direction.

## 5 Extended Automated Threshold Selection Technique

We have seen that the Selsey Bill data set comprises information about wave direction as well as wave height. So far we have worked only with wave height. It is clear from Figure 12 that the behaviour of wave height varies with wave direction. It therefore makes sense to include the directional effect in our automated threshold selection procedure, rather than to have a threshold that is constant over wave direction.

In extreme wave analysis directional effects are usually dealt with using one of two methods: either the data are split according to different directions with each separate data set being modelled independently, or the wave direction is included as a covariate as in Ewans & Jonathan [6] and Jonathan & Ewans [14], for example. In this section we propose a new approach to blocking the data.

Our approach is based on the automated threshold selection procedure that we have already presented and is as follows:

(1) First the data set is blocked according to the cosine of wave direction. The number of blocks is initially defined by the user; see Figure 14(a) for example where the covariate axis is split into 40 equal width blocks.

Each block is then altered iteratively to its optimum size as described in (2).

(2) The constant automated threshold selection procedure is applied to the data in each block. The block size can then be altered in order to achieve a satisfactory GPD fit in each block. If there is not a sufficient number of observations within the block or if the block's optimal threshold choice does not define enough exceedances to achieve a good GPD fit, then the block is merged with the next consecutive block and the process is repeated. Through a simulation study we found that a sufficient number of observations would be the larger of 5% of the total number of observations and 500, and a sufficient number of exceedances would be the maximum of 1% of the total number of observations and 50. The simulation study involved fitting a number of GPD models to different data sets and assessing the dependence of model fit quality on the number of observations and the number of exceedances. The merging of consecutive blocks is continued until the required minimum values for the number of observations and the number of exceedances for the merged block allows is reached. Our optimal blocks are shown in Figure 14(b).

(3) Each block now has a constant optimal threshold associated with it. A separate GPD can be fitted to the wave height data within each block, and associated direction specific inferences about return levels can be made.

If the individual block thresholds shown in Figure 14(b) are considered together, a piecewise constant threshold function is defined. A threshold that is continuous in the cosine of wave direction covariate can be obtained by applying a smoothing spline, for example. We did this using `R`[16]'s `smooth.spline` function; see Green & Silverman [8], for example. The resulting smoothed direction varying threshold function is shown in Figure 15.

In order to justify further the choice of these direction varying thresholds we also show in Figure 15 probability density contours for a bivariate kernel density estimate (calculated using the `kde2d` function of the `MASS` library; see Venables & Ripley [21]) based on wave height and the cosine of wave direction. We see that the chosen thresholds aline well with the tail of this probability density function across the range of cosine wave direction, supporting our direction varying threshold choice procedure. We conclude by remarking that, as mentioned, the more appropriate thresholds that this extended automated threshold selection technique provides can yield more accurate direction specific return level estimates. These in turn can lead to improved coastal defence designs that account for directional variations in extreme wave heights.

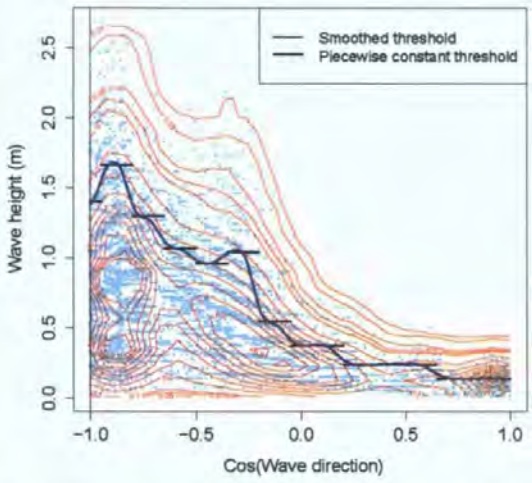



Figure 15: The bivariate wave data with piecewise constant and smoothed covariate varying thresholds. Bivariate probability density estimate contours are overlaid on the scatter plot.

## 6 Concluding Comments

In this paper, we have presented a new, automated, simple and computationally inexpensive method for selecting the threshold for the GPD in extreme value modelling. Our pragmatic method uses a series of normality tests to find an appropriate threshold choice for a given data set. We have contrasted our methodology with one of the currently available subjective approaches. We have shown the practical applicability of our method using an example from coastal engineering. We have demonstrated that our automated technique can recover a known threshold from a simulated data set to a good degree of accuracy. We have assessed the effect of the uncertainty associated with threshold selection on return level estimation using the bootstrap procedure. We have also provided comparisons of our new approach with the existing JOINSEA technique, pointing out improvements of our method over the existing one. In practice, our method can be seen as a additional tool that complements existing threshold selection methods.

We have extended our methodology to incorporate a direction covariate dependant threshold. This extension uses our automated threshold selection technique to segregate the data into optimal blocks based on goodness of fit and sample size requirements.

Our methodology can lead to more accurate return level estimates, with their uncertainty properly qualified, which can inform and enhance the coastal design process.

## 7 Acknowledgement

The authors acknowledge the support of a doctoral scholarship from the University of Plymouth and funding from the EPSRC projects RF-PeBLE (grant No. EP/C005392/1), LEACOAST 2 (grant No: EP/C013085/1) and BVANG (grant No: EP/C002172/1). We would also like to thank Dr. Peter Hawkes from HR Wallingford for the data set and supplementary information provided. We warmly thank the referees and the Editor-in-Chief for their thorough reading of the manuscript and suggestions that have led to considerable improvements in this paper.

## References

[1] Coles, S., 2001. An Introduction to Statistical Modelling of Extreme Values. Springer, London.

[2] Coles, S. (Original S functions) and Stephenson, A. (R port and R documentation files), 2006. `ismev`: An Introduction to Statistical Modeling of Extreme Values. http://www.maths.lancs.ac.uk/stephena/, R package version 1.2.

[3] Davidson, A.C. and Smith, R.L., 1990. Models for exceedances over high thresholds. J. R. Stat. Soc. C, 52(3) 393–442.

[4] Dupuis, D.J., 1999. Exceedances over high thresholds: A guide to threshold selection. Extremes, 1(3) 251–261.

[5] Efron, B. and Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, London.

[6] Ewans, K. and Jonathan, P., 2006. Estimating Extreme Wave Design Criteria Incorporating Directionality. 9th International Workshop on Wave Hindcasting and Forecasting, Victoria, Canada.

[7] Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82 711–732.

[8] Green, P.J. and Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London.

[9] Greenwood, P.E. and Nikulin, M.S., 1996. A Guide to Chi-Squared Testing. Wiley (Probability & Statistics Series), New York.

[10] Guillou, A. and Hall, P., 2001. A diagnostic for selecting the threshold in extreme value analysis. J. R. Stat. Soc. B, 63(2) 293–305.

[11] Davison, A.C. and Hinkley, D.V., 1997. Bootstrap Methods and their Application. Cambridge University Press, Cambridge.

[12] HR Wallingford, 1998a. The Joint Probability of Waves and Water Levels: JOINSEA. Report: TR71, HR Wallingford.

[13] HR Wallingford, 1998b. The Joint Probability of Waves and Water Levels: JOINSEA. Report: SR 537, HR Wallingford.

[14] Jonathan, P. and Ewans, K., 2007. The effect of directionality on extreme wave design criteria. Ocean Engineering, 34 1977–1994.

[15] Mooney, C.Z. and Duval, R.D., 1993. Bootstrapping: A Nonparametric Approach to Statistical Inference. Sage University Paper, London.

[16] R Development Core Team, 2009. R: A language and environment for statistical computing. http://www.R-project.org, ISBN 3-900051-07-0. Vienna, Austria.

[17] Reeve, D., Chadwick, A. and Fleming, C., 2004. Coastal engineering: processes, theory and design practice. SPON, London.

[18] Smith, R., 1985. Maximum likelihood estimation in a class of nonregular cases. Biometrika, 72(1) 67–90.

[19] Tancredi, A., Anderson, C. and O'Hagan, A., 2006. Accounting for threshold uncertainty in extreme value estimation. Extremes, 9 86–106.

[20] Tawn, J.A. and Coles, S.G., 1994. Statistical methods for multivariate extremes: an application to structural design. J. R. Stat. Soc. C, 43(1) 1–48.

[21] Venables, W.N. and Ripley, B.D., 2002. Modern applied statistics with S (4th Ed). Springer-Verlag, New York.

# Bayesian nonparametric quantile regression using splines

Paul Thompson [a,*], Yuzhi Cai [b], Rana Moyeed [b], Dominic Reeve [a], Julian Stander [b]

[a] *C-CoDE, School of Engineering, Reynolds Building, University of Plymouth, Devon, PL4 8AA, UK*

[b] *School of Mathematics & Statistics, 2-5 Kirby Place, University of Plymouth, Devon, PL4 8AA, UK*

**Abstract**

A new technique based on Bayesian quantile regression that models the dependence of a quantile of one variable on the values of another using a natural cubic spline is presented. Inference is based on the posterior density of the spline and an associated smoothing parameter and is performed by means of a Markov chain Monte Carlo algorithm. Examples of the application of our technique to two real environmental data sets and to a simulated data for which polynomial modelling is inappropriate are given. An aid for making a good choice of proposal density in the Metropolis-Hastings algorithm is discussed. Our nonparametric methodology provides more flexible modelling than the currently used Bayesian parametric quantile regression approach.

*Key words:* Acceptance rate, coastal wave data, inference about smoothing parameter, Markov chain Monte Carlo, motorcycle accident data, proposal density choice

## 1 Introduction

Quantile Regression can be described as a method that provides a more complete inferential picture than ordinary least-squares regression. The latter technique estimates the conditional mean of some response variable $Y$ given the value $t$ of a covariate, while quantile regression takes a different approach by

* Corresponding author.
*Email address:* p1thompson@plymouth.ac.uk (Paul Thompson).

estimating the conditional quantiles of $Y$ given $t$. More precisely, in quantile regression we are interested in estimating quantile functions $q_p(t)$, $0 < p < 1$, such that $\Pr(Y \leq q_p(t)$ given that the covariate takes the value $t) = p$. This allows the full range of the data to be modelled and so can be beneficial when large values are of particular interest. It also means that quantile regression can be viewed as a data exploration technique. Koenker et al. [14] present a wide ranging discussion about the use of quantile regression. Quantile regression can be implemented in a range of different forms, and Yu et al. [27] provide an overview of some commonly used quantile regression techniques.

Bayesian inference for quantiles has been considered by several authors, including Yu & Moyeed [26] and Dunson & Taylor [6]. Dunson & Taylor [6] discuss appropriate Bayesian inference for quantiles when the likelihood function is not fully specified. They present an example based on linear quantile regression function. Kottas & Gelfand [17] consider Bayesian semiparametric median regression modelling under a Dirichlet process mixing framework. Kottas & Krnjajić [18] extend this approach to general quantiles.

The Bayesian quantile regression (BQR) methodology developed in Yu & Moyeed [26] adopts a parametric approach based on a polynomial quantile regression function. Inference about the posterior distribution of the parameters of this regression function is made by means of a Markov chain Monte Carlo (MCMC) algorithm. Although Yu & Moyeed [26] present excellent results, there are certain drawbacks associated with using polynomials. These include the influence of outliers and the need to choose the degree of the polynomial, possibly for each quantile considered. Also, the data may have a limited local effect on the shape of a polynomial regression curve especially when modelling high quantiles. Yu & Moyeed [26] work with low order polynomials; problems associated with using very high order polynomials may include over-fitting and poor MCMC convergence.

In this paper we present a nonparametric alternative to the parametric approach of Yu & Moyeed [26] based on using natural cubic splines rather than polynomials. Our approach provides a more versatile and flexible method of fitting a quantile regression curve. Section 2 of the paper provides an introduction to natural cubic splines. It then presents our Bayesian nonparametric quantile regression methodology by describing the posterior density of the spline and an associated smoothing parameter and outlining a MCMC algorithm for making inferences from this posterior. Section 3 presents applications of our methodology to two real environmental data sets and to simulated data. The first data set comprises coastal wave conditions from near the Selsey Bill area and were generated using a hindcasting technique (see Reeve et al. [22]) using wind records. The data comprise of hourly hindcast measurements of the variables significant wave height, wave period and wave direction over an approximate time span of 27 years. A good understanding of this type

of data is important for the coastal design process, as illustrated by Thompson et al. [24]. Here, our variable $Y$ of interest will be wave height, while the covariate $t$ will be the cosine of wave direction. In this example we take a random sample of 10,000 observations for computational and presentational reasons. The resulting data set is shown later in Figure 1 and will be denoted $(t_1, y_1), \ldots, (t_n, y_n)$, where sample size $n = 10,000$. Our second application is similar, but consists of offshore wave data. The simulated data for our third application are based on a well known published data set for which polynomial modelling is inappropriate. Our applications allow us to illustrate learning about model parameters from data. We also discuss the advantages offered by our nonparametric methodology. In Section 4 we discuss the performance of our MCMC procedure and presents an aid for making a good choice of proposal density in the Metropolis-Hastings algorithm so improving its efficiency. Finally Section 5 is a short conclusion.

## 2 Bayesian Nonparametric Quantile Regression Methodology

When fitting a curve through a bivariate data set, one important consideration is the roughness of the curve, i.e. how "wiggly" it is. More specifically, we tend to prefer smooth curves that have a reduced amount of rapid fluctuation. We are able to quantify the roughness of a curve $g$ with continuous second derivative on the interval $[a, b]$ by means of a roughness penalty which is defined in this paper as the integrated squared second derivative $\int_a^b g''(t)^2 dt$; see Green & Silverman [11]. A standard approach to curve fitting is based on a trade-off between the lack-of-fit of a curve to the data and its roughness, or, equivalently, between goodness-of-fit and smoothness, as discussed in Green & Silverman [11]. These authors also shown how this approach can be formalized within the Bayesian framework (see Gamerman [7]) by having a prior distribution which quantifies probabilistically the roughness of the fitted curve; we describe this in detail in Section 2.2.

In Section 2.1 we define natural cubic splines by following the standard approach given by [11]. The aim of this paper is to include the natural cubic spline in the Bayesian quantile regression methodology of Yu & Moyeed [26] so extending and making their parametric technique more flexible. Full details of our proposed methodology are given in Section 2.2.

### *2.1 The Natural Cubic Spline and Associated Results*

We say that a curve $g$ is a cubic spline with $N \geq 2$ knots $\tau_1 < \cdots < \tau_N$ if $g$ is a cubic polynomial between knots $\tau_{i-1}$ and $\tau_i$, $i = 2, \ldots, N$, and if $g$ has

continuous first and second derivatives at $\tau_i$, $i = 2, \ldots, N-1$. Let $a < \tau_1$ and $b > \tau_N$. The cubic spline $g$ is said to be a natural cubic spline (NCS) on $[a, b]$ if it is linear on the intervals $[a, \tau_1]$ and $[\tau_N, b]$ and if it has continuous first and second derivatives at $\tau_1$ and $\tau_N$. This definition of a NCS is equivalent to the one given by Green & Silverman [11, pages 11–12]; see also Hastie et al. [12, Section 5.2].

We now introduce notation and present results that we will use later. Let $\boldsymbol{g} = (g_1, \ldots, g_N)^T$ be a column vector of values $g_i = g(\tau_i)$, $i = 1, \ldots, N$, of a NCS $g$ at its knots $\tau_1, \ldots, \tau_N$. Further, let $h_i = \tau_{i+1} - \tau_i$, $i = 1, \ldots, N-1$, let $Q$ be the $N \times (N-2)$ banded matrix with entries $q_{ij}$, $i = 1, \ldots, N$ and $j = 2, \ldots, N-1$, given by $q_{j-1,j} = 1/h_{j-1}$, $q_{jj} = -1/h_{j-1} - 1/h_j$, $q_{j+1,j} = 1/h_j$ and $q_{ij} = 0$ for $|i-j| \geq 2$, and let $R$ be the $(N-2) \times (N-2)$ banded positive definite symmetric matrix with entries $r_{ii} = (h_{i-1} + h_i)/3$, $i = 2, \ldots, N-1$, $r_{i,i+1} = r_{i+1,i} = h_i/6$, $i = 2, \ldots, N-2$, and $r_{ij} = 0$ for $|i-j| \geq 2$. We can now define the $N \times N$ symmetric matrix $K$ with rank $N-2$ as $K = QR^{-1}Q^T$. We will make use of Theorem 2.1 of Green & Silverman [11] that tells us that the roughness penalty satisfies

$$\int_a^b g''(t)^2 dt = \boldsymbol{g}^T K \boldsymbol{g}. \tag{1}$$

We will also use Theorem 2.2 of [11] that establishes that, given any values $g_1, \ldots, g_N$, there is a unique NCS $g$ with knots at $\tau_1, \ldots, \tau_N$ satisfying $g(\tau_i) = g_i$, $i = 1, \ldots, N$.

### *2.2 Bayesian Nonparametric Quantile Modelling and Inference*

In this section we present a framework for Bayesian nonparametric quantile regression using splines rather than polynomials as in Yu & Moyeed [26]. In our approach we model a quantile function of a covariate $t$ using a NCS with $N$ fixed knots at points $\tau_1, \ldots, \tau_N$ that cover the range of $t$. The NCS is uniquely determined by its values $\boldsymbol{g} = (g_1, \ldots, g_N)^T$ at these knots, since, as explained in Section 2.1, there is a unique NCS that can be drawn through the points $(\tau_i, g_i)$, $i = 1, \ldots, N$. As our approach is Bayesian, we begin by defining the prior density for $\boldsymbol{g}$ as multivariate normal; see Green & Silverman [11, page 51] for a discussion about the use of the multivariate normal density as a prior in this context.

Our prior for $\boldsymbol{g}$ is defined by means of the multivariate normal density

$$\pi(\boldsymbol{g}|\lambda) = \frac{\lambda^{(N-2)/2}}{(2\pi)^{(N-2)/2}(\mu_1 \cdots \mu_{N-2})^{1/2}} \exp\left(-\frac{1}{2}\lambda\, \boldsymbol{g}^T K \boldsymbol{g}\right), \tag{2}$$

in which $\mu_1, \ldots, \mu_{N-2}$ are the inverses of the $N-2$ non-zero eigenvalues of

$K$ and $\lambda > 0$ is an unknown parameter. More detail about this multivariate normal distribution can be found in Rao [21, page 528]. Note that (2) depends through (1) on the roughness $\int_a^b g''(t)^2 dt = \boldsymbol{g}^T K \boldsymbol{g}$ of the NCS $g$ uniquely defined by $\boldsymbol{g}$. As larger values of $\lambda$ result in more probability density being given to less rough curves $g$, we will refer to $\lambda$ as a smoothing parameter.

We next require a prior on the smoothing parameter $\lambda$ which is constrained by a lower limit of zero. Hence, we follow standard practice by using the gamma density as our prior for $\lambda$ which takes the form

$$\pi(\lambda) = \frac{\lambda^{\alpha-1} \exp(-\lambda/\beta)}{\Gamma(\alpha)\beta^{\alpha}}, \quad \lambda > 0, \tag{3}$$

in which $\Gamma$ is the usual gamma function. The user is able to specify the hyperparameters $\alpha$ and $\beta$. Under this prior $\mathrm{E}[\lambda] = \alpha\beta$ and $\mathrm{Var}[\lambda] = \alpha\beta^2$, results that can be used to guide hyperparameter choice.

The final step in our Bayesian approach is to define the likelihood of the data $(t_i, y_i)$, $i = 1, \ldots, n$, given $\boldsymbol{g}$. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. We proceed in accordance with the BQR approach of Yu & Moyeed [26] by substituting our NCS $g$ for their polynomial. The resulting likelihood takes the form:

$$L(\boldsymbol{y}|\boldsymbol{g}) = p^n(1-p)^n \exp\left\{-\sum_{i=1}^{n} \rho_p(y_i - g(t_i))\right\} \tag{4}$$

where $p$ is the probability corresponding to the quantile of interest, $0 < p < 1$, and $\rho_p$ is the standard quantile regression loss function

$$\rho_p(u) = u(p - I(u < 0)) \tag{5}$$

in which $I$ is the usual indicator function. The values of $g(t_i)$, $i = 1, \ldots, n$, in (4) are uniquely determined by $\boldsymbol{g}$. We note that the likelihood is not dependent on $\lambda$. Combining $\pi(\lambda)$, $\pi(\boldsymbol{g}|\lambda)$ and $L(\boldsymbol{y}|\boldsymbol{g})$, we can write the posterior density function of $\boldsymbol{g}$ and $\lambda$ as

$$\pi(\boldsymbol{g}, \lambda|\boldsymbol{y}) \propto L(\boldsymbol{y}|\boldsymbol{g})\pi(\boldsymbol{g}|\lambda)\pi(\lambda). \tag{6}$$

We now simulate realizations of $\boldsymbol{g}$ and $\lambda$ from this posterior density using an MCMC approach implemented via the Metropolis-Hastings algorithm; see Gamerman [7]. Our inferences will be based on these posterior realizations. In particular, we shall use the posterior mean $(\overline{g_1}, \ldots, \overline{g_N})$ of $\boldsymbol{g}$ to produce our estimated quantile. Our algorithm can be summarized as follows:

(i) Assign initial values $\boldsymbol{g}^{(0)}$ and $\lambda^{(0)}$ to $\boldsymbol{g}$ and $\lambda$. We set $\boldsymbol{g}^{(0)}$to be the values at $\tau_1, \ldots, \tau_N$ of the posterior mean cubic quantile regression curve obtained using the methodology of Yu & Moyeed [26]. The cubic quantile

regression curve was chosen as this is also an example of a cubic spline, although a very constrained one. We obtain the value of $\lambda^{(0)}$ by applying generalized cross validation (GCV) to the usual mean smoothing spline; see Green & Silverman [11]. We chose this value, which we shall refer to as GCV(mean spline), because it can be found easily and quickly using R's [20] `smooth.spline` function (see Venables & Ripley [25], for example); Section 4.3 provides brief further discussion. We set iteration number $j = 1$.

(ii) We generate a candidate vector $\boldsymbol{g}^\star$ from the multivariate normal distribution

$$\boldsymbol{g}^\star | \boldsymbol{g}^{(j-1)} \sim MVN(\boldsymbol{g}^{(j-1)}, \Sigma) \tag{7}$$

with mean $\boldsymbol{g}^{(j-1)}$ and variance-covariance matrix $\Sigma = \sigma^2 K^- / \lambda$, where $K^-$ is the generalized inverse of $K$. The constant $\sigma^2$ is specified by the user; see Section 4.2.

(iii) We set $\boldsymbol{g}^{(j)}$ to $\boldsymbol{g}^\star$ with probability:

$$\alpha(\boldsymbol{g}^{(j-1)}, \boldsymbol{g}^\star) = \min\left\{1, \frac{L(\boldsymbol{y}|\boldsymbol{g}^\star)\pi(\boldsymbol{g}^\star|\lambda^{(j-1)})q(\boldsymbol{g}^{(j-1)}|\boldsymbol{g}^\star)}{L(\boldsymbol{y}|\boldsymbol{g}^{(j-1)})\pi(\boldsymbol{g}^{(j-1)}|\lambda^{(j-1)})q(\boldsymbol{g}^\star|\boldsymbol{g}^{(j-1)})}\right\} \tag{8}$$

where the proposal density $q(\boldsymbol{g}^\star|\boldsymbol{g}^{(j-1)})$ is the probability density function of the multivariate normal specified in (7). Because $q$ is symmetric in its arguements, it cancels out of (8). Otherwise, $\boldsymbol{g}^{(j)} = \boldsymbol{g}^{(j-1)}$.

(iv) We now generate a candidate $\lambda^\star$ from the log-normal distribution as follows:

$$\mu^\star \sim N(\log(\lambda^{j-1}), \sigma_\lambda^2), \quad \lambda^\star = \exp(\mu^\star) \tag{9}$$

where the normal distribution has mean $\log(\lambda^{(j-1)})$ and variance $\sigma_\lambda^2$, which can be specified by the user; see Section 4.2.

(v) We set $\lambda^{(j)}$ to $\lambda^\star$ with probability:

$$\alpha(\lambda^{(j-1)}, \lambda^\star) = \min\left\{1, \frac{\pi(\boldsymbol{g}^{(j)}|\lambda^\star)\pi(\lambda^\star)q(\lambda^{(j-1)}|\lambda^\star)}{\pi(\boldsymbol{g}^{(j)}|\lambda^{(j-1)})\pi(\lambda^{(j-1)})q(\lambda^\star|\lambda^{(j-1)})}\right\} \tag{10}$$

where $q$ is the log-normal probability density function specified through (9). In this case cancellation of the $q$ terms in (10) is not possible as $q$ is not symmetric in its arguements. Otherwise, $\lambda^{(j)} = \lambda^{(j-1)}$.

(vi) We now increment $j$ by 1, and repeat steps (ii)-(vi) for a total of $d$ iterations.

Whilst the methodology of Yu & Moyeed [26] updates the parameters of a fixed degree regression polynomial at each iteration of the Metropolis-Hastings algorithm, our methodology updates both the entire vector of values $\boldsymbol{g}$ at the fixed knots of the NCS and the smoothing parameter $\lambda$. We set the number of iterations $d$ to 500,000. We allow a burn-in of 50,000 iterations. Inference is based on thinned values of $\boldsymbol{g}$ and $\lambda$ produced by the Metropolis-Hastings algorithm after burn-in. Convergence issues are discussed in detail in Section 4.2. All code was written in R [20], using R's random number generating functions.

A considerable advantage of the Bayesian approach is that we can calculate associated credible intervals to provide an idea of the associated posterior uncertainty. These credible intervals are obtained by ordering the thinned $g^{(j)}(\tau_i)$ sequence over $j > 50,000$ and extracting the values which correspond to, for example, the 2.5% and 97.5% quantiles. A 95% posterior credible interval for $\lambda$ can be obtained in a similar way. In the next section an example of this methodology applied to the coastal wave condition data set discussed in Section 1 is presented.

Although we have adopted the Metropolis-Hastings algorithm to simulate realizations from our posterior, we note that due to its multivariate nature, potentially more efficient samplers may be available. Neal [19] presents an alternative sampling technique called slice sampling, based on the principle that we can simulate from a distribution by sampling uniformly from the area below its plotted density function. The algorithm proceeds by alternating two steps: uniform sampling in the 'vertical' direction at the current 'horizontal' point, and uniform sampling from the 'horizontal' slice defined by the the current 'vertical' position. This latter step can be computationally very demanding with the consequence that the computational expense of slice sampling may outweigh any potential advantages over our more simple Metropolis-Hastings algorithm.

We finish this section by remarking that another approach to quantile regression is based on the minimization over curves $g$ of

$$\sum_{i=1}^{n} \rho_p(y_i - g(t_i)). \tag{11}$$

Often $g$ is taken to be a B-spline (Hastie et al. [12]) or a NCS with prespecified knots and hence smoothness. The minimizing $g$ can be found using the `quantreg` package [16] running under `R` [20]; see Koenker [14] for an example. Some other authors have considered the problem of minimizing over curves $g$ belonging to a suitable space a version of (11) penalized for roughness such as

$$\sum_{i=1}^{n} \rho_p(y_i - g(t_i)) + \lambda \int_a^b g''(t)^2 dt; \tag{12}$$

see Bosch et al. [2] and reference therein, and Koenker et al. [15] for further discussion. Koenker et al. [15] also describe a similar minimization approach based on a total variation roughness penalty; software for this is again available in [16]. As far as we know, none of these approaches routinely yield confidence envelopes for the estimated curve.

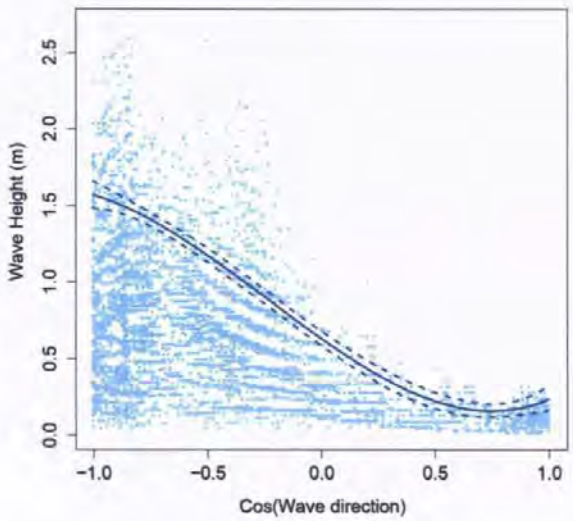



Fig. 1. Scatter plot of the coastal wave data showing the $p = 0.9$ Bayesian quantile regression curve using a cubic polynomial. A 95% credible envelope is also presented.

## 3 Applied Examples

### *3.1 Application to Coastal Wave Data*

To illustrate the practical effectiveness of the approach described in Section 2 we present results obtained from applying our methodology to the hindcast coastal wave data discussed in Section 1 and plotted in Figure 1. This plot also shows the parametric Bayesian quantile cubic regression curve of Yu & Moyeed [26] for $p = 0.9$ together with a 95% credible envelope. For our spline based approach we used a fixed grid $\tau_1 < \cdots < \tau_{30}$ of $N = 30$ equally spaced knots over the range of covariate values $t_1, \ldots, t_n$. We found that such a grid of knots allows flexible modelling without imposing a very high computational burden. We remark that in the context of mean regression some authors such as Denison et al. [4] and Dias & Gamerman [5] have also made inference about the number and position of knots. The resulting algorithm is based on the reversible jump Markov chain Monte Carlo method of Green [10] and can be computationally highly demanding.

We set the gamma prior hyperparameters $\beta = 0.1/\text{GCV(mean spline)} \approx 10^5$ and $\alpha = \text{GCV(mean spline)}/\beta \approx 10^{-11}$ in which $\text{GCV(mean spline)} \approx 10^{-6}$. With these hyperparameters the prior mean and variance of $\lambda$ are $\text{E}[\lambda] \approx 10^{-6}$ and $\text{Var}[\lambda] \approx 0.1$, representing a large amount of prior uncertainty about $\lambda$. We set the hyperparameters to yield an sensible expected $\lambda$ which is comparable to the GCV value for $\lambda$ from the usual mean smoothing spline. Figure 2 displays

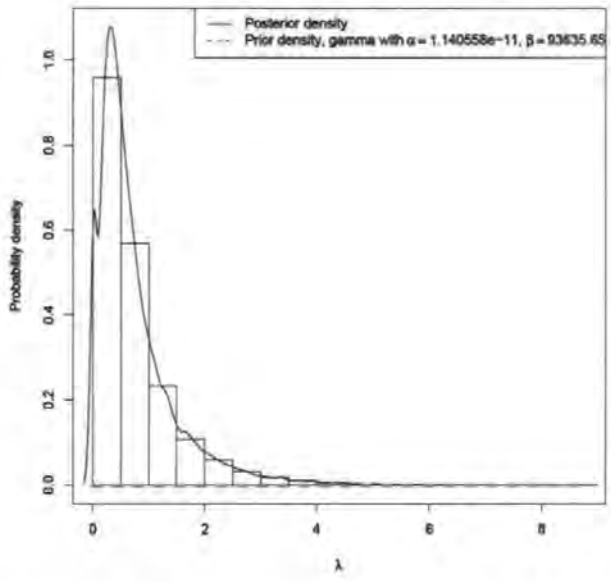



Fig. 2. Plot comparing the prior and posterior densities of $\lambda$ given the coastal wave data. The prior is a gamma density with parameters $\alpha \approx 10^{-11}$ and $\beta \approx 10^5$ and is effectively flat over a large range of $\lambda$ values. The posterior density is very different from the prior, clearly showing that learning about $\lambda$ has taken place.

the prior and posterior densities of $\lambda$ for this example. The difference between the prior and the posterior of $\lambda$ clearly shows that learning about $\lambda$ has been achieved. This is to be expected for such a diffuse prior. Learning about $\lambda$ can still be achieved with a much more informative prior as shown in Figure 3, with hyperparameters $\alpha = 160$ and $\beta = 0.025$. Here the marginal posterior of $\lambda$ lies between the prior and the posterior shown in Figure 2 based on much larger uncertainty about $\lambda$.

Figure 4 presents the resulting Bayesian nonparametric quantile regression curve and 95% credible envelope for our first choice of gamma prior hyperparameters. To obtain the regression curve shown in Figure 4, we drew the unique NCS through the points $(\tau_i, \overline{y_i}), i = 1, \ldots, N$, using the R's `spline` function [20]. Similarly, we produce our 95% credible envelope by drawing NCSs through the 2.5% and 97.5% posterior quantiles found in Section 2.2. The more local nature of the fitting procedure is easily seen from Figure 4. In order to judge the goodness-of-fit of both approaches we found empirical and fitted quantiles on a grid of 100 sections along the covariate and calculated 'residuals' as:

$$\text{residual} = \text{empirical quantile} - \text{fitted quantile}, \tag{13}$$

in which for each grid section the empirical quantile is the $p^{\text{th}}$ quantile of the data values in the section and the fitted quantile is the value produced by

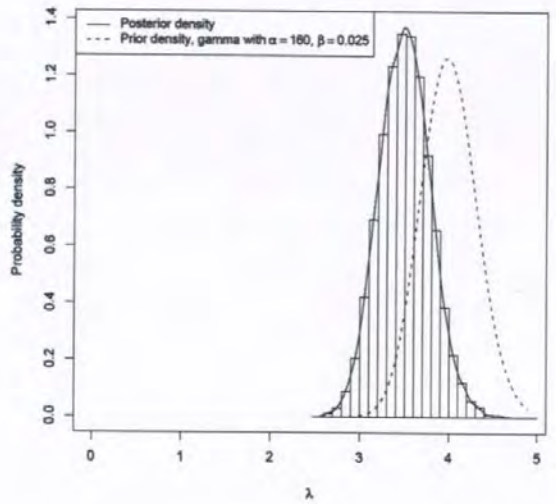



Fig. 3. Plot comparing the prior and posterior densities of $\lambda$ given the coastal wave data for a stronger prior with $\alpha = 160$ and $\beta = 0.025$. The posterior density is different from the prior, showing that learning about $\lambda$ has taken place.

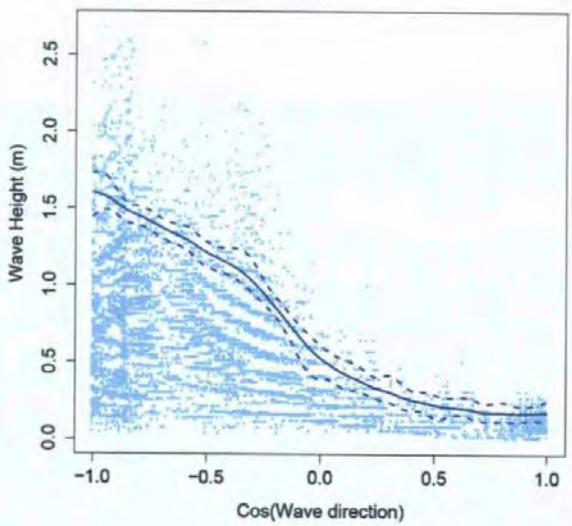



Fig. 4. Scatter plot of the coastal wave data showing the $p = 0.9$ Bayesian non-parametric quantile regression curve using splines. A 95% credible envelope is also presented.

our model at the centre of the section. As usual, smaller residuals in absolute value are associated with better fits. Figure 5 shows the absolute value of the residuals from both the cubic polynomial quantile regression curve shown in Figure 1 and the spline based curve shown in Figure 4 against the cosine of wave direction. A robust locally linear smoother provided by R's [20] `loess`

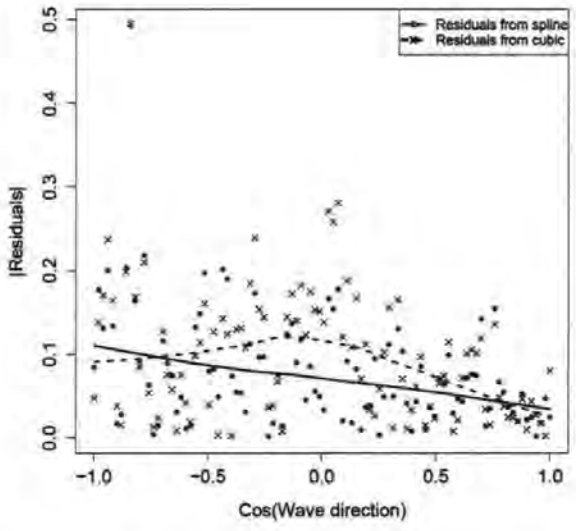



Fig. 5. The absolute values of the residuals against the cosine of wave direction with associated `loess` smoother from both the spline (dots, unbroken line) and the cubic (crosses, dashed line) quantile regressions. A grid of 100 sections along the covariate was used in the calculation of the residuals.

function (see Venables & Ripley [25], for example) was added through each set of (covariate, |residual|) points. These smoothers indicate that the spline based quantile curve gives a better quality of fit through almost the full covariate range than the cubic polynomial quantile. This is due to the more local nature of the spline based fitting procedure. We also calculated the mean square error based on the residuals for each model as a further method of assessing goodness-of-fit. We obtained mean square error values of 0.010 and 0.016 for the spline and polynomial based approach respectively. This is a further indication of the improvement that the nonparametric approach provides over its parametric counterpart. We should, however, bear in mind that in general goodness-of-fit and smoothing are competing aims in curve fitting.

Finally, we remark that the $p = 0.9$ Bayesian nonparametric quantile regression curve using splines obtained with the gamma hyperparameters $\alpha = 160$ and $\beta = 0.025$ was very similar to that shown in Figure 4. The credible envelope was, however, somewhat smoother. Our experience is that an estimated quantile is relatively insensitive to hyperparameter choice.

### *3.2 Application to Offshore Wave Data*

In our second example we use offshore wave data to further illustrate and validate our approach. These data refer to an offshore location in Poole Bay,

UK. There are three variables: wave height, wave period and wave direction, each having 86,384 observations at 3 hourly intervals, which amounts to just over 29 years of data. The data are shown in Figure 6. We can see that this data set has a different underlying structure from the coastal wave data as there is less variation in the magnitude of values (including high values) over the direction covariate. We also show in Figure 6 the same $p = 0.9$ Bayesian quantile regression curves and associated credible intervals as before. Our nonparametric quantile regression curve using splines provides us with a better understanding of the fine features of the $p = 0.9$ quantile than the cubic quantile regression curve, which is also shown. This advantage can be particularly helpful with data sets of this size and visual complexity.

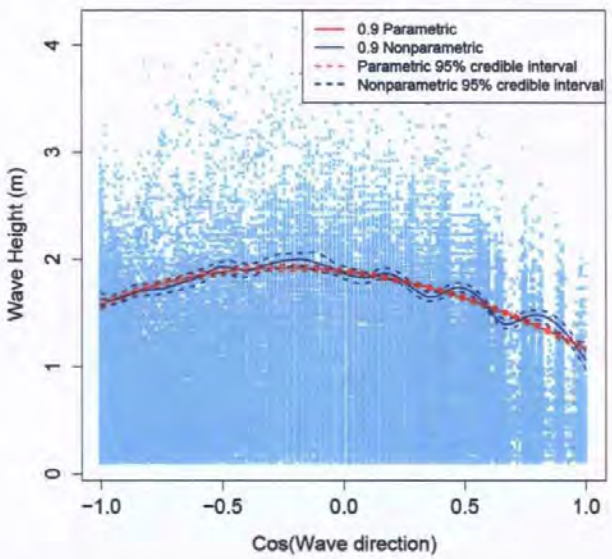



Fig. 6. Scatter plot of the offshore wave data showing the $p = 0.9$ Bayesian nonparametric quantile regression curve using splines and $p = 0.9$ parametric Bayesian quantile regression curve. 95% credible envelopes are also presented.

Figure 7 shows the absolute value of the residuals from both the cubic polynomial quantile regression curve and the spline based curve against the cosine of wave direction. We can clearly see that our spline based approach again provides a better quality of fit through the full covariate range than the cubic polynomial quantile curve. Again this is as a result of the more local nature of the spline based fitting procedure. This is apparent in this example as we have a large amount of data to work with meaning that local variation can be better identified than in smaller data sets.

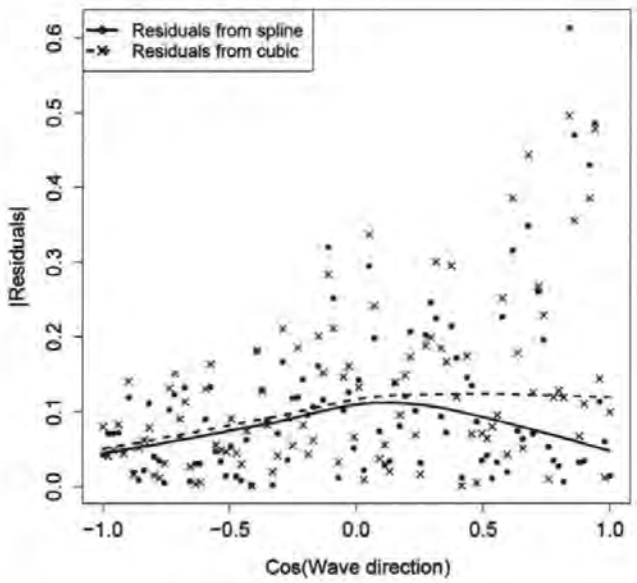



Fig. 7. The absolute values of the residuals against the cosine of wave direction with associated loess smoother from both the spline (dots, unbroken line) and the cubic (crosses, dashed line) quantile regressions. A grid of 100 sections along the covariate was used in the calculation of the residuals.

### *3.3 Application to a Simulated Data Set Bases on the Motorcycle Accident Data*

In our third example we apply our Bayesian nonparametric quantile regression spline based methodology to a simulated data set generated from the famous motorcycle accident data, discussed by Silverman [23] and presented in Figure 8. The data set comprises the head acceleration in multiples of the acceleration due to gravity $g$ at 133 times in milliseconds after a simulated motorcycle accident used to test crash helmets. This well known data set has been used frequently to motivate and demonstrate spline based methodology, since the nature of the underlying process makes polynomial modelling inappropriate. It provides a suitable test for our methodology.

Figure 8 also shows a smoothing spline found using the R's `spline` function [20]; see Green & Silverman [11] for a detailed discussion about the definition and calculation of smoothing splines. We simulated 100 values of acceleration at each of 30 equally spaced time points from a normal distribution with mean equal to the value of the smoothing spline at the time point, as shown in Figure 8 using filled dots, and standard deviation set to 20. We present the simulated data together with the smoothing spline in Figure 9. It is straightforward to calculate the true $p = 0.95$ quantile using mean + 1.96

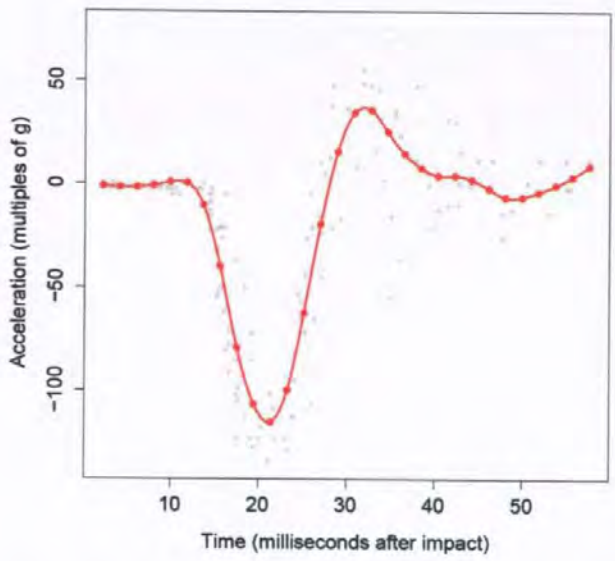



Fig. 8. Scatter plot of the motorcycle accident data from [23]. A smoothing spline has been added using the R package `splines` [20]. The values of the spline at 30 equally spaced times are shown using filled dots.

standard deviation; this is also shown in Figure 9, together with the empirical 0.95 quantile at each time point. We now apply our Bayesian nonparametric quantile regression spline based methodology to this simulated data set. We used $N = 30$ knots, one at each of the time points at which the data are generated. The resulting curve is shown in Figure 10. It recovers the true quantile function well, so confirming the effectiveness of our methodology.

## 4 Markov chain Monte Carlo Performance

### *4.1 Choosing the Proposal Density and Acceptance Rate*

In step (ii) of the Metropolis-Hastings algorithm presented in Section 2.2 the candidate vector $\boldsymbol{g}^\star$ was drawn from a multivariate normal distribution with variance-covariance matrix $\Sigma = \sigma^2 K^-/\lambda$. In this way a candidate $\boldsymbol{g}^\star$ has similar structure to a $\boldsymbol{g}$ from the prior term $\pi(\boldsymbol{g}|\lambda)$. We also considered generating $\boldsymbol{g}^\star$ from a multivariate normal distribution with $\Sigma = \sigma^2 I_N$ where $I_N$ is the $N \times N$ identity matrix. As a third possibility we updated a random subset of $g_1, \ldots, g_N$ again using independent normal distributions with variance $\sigma^2$. All three possibilities of generating $\boldsymbol{g}^\star$ performed similarly, with the choice of $\sigma^2$ having the greatest effect on the convergence of the Metropolis-Hastings

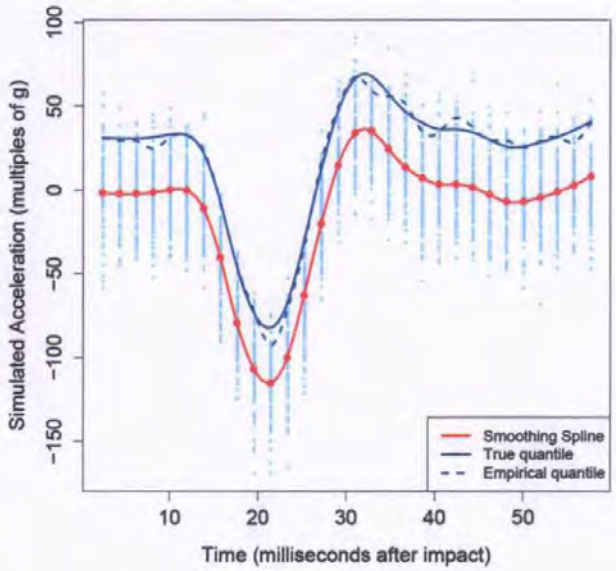



Fig. 9. Scatter plot of simulated data based on the smoothing spline shown in Figure 8. The true $p = 0.95$ quantile function is shown together with the empirical $p = 0.95$ quantile at each time point.

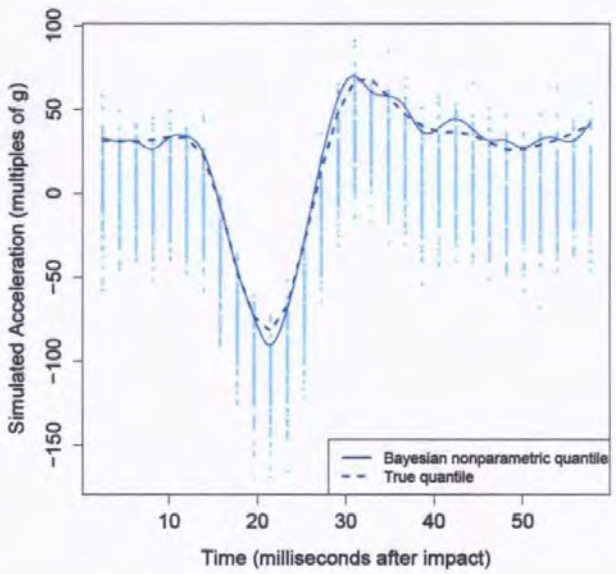



Fig. 10. Scatter plot of simulated data based on the smoothing spline shown in Figure 8. The true $p = 0.95$ quantile function is shown together with the $p = 0.95$ Bayesian nonparametric quantile regression curve using splines.

algorithm.

Bédard [1a] introduced a technique that can be applied here to optimally choose the parameter $\sigma^2$ that controls the variance $\Sigma = \sigma^2 K^-/\lambda$ of the proposal density $q$ for $\boldsymbol{g}$ in the Metropolis-Hastings algorithms. The technique plots an efficiency criterion against acceptance rates from the Metropolis-Hastings algorithm or against $\sigma^2$. The acceptance rate or value of $\sigma^2$ that corresponds to the maximum efficiency can then be chosen.

The key to this procedure is the use of the first order efficiency criterion which measures the average squared jumping distance for each parameter from one iteration to the next. In the case of the polynomial model of Yu & Moyeed [26] in which the parameters $\beta_0, \beta_1, \beta_2$ and $\beta_3$ are updated individually, Bédard [1a] would define the first order efficiency criterion (FOE) for the $i^{\text{th}}$ parameter as

$$\text{FOE}_i = E\left[\left(\beta_i^{(j+1)} - \beta_i^{(j)}\right)^2\right], \tag{14}$$

where the expectation is over iterations $j$.

The definition can be easily extended to the case of the spline, in which all the parameters $\boldsymbol{g} = (g_1, \ldots, g_N)^T$ are updated simultaneously, by using squared Euclidean distance as follows:

$$\text{FOE} = E\left[\sum_{i=1}^{N}\left(g_i^{(j+1)} - g_i^{(j)}\right)^2\right], \tag{15}$$

where again the expectation is over iterations $j$.

Figures 11 and 12 show plots of FOE against acceptance rate and against $\sigma^2$ for updating $\boldsymbol{g}$. These plots allow the user to choose the acceptance rate or $\sigma^2$ corresponding to the highest value of FOE. From Figure 11 it can be seen that an acceptance rate of about 0.24 is most appropriate. This may seem rather low, but is due to the fact that we are updating a whole vector of parameters $\boldsymbol{g}$ and not just an individual parameter. It is also in agreement with some of the literature about optimal acceptance rates; see Bédard [1a–b] and references therein for example. A relatively low acceptance rate corresponds to a relatively high proposal variance which itself allows larger possible jumps for the vector of parameters $\boldsymbol{g}$. A similar approach can be used to choose the value of $\sigma_\lambda^2$ for updating the smoothing parameter $\lambda$ in step (v) of the Metropolis-Hastings algorithm. In our application we fixed a value for $\sigma_\lambda^2$ and tuned $\sigma^2$. We then fixed our chosen $\sigma^2$ and tuned $\sigma_\lambda^2$. Finally, we fixed our chosen $\sigma_\lambda^2$ and re-tuned $\sigma^2$. We found that we were able to achieve good convergence for both $\boldsymbol{g}$ and $\lambda$ with these tuned values of $\sigma^2$ and $\sigma_\lambda^2$, as we will discuss in Section 4.3. We also found that this approach yielded a value of $\sigma_\lambda^2$ that was relatively insensitive to the value of $\sigma^2$.

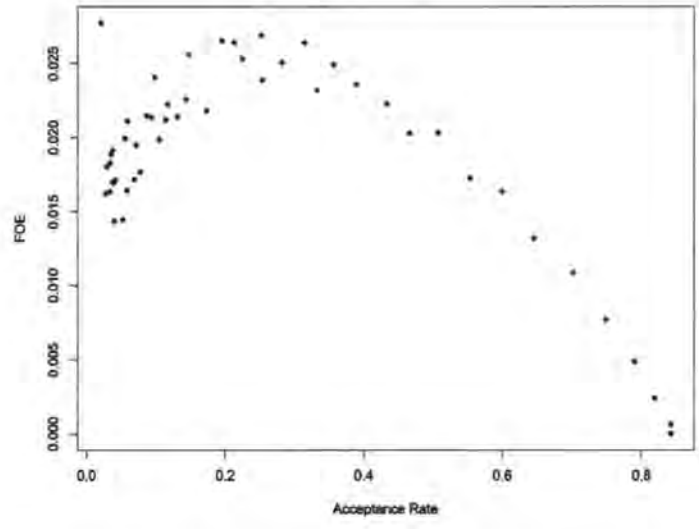



Fig. 11. First order efficiency criterion (FOE) against acceptance rate when updating $\boldsymbol{g}$ in the Metropolis-Hastings algorithm.

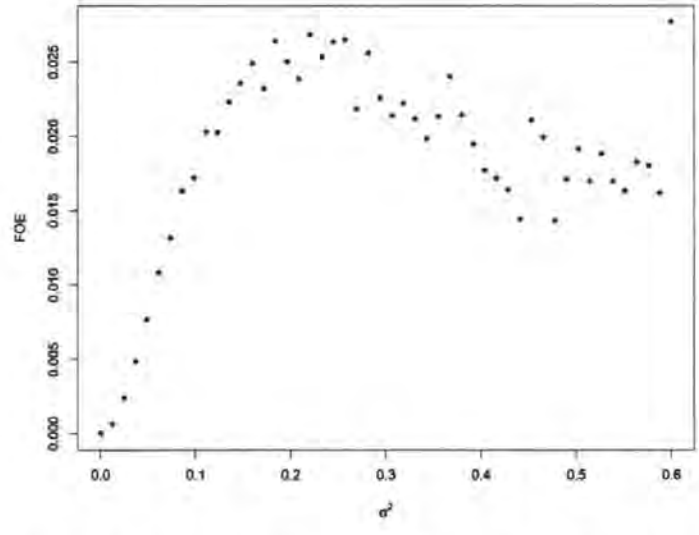



Fig. 12. First order efficiency criterion (FOE) against $\sigma^2$ for updating $\boldsymbol{g}$ in the Metropolis-Hastings algorithm.

### *4.2 Assessing Markov Chain Monte Carlo Convergence*

Visual assessment of the convergence of the Metropolis-Hastings algorithm was found to be difficult as the simulated elements included $N = 30$ points along the spline rather than just a few model parameters. We found that the

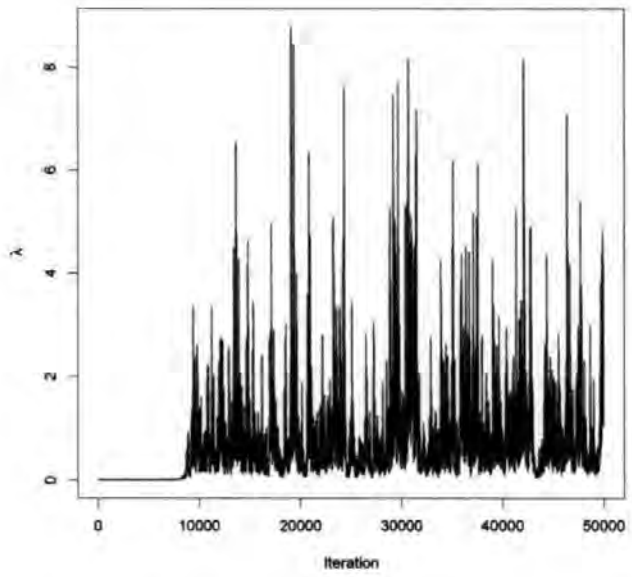



Fig. 13. Plot of $\lambda^{(j)}$ against thinned iteration $j$.

combination of a large number of sub-chains and an acceptance step based on a vector of points rather than a individual parameter could cause some convergence issues, although these could be overcome with good choices of $\sigma^2$ and $\sigma_\lambda^2$ as discussed in Section 4.1. Convergence for our nonparametric approach is generally slower than for parametric models. However this computational cost is balanced by the improved localized fitting of the model that we have seen. The visual assessment of convergence of $\lambda$ was also difficult as the parameter took a wide range of values as shown in Figure 13, where we can see that the time series converges around a lower value, with a tendency to jump to higher values (corresponding to smoother curves). We see that the time series has moved away from the low initial value of $\lambda^{(0)} = 10^{-6}$. In fact, values of $\lambda$ as low as $10^{-6}$ produce curves (not shown) that are visually far too rough.

After initially examining time series plots of $\lambda^{(j)}$ and of the individual $g_i^{(j)}$ sub-chains as shown in Figure 14, we used the more formal Gelman-Rubin statistic, discussed in Gelman & Rubin [9], Gelman [8] and Brooks & Gelman [3], to assess convergence of $\boldsymbol{g}$ and of $\lambda$. The Gelman-Rubin procedure compares the variances between and within chains to monitor convergence and is based on the 'estimated potential scale reduction factor' $\hat{R}^{1/2}$, which represents the estimated factor by which a credible interval for a parameter of interest may shrink if further simulation is carried out. Good performance is indicated by values of $\hat{R}^{1/2}$ close to 1. The value of $\hat{R}^{1/2}$ should certainly not exceed 1.2 as suggested in Kass et al. [13]. We calculated $\hat{R}^{1/2}$ for each sub-chain $g_i$, $i = 1, \ldots, N$, and for $\lambda$ and found that $\hat{R}^{1/2}$ took values between 1.0006 and 1.0152. Thinning was applied by taking every tenth value as particular sub-chains

showed strong autocorrelations. Thinning also reduced storage requirements.

Our examination of time series plots together with satisfactory values of the Gelman-Rubin statistic gave us confidence that the Metropolis-Hastings algorithm was producing realizations from the posterior distribution $\pi(\boldsymbol{g}, \lambda|\boldsymbol{y})$.

## 5 Conclusions

In this paper we have developed a methodology within the Bayesian framework to extend fixed degree polynomial based quantile regression to nonparametric quantile regression by using a spline based approach. We sampled from the posterior density of a NCS and an associated smoothing parameter by means of a specially tuned Metropolis-Hastings algorithm and used our sample to make inferences that include the quantification of uncertainty.

We have presented applications of our Bayesian nonparametric quantile regression methodology to two real environmental data sets, providing favourable comparisons with an existing parametric method and illustrating that learning about model parameters from data has taken place. We have confirmed the effectiveness of our methodology using simulated data based on a well known published data set for which polynomial modelling is inappropriate.

## Acknowledgement

The authors acknowledge the support of a doctoral scholarship from the University of Plymouth and funding from the EPSRC projects RF-PeBLE (grant No. EP/C005392/1), LEACOAST 2 (grant No: EP/C013085/1) and BVANG (grant No: EP/C002172/1). The authors would also like to thank Dr. Peter Hawkes from HR Wallingford for the coastal wave data set, and Dr. Anna Zacharoudaki from the University of Plymouth for the offshore wave data set, and supplementary information provided. We are very grateful to the editor and two reviewers for helpful suggestions that have substantially improved this paper.

## References

[1a] Bédard, M, 2006. Efficient sampling using algorithms: applications of optimal scaling results. http://probability.ca/jeff/ftpdir/mylene3.pdf, 2006.

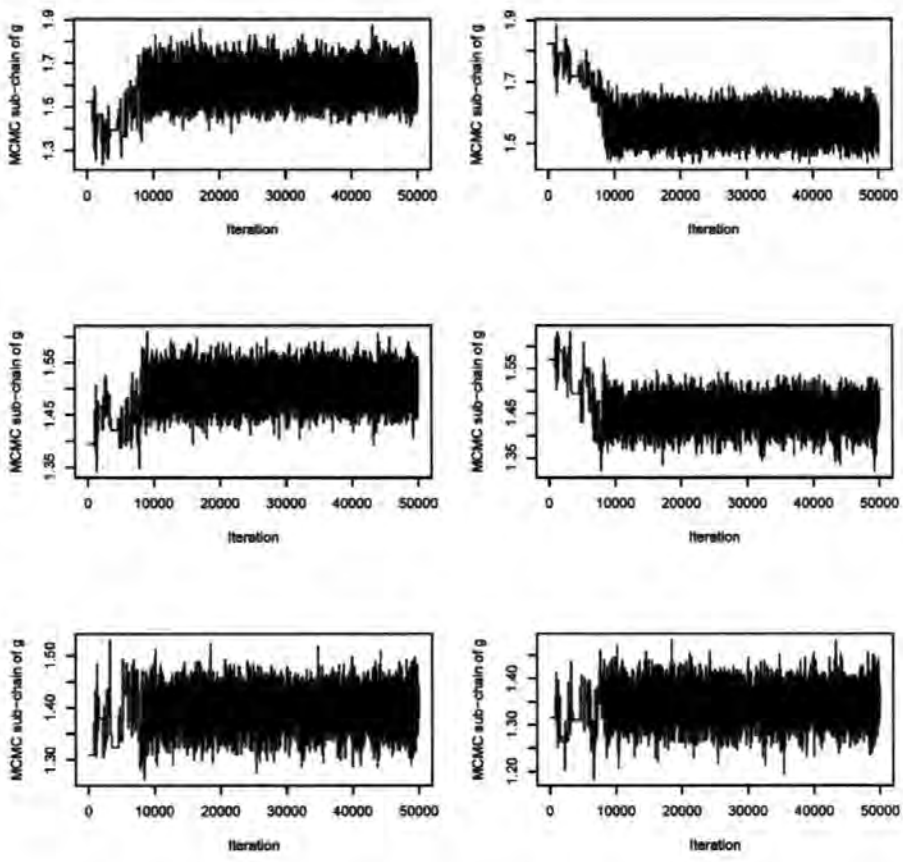



Fig. 14. Plots of $g_i^{(j)}$ against thinned iteration $j$ for six values of $i$.

[1b] Bédard, M, 2006. Optimal acceptance rates for Metropolis-Hastings algorithms: moving beyond 0.234. Submitted to the Annals of Statistics, 2006.

[2] Bosch, R., Ye, Y. and Woodworth, G.G., 1995. A convenient algorithm for quantile regression with smoothing splines. Computational Statistics and Data Analysis, 19 613–630.

[3] Brooks, S.P. and Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7(4) 434–455.

[4] Denison, D.G.T., Mallick, B.K. and Smith, A.F.M., 1998. Automatic Bayesian curve fitting. J.R. Stat. Soc.: Series B, 60(2) 333–350.

[5] Dias, R. and Gamerman, D., 2002. A Bayesian approach to hybrid splines non-parametric regression. J. Statist. Comput. Simul., 72(4) 285–297.

[6] Dunson, D.B. and Taylor, J.A., 2005. Approximate Bayesian inference for quantiles. Nonparametric Statistics, 17(3) 385–400.

[7] Gamerman, D., 1997. Markov chain Monte Carlo: stocastic simulation for Bayesian inference. Chapman and Hall, London.

[8] Gelman, A., 1996. Inference and monitoring convergence. In: W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.), Markov chain Monte Carlo in practice. Chapman and Hall, London, 131–143.

[9] Gelman, A. and Rubin, D., 1992. Inference from iterative simulation using multiple sequences (with discussion). Statistical Science, 7 457–511.

[10] Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82 711–732.

[11] Green, P.J, and Silverman, B.W., 1994. Nonparametric regression and generalized linear models. Chapman and Hall, London.

[12] Hastie, T., Tibshirani, R. and Friedman, J., 2001. The elements of statistical learning: Data mining, inference, and prediction. Springer, New York.

[13] Kass, R.E., Calin, B.P., Gelman, A. and Neal, R.M., 1998. MCMC in practice: A roundtable discussion. The American Statistician: Statistical Practice, 52(2) 93–100.

[14] Koenker, R., 2005. Quantile regression. Cambridge University Press, New York.

[15] Koenker, R., Ng, P. and Portnoy, S., 1994. Quantile smoothing splines. Biometrika, 81(4) 673–680.

[16] Koenker, R., 2008. `quantreg`: Quantile Regression (version 4.17). http://www.r-project.org. Vienna, Austria.

[17] Kottas, A. and Gelfand, A.E., 2001. Bayesian Semiparametric Median Regression Modeling. Journal of the American Statistical Association, 96(456) 1458–1468.

[18] Kottas, A. and Krnjajić, M., 2009. Bayesian Semiparametric Modelling in Quantile Regression. Scandinavian Journal of Statistics, 36 297-319.

[19] Neal, R.M., 2003. Slice sampling. Annals of Statistics, 31(3) 705–767.

[20] R Development Core Team, 2009. R: A language and environment for statistical computing. http://www.R-project.org, ISBN 3-900051-07-0. Vienna, Austria.

[21] Rao, R.C., 2002. Linear statistical inference and its applications (2nd Ed). John Wiley & Sons, New York.

[22] Reeve, D., Chadwick, A. and Fleming, C., 2004. Coastal engineering: processes, theory and design practice. SPON, London.

[23] Silverman, B.W., 1985. Some aspects of the spline smoothing approach to nonparametric curve fitting. J.R. Stat. Soc.: Series B, 47, 1–52.

[24] Thompson, P., Reeve, D., Cai, Y., Stander, J. and Moyeed, R., 2008. Bayesian nonparametric quantile regression using splines for modelling wave heights. FloodRisk 2008 conference. Keble College, University of Oxford, UK.

[25] Venables, W.N. and Ripley, B.D., 2002. Modern applied statistics with S (4th Ed). Springer-Verlag, New York.

[26] Yu, K. and Moyeed, R., 2001. Bayesian quantile regression. Statistics and Probability Letters, 54 437–447.

[27] Yu, K., Lu, Z. and Stander, J., 2003. Quantile regression: application and current research areas. J. R. Stat. Soc: The Statistician, 52(3) 331–350.