Methods for extreme value threshold selection and uncertainty quantification with application to induced seismicity

Conor Murphy, B.Sc.(Hons.), M.Res

Lancaster University

Submitted for the degree of Doctor of Philosophy at Lancaster University.

January 2025

STOR-i
excellence with impact

Abstract

Designing protection mechanisms to safeguard against ever-changing environmental processes requires the accurate estimation of future extreme hazards combined with reliable measures of their uncertainty. This thesis uses the peaks-over-threshold (POT) framework of extreme value modelling to provide high-quality tail inferences. The fundamental problem with POT analyses is selecting the threshold to characterise extreme values, with inferences sensitive to the choice. We develop improved methodology for threshold selection and for quantifying the uncertainty of this selection in tail inferences.

Even for independent and identically distributed data, threshold selection is a difficult task. Existing approaches can be subjective, sensitive to tuning parameters, or rely on asymptotics, resulting in suboptimal performance in practice. We develop a novel, objective, and effective methodology to automate threshold selection and propagate its uncertainty through to high quantile inference. We extend these approaches to handle non-identically distributed data with smooth generalised additive model formulations for the threshold and excess distribution parameters.

We adapt the methodology to address requirements for important applications. For coastal flooding, we focus the goodness-of-fit metric on the upper tail to ensure that the selected thresholds lead to accurate fitting to the most extreme observations. For modelling induced earthquakes, we use geophysical covariates regarding the measurement network and stresses induced by gas extraction, to form spatio-temporal threshold and excess distribution parameter functions. We develop a powerful estimator for a key quantity in seismicity modelling, the magnitude of completion. This estimator reduces the uncertainty and provides stronger evidence of a finite upper-endpoint than in previous research. We expand our uncertainty algorithms to account for the unknown model-covariate formulation and incorporate this uncertainty in inference for future endpoint summaries and quantile estimates relevant for design standards. Our methods have much wider applicability for inference for other induced seismicity contexts and wider environmental hazards.

Acknowledgements

Firstly, I would like to acknowledge my supervisors, Jon Tawn, Zak Varty, Ross Towe, and Pete Atkinson.

I feel so overwhelmingly grateful to have gone through this PhD with the support of Jon Tawn. During my MRes year, I had a short research project supervised by Jon, and in that short time, I knew Jon was the only PhD supervisor for me. I was then lucky enough to be allocated the project with him and it's safe to say, I could not have completed this PhD with anyone else. Jon, thank you so much for your unwavering support, seemingly endless patience and your expert guidance. Whether it be some early morning simple concept that wasn't clicking with me or a highly complex topic, your ability to break things down, draw a picture, and make everything clear is a talent that still baffles me. And that's only from the academic side! For the silly chats, funny stories, music recommendations, and all the rest of it, thank you so much for your light-hearted, welcoming, friendly spirit you bring everytime we chat. It's been a true privilege to work with someone I respect so highly and it's been an absolute pleasure every step of the way.

Zak Varty is an amazingly committed, patient and supportive supervisor. Thank you for being so generous with your time, especially when we use double the allocated time in our meetings! Those in-depth discussions, coding sessions and mini-workshops you have provided have helped so much in building my knowledge and skills, and importantly, my confidence, over the years. I look forward to over-running more meetings with you in the future!

Thank you to Ross for always being so welcoming on visits to Shell, and for providing so much guidance during my internship (and for all the coffees!). I'm very grateful for your support and willingness to delve into specific aspects when questions came up. For your useful comments, help with presentations and, your enthusiasm to organise meetings with the wider Shell community, thank you.

Pete, thank you for the interesting discussions and useful comments on work throughout the years.

Thank you also to the wider community of people I have had the pleasure to get to know and work alongside both within STOR-i and Shell.

Secondly, to STOR-i, I feel so privileged to have worked through my PhD in this centre. A PhD can be an isolating few years but with STOR-i, it was the opposite for me. The chats in the hub, the social activities and the network of people tackling the same challenge as you, was a huge comfort to me over the years. In particular, the experience would not have been the same without the Extremes Reading Group. Being able to talk through problems, tackle data challenges, and sometimes, to just have a bit of fun, has been truly beneficial to me. Thank you all.

I must also say a thank you to Kevin Hayes for pointing me in the direction of STOR-i. Thank you to the STOR-i admin team for the sheer volume of organising and chasing PhD students you do (me included!) and for the friendly guidance you provide any time a question has come up.

During my time at Lancaster, I have had the pleasure of living with multiple different groups of people, both from within and outside of STOR-i. Thank you to you all for the relief, fun and endless chats. I will always look back fondly on my time in Lancaster and you all are a big part of that. To the friends with whom I've hiked, had movie nights, cooked dinners, had big kitchen chats, Rebecca, Lídia, Robyn, Owen, Thomas, Ben, Josh, Jacob, Callum, Ethan, thank you for making my time in Lancaster all the more enjoyable.

I must mention the particularly close friends I've made throughout my time in Lancaster. Josh, Jacob, Owen, Callum and Thomas, I'm so grateful to have gotten to know you all. You are some of the most kind, supportive and lovely people I have ever met. You have made my time in Lancaster truly worthwhile and I'm so thankful to call you my friends.

Now, onto my amazing family and friends at home.

To Kev & Jess, thank you for your supporting chats, humbling mockery, all the laughs in the times of stress, and for having the most amazing kids whom I absolutely adore. Love you all!

To my oldest friend, Megan, thank you for still putting in effort after 28 years of friendship, for the catch-ups I look forward to and for allowing me to complain about my big work to-do list every time we see each other!

To Aidan & Paddy, my closest friends in the world. There's a long list of things I could say I wouldn't have done without your support, and this is certainly one of them. It still baffles me sometimes how soon after saying hello to either of you, I'll be laughing. Our childish sense of humour, unchanged from secondary school, is something I cherish. What a trio! Thank you for the pep talks, the aggressive debates and relieving laughter. I love you both.

Finally, to my truly amazing parents, thank you for your unending support through this PhD and all through my life. Thank you for the endless love you show all of your family (it's heartwarming to both observe and be a part of). Thank you for pushing me when I needed it, for listening to me when I needed support and for your belief in my ability. Thank you for your perspective, your reassurance, for all the pep talks, the handshakes, the hugs and the comforting chats. Thank you for celebrating every little milestone. If I keep going with this list, it will be longer than the thesis. I could not have done this without you. I love you both so very much.

Declaration

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

Chapter 3 has been published as Murphy, C., Tawn, J. A., and Varty, Z. (2025). Automated threshold selection and associated inference uncertainty for univariate extremes. Technometrics, 67(2):215–224. https://doi.org/10.1080/00401706.2024.2421744

Chapter 4 is linked strongly to work which has been submitted for publication as Collings, T. P., Murphy-Barltrop, C. J. R., Murphy, C., Haigh, I. D., Bates, P. D., and Quinn, N. D. (2025). Automated tail-informed threshold selection for extreme coastal sea levels.

Chapter 5 has been submitted for publication as Murphy, C., Tawn, J. A., and Varty, Z., Towe, R., Atkinson, P. M. (2025). Spatio-temporal modelling of extreme induced seismicity in the presence of an evolving measurement network.

Chapter 6 is a result of the Lancaster and Maynooth University contribution for a data challenge competition as part of the 2023 Extreme Value Analysis conference at the University of Bocconi, Italy. This has been published as André, L. M., Campbell, R., D'Arcy, E., Farrell, A., Healy, D., Kakampakou, L., Murphy, C., Murphy-Barltrop, C. J. R., and Speers, M. (2025). Extreme value methods for estimating rare events in Utopia: EVA (2023) conference data challenge: team Lancopula Utopiversity. Extremes, 28:23–45. https://doi.org/10.1007/s10687-024-00498-w. My primary contributions are in Sections 6.1-6.3.

The word count for this thesis is approximately 55071.

Conor Murphy

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Accurate modelling and forecasting of extreme values is important for a wide variety of applications in areas such as finance, actuarial science, engineering, and environmental science. In particular, with the continued escalation in the volatility of weather and environmental phenomena due to the effects of climate change and global warming, providing accurate, informative estimates of expected hazards, along with meaningful quantifications of the uncertainties in such estimates, has become absolutely essential. Natural hazard events such as storms, wildfires, coastal and river flooding, and earthquakes can pose severe threats to infrastructure and populated areas, depending on the quality of defences in place and the magnitude, frequency and type of the hazard.

Earthquakes pose a significant threat to infrastucture and populations with inadequate defences in place. For example, on 28th March 2025, central Myanmar was struck by a magnitude $7.7\mathrm{M_L}$ earthquake, measured in local magnitude, a logarithmic scale used to measure earthquake severity. The earthquake was followed by an aftershock of magnitude $6.4\mathrm{M_L}$ 12 minutes later. The combined impact left the city of Mandalay with significant damage including numerous collapsed buildings and bridges and unfortunately, caused multiple fatalities and injuries in the local populations. The underlying physical processes that generate earthquakes are highly complex. Modelling such processes is complicated further by limitations in the measurement equipment. Hence, advanced statistical approaches are necessary to provide informative estimates of future hazards. The area of statistical seismology has a large focus on modelling tectonic earthquakes resulting from movement in the tectonic plates deep below the Earth's surface. For such modelling, there is usually a wealth of large magnitude earthquakes observed across a large space and time window (Panakkat and Adeli, 2007; Zhuang et al., 2012; Kagan and Jackson, 2016).

While the magnitude and frequency of environmental hazards are continuously affected by the changing climate, anthropogenic processes can also contribute to increased risk more directly for certain environmental phenomena. We are interested in modelling induced earthquakes, the generating processes of which are directly affected by human activity. While these earthquakes have similarities with tectonic earthquakes, they present a variety of unique modelling challenges and opportunities. Induced earthquake catalogues typically do not contain the quantity of large magnitude events observed in tectonic datasets, in fact, the catalogues overall are generally significantly smaller. Furthermore, the events occur with much smaller magnitudes which can pose an extra difficulty for accurate detection, location and measurement, especially for the smaller magnitudes in the catalogue. Despite the smaller magnitudes, the events occur at much shallower depths relative to tectonic earthquakes and so, for an equivalent magnitude, induced earthquakes can cause much more damage in a more localised region.

A key quantity for induced earthquake modelling is the *magnitude of completion*, m_c, which is the smallest earthquake magnitude which can be detected and recorded with certainty if it occurs at a particular location and time. The value of m_c varies in space and time and relates directly to the density and sensitivity of the geophone network used to detect them. For the Groningen gas field in the Netherlands, the specific induced earthquake catalogue of interest in this thesis, it is conventionally accepted that since 1995, a conservative estimate of the magnitude of completion is 1.5 $\mathrm{M_L}$ across the whole region and period. However, with investment in the geophone network, the ability to detect smaller magnitudes events has improved, i.e., the magnitude of completion has reduced. Thus, an estimate of m_c is needed which allows for spatial and temporal variability, ensuring that this investment is not made redundant. We discuss the Groningen gas field further in the next section and detail our approaches for estimation of the magnitude of completion in Chapter 5.

The fewer observations in the induced earthquakes context also poses difficulties for parameter estimation and comparison of models and so, exploiting as much of the available data and information in a statistical model for induced earthquakes is of paramount importance. There are also unique opportunities, in this context, due to the direct relation between human activity and induced seismicity. As a result, for induced seismic catalogues, detailed records are typically kept of the human activity relating to such processes. Geophysicists combine such records with geological knowledge and geophysical models to estimate relevant physical covariates with good descriptive potential which we can utilise to improve statistical models. Furthermore, the geophysical models can be used to provide estimates of these covariates into the future under different scenarios - these can be utilised to provide useful future hazard estimates from statistical models. Induced seismicity could potentially be reduced or stopped with the right adjustments to the activity causing it, e.g., in Groningen, prior to the termination of extraction, in an attempt to reduce the seismic hazard, changes were made to the method of gas extraction to extract more evenly across the field and throughout the year. Failing this, improved infrastructure, designed to the standards derived from the resulting estimates of such models, can reduce the hazard to a safe level.

The ability to alter the process generating such earthquakes also presents a further modelling challenge in the form of covariate dependence which must be incorporated into the model, in this low-data setting. Moreover, improvements to the networks of geophones used to detect, locate and measure the magnitude of induced seismic events enables the detection of smaller magnitude earthquakes. This leads to a problem of missing data from times and locations where the network was too sparse or insensitive to detect, with probability one, the magnitudes of the missing events, i.e., events below the magnitude of completion. Accounting for such changes in both the earthquake generating process and the detection systems poses the major challenge in modelling induced seismicity, which we address in Chapter 5.

The main goal with modelling induced seismicity, and with the environmental phenomena mentioned above, is to provide meaningful estimates of future extreme hazards to contribute to more informed decision-making on the actions or interventions necessary to ensure the risk to infrastructure or population is at an acceptable level. Typical approaches for induced seismicity modelling rely on the Gutenberg-Richter law for earthquakes (Gutenberg and Richter, 1956) which states that earthquake magnitudes are exponentially distributed. This tends to lead to over-estimation of the upper-tail of induced magnitudes and also, assumes an identical distribution which is unlikely to be accurate for induced earthquake catalogues with known dependence on covariates. Across a variety of applications, when interested in estimates of extreme hazards beyond what we have already observed, we tend to focus on accurately modelling the most extreme observations. In any setting, this poses a challenge: By definition, such rare events do not occur often and so, within the observational window, usually only a small number of extreme observations are available. Extreme value theory (Coles, 2001; Leadbetter et al., 2012; Dey and Yan, 2016) is an area of statistics focussed on this particular problem and provides a powerful and robust framework which provides models focussed on the most extreme observations of a dataset and importantly, enables a mathematically-justified basis for the estimation of hazard levels beyond what we have previously observed. As a result, extreme value analysis has become instrumental in guiding the design of hazard protection infrastructure.

The most widely-used extreme value approach is the peaks-over-threshold (POT) framework (Davison and Smith, 1990) which fits a model to the observations in a dataset which exceed a suitably high threshold. Within this framework, the most fundamental challenge is the selection of the threshold that defines which observations are deemed extreme and thus, used in model fitting. This problem is fundamental to any threshold-based extreme value modelling approach regardless of the hazard of interest. Developing improved generic methodology for the estimation of this fundamental quantity is the main focus of this thesis, and features across all chapters. Although induced seismicity is a key driver of the methodological development in this thesis, we demonstrate the versatility of our techniques by applying them to other environmental applications. In Chapter 4, we adapt the methodology for the specific requirements of risk assessments for coastal flooding. In Chapter 6, we apply our methodology with key adjustments to simulated data from a global data challenge (where the truth is unknown).

When utilising such threshold models, parameter estimates and the subsequent inferences or hazard estimates, as well as the uncertainties of such quantities, can be highly sensitive to the choice of threshold. In this thesis, we develop methods to select this threshold which improve upon the leading existing approaches in the context of independent and identically distributed (IID) data in Chapter 3.

For the context of induced seismicity, with appropriate model setup, the extreme value threshold can be used as an estimator for the magnitude of completion. With the exponential distribution a special case of the POT framework, there is a clear link which can be drawn. In this thesis, in Chapter 5, we also develop a specialised method to select a spatio-temporal threshold function of unknown formulation for the more complex context of induced seismicity. In this context, selecting the threshold at the appropriate level is of even more importance. Making use of as much of the available information without overly biasing the model due to the missing data below the magnitude of completion is key in providing useful future hazard assessments. The methodology developed for this specialised context is also easily applicable to other datasets exhibiting covariate dependence and/or data missing-not-at-random. In the following section, we provide some background on the specific catalogue of induced earthquakes for which we develop our methodology.

Typically, when applying extreme value methods in practice, uncertainty is quantified by treating the threshold as fixed and propagating parameter estimation uncertainty through to inference using standard approaches. However, given the fundamental relationship of the unknown threshold to the resulting inference, omitting this aspect of uncertainty can lead to misleading and potentially, underestimated hazard forecasts. In this thesis, in Chapter 3, we develop methods to incorporate the uncertainty in the threshold estimation through to quantile inference. This aspect of uncertainty is also vital to include in the induced seismicity context due to the unknown formulation of the magnitude of completion across space and time. We also detail approaches to account for the uncertainty in the estimation of the magnitude of completion and its spatio-temporal formulation. These methods again are widely applicable to other contexts as a way to account for threshold estimation uncertainty and the uncertainty in the form of covariate dependence within the model.

1.2 The Groningen gas field

In 1959, the Groningen gas field in the Netherlands was discovered. It is estimated that the reservoir contained nearly 3000 billion cubic metres (bcm) of gas when discovered and today, Groningen still remains one of the largest natural gas fields in the world. Extraction from the gas field began in 1963 operated by the Dutch Petroleum Society (Nederlandse Aardolie Maatschappij, NAM), a joint operation between Royal Dutch Shell and Exxon Mobil. The gas field is located in the north-east of the Netherlands spanning numerous populated towns and villages. With this in mind, the region has been monitored for low seismic events resulting from gas extraction since 1986. The network of geophones used to detect, locate and record the magnitude of such events is owned by the Royal Netherlands Meteorological Institute (KNMI). This network was not sufficient to detect and locate events to a meaningful degree of accuracy until April 1995, the start date of the earthquake catalogue we utilise in our analysis. Continued investment in this network beyond this date has improved detection ability over time and allowed for deeper understanding of induced seismicity. The region now has the highest resolution geophone network on the planet.

While magnitudes of induced earthquakes are typically significantly lower than tectonic earthquakes, e.g., the largest earthquake recorded in the Groningen gas field was $3.6\mathrm{M_L}$ in August 2012, induced earthquakes occur at shallow depths and thus, can lead to significant damage. As of now, despite much of the gas reserves remaining, gas extraction has ceased at Groningen, partially due to similarly large events of magnitude $3.4\mathrm{M_L}$ occurring in January 2018 and May 2019, after steps had been taken to mitigate induced seismicity. However, despite the cessation of extraction, accurate modelling and forecasts of induced seismicity are still vital in maintaining the safety of the surrounding areas as seismic events will still occur.

The Groningen region is fault-closed and tectonically stable meaning that all earthquakes can be attributed to the physical characteristics of the reservoir and its relationship with the human activity. The gas reservoir is located at a depth of 2.6-3.2km below the surface and is composed of a porous rock structure filled with gas from the carboniferous layer below. Combined with the pre-existing fault structure of the reservoir, pore pressure depletion due to gas extraction and the resulting compaction of the reservoir under the additional stresses are considered the key factors in inducing seismic events in the Groningen region. As gas is removed through wells, the remaining gas redistributes slowly across the reservoir to equalise areas of high and low pressure. As this pressure in the pore space redistributes, it can put additional stresses on the rock structure which can lead to compaction. This compaction can place additional stresses on faults and if these stresses reach a limiting stress, the fault can slip and cause seismic activity. Even after the cessation of extraction, the gas continues to slowly redistribute, changing stresses on the rock structure and potentially inducing further seismic events.

This complex geophysics is beyond the remit of a statistical model, but is captured in numerical models used by geophysicists which provide an informative, spatio-temporal covariate, known as the Kaiser stress, which has been demonstrated in statistical approaches to be useful in describing the intensity of earthquake occurrences and their magnitudes. Physically-motivated projections of this covariate into the future can also be used in statistical models for future hazard estimates under extraction scenarios.

1.3 Overview of thesis

This thesis has three main objectives; firstly, to develop improved methods for the selection of an appropriate modelling threshold when utilising a threshold-based extreme value model in a variety of contexts; secondly, to quantify the uncertainty in the threshold estimation procedure and to propagate this through to inference; thirdly, to improve the statistical modelling of induced earthquake magnitudes and provide useful future hazard estimates for the specific application of the Groningen gas field.

This thesis aims to preserve the integrity of the papers that have been submitted for publication. As a result, each of the main chapters of novel research corresponds to a submitted paper. Hence, each of those chapters has overlap with the broader literature review of the thesis.

Chapter 2 provides an overview of the existing methods for modelling univariate extreme values. We introduce the two key extreme value models of maxima and threshold exceedances under the IID assumption and then, outline how such models can be adjusted when the IID assumption is relaxed.

Chapter 3 details our novel methodology for the automated selection of a threshold for univariate extreme value modelling for IID random variables. Threshold selection is a fundamental problem in any threshold-based extreme value analysis. While models are asymptotically motivated, selecting an appropriate threshold for finite samples is difficult and highly subjective through standard methods. Inference for high quantiles can also be highly sensitive to the choice of threshold. Too low a threshold choice leads to bias in the fit of the extreme value model, while too high a choice leads to unnecessary additional uncertainty in the estimation of model parameters. We develop a novel methodology for automated threshold selection that directly tackles this bias-variance trade-off. We also develop a method to account for the uncertainty in the threshold estimation and propagate this uncertainty through to high quantile inference. Through a simulation study, we demonstrate the effectiveness of our method for threshold selection and subsequent extreme quantile estimation, relative to the leading existing methods, and show how the method's effectiveness is not sensitive to the tuning parameters. We apply our method to the well-known, troublesome example of the River Nidd dataset. This chapter is published as Murphy et al. (2025) with an accompanying Github repository (Murphy et al., 2023) detailing the implementation of our methodology. Alongside the published paper, there is a large online supplementary material including more detailed description of methods and additional simulation experiments, analyses and results. The description of methods in this supplementary has been used in the literature review of this thesis while the additional analyses are shown in Appendix A.

Chapter 4 builds upon the work of Chapter 3 to develop an extension of the automated threshold selection method for a systematic application to coastal flood risk. Peaks over threshold techniques are commonly used in practice to assess coastal flood risk, with the threshold often still selected through rule of thumb or subjective methods.

Using the data-driven method of Chapter 3 for this specific application led to threshold choices with resulting fits inadequate at the most extreme observations. We adapt our methodology of Chapter 3 to focus the threshold selection on capturing the most extreme values in the upper tail, at the cost of additional uncertainty in subsequent inference. We apply our method to a global data set of coastal observations, where we illustrate the robustness of our approach and compare it to the methods developed in Chapter 3. Material related to this chapter has been submitted for publication as Collings et al. (2025).

Chapter 5 extends the methodology developed in Chapter 3 and builds upon the work of Varty et al. (2021) to develop an automatic selection procedure for a spatio-temporal magnitude of completion for induced earthquake modelling. Seismic activity arising from gas injection/extraction underground poses a significant hazard to the surrounding infrastructure and populations. Efficiently estimated models of the upper tail of the earthquake magnitude distribution, that can vary with intervention strategies, are vital for understanding such hazards into the future. To this end, we utilise an extreme value model by employing a new procedure for the automatic selection of a parametric spatial-temporal threshold function, utilising knowledge of changes in the measurement network. This threshold function choice excludes data from the analysis that are subject to sampling bias. We propose novel methods to propagate the uncertainty in the threshold estimation and the choice of covariate formulation through to tail inference. We apply our methodology to the catalogue of induced earthquakes from the Netherlands' Groningen gas field, with our methods delivering clear improvements over existing analyses and providing the first quantification of the different sources of uncertainty in such estimates. The methodology has the potential to be useful for a range of novel extreme value contexts where data are missing due to measurement equipment limitations, where parametric models are used for the threshold, and in accounting for threshold uncertainty in subsequent inference.

Chapter 6 introduces a variety of methods to capture the extremal behaviour of complex environmental phenomena where flexible techniques for modelling tail behaviour are required. We introduce a variety of such methods, which were used by the Lancopula Utopiversity team to tackle the EVA (2023) Conference Data Challenge. This data challenge was split into four challenges, labelled C1-C4. Challenges C1 and C2 comprise univariate problems, where the goal was to estimate extreme quantiles for a non-stationary time series exhibiting several complex features. For these, we propose a flexible modelling technique, based on generalised additive models, with diagnostics indicating generally good performance for the observed data. Challenges C3 and C4 concern multivariate problems where the focus was on estimating joint probabilities. For challenge C3, we propose an extension of available models in the multivariate literature and use this framework to estimate joint probabilities in the presence of non-stationary dependence. Finally, for challenge C4, which concerns a 50-dimensional random vector, we employ a clustering technique to achieve dimension reduction and use a conditional modelling approach to estimate extremal probabilities across independent groups of variables. This chapter has been published as André et al. (2025).

Chapter 7 concludes with a summary of the contributions of this thesis and discussing the limitations of the presented methods and potential avenues for future work.

Chapter 2

Literature review

2.1 Introduction

In this chapter, we outline the extreme value methods relevant throughout the rest of the work in this thesis. In a variety of applications, such as hydrology or sports (Coles et al., 2003; Spearing et al., 2023), interest lies in the behaviour of data lying in the tails of a distribution. In such contexts, data values can be quite sparse and thus, standard statistical approaches are not particularly useful. Extreme value theory is a growing branch of statistics which provides asymptotically-justified frameworks for modelling the most extreme values of a distribution. It is applicable and quite reliable when events are scarce (Reiss and Thomas, 2007; Coles, 2001). Most importantly, it provides convenient methods for extrapolation beyond levels which have already been observed, making such frameworks invaluable when hazard and risks assessment of future extreme values is of paramount importance, e.g., in finance (Smith, 2003), flood risk analysis D'Arcy et al. (2023), or nuclear regulation Murphy-Barltrop and Wadsworth (2024).

Below, we describe the two most common approaches for univariate extreme value analysis. We first look at these approaches in the context of independent and identically distributed (IID) random variables. Then, we focus on the peaks-over-threshold framework which is the basis for all of the work in this thesis. We present the existing approaches used for extending this framework for identically-distributed variables changing with covariates. We explore existing approaches for the selection of the threshold in this framework, which is the most fundamental challenge of this area and the main focus of this thesis.

All of the developed methods in this thesis are concerned with modelling a univariate process. There is also a growing body of research in multivariate and spatial extreme value analysis (Wan and Davis, 2019; Heffernan and Tawn, 2004; Wadsworth and Tawn, 2013; Mhalla et al., 2019; Murphy-Barltrop et al., 2024; Schlather and Tawn, 2003; Shaby and Reich, 2012; Richards et al., 2022); we foresee our methodologies being useful for these contexts but we do not explore them in this thesis.

2.2 Extreme value theory

In general, statistical modelling techniques aim to describe and predict the typical values of a process. For a wide range of contexts, interest lies in values which are not typical, values which lie far from the central tendencies of a process, values about which we usually have far less information. Values such as this lie in the lower or upper tails of a process/distribution and are usually termed *extreme values*. To describe the behaviour of extreme values, statistical models focused on the tails of a distribution, rather than on the main body, are required. Usual statistical techniques rely on a large number of typical observations and so are not appropriate when concern lies with small numbers of extreme observations.

Extreme value theory provides asymptotically justified models for the tails of a probability distribution. These models are focused on the extreme observations lying in the tail of the distribution and thus are usually fitted solely to these observations so as not to be affected by the central values. Most importantly, extreme value models allow for extrapolation beyond observations and so are vital for hazard protection against levels of a process which have not already been observed.

This section provides background for the asymptotic justification of the two most widely-used univariate extreme value models; the *block-maxima* approach and the *peaks-over-threshold* approach.

2.2.1 Block maxima approach

The block maxima approach (Coles, 2001) splits data into predefined blocks, often taken as a year, extracts the maxima from each block and models these maxima using a generalised extreme value distribution. The probabilistic justification for this, summarised in Leadbetter et al. (2012), is as follows.

Consider a sequence of independent and identically distributed random variables $X_1, ..., X_n$ with distribution function F and let $M_n = \max\{X_1, ..., X_n\}$. To obtain the distribution for M_n, theoretically, we can derive this simply as:

$$\Pr(M_n \leq x) = \Pr(X_1 \leq x, \dots, X_n \leq x) = (F(x))^n.$$

As F is unknown, problems arise when using standard techniques for the estimation of F with the goal of estimating M_n. Any small discrepancies in any parametric model choice for F lead to large deviations in F^n and thus, in the distribution of M_n.

To avoid this issue, we instead approximate F^n directly with a flexible family models for F^n as $n \to \infty$. To avoid convergence of the distribution of M_n to a point mass on the upper end point of F, we first use sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ to renormalise M_n as :

$$M_n^* = \frac{M_n - b_n}{a_n}.$$

The Extremal Types Theorem (Fisher and Tippett, 1928) states that if there exists sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr(M_n^* \leq x) = [F(a_n x + b_n)]^n \to G(x) \quad \text{as}\, n \to \infty, \tag{2.2.1}$$

with G a non-degenerate limit distribution, then G belongs to one of the Fréchet, Gumbel or negative Weibull families which, under the Unified Extremal Types Theorem (Jenkinson, 1955), make up special cases of a single flexible family of models known as the generalised extreme value (GEV) distribution such that:

$$G(x) = \exp\left(-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right), \tag{2.2.2}$$

with $w_+ = \max(w, 0)$, $\mu \in \mathbb{R}$ the location parameter, $\sigma > 0$ the scale parameter, and $\xi \in \mathbb{R}$ the shape parameter.

A justification of the GEV is given in Leadbetter et al. (2012) while estimation and uncertainty quantification of the GEV parameters via maximum likelihood estimation are covered in Coles (2001). The shape parameter ξ defines the tail behaviour of the GEV. For $\xi < 0$, the GEV corresponds to the light-tailed negative Weibull distribution which has a finite upper end point. Whereas for $\xi \geq 0$, the distribution is unbounded with $\xi = 0$ corresponding to a Gumbel distribution which has an exponential upper tail and $\xi > 0$ leading to the heavy-tailed Fréchet distribution.

A critical property of the GEV distribution is the max-stability property which states that, for any $m \in \mathbb{N}$, there exists $\alpha_m > 0$ and β_m such that $G^m(\alpha_m x + \beta_m) = G(x)$, $x \in \mathbb{R}$. Hence, taking the maxima of a variable that follows the GEV distribution, results in a random variable which is itself also GEV, subject to a change in location and scale parameter values, but with the same shape parameter. The GEV family of distributions is the only family which satisfies this useful property. More specifically, given that distribution (2.2.2) holds exactly for large n, then we have,

$$\Pr(M_n \leq x) = G\left(\frac{x-b_n}{a_n}\right) = \tilde{G}(x)$$

where $\tilde{G}$ corresponds to the GEV distribution with adjusted location and scale parameter values. The max-stability property is analogous to the sum-stability property of the central limit theorem, which motivates the wide use of the Gaussian distribution as a model for finite sample means. The max-stability property allows for the modelling of maxima in practice where data are split into blocks of equal length and the maxima within each block are then considered as realisations from a GEV, given by $\tilde{G}$.

Careful consideration must be given to the choice of block size as there is a trade-off between bias and variance; if the block size is chosen to be too small, approximation by the limit result (2.2.2) is likely to be poor, leading to bias in estimation of parameters and more importantly, in extrapolation; while a large block size could lead to a significant wastage of data, very few block maxima for the model fit, and thus large estimation variance. Such considerations often lead to block sizes being chosen to be one year in length. In some situations, only annual maxima may have been recorded, making this choice of block size the only option. Once parameters are estimated, the $(1-p)$-quantile of the distribution can be calculated by letting $G(x_p) = 1-p$ and solving for x_p such that:

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - \{-\log(1-p)\}^{-\xi}], & \text{for } \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\}, & \text{for } \xi = 0. \end{cases}$$

The level x_p is exceeded on average once every $1/p$ years and is termed a *return level*, in this case, corresponding to a *return period* $1/p$. Assuming block length of one-year, x_p refers to the level we expect to be exceeded by the annual maximum in any year with probability p.

Note that if interest lies in the extremes of the smallest values of a dataset, for example, when applying extreme value methods to low temperatures (Fyodorov and Bouchaud, 2008), the block maxima approach may still be used with an adjustment by considering the minima of the process as maxima such that:

$$\min(X_1, \ldots, X_n) = -\max(-X_1, \ldots, -X_n).$$

A common problem with the block maxima approach is that as some blocks may have larger numbers of extreme observations than others, by taking only the maximum within each block, this may lead to a significant wastage of the useful extreme data. Another problem may be encountered with this approach when wanting to model the distribution of the extreme values in a situation where the extremal behaviour of the series of interest may be related in some way to a covariate. The GEV is only appropriate in this scenario if the covariate is changing slowly relative to the chosen block size. If covariates vary significantly within blocks, the GEV's within-block identical distribution assumption will be violated. These two problems can be overcome by instead using threshold models to model extreme tail behaviour.

2.2.2 Peaks over threshold

The more widely-used approach for modelling extremal behaviour is to define extreme observations as those which exceed some appropriately high threshold u and to then model both the rate at which exceedances of u occur and to model the distribution of excesses of u. This method makes better use of more of the available data, avoids wastage when multiple extreme events occur within a single block, as shown in Figure 2.2.1, and can be easily adapted to incorporate covariates.

To derive an asymptotic model for peaks over threshold, let X be a random variable

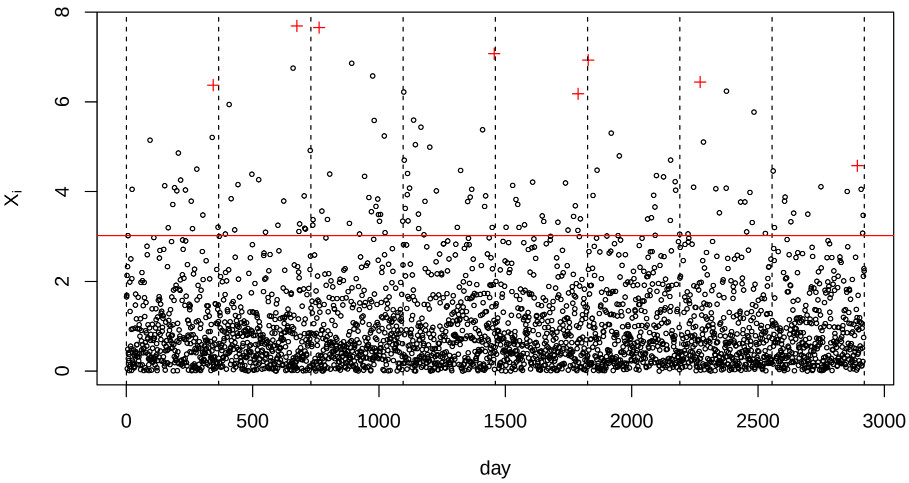

Figure 2.2.1: Simulated daily data. Red crosses show the values in an annual maxima extreme value analysis. Red line indicates threshold value $u = 3.0$, all exceedances of which are used in a peaks over threshold or point process approach to extreme value analysis.

with distribution function F. For any threshold u, we have:

$$\Pr(X > u + y | X > u) = \frac{1 - F(u + y)}{1 - F(u)} \quad \text{for } y > 0.$$

If the limiting behaviour of the block maxima of IID variables $X_1, \ldots, X_n$ with distribution function F (i.e., the same distribution as X) can be characterised by G (as in (2.2.2)), then the scaled excesses of a threshold u, as u tends to the upper endpoint of F, show corresponding limiting behaviour characterised by the family of distributions known as the generalised Pareto distribution (GPD). For $u_n(u) = b_n + a_n u$, with a_n, b_n satisfying limit (2.2.1), then, as $n \to \infty$, $u_n(u) \to x^F$, where $x^F := \sup\{x : F(x) < 1\}$ is the upper endpoint of the distribution. Also, as $n \to \infty$, for $x > u$,

$$\Pr(X > u_n(x) | X > u_n(u)) \to 1 - H_u(x) \tag{2.2.3}$$

where

$$H_u(x) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi} & \xi \neq 0 \\ 1 - \exp\left(\frac{x-u}{\sigma_u}\right) & \xi = 0 \end{cases} \tag{2.2.4}$$

with $(\sigma_u, \xi) \in \mathbb{R}_+ \times \mathbb{R}$ being the scale and shape parameters respectively and $w_+ = \max(w, 0)$. The threshold-invariant shape parameter is equal to that of the corresponding GEV shape parameter for modelling block maxima while the threshold-dependent scale parameter satisfies the property $\sigma_u = \sigma + \xi(u - \mu)$ with μ and ξ denoting the GEV location and shape as in distribution (2.2.2). Smith (1989) provides explanation of the relationship between the GEV and GPD, Coles (2001) provides outline justification for the GPD in approximating threshold excesses while Pickands (1975) provides a formal justification of the asymptotic model under weak conditions. Davison and Smith (1990) overview the properties of the GPD.

The GPD shape parameter defines the tail behaviour of the distribution leading to three sub-classes depending on the value of ξ:

1. For $\xi > 0$, the GPD is heavy-tailed and the distribution of the excess $(X-u)|(X > u)$ is Pareto with tail index $1/\xi$.

2. For $\xi = 0$ (interpreted as the limit $\xi \to 0$), $(X - u)|(X > u)$ has an exponential distribution with expectation $1/\sigma_u$.

3. For $\xi < 0$, the GPD is light-tailed and has a finite upper end point at $u - \sigma_u/\xi$.

In practice, we assume that the above limit (2.2.3) holds for a suitably high threshold $u_n(u) = u$ with $u_n(x) = x$ with $x > u$. We can then model the excesses of this threshold as approximately GPD under this limiting tail model. We can then provide a model for the $X > u$ given by:

$$\Pr(X > x) = \lambda_u \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi} \tag{2.2.5}$$

for $x > u$, where $\lambda_u = \Pr(X > u)$ is the threshold-exceedance rate.

Return levels for the GPD can be obtained similarly to the block maxima procedure. To estimate the m-observation return level x_m, for $\frac{1}{m} < \lambda_u$, we utilise equation (2.2.5) and let $\Pr(X > x_m) = \frac{1}{m}$. Solving this equation for x_m leads to:

$$x_m = \begin{cases} u + \frac{\sigma_u}{\xi}\left[(m\lambda_u)^{\xi} - 1\right], & \text{for } \xi \neq 0, \\ u + \sigma_u \log(m\lambda_u), & \text{for } \xi = 0, \end{cases}$$

which provides the value expected to be exceeded once every m observations. This procedure may be adjusted to obtain return levels with return periods in terms of years, i.e., to obtain the T-year return level, we set $m = T\bar{n}$ where $\bar{n}$ is the average number of observations occurring in a year. Return levels of this type are typically more useful in practice (Coles, 2001).

2.2.3 Inference for GEV and GPD

Based on the equations above, it is clear that estimating return levels requires the estimation of parameter values. Once the appropriate block length is determined for the GEV or the appropriate threshold is selected for the GPD (discussed in Section 2.3), either of the above extreme value frameworks can be fitted using likelihood-based or Bayesian methods (Davison and Smith, 1990; Coles and Tawn, 1996). Numerical optimisation techniques are required to estimate the parameters for a likelihood-based approach as closed forms are not available for the maximum likelihood estimators for either model (2.2.2) or (2.2.4). This is also the case for the posteriors of the distribution in a Bayesian framework where some form of Markov-Chain Monte-Carlo (MCMC) methods would be needed. Thus, Bayesian procedures for extreme value analysis can be computationally intensive, however they do allow for expert knowledge to be incorporated into the modelling framework (Yue et al., 2025b). Bayesian methods also allow for convenient estimation of parameter uncertainty.

A potential difficulty with likelihood methods concerns the regularity conditions which are required for the usual asymptotic properties associated with the maximum likelihood estimators. These conditions do not hold for cases where the end-points of the distribution are functions of the parameter values, i.e. $\mu - \sigma/\xi$ is the upper end point when $\xi < 0$ for the GEV, and the lower end point when $\xi > 0$. For the GPD, $u - \sigma_u/\xi$ is the upper endpoint for $\xi < 0$. As a result, standard asymptotic likelihood methods are not automatically applicable for $\xi \in \mathbb{R}$. According to Smith (1985), there are three cases to note:

- if $\xi > -0.5$, MLEs have the usual asymptotic properties;
- if $-1 < \xi < -0.5$, MLEs may be obtainable but do not have the standard asymptotic properties;
- if $\xi < -1$, MLEs are unlikely to be obtainable, with the best estimator of the upper endpoint being the sample maximum.

Typically, for most environmental applications, we see values of ξ in the range $-0.4 < \xi < 0.4$, so these limitations of the MLEs are not a problem in practice (Coles, 2001).

For $\xi > -0.5$, standard confidence intervals can be generated under the assumption of asymptotic normality (Smith, 1985). However, due to the typically small sample sizes of datasets relevant to extreme value analysis, bootstrapping methods can be more useful for uncertainty quantification (Healy et al., 2025). For the GPD, the natural estimator for λ_u, the probability of an observation exceeding the threshold u, is $\hat{\lambda}_u = n_u/n$ where n is the total number of observations and n_u is the number of observations exceeding the threshold u, i.e., this is the sample proportion of points exceeding u. Furthermore, the number of exceedances of u follows a binomial distribution, $\text{Bin}(n, \lambda_u)$, so $\hat{\lambda}_u$ can also be estimated as the MLE for λ_u.

Standard diagnostic approaches are applicable for assessing the fit of the GEV and GPD models. While there are no methods outside of the asymptotic justification to verify the model's ability for extrapolation, goodness-of-fit assessments for the already-observed levels of the data provide an indication of whether the model will be appropriate further into the tail. Standard PP- and QQ-plot assessments can provide useful information, with the former focussing on the bulk of the distribution and providing little information about the fit for the largest values, whereas the latter focusses on the areas of most interest for extreme value modelling, showing how the model deviates from the observed data with a focus on the largest values (Heffernan and Tawn, 2003).

A detailed description of the frequentist procedure for inference and model assessment for the above extreme value methods is given in Coles (2001). Coles and Tawn (1996) and Sharkey and Tawn (2017) provide details on Bayesian inference procedures for extreme value analysis.

2.3 Threshold selection

Section 2.2.2 motivates the use of the GPD as a model for the excesses of a high threshold u, as $u \to x^F$. In practice, a suitably high threshold must be chosen such that the excesses may be taken as approximate samples from a GPD and parameters may be estimated accurately. A fundamental challenge with the use of such a model in practice is choosing an appropriate threshold value. This choice is analagous to the choice of block size under the GEV model (2.2.2) in that it involves a bias-variance trade-off: selecting a threshold too low is likely to violate the asymptotic basis of the GPD model, incorporating bias into the estimation of the parameters and resulting inference, whereas too high a threshold results in an unnecessarily small number of excesses with which to fit the model and thus, large estimation uncertainty. Ideally, we must choose the threshold as low as possible provided the GPD shows an adequate fit to the excesses.

A plethora of methods have been developed to tackle this problem - see Scarrott and MacDonald (2012); Belzile et al. (2023) for extensive reviews of existing approaches. Despite the variety of methods, the most commonly-used are graphical goodness-of-fit diagnostics, demonstrated in Sections 2.3.1-2.3.2, which suffer from subjectivity due to the visual nature of the selection. The leading existing automated approaches, discussed in detail Section 2.3.5 & 2.3.5, remove the subjectivity problem but suffer from other shortcomings. Wadsworth (2016) exhibits strong reliance on asymptotic theory leading to problems with the small samples typical of extreme value analysis; this is particularly troublesome if considering a fine grid of candidate thresholds. Northrop et al. (2017) show significant sensitivity to tuning parameters leading to unpredictable results if time is not spent selecting such tuning parameters carefully for each sample, a setback if using methods for repeated estimation of thresholds across a variety of datasets, e.g. for widespread flood risk analysis (Keef et al., 2013a).

2.3.1 Threshold stability property

The *threshold stability property* of the GPD (Davison and Smith, 1990) is instrumental to some of the most widely-used threshold selection approaches. It states that if excesses of a threshold u follow a GPD, then excesses of a higher threshold v $(u < v < x^F)$ also follow a GPD, with the same shape parameter and an adjusted scale parameter, i.e., if $(X - u)|(X > u) \sim \text{GPD}(\sigma_u, \xi)$, then $(X - v)|(X > v) \sim \text{GPD}(\sigma_u + \xi(v - u), \xi)$. This implies that the GPD shape parameter ξ should have the same value for all valid choices of threshold, a useful property to exploit when selecting an appropriate threshold, as a modelling threshold may be selected as the lowest candidate value for which the shape parameter shows adequate stability, accounting for the sampling variability in the parameter estimation. Although, in practice, this can pose problems, which we will discuss later in this section.

The threshold stability property is derived as follows: Suppose that we have threshold exceedances above a threshold u which follow a GPD such that $(X-u)|(X>u) \sim \text{GPD}(\sigma_u, \xi)$. We want to find the distribution of exceedances of a higher threshold $v > u$. Thus, to find the distribution of $(X-v)|(X>v)$ for $v > u$ and $x > 0$, we have:

$$\begin{aligned}\mathbb{P}(X - v > x | X > v) &= \frac{\mathbb{P}(X > v + x)}{\mathbb{P}(X > v)} \\ &= \frac{\mathbb{P}(X > v + x | X > u)\mathbb{P}(X > u)}{\mathbb{P}(X > v | X > u)\mathbb{P}(X > u)}\end{aligned}$$

since $v + x > v > u$. Now, given that $(X-u)|(X>u) \sim \text{GPD}(\sigma_u, \xi)$, we have:

$$\begin{aligned}\frac{\mathbb{P}(X > v + x | X > u)}{\mathbb{P}(X > v | X > u)} &= \frac{\left[1 + \frac{\xi(v+x-u)}{\sigma_u}\right]^{-1/\xi}}{\left[1 + \frac{\xi(v-u)}{\sigma_u}\right]^{-1/\xi}} \\ &= \left[\frac{\tilde{\sigma}_u + \xi(v-u) + \xi x}{\sigma_u + \xi(v-u)}\right]^{-1/\xi} \\ &= \left[1 + \frac{\xi x}{\sigma_u + \xi(v-u)}\right]^{-1/\xi}.\end{aligned}$$

This is the survivor function of a GPD with shape and scale parameters of ξ and $\sigma_u + \xi(v-u)$ respectively. Thus, $(X-v)|(X>v) \sim \text{GPD}(\sigma_v, \xi)$ where $\sigma_v = \sigma_u + \xi(v-u)$.

2.3.2 Parameter stability plots

The most widely used approach for threshold selection relies on visual inspection of the stability of the shape parameter estimates. Figure 2.3.1 provides three examples of parameter stability plots. The results come from three simulated datasets used in Chapter 3, specifically the first simulated sample from each of Cases 1-3. These datasets all have true underlying threshold at $u = 1$, i.e., with GPD above this level, plotted as a vertical green dashed line. The sample sizes of the data analysed in the plots are $n = 1200, 480, 2400$ from left to right. The three datasets were assessed for stability in the shape parameter estimates $\hat{\xi}$ for a grid of candidate threshold choices at sample quantile levels of $0\%, 5\%, \ldots, 95\%$. The first and third plot seem to show approximate stability above the true threshold. However, the increasing uncertainty as the threshold choice is raised, shown by the 95% confidence interval calculated using the delta-method, demonstrates the difficulty in objectively assessing this stability. This difficulty is more evident in the centre plot of Figure 2.3.1 due to the smaller sample size for these data. The smaller sample size, which is certainly not unusually small for an extreme value analysis, results in highly variable parameter estimates across candidate thresholds and a level of uncertainty which makes the visual assessment of stability difficult. These simulated examples are cases where the underlying true threshold should be fairly clear, and yet, it is not a straightforward task to make appropriate inferences from the parameter stability plots.

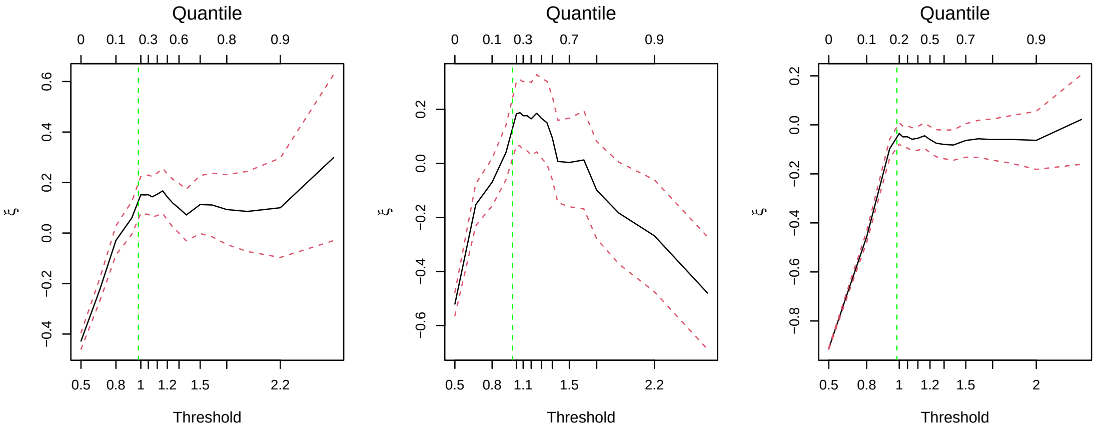

Figure 2.3.1: Examples of parameter stability plots based on three simulated datasets used in simulation study in Chapter 3. Solid lines are point estimates (interpolated), dotted lines are pointwise 95% confidence intervals (interpolated) calculated by the delta-method, vertical dashed line is the true threshold.

The major criticism of the parameter stability plots is their lack of interpretability, since pointwise confidence intervals are highly dependent across the set of candidate thresholds and, thus, are difficult to account for when assessing stability (Wadsworth and Tawn, 2012; Northrop and Coleman, 2014; Wadsworth, 2016). Estimates of the shape parameter and confidence intervals are only evaluated at each candidate threshold choice in the grid, meaning that interpretation of a parameter stability plot and the threshold choice itself can be sensitive to the grid of candidate thresholds.

2.3.3 Existing methods

There have been a variety of methods developed to improve upon the shortcomings of parameter stability plots and more appropriately tackle the bias-variance trade-off. Wadsworth and Tawn (2012) and Northrop and Coleman (2014) utilise penultimate models and hypothesis testing to provide approaches more robust to human subjectivity, see Section 2.3.4 for discussion of Northrop and Coleman (2014). Based on a quantifiable criterion, Bader et al. (2018) and Danielsson et al. (2019) use goodness-of-fit metrics to automate the selection of the threshold, with the former employing the Anderson-Darling test and a stopping rule to control the false-discovery rate of multiple hypothesis tests, the latter is discussed in Section 2.3.5. Wadsworth (2016) automate threshold selection through a sequential changepoint approach, while Northrop et al. (2017) employ a Bayesian procedure with a measure of predictive performance - both methods are discussed in detail in Section 2.3.5.

Mixture-model approaches aim to remove the preceding threshold selection and estimate a model for data lying in the body and tail; Tancredi et al. (2006) compose a model of a piece-wise constant density from a low value up to a threshold above which a GPD is used to model the tail with the threshold estimated as part of the parameter estimation; Naveau et al. (2016) use extreme value models on both the upper and lower tail of rainfall data and allow a smooth transition between the two tails.

Semi-parametric approaches such as Danielsson et al. (2001) and Danielsson et al. (2019) are discussed in Section 2.3.5. Scarrott and MacDonald (2012) and Belzile et al. (2023) review the extensive literature of threshold selection. There is a large body of applied literature with a variety of threshold methods applied to particular data contexts. For the hydrological setting, Durocher et al. (2018) and Curceac et al. (2020) compare several goodness-of-fit approaches for automatic selection of the threshold. Furthermore, Choulakian and Stephens (2001), Li et al. (2005) and Solari et al. (2017) automate goodness-of-fit procedures and apply these techniques to a range of precipitation and river flow data sets.

2.3.4 Northrop and Coleman (2014)

In this section, we describe the Northrop and Coleman (2014) method which aimed to improve upon the subjectivity problems of the parameter stability plot by testing the stability of shape parameter formally through a hypothesis test. We describe some of the inadequacies which both limit the applicability of this method and complicate the interpretation of the associated threshold selection plots.

Northrop and Coleman (2014) test the hypothesis that the underlying shape parameter is constant for any threshold above a selected candidate threshold. They extend the piecewise-constant model for the shape parameter of Wadsworth and Tawn (2012) to allow for an arbitrary number of thresholds and avoid the multiple-testing issue. For their model, they derive likelihood ratio and score methods to test for equality of the shape parameter estimates above a candidate threshold. The method results in a plot of p-values against each candidate threshold u. An example of the plot of p-values obtained using from the Northrop and Coleman (2014) method is given in Figure 2.3.2.

Figure 2.3.2 shows two plots of the p-values derived from the first samples generated in Cases 1 and 2 respectively, described in Chapter 3, with candidate thresholds given at the sample $0\%, 5\% \ldots, 95\%$-quantiles (black points) and the true threshold of $u = 1.0$ (16.67% quantile), plotted as a green dashed line. While the Northrop and Coleman (2014) method was aimed to improve upon the parameter stability plot in terms of interpretability, it still suffers from problems with subjectivity. For example, there is a subjectivity in the choice of whether to select the threshold as the lowest candidate threshold for which the p-value rises above the significance level, of say 0.05, or as the candidate threshold which causes the largest increase in the p-value. In the second plot of Figure 2.3.2, both of these approaches would lead us to choose a threshold at the 20%-quantile which lies near the true value. However, the variability of the p-values above casts doubt on this choice. We want to select a threshold where the p-values indicate strong evidence for parameter stability for all higher thresholds, excluding very high quantiles where uncertainty may become large. In the second plot of Figure 2.3.2, beyond the true threshold, the p-values decrease and remain at relatively low levels until another spike at the 65%-quantile level. Hence, a user of the Northrop and Coleman (2014) method may choose a threshold at the 65%-quantile for this dataset, but even above this value, there is another drop in the p-values which may lead the user to an even higher choice, leaving very few exceedances. While the choice is not clear-cut, the outputted p-values at least could lead the reader to conduct a more detailed investigation of candidate thresholds near the true threshold in this case. The same cannot be said for the first plot, where the p-values give no indication of a good choice of threshold for this sample. The one value which shows a slight rise in the p-value would lead to a threshold choice far from the truth. Whatever the reason for the poor performance in this specific case, while the Northrop and Coleman (2014) method tackles some of the inadequacies of parameter stability plots, it suffers from similar shortcomings to the parameter stability plotting method due to the difficulty of interpretation of the resulting plot of p-values.

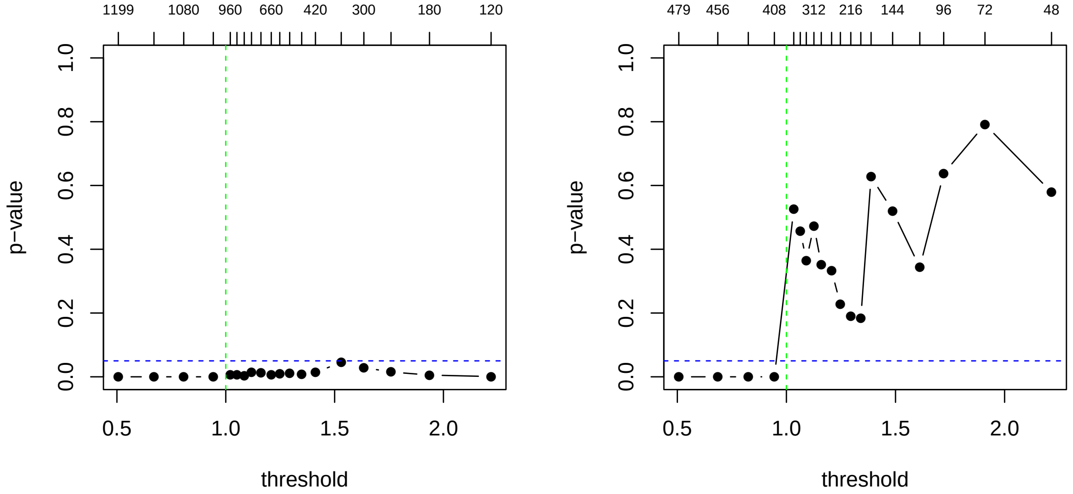

Figure 2.3.2: p-values derived from Northrop and Coleman (2014) method for threshold selection applied to the first simulated samples from Case 1 and 2. Vertical dashed line is the true threshold and the horizontal dashed line shows a p-value of $p = 0.05$. The numbers above the plot correspond to the numbers of exceedances for each candidate threshold.

2.3.5 Core existing automated methods

In this section, we discuss the leading existing methods identified in Belzile et al. (2023) and two semi-parametric approaches, one of which was not mentioned in this review. In an extensive simulation study in Chapter 3, we assess the performance of this set of threshold selection methods by comparison to the method we develop in Chapter 3.

Semi-parametric methods

Here, we discuss the semi-parametric procedures of Danielsson et al. (2001, 2019) utilised for performance comparison in the supplementary material of Chapter 3. Firstly, a key obstacle to the wide use of these methods is that both assume that $\xi > 0$, which is problematic when typically, $\xi < 0$ for a wide range of environmental applications such as wind speeds (Fawcett and Walshaw, 2006), wave heights (Jonathan and Ewans, 2007b), sea levels (D'Arcy et al., 2023) and earthquakes (Yue et al., 2025b). Now, the Danielsson et al. (2001) method selects the threshold by minimising the asymptotic mean squared error (MSE) of the Hill estimator (Hill, 1975) through a double-bootstrap procedure. The first bootstrap stage computes the optimal size n_1 for their second bootstrap stage, where $n_1 < n$ and n is the data sample size. To reduce computations, the *tea* package (Ossberger, 2020) fixes $n_1 = 0.9n$. The reliance on asymptotic theory leads to inadequate finite sample performance. The Danielsson et al. (2019) method picks the threshold to minimise the *maximum* distance between the empirical and modelled quantiles, i.e., the distance from the diagonal of a QQ-plot. As the largest such deviations occur at the highest quantiles and the method fails to account for uncertainty, which changes across candidate thresholds, this method over-estimates the threshold.

Wadsworth (2016)

Here, we detail the procedure of Wadsworth (2016) which is utilised in Chapter 3 for comparison of performance. This method aims to address the subjective nature of the standard parameter stability plots by utilising the asymptotic distributional theory of the joint distribution of maximum likelihood estimators (MLEs) from samples of exceedances over a range of thresholds. By construction, the exceedances of threshold v are a subset of that of any candidate threshold u, whenever $v > u$. Thus, due to this data-overlap, non-standard statistical testing is required, as this induces dependence between estimates at different thresholds. The method outputs more interpretable diagnostic plots to improve standard parameter stability plots, primarily by removing dependence between estimates at different candidate thresholds. A simple likelihood-based testing procedure is suggested to allow automated selection of the threshold.

Wadsworth (2016) used the point process representation of extremes, derived in Pickands (1971), which considers exceedances of a high threshold u as a realisation from a non-homeogeneous Poisson process (NHPP). The representation is outlined as follows. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sequence of independent and identically distributed random variables with common distribution function F. Suppose there exists normalising sequences $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ such that the sequence of point processes $\{P_n : n = 1, 2 \ldots,\}$ defined by

$$P_n = \left\{ \frac{X_i - b_n}{a_n} : i = 1, \ldots, n \right\},$$

has the property that $P_n \xrightarrow{d} \mathcal{P}$ as $n \to \infty$ with $\mathcal{P}$ non-degenerate, on the interval $(b_l = \lim_{n\to\infty}(x_F - b_n)/a_n, \infty)$ where $x_F := \inf\{x : F(x) > 0\}$ is the lower end-point of F. Then, $\mathcal{P}$ is a NHPP with intensity $\lambda_{\boldsymbol{\theta}}(x)$, for $x > b_l$, and integrated intensity $\Lambda_{\boldsymbol{\theta}}(A)$ on $A = (u, \infty)$, with $u > b_l$, where

$$\lambda_{\boldsymbol{\theta}}(x) = \frac{1}{\sigma}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1-1/\xi} \text{ and } \Lambda_{\boldsymbol{\theta}}(x) = \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}.$$

with $\boldsymbol{\theta} = (\mu, \sigma, \xi)$, where $\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R}$ corresponding to the location, scale and shape parameter respectively, with ξ as in the GPD (2.2.4). Hereafter, we let $\boldsymbol{\theta_0}$ denote the true value of $\boldsymbol{\theta}$.

Wadsworth (2016) considers $\boldsymbol{x}_N = (x_1, \ldots, x_N)$ as a realisation from a NHPP with a random count N on some region $R = [u_1, \infty)$. It is assumed that $\boldsymbol{x}_N$ are sorted such that x_i is the i^{th} largest value, i.e., $x_N < \cdots < x_i < \cdots < x_1$. A set of candidate threshold choices $(u_1, \ldots, u_k)$ with $b_l \leq u_1 < u_2 \cdots < u_k$ which define nested regions $R_1, R_2, \ldots, R_k$ in R such that $R_j = (u_j, \infty)$ for $j = 1, \ldots, k$, so $R_k \subset R_{k-1} \subset \cdots \subset R_1 = R$. Thus, if $x_1, \ldots, x_{N_j}$ are all the observations which lie in the region R_j, then, there are N_j observations in the region R_j. The likelihood of the process over the region R_j is then given, up to a constant of proportionality, by:

$$L_{R_j}(\boldsymbol{\theta}) := \left(\prod_{i=1}^{N_j} \lambda_{\boldsymbol{\theta}}(x_i)\right) \exp[-\Lambda_{\boldsymbol{\theta}}(R_j)].$$

Now, we denote the MLE of $\boldsymbol{\theta}$ based only on the data in region R_j by $\hat{\boldsymbol{\theta}}_j$ and the

3×3 Fisher information matrix for this likelihood as I_j with I_j^{-1} its inverse. Wadsworth (2016) considers a superposition of m replicate Poisson processes as $m \to \infty$ giving the limit result:

$$m^{1/2}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0, \ldots, \hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0\right)^T \xrightarrow{d} N_{3k}(\mathbf{0}, \Sigma),$$

with Σ the $3k \times 3k$ covariance matrix given by $\Sigma = (I^{-1}_{\min(i,j)})_{1\le i\le k, 1\le j\le k}$.

Wadsworth (2016) uses the above result to construct a threshold selection procedure. Isolating the shape parameter ξ in the inference, Wadsworth (2016) denote $m^{-1}\{(I_{j+1}^{-1} - I_j^{-1})_{\xi,\xi}\}$ as the asymptotic variance of the estimated increment $\hat{\xi}_j - \hat{\xi}_{j+1}$ where $\hat{\xi}_j$ is the MLE of the shape parameter on the region R_j. As these increments have changing variance with j, consider instead the standardised increments $\hat{\boldsymbol{\xi}}^* = (\hat{\xi}_1^*, \ldots, \hat{\xi}_{k-1}^*)^T$ given by:

$$\left(\hat{\xi}_1^*, \hat{\xi}_2^*, \ldots, \hat{\xi}_{k-1}^*\right)^T := m^{1/2}\left(\frac{\hat{\xi}_1 - \hat{\xi}_2}{((I_2^{-1} - I_1^{-1})_{\xi,\xi})^{1/2}}, \frac{\hat{\xi}_2 - \hat{\xi}_3}{((I_3^{-1} - I_2^{-1})_{\xi,\xi})^{1/2}}, \ldots, \frac{\hat{\xi}_{k-1} - \hat{\xi}_k}{((I_k^{-1} - I_{k-1}^{-1})_{\xi,\xi})^{1/2}}\right)^T, \tag{2.3.1}$$

which have common unit variances over all components. It follows that if the excesses of u_1 follow a GPD then, as $m \to \infty$, $\hat{\boldsymbol{\xi}}^* \to \boldsymbol{Z}$ where $\boldsymbol{Z} \sim N_{k-1}(\mathbf{0}, \mathbf{1}_{k-1})$ with $\mathbf{1}_n$ denoting the $n \times n$-dimensional identity matrix. Given these properties, Wadsworth (2016) term $\hat{\boldsymbol{\xi}}^*$ as a white-noise process.

As a result of the penultimate theory of extremes, described by Smith (1987) and Gomes (1994), Wadsworth (2016) explains that departures from the null assumption of the white-noise process (2.3.1) are a direct consequence of too many values from the body of the data (where the GPD is not appropriate) being included in the estimation. This logic suggests that below the lowest appropriate candidate threshold, say u_j, the variables $\hat{\xi}_i^*$, $i = 1, \ldots, j-1$ might be better approximated by a $N(\beta, \gamma^2)$, where at least one of $\beta \neq 0$ and $\gamma \neq 1$ holds, than by a standard normal distribution which is the limit distribution if the threshold u_j was correct. Formally, this gives a changepoint model as:

$$\xi_i^* \sim N(\beta, \gamma^2) \quad IID, \quad i = 1, \ldots, j-1, \qquad \xi_i^* \sim N(0,1) \quad IID, \quad i = j, \ldots, k-1,$$

where j, β, γ are unknown.

Wadsworth (2016) maximises the profile likelihood for β and γ across j and uses a likelihood ratio test to assess if this gives a significantly better fit to $\hat{\boldsymbol{\xi}}^*$ than the standard normal distribution. A threshold is automatically selected as the candidate threshold u_j which provides the best fit. If there is no evidence of $\hat{\boldsymbol{\xi}}^*$ deviating from white-noise, then the lowest candidate threshold is selected, i.e., u_1.

Northrop et al. (2017)

Now, we detail the procedure of the last approach compared in Chapter 3, the Northrop et al. (2017) method. Consider $\boldsymbol{X} = (X_1, \ldots, X_n)$ where X_i are IID with associated realisations $\boldsymbol{x} = (x_1, \ldots, x_n)$, where $x_1 < \ldots < x_n$. This contrasts with the notation for the Wadsworth (2016) method, however, we keep this to stay aligned with Northrop et al. (2017) in our explanation.

Northrop et al. (2017) consider u as a training threshold in the cross-validation scheme and allow for the threshold exceedance rate, denoted by $\lambda_u = \mathbb{P}(X > u)$, to be incorporated with the GPD parameters (σ_u, ξ) into the fit. Thus, $\boldsymbol{\theta} = (\lambda_u, \sigma_u, \xi)$ and subsequently, in this section, we refer to the tail model as the GPD$(\lambda_u, \sigma_u, \xi)$. Let $\pi_u(\boldsymbol{\theta})$ be a prior density for $\boldsymbol{\theta}$. Let $\boldsymbol{x}_{(-r)} = \{x_i : 1 \leq i \leq n, i \neq r\}$. The posterior density for $\boldsymbol{\theta}$ given data $\boldsymbol{x}_{(-r)}$ is denoted $\pi_u(\boldsymbol{\theta}|\boldsymbol{x}_{(-r)})$ with $\pi_u(\boldsymbol{\theta}|\boldsymbol{x}_{(-r)}) \propto L(\boldsymbol{\theta}; \boldsymbol{x}_{(-r)}, u)\pi_u(\boldsymbol{\theta})$ with the likelihood L assumed to take the form:

$$\begin{gathered} L(\boldsymbol{\theta}; \boldsymbol{x}_{(-r)}, u) = \prod_{i : x_i \in \boldsymbol{x}_{(-r)}} f_u(x_i|\boldsymbol{\theta}), \\ f_u(x|\boldsymbol{\theta}) = (1-\lambda_u)^{I(x \leq u)}[\lambda_u h(x-u; \sigma_u, \xi)]^{I(x>u)}, \end{gathered} \tag{2.3.2}$$

where $I(w)$ is an indicator function giving 1 if w is true and 0 otherwise, and $h(x;\sigma_u,\xi) = \sigma_u^{-1}[1+\xi x/\sigma_u]_+^{-(1+1/\xi)}$ is the density of a GPD tail. In the case of $\lambda_u = 0$, $f_u(x|\boldsymbol{\theta}) = I(x \leq u)$. Note that, as defined, $f_u(x|\boldsymbol{\theta})$ is not a valid density function as it integrates to ∞ and it is discontinuous at $x = u$. The use of the term "density" is identified in Northrop et al. (2017) as an abuse of terminology.

Northrop et al. (2017) aim to compare a set of candidate values for u, denoted $(u_1,\ldots,u_k)$ with $u_1 < \cdots < u_k$, by introducing a fixed validation threshold $v \geq u$ and quantifying the predictive ability of the implied GPD(λ_v,σ_v,ξ) using each candidate threshold $u_i, i = 1,\ldots,k$. They select $v = u_k$. Since v is fixed, the performance of each of the candidate thresholds is compared based on the same validation data.

To undertake comparisons of fit over different candidate thresholds, a slight extension of the threshold stability property, stated in Section 2.3.1, is required, i.e., if a GPD(λ_u,σ_u,ξ) tail model applies at u, this implies a GPD(λ_v,σ_v,ξ) tail model above v where $\sigma_v = \sigma_u + \xi(v-u)$ and $\lambda_v = \lambda_u[1+\xi(v-u)/\sigma_u]^{-1/\xi}$ assuming that v is such that $1+\xi(v-u)/\sigma_u > 0$.

For the cross-validation scheme, the data $\boldsymbol{x}_{(-r)}$ are the training data with x_r the validation data, and this is repeated for each $r = 1,\ldots,n$. To assess the threshold choice performance above v, they use leave-one-out cross-validation. The cross-validation predictive density for exceedances of the validation level v under model (2.3.2), using the candidate threshold u_j, $j = 1,\ldots,k-1$, is then given by:

$$f_v(x_r|\boldsymbol{x}_{(-r)},u_j) = \int f_v(x_r|\boldsymbol{\theta})\pi_{u_j}(\boldsymbol{\theta}|\boldsymbol{x}_{(-r)})\,\mathrm{d}\boldsymbol{\theta}, \quad r = 1,\ldots,n.$$

A Monte Carlo estimator for approximating $f_v(x_r|\boldsymbol{x}_{(-r)},u)$ uses a MCMC generated sample of realisations $\boldsymbol{\theta}_j^{(-r)}, j = 1,\ldots,m$ (where m is a user choice for the run length of the MCMC after convergence has been deemed to have been achieved) from the posterior distribution $\pi_u(\boldsymbol{\theta}|\boldsymbol{x}_{(-r)})$ through:

$$\hat{f}_v(x_r|\boldsymbol{x}_{(-r)}, u) = \frac{1}{m}\sum_{j=1}^{m} f_v(x_r|\boldsymbol{\theta}_j^{(-r)}), \quad r = 1, \ldots, n.$$

This leads to a measure of predictive ability at v given by:

$$\hat{T}_v(u) = \sum_{r=1}^{n} \log\{\hat{f}_v(x_r|\boldsymbol{x}_{(-r)}, u)\},$$

which is evaluated over all candidate thresholds choices of $u_1, \ldots, u_k$. Out of these candidate thresholds, Northrop et al. (2017) select the one which maximises the measure, $\hat{T}_v$.

To improve the computational efficiency of the estimator $\hat{f}_v(x_r|\boldsymbol{x}_{(-r)}, u)$ for $r = 1, \ldots, n$, Northrop et al. (2017) use importance sampling (Gelfand, 1996). This allows for estimation over r using a single sample from the posterior distribution $\pi_u(\boldsymbol{\theta}|\boldsymbol{x})$. Specifically, for a single sample $\{\boldsymbol{\theta}_j, j = 1, \ldots, m\}$ from the posterior $\pi_u(\boldsymbol{\theta}|\boldsymbol{x})$,

$$\hat{f}_v(x_r|\boldsymbol{x}_{(-r)}, u) = \frac{\sum\limits_{j=1}^{m} f_v(x_r|\boldsymbol{\theta}_j) q_r(\boldsymbol{\theta}_j)}{\sum\limits_{j=1}^{m} q_r(\boldsymbol{\theta}_j)} = \frac{\sum\limits_{j=1}^{m} f_v(x_r|\boldsymbol{\theta}_j)/f_u(x_r|\boldsymbol{\theta}_j)}{\sum\limits_{j=1}^{m} 1/f_u(x_r|\boldsymbol{\theta}_j)},$$

by taking $q_r(\boldsymbol{\theta}) = \pi_u(\boldsymbol{\theta}|\boldsymbol{x}_{(-r)})/\pi_u(\boldsymbol{\theta}|\boldsymbol{x}) \propto 1/f_u(x_r|\boldsymbol{\theta})$.

2.4 Non-identically distributed extremes

In this section, we discuss the well-established methods for extreme value modelling of non-identically distributed variables, when such variables vary with covariates under the assumption of independence between observations. The standard extreme value methods discussed so far are not directly applicable in this context as further steps are required to capture the covariate effects. Here, we focus on the extension of the GPD for modelling such processes as this is more commonly used in practice and is the focus throughout this thesis.

Even in the case of an identically distributed process, if we impose a varying threshold, this will result in an excess distribution which is non-identical. Specifically, consider X as an arbitrary variable in the sequence of random variables $X_1, \ldots, X_n$ all with common distribution function. If $(X-u)|(X>u) \sim \text{GPD}(\sigma_u, \xi)$, then, for any set of thresholds $(v_1 \ldots, v_n)$, with $u \leq v_i < x^F$, we have that

$$(X_i - v_i)|(X_i > v_i) \sim \text{GPD}(\sigma_{v_i}, \xi), \text{ with } \sigma_{v_i} = \sigma_u + \xi(v_i - u),$$

for $i = 1, \ldots, n$. This implies that the GPD shape parameter is stable with respect to the varying threshold and the GPD scale can be simply adjusted as before for each value of the varying threshold. This leads to a non-identically distributed excess of v_i, for $i = 1, \ldots, n$, even though the original X variable is identically distributed.

Now, consider a non-identically distributed process, such that $X_1, \ldots, X_n$ now vary according to some covariate variable $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ where $\boldsymbol{Z} = (Z_1, \ldots, Z_q)$. We are interested in modelling the extremes of the conditional variable $X \mid (\boldsymbol{Z} = \boldsymbol{z})$ for observed realisations of the covariates $\boldsymbol{z} \in \mathbb{R}^q$. Davison and Smith (1990) proposed an extension of the standard GPD framework which allowed the GPD scale and shape parameters to vary according to covariates. Kyselý et al. (2010) and Northrop and Jonathan (2011) provide further extensions additionally allowing the threshold to be a function of covariates. If we allow the covariates $\boldsymbol{Z}$ to affect the parameters of the GPD and the threshold through functional parameterisations, we have, for threshold function $u(\boldsymbol{z})$, that:

$$X - u(\boldsymbol{z}) \mid (X > u(\boldsymbol{z}), \boldsymbol{Z} = \boldsymbol{z}) = \text{GPD}(\sigma_u(\boldsymbol{z}), \xi(\boldsymbol{z})),$$

such that, for $x > u(\boldsymbol{z})$,

$$\Pr(X > x \mid \boldsymbol{Z} = \boldsymbol{z}) = 1 - \lambda_u(\boldsymbol{z}) \left(1 + \xi(\boldsymbol{z}) \frac{x - u(\boldsymbol{z})}{\sigma_u(\boldsymbol{z})}\right)_+^{-1/\xi(\boldsymbol{z})},$$

with $\lambda_u(\cdot), \sigma_u(\cdot)$ and $\xi(\cdot)$ the respective covariate-dependent threshold exceedance rate, scale and shape parameters functions. In Davison and Smith (1990), each parameter function $\phi(\cdot) \in \{u, \lambda_u, \sigma_u, \xi\}$ has the form $h(\phi(\boldsymbol{z})) = \boldsymbol{z}^T\boldsymbol{\beta}$ where $\boldsymbol{z}^T$ is the transpose of the covariate vector, $\boldsymbol{\beta} \in \mathbb{R}^q$ is the vector of coefficients, and $h(\cdot)$ is a link function which transforms the feasible parameter space onto that of the linear combination $\boldsymbol{z}^T\boldsymbol{\beta} \in \mathbb{R}$. Link functions are typically chosen to constrain a parameter to the suitable domain, e.g., to ensure positivity of the scale parameter function $\sigma_u(\boldsymbol{z})$ for all $\boldsymbol{z}$, a log-link function can be used such that $\log \sigma_u(\boldsymbol{z}) = \boldsymbol{z}^T\boldsymbol{\beta}$. However, Eastoe and Tawn (2009) show that this choice of link function violates the threshold stability property. In particular, they show that to ensure this property holds, across covariate values, the identity link function is required for the scale parameter.

While all the parameters can be allowed to vary with covariates, it is typical to assume that $\xi(\boldsymbol{z}) = \xi$ is constant across $\boldsymbol{z}$, with ξ still needing to be estimated. This has been shown to be a reasonable assumption across a range of applications (Healy et al., 2025) and avoids incorporating additional uncertainty into the model for a parameter which is already difficult to estimate (Chavez-Demoulin and Davison, 2005). Of course, for any application, adequate checks should be conducted to ensure the assumption of a constant shape parameter is reasonable.

In the above framework, the GPD threshold, whether constant or a function of covariates, is estimated in advance of the GPD parameter function estimation, and then, treated as fixed. As a result, the uncertainty in the threshold choice is not taken into account in the subsequent inference. In Chapters 3 & 5, we develop methods to account for this aspect of uncertainty for IID and non-identical contexts, respectively.

For the above framework, one can either specify a parametric form for each of the functions in advance of fitting or use a non-parametric regression procedure where the covariate effects are estimated by assuming that the parameter function $\phi(\boldsymbol{z})$ is smooth across $\boldsymbol{z}$. Under the parametric framework, the advance choice of functional form can be limiting but should be informed by exploratory analysis of the process of interest. Smith (1989) first used parametric regression techniques within an extreme value analysis. Davison and Smith (1990) take the simplest approach for the parameter function by using linear models, however they keep the threshold constant. Piecewise constant functions provide another simple choice, where parameters are assumed constant over consecutive subsets of the covariate space, for example, (Varty et al., 2021) take piecewise constant forms for the threshold and scale parameter and extend this further to a smooth sigmoid function. Inference procedures in these cases are easily adjusted by replacing constant parameters by their parametric functional forms in the likelihood and estimating coefficients through maximum likelihood estimation. A potential shortcoming of this approach is that, while the choice of parametric form for the parameters may be suitable within the observed range of the data and covariates, this does not imply that this choice will be appropriate for future values.

Chavez-Demoulin and Davison (2005) propose a more flexible framework by capturing relationships with covariates through the use of generalised additive models (GAMs). GAMs are a semi-parametric model consisting of a linear predictor which incorporates a sum of smooth functions of the covariates, with a variety of suitable choices for the smooth functions (Wood, 2017). This provides more flexibility as the parameter functions do not follow a pre-specified functional form. Youngman (2019) provides a framework for fitting GAMs with extreme value models, with a corresponding R package `evgam` (Youngman, 2022).

As mentioned previously, Kyselý et al. (2010) extend the Davison and Smith (1990) framework by allowing the threshold to also be time-dependent at a fixed high quantile, using quantile regression. Northrop and Jonathan (2011) also use quantile regression to estimate a varying threshold across a spatially-dependent covariate range such that the rate of exceedance of the threshold is constant across the range. Quantile regression can be implemented as part of the `evgam` package (Youngman, 2022). Again, uncertainty in the threshold choice is not incorporated in inference here. Once an appropriate quantile is chosen through quantile regression, this is treated as fixed and the GPD parameters estimation uncertainty is propagated through to inference. This is problematic even in the IID context but with non-identical distributions, there is additional uncertainty both in the value of the threshold selected and in the formulation of the threshold function over the covariate range. We explore this aspect of uncertainty in Chapter 5.

Eastoe and Tawn (2009) propose an alternative preprocessing approach to handle non-identically distributed variables. This approach is more in line with the common techniques used for handling time-varying stochastic properties in time series analysis. The key difference in this approach is that the covariate-dependence is modelled for the entire dataset and then removed so that the bulk of the sequence can then be treated as identical. However, although the preprocessing should account for the most complex covariate dependence, this dependence may vary between the body and tail and so, non-identical extreme value models are used in the tail to account for any residual covariate dependence with these models now being simpler in formulation due to the preprocessing step.

There is debate in the literature about whether it makes more sense to model covariate effects in the threshold or model parameters, especially when such effects are small in comparison to the variability of the data (Healy et al., 2025). Model selection techniques which can allow for different functional forms of parameters and thresholds can be useful in this context. This is discussed further in Chapter 5.

2.5 Dependence in extremes

So far and throughout the work in the thesis, we assume the univariate process of interest $\{X_t; t \in \mathbb{Z}\}$ is serially independent. However, in practice, there are cases where this assumption becomes unrealistic and impractical. In this section, we discuss how extreme value methods can be used when the assumption of temporal independence is relaxed and the focus lies on a stationary sequence of random variables $\{X_t : t \in \mathbb{Z}\}$. In many applications, this is a more realistic assumption than IID as it implies that variables may be serially dependent but that the stochastic properties of the process do not change over time (Coles, 2001). More specifically, $\{X_t; t \in \mathbb{Z}\}$ is a stationary process if the joint probability density function f of any set of values in the series is the same as if they were all shifted in time by the same lag $\tau \in \mathbb{Z}$, so that $f_{X_{i_1},\ldots,X_{i_n}}(x_{i_1},\ldots,x_{i_n}) = f_{X_{i_1+\tau},\ldots,X_{i_n+\tau}}(x_{i_1},\ldots,x_{i_n})$ for all n, for any subsequence $(i_1,\ldots,i_n) \in \mathbb{N}^n$, with $i_1 \leq \cdots \leq i_n$, and for any value $(x_{i_1},\ldots,x_{i_n}) \in \mathbb{R}^n$ (Chatfield, 2013).

When accounting for dependence in extremes, the usual approach is to assume that over a long-range, the strongest possible dependence of extreme events is very near-independent and then, focus on how to account for the effects of short-range dependence. Leadbetter et al. (2012) provide a detailed development of the required long-range condition and dependence modelling in extremes. Specifically, this condition is termed the $D(u_n)$ condition, with $u_n = a_n x + b_n$, for $x \in \mathbb{R}$, where $\{a_n > 0\}$ and $\{b_n\}$ are the sequences used to normalise the maximum to give a non-degenerate limit (2.2.1) of IID random variables with the same marginal distribution as X_t. A series is said to satisfy this condition if, for all $i_1 < \cdots < i_p < j_1 < \cdots < j_q$ and $j_1 - i_p > l$,

$$\begin{aligned} &|\Pr(X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n) - \\ &\Pr(X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n) \Pr(X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n) | \leq \alpha(n, l), \end{aligned}$$

where $\alpha(n, l_n) \to 0$ for some sequences l_n such that $l_n/n \to 0$ as $n \to \infty$. This condition allows the same extremal type limit laws (see Section 2.2.1 & 2.2.2) for independent series to be applied to the maxima of a stationary series, however the parameters (other than the shape) of the limit distribution will be affected by the dependence (Coles, 2001).

While the GPD is still appropriate for modelling the marginal distribution of exceedances in this case, the dependence in the series makes using the usual product likelihood for inference for the joint modelling of neighbouring threshold exceedances inappropriate, as it violates the IID assumption. The typical approach for accounting for the dependence of threshold exceedances is to decluster. This requires the use of techniques to filter out dependent realisations from the series and obtain a set of near-independent threshold exceedances. Following the spirit of the $D(u_n)$ condition, it is typically assumed that clusters of short-range dependent exceedances separated by a significant gap are independent in the limit and so, individual values (such as the cluster maxima) extracted from clusters are also considered independent as the threshold gets sufficiently large. Leadbetter (1991) showed that under the long-range near-independence assumption, cluster maxima can be modelled using a GPD. As a result, the standard approach when modelling a stationary sequence using the GPD is to identify independent clusters of threshold exceedances and take the maxima within each cluster as an IID sample of exceedances and apply the classic POT methodology (Davison and Smith, 1990).

A key quantity used to account for dependence in extreme observations is the extremal index θ, with $0 < \theta \leq 1$ (Leadbetter et al., 2012), which can be loosely defined as the reciprocal of the limiting mean cluster size of the exceedances, with limiting here meaning increasingly high thresholds. Values of the extremal index provide information about the dependence at asymptotically high levels and so, care must be taken when interpreting values of this quantity for extreme but relevant levels of a dataset.

There are a variety of methods available to identify clusters. Walshaw (1994) develops an approach that selects the most appropriate combination of the threshold and separation length between clusters. Smith and Weissman (1994) develop the runs estimator for the extremal index which assumes that neighbouring exceedances, separated by less than a specified number of consecutive below-threshold observations (known as the run length), belong to the same cluster. The choice of the run length is complex and subjective, a major shortcoming of this approach. Ferro and Segers (2003) define the intervals estimator for the extremal index and propose an automatic declustering scheme which chooses the run length using the limiting distribution for the inter-exceedance times of stationary sequences. Fawcett and Walshaw (2007) ignore the cluster identification problem and instead, treat stationary sequences as independent, making an adjustment to inflate the standard errors of the parameter estimates to represent the true uncertainty more accurately.

Ledford and Tawn (2003) develop a measure of extremal dependence in a series as a diagnostic for whether values separated by a certain lag τ are asymptotically dependent or asymptotically independent. Loosely, X_t and $X_{t+\tau}$ are asymptotically dependent if given X_t is extreme, there is a positive probability that $X_{t+\tau}$ is extreme. Perfect extremal dependence occurs if this probability is one and when this probability is zero, we say X_t and $X_{t+\tau}$ are asymptotically independent. Extremal dependence is entirely separate to generic dependence, a series can be asymptotically independent with strong correlation (bivariate Gaussian with large ρ is a classic example) or exhibit asymptotic dependence and show little correlation. Thus, using diagnostics such as that of Ledford and Tawn (2003) to identify a process as asymptotically dependent or asymptotically independent is vital as it provides information about whether we can expect clustering in extreme values of a series.

The methods mentioned so far aim to identify clustering, account for it and apply standard extreme value methods. Other approaches aim to model the dependence structure of the series explicitly, with the focus being to model within-cluster dependence. Smith et al. (1997) propose a general framework for modelling a dependent series by assuming the series is a stationary first-order Markov chain and applying bivariate extreme value methods, with their methods only applicable if the series exhibits asymptotic dependence. Winter and Tawn (2017) use time series copula methods from conditional multivariate extreme value theory to apply a k^{th}-order Markov model to extreme heatwaves, with their methods covering both asymptotic dependence and asymptotic independence cases. Eastoe and Tawn (2012) utilise subasymptotic theory of extremes of a stationary series to compose a model for the distribution of cluster maxima which comprises two components; the marginal distribution of exceedances and the dependence structure of within-cluster exceedances using the Heffernan and Tawn (2004) and Ledford and Tawn (1996) frameworks. This approach avoids the need to make specific Markov assumptions and also allows asymptotic dependence at some lags and asymptotic independence at other lags within the cluster.

Chapter 3

Automated threshold selection and associated inference uncertainty for univariate extremes

3.1 Introduction

An inherent challenge in risk modelling is the estimation of high quantiles, known as *return levels*, beyond observed values. Such inference is important for designing policies or protections against future extreme events, e.g., in finance or hydrology (Smith, 2003; Coles et al., 2003). Extreme value methods achieve this extrapolation by using asymptotically exact models to approximate the tail of a distribution above a high, within-sample, threshold u. The choice of threshold is fundamental in providing meaningful inference. Here, we develop novel methods for automatic selection of the threshold and for propagating the uncertainty in this selection into return level inferences.

Throughout, we assume that all data are realisations of an independent and identically-distributed (IID) univariate continuous random variable X with unknown distribution function F, with upper endpoint $x^F := \sup\{x : F(x) < 1\}$. Under weak conditions, Pickands (1975) shows that for $X > u$, with $u < x^F$, the distribution of the rescaled excess $Y = X - u$, converges to the generalised Pareto distribution (GPD) as $u \to x^F$. To use this limit result in practice, a within-sample threshold u is chosen, above which this limit result is treated as exact. Specifically, whatever the form of F, the excesses Y of u are modelled by the single flexible GPD(σ_u, ξ) family, with distribution function

$$H(y; \sigma_u, \xi) = 1 - (1 + \xi y/\sigma_u)_+^{-1/\xi}, \tag{3.1.1}$$

with $y > 0$, $w_+ = \max(w, 0)$, $(\sigma_u, \xi) \in \mathbb{R}_+ \times \mathbb{R}$ being scale and shape parameters. The exponential distribution arises when $\xi = 0$, i.e., as $\xi \to 0$ in distribution (3.1.1), whereas for $\xi > 0$, the distribution tail decay is polynomial. For $\xi < 0$, X has a finite upper end-point at $u - \sigma_u/\xi$ but is unbounded above for $\xi \geq 0$. To estimate the $(1-p)^{\text{th}}$ quantile, x_p, of X, for $p < \lambda_u := \mathbb{P}(X > u)$, we can solve $\hat{F}(x_p) = 1 - p$, where $\hat{F}(x_p) = 1 - \hat{\lambda}_u[1 - H(x_p - u; \hat{\sigma}_u, \hat{\xi})]$, $\hat{\lambda}_u$, the MLE of the threshold-exceedance rate parameter, is the proportion of the realisations of X exceeding u and $(\hat{\sigma}_u, \hat{\xi})$ are maximum likelihood estimates (MLEs) obtained by using realisations of the threshold excesses. Davison and Smith (1990) overview the properties of the GPD.

Threshold selection involves a bias-variance trade-off: too low a threshold is likely to violate the asymptotic basis of the GPD, leading to bias, whilst too high a threshold results in very few threshold excesses with which to fit the model, leading to large parameter and return level uncertainty. Thus, we must choose as low a threshold as possible subject to the GPD providing a reasonable fit to the data. There are a wide variety of methods aiming to tackle this problem (Scarrott and MacDonald, 2012; Belzile et al., 2023) with the most commonly used methods suffering from subjectivity and sensitivity to tuning parameters.

A novel automated approach to threshold selection is introduced by Varty et al. (2021) specifically for modelling large, human-induced earthquakes. These data are complex due to improvements in measurement equipment over time. The major implication of such change is that data are missing-not-at-random, with the dataset appearing to be realisations of a non-identically distributed variable, requiring a threshold $u(t)$ which varies with time t, even though the underlying process is believed to be identically distributed over t. Since excesses of $u(t)$ do not have the same GPD parameters over time, Varty et al. (2021) transform these to a common standard exponential distribution via the probability integral transform, using estimates of $(u(t), \sigma_{u(t)}, \xi)$. They then quantify the model fit using a metric based on a QQ-plot and select a time-varying threshold that optimizes this metric. The key novel aspect of their assessment is the use of bootstrapping methods in the metric evaluation which fully accounts for the uncertainty in the GPD fit, which varies across threshold choices.

Due to the lack of existing threshold selection methods designed for the context of Varty et al. (2021), that paper focuses on the data analysis rather than investigating the performance of the threshold selection method. We explore how their ideas can be best adapted to threshold selection in a univariate, IID data context. We find that a variant of the Varty et al. (2021) metric improves the performance and leads to substantially better results than existing automated methods, including greater stability with respect to tuning parameters.

We differ from Varty et al. (2021) as we study both threshold selection and return level estimation when the truth is known. We also address an entirely different problem of how to incorporate the uncertainty resulting from threshold selection into return level estimation. Existing methods typically treat the threshold, once it has been selected, as known, for subsequent return level inference. The available data above candidate threshold choices are often few and so inference can be highly sensitive to the chosen threshold. Reliance on a single threshold leads to poor calibration of estimation uncertainty and as a result, can mislead inference. In particular, we show that the resulting confidence intervals for such an approach considerably under-estimate the intended coverage. We propose a novel and simple method, based on a double-bootstrap procedure, that incorporates the uncertainty in the selected threshold during inference. We show that the coverage probabilities of confidence intervals from our approach are close to the required nominal levels, thus ensuring our inferences provide meaningful information for design policies.

Ultimately, our aim is to provide a threshold selection method that does not require any user decisions to achieve adequate results. The method should not be sensitive to the choice of candidate threshold grid, it should not require the estimation of a mode to select this grid, it should not have a limit on the number of candidate thresholds for a given sample size, nor should it exclude the possibility that the available data have been pre-processed, such as containing only the exceedances of some arbitrary level.

In Section 3.2, we illustrate problems with threshold selection and outline existing strategies. Section 3.3 describes the core existing automated methods while Section 3.4 introduces our procedure for the selection of a threshold, contrasting it with that of Varty et al. (2021). Section 3.5 presents our proposed method for incorporating threshold uncertainty into return level inference. In Section 3.6, the proposed methods are compared against existing methods on simulated data. In Section 3.7, we apply our methodology to the widely studied troublesome dataset of the River Nidd, first analysed by Davison and Smith (1990).

3.2 Background

The *threshold stability property* of the GPD is key in many threshold selection approaches: if excesses of a threshold u follow a GPD then excesses of a higher threshold v ($u < v < x^F$) will also follow a GPD, with adjusted parameter values, i.e., if $X - u|(X > u) \sim \text{GPD}(\sigma_u, \xi)$, then $X - v|(X > v) \sim \text{GPD}(\sigma_u + \xi(v - u), \xi)$, see Section 2.3.1. By this property, the GPD shape parameter ξ should have the same value for all valid choices of threshold. A modelling threshold can be selected as the lowest value for which this property holds, accounting for the sampling variability in the estimates of ξ. The conventional method for this assessment is known as a *parameter stability plot* (Coles, 2001). This plot displays the estimates of ξ and their associated confidence intervals (CIs) for a set of candidate thresholds. The threshold is selected as the lowest value for which the estimate of ξ for that level is consistent with estimates of ξ at all higher thresholds. Throughout the paper, we use maximum likelihood estimation and parametric bootstrap-based CIs.

Figure 3.2.1 shows two parameter stability plots, with the left plot for a simulated dataset of 1000 values generated from the Case 4 distribution, described in Section 3.6, where excesses of the threshold $u = 1.0$ follow a GPD(0.6, 0.1); and the right plot for 154 measurements from the River Nidd. Each plot has 95% CIs of two types; the delta method and the bootstrap. Profile log-likelihood based CIs were also evaluated but resulted in very similar intervals to the bootstrap method, so they were omitted. The delta method gives narrower CIs, though close to the bootstrap intervals for the larger dataset. Selecting an appropriate threshold using this method is challenging and subjective as the parameter estimates are dependent across threshold choices, there is a high level of uncertainty due to the small sample sizes that characterise extreme value analyses, and the uncertainty increases with threshold choice.

For the Case 4 data, the plot shows that candidate thresholds above (below) 0.3 are possibly appropriate (not appropriate) as CIs for higher candidate thresholds include (exclude) the corresponding shape parameter estimates, and above 0.8 the point estimates appear more stable. Here $(u, \xi) = (1, 0.1)$, so we can see that candidates below 0.3 are not suitable as ξ is outside their CIs, but the true threshold is higher than may be selected using this plot. For the River Nidd, lower candidate threshold values imply a very heavy-tailed distribution ($\hat{\xi} \approx 0.5$), whilst high candidate thresholds imply a very short tail, with estimates ($\hat{\xi} \approx -0.5$). As a result of this unusual behaviour, the Nidd data has become the primary example for non-trivial threshold selection (Davison and

Smith, 1990; Northrop and Coleman, 2014). We apply our new method to this dataset in Section 3.7. Further examples of the problems encountered when using parameter stability plots are given in Section 2.3.2.

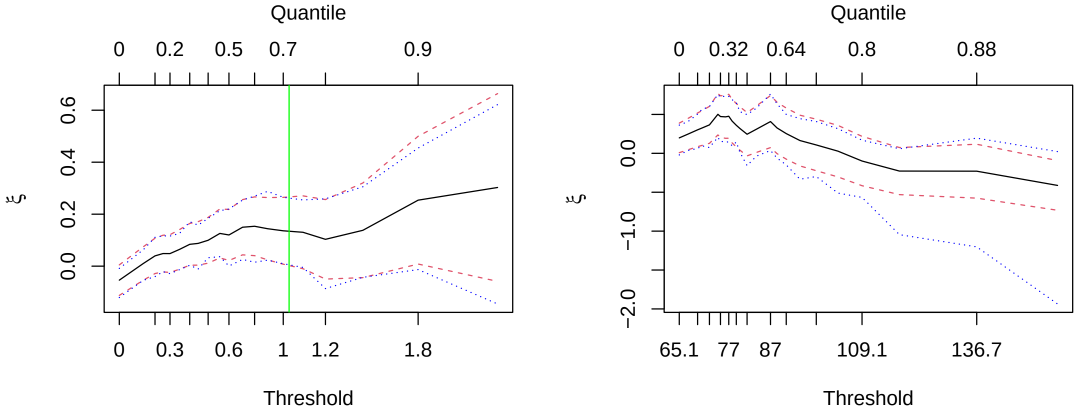

Figure 3.2.1: Examples of parameter stability plots with pointwise CIs using the delta-method [dashed] and bootstrapping [dotted] for [left] a simulated dataset with true threshold $u = 1.0$ following Case 4 distribution [green-vertical] and [right] the River Nidd dataset.

Scarrott and MacDonald (2012) and Belzile et al. (2023) review the extensive literature that aims to improve upon parameter stability plots. The latter characterises these methods, with a core reference, as follows: penultimate models (Northrop and Coleman, 2014), goodness-of fit diagnostics (Bader et al., 2018), sequential-changepoint approaches (Wadsworth, 2016), predictive performance (Northrop et al., 2017), and mixture models (Naveau et al., 2016). It also discusses semi-parametric inferences (Danielsson et al., 2001), but it excludes the development by Danielsson et al. (2019), with similarities to the goodness-of-fit approaches.

In Section 3.3, we outline the key aspects of the core automated approaches with which we compare our proposed method. Supplementary material 2.3.4 and 2.3.5 describe Northrop and Coleman (2014) and Danielsson et al. (2001, 2019) respectively, finding that the former suffers from subjectivity of interpretation similar to the parameter stability plots. We do not describe any mixture methods in this paper as although they benefit from accounting for threshold uncertainty, their inferences are strongly dependent on the choice of model for below the threshold, which we feel is inconsistent to the strategy of extreme value modelling and can induce bias in the threshold selection and subsequent quantile estimation.

3.3 Existing automated methods

Automated threshold selection methods aim to remove subjectivity from the choice of threshold by selecting an optimal threshold from a set of user-defined candidate thresholds based on optimising some criterion. We outline and compare the approaches of Wadsworth (2016) and Northrop et al. (2017), which we find to perform best of the considered existing methods. Further details of these methods are given in Section 2.3.3.

Wadsworth (2016) addresses the dependence between MLEs of ξ, denoted by $\hat{\boldsymbol{\xi}}$, over candidate thresholds. Using asymptotic theory for the joint distribution of MLEs from overlapping samples of data, $\hat{\boldsymbol{\xi}}$ are transformed to the vector $\hat{\boldsymbol{\xi}}^*$ of normalised increments between successive $\hat{\boldsymbol{\xi}}$ values. For GPD data, asymptotically, $\hat{\boldsymbol{\xi}}^*$ would be IID realisations from a standard normal distribution, whereas if the data above any candidate threshold were not from a GPD, the associated elements of $\hat{\boldsymbol{\xi}}^*$ would be better approximated by a non-standard normal. This changepoint behaviour is used to select the threshold. The underlying asymptotic arguments can cause considerable threshold sensitivity and the failure of the method to converge. Both issues are exacerbated by small samples and we identify systematic failures of the associated open source software when $\xi < 0$. To reduce such problems, Wadsworth (2016, Table 1) provides guidance on the number of candidate thresholds for a given sample size.

Northrop et al. (2017) model data using the binomial-GPD (BGPD) model, which is GPD above u, with $\lambda_u = \mathbb{P}(X > u)$ a model parameter, and an improper uniform density, of value $1 - \lambda_u$, below u. They use Bayesian inference and, for each candidate threshold, assess the predictive density of GPD fits above a fixed validation threshold v, where v is the largest candidate threshold. The selected threshold maximises the predictive ability of this model, above v, using leave-one-out cross-validation. The method is sensitive to the validation and candidate threshold set and to the prior joint density of the BGPD parameters.

3.4 Novel metric-based constant threshold selection

3.4.1 Metric choice

We propose an adaptation of the Varty et al. (2021) approach to identify the threshold u for which the sample excesses, arising from IID and non-missing realisations of a continuous random variable, are most consistent with a GPD model. Both methods use a QQ-plot-based metric to approximate the integrated absolute error (IAE) between the quantiles of the model and the data-generating process. Our method, the *expected quantile discrepancy* (EQD), uses the data on the original scale. In contrast, the method of Varty et al. (2021) transforms the data to an Exponential(1) marginal scale and will be termed the *Varty method*. This transformation is beneficial for assessment of non-identically distributed variables but we assess its merit in the IID case in this chapter. We revisit this transformation for non-identical settings in Chapter 5.

The following makes the difference between the two methods precise. Let $\boldsymbol{x}_u = (x_1, \ldots, x_{n_u})$ be the sample of n_u excesses of candidate threshold u and $\boldsymbol{q} = \{q_i = (i-1)/(n_u-1) : i = 1, \ldots, n_u\}$ be the vector of probability plotting points corresponding to the sample size of $\boldsymbol{x}_u$. The sample quantile function $Q(\cdot; \boldsymbol{x}_u, \boldsymbol{q}) : [0, 1] \to \mathbb{R}^+$ is defined as the linear interpolations of the points $\left\{(q_i, x_u^{(i)}) : i = 1, \ldots, n_u\right\}$, with $x_u^{(i)}$ denoting the i^{th} order statistic of $\boldsymbol{x}_u$ (increasing with i), where any ties are handled similarly through linear interpolation. The transformation to Exponential(1) margins is defined by $T(x; \sigma, \xi) = F_{\text{Exp}}^{-1}\{H(x; \sigma, \xi)\}$ where F_{Exp}^{-1} is the inverse distribution function of an Exponential(1) variable, H is the GPD function (3.1.1), and let $T(\boldsymbol{x}_u; \sigma_u, \xi) = \{T(x_1; \sigma_u, \xi), \ldots, T(x_{n_u}; \sigma_u, \xi)\}$. To incorporate the effect of sampling variability in the data into the threshold choice, the expected (average) deviation over the QQ-plot, calculated for the probabilities $\{p_j = j/(m+1) : j = 1, \ldots, m\}$, is calculated across bootstrapped samples of $\boldsymbol{x}_u$, denoted $\boldsymbol{x}_u^b$ for the b^{th} bootstrap sample, $b = 1, \ldots, B$. For both methods, this results in the overall measure of fit $\hat{d}_E(u) = \sum_{b=1}^{B} d_b(u)/B$, where

$$d_b(u) = \begin{cases} \frac{1}{m}\sum\limits_{j=1}^{m} \mid \frac{\hat{\sigma}_u^b}{\hat{\xi}_u^b}[(1-p_j)^{-\hat{\xi}_u^b} - 1] - Q(p_j; \boldsymbol{x}_u^b, \boldsymbol{q}) \mid & \text{EQD} \\ \frac{1}{m}\sum\limits_{j=1}^{m} \mid -\log(1-p_j) - Q(p_j; \hat{T}(\boldsymbol{x}_u^b; \hat{\sigma}_u^b, \hat{\xi}_u^b), \boldsymbol{q}) \mid & \text{Varty,} \end{cases} \tag{3.4.1}$$

and $(\hat{\sigma}_u^b, \hat{\xi}_u^b)$ are the estimated GPD parameters fitted to the bootstrapped sample $\boldsymbol{x}_u^b$. The selected threshold minimises the estimated IAE, $\hat{d}_E$, over a set of candidate thresholds. In Sections 3.4.2 and 3.4.3 respectively, we justify the choices made in the formulation of the EQD metric and discuss our recommendation for default values for the tuning parameters (B, m).

In supplementary material A.3.2, we compare the EQD and Varty methods through an extensive simulation study to assess which version of metric (3.4.1) performs better for threshold selection and quantile estimation. For threshold selection, the methods perform similarly; each method achieves the smallest root-mean-square error (RMSE) in two of Cases 1-4, discussed in Section 3.6. However, for the estimation of high quantiles, the EQD outperforms the Varty method obtaining the lowest RMSE in the majority of cases and quantiles, due to the smaller variance of estimates. We ultimately aim to estimate high quantiles accurately following threshold selection. Given that this study indicates that the EQD should be preferred for this aim and to avoid unnecessary repetition, we omit the results for the Varty method for the remainder of the chapter.

3.4.2 Investigation of the EQD metric choice

For a given u, $d_b(u)$ evaluates the mean-absolute deviation between the b^{th} bootstrap sample quantiles and the fitted model-based quantiles, i.e., the mean-absolute deviation from the line of equality in a QQ-plot for that particular bootstrap sample. This type of assessmen is not radical, as for any observed sample data, QQ-plots are a standard method of assessing model fit (Coles, 2001). The novelty for assessing the validity of a candidate threshold u comes from the way that the EQD metric is constructed.

There are a number of novel choices which we have made in the EQD metric that require justification, in particular; the use of the mean-absolute deviation; the choice of quantiles and their interpolation in the QQ-plot; the use of bootstrap samples; and that the observed data are not explicitly used in the metric. We examine each of these features in the supplementary material, through simulation studies involving the case studies of Section 3.6. For each feature, we find positive evidence for our selections. Below, we explain why we made these choices and outline how they performed relative to other alternative formulations.

We focus on the mean-absolute deviation on the QQ scale as Varty et al. (2021) found that this was more effective than using the mean-squared deviation on that scale and either metric on the PP scale. Our simulation studies found this to be a more robust measure of fit than the maximum deviation proposed by Danielsson et al. (2019).

We choose to take $\{p_j\}$ to be equally-spaced and to weight contributions to $d_b(u)$ equally across the corresponding quantiles. Although higher (lower) sample quantiles exhibit greater (less) sampling variability, equal weighting is appropriate when taking the $\{p_j\}$ values to be equally-spaced because for any $\xi > -1$, the GPD density is monotonically decreasing. This leads to dense evaluation for lower sample quantiles and more sparse evaluation in the upper tail. The choice of equal weighting on this scale is motivated and supported by empirical evidence in Varty et al. (2021). Our choice for p_j is based on the expression for plotting points in a QQ-plot assessment (Coles,

2001) and the choice for q_i is the default option for the R `quantile` function. As these choices are subjective, we also consider alternative definitions but find that there is no systematic ordering of the performance over these definitions and any differences in RMSE for the thresholds selected by the EQD are minimal, especially when compared to the differences between the EQD and existing methods. However, the effect on smaller samples, e.g. $n < 100$, could be explored. Furthermore, there are approaches for choosing optimal plotting positions which also could have been considered, e.g., Jones (1997).

The average over bootstrapped samples in metric (3.4.1) is not a standard use of bootstrapping in the extreme value community, i.e., we utilise the bootstraps in a measure of fit rather than to describe the uncertainty in some estimated quantity. However, the approach of bootstrap aggregating is often used in machine learning classification algorithms and regression trees to reduce variance and overfitting (Breiman, 1996). Our aim is to account for the sampling variability in the observed data, thus avoiding overfitting of the GPD model to the observed dataset which could lead to higher threshold choices than necessary, reduced numbers of exceedances, and extra uncertainty in parameter and quantile estimates. To confirm this, we considered using only the observed sample values in the metric. This leads to higher and more variable thresholds choices in a variety of cases and an overall performance which is either noticeably worse or at best, comparable to our approach.

One may also be concerned that $\boldsymbol{x}_u$ is not included directly in metric (3.4.1). We additionally explored the effect of using $Q(p_j; \boldsymbol{x}_u, \boldsymbol{q})$ instead of $Q(p_j; \boldsymbol{x}_u^b, \boldsymbol{q})$ within the EQD metric, despite it being unconventional to compare sample quantiles to those of a model fitted using a different (bootstrapped) sample. We found no benefit to doing so. Moreover, using only $\boldsymbol{x}_u$ to estimate the IAE ignores that this estimate would change for another realisation of the same data generating process and that variability in this estimate increases with u. Our approach utilises the bootstrap resamples in the measure of fit to provide more stability in the threshold choice and allow us to account for the increasingly uncertain parameter and quantile estimates as the threshold increases.

3.4.3 Choice of tuning parameters

An in-depth study in supplementary material A.3.3 demonstrates that the EQD method is robust to the choice of the tuning parameters B and m. Consequently, we take $(B, m) = (100, 500)$ throughout the paper and in the supplementary material, unless stated otherwise.

The number of bootstrapped samples B controls the level of sampling variability incorporated into the threshold choice and so we expect higher values of B to lead to more stable threshold choices. The RMSE values for threshold estimation reflect this but also show that computation time increases linearly with B. For a one-off analysis, there is certainly merit in taking as large a value for B that is computationally feasible. For simulation studies, when the computational implications of the choice of B are more important, we find that $B = 100$ balances accuracy and computation time sufficiently.

The tuning parameter m gives the number of equally-spaced evaluation probabilities used in expression (3.4.1). The EQD metric aims to approximate the IAE between model quantiles and quantiles of the data generating process (i.e., not for a particular sample) and a larger choice of m improves this approximation. To compare fairly across a range of candidate thresholds, we choose to keep the quality of the approximation of the IAE fixed across thresholds and bootstraps by fixing the number m of points in the quantile interpolation grid.

For a particular bootstrap sample, this choice of fixed m can lead to under- or over-sampling of the upper tail depending on whether $m < n_u$ or $m > n_u$. We explore the sensitivity of the EQD method to m with $m = cn$ and $m = cn_u$, with $c = 0.5, 1, 2, 10$. For both strategies, we find that increasing m beyond 500 essentially wastes the increased computation time as the RMSE values for threshold estimates showed little sensitivity to m. We also explore the effect of the interpolation grid on the sampling distribution of $d_b(u)$ values, over different thresholds, when evaluated using $m = 500$ or $m = n_u$. We find that there is little effect from the choice of interpolation grid outside of the very highest candidate thresholds, but these differences have no effect on the selected threshold in our examples. We conclude that $m = 500$, is suitable as a default value in practice but we can see merits in also ensuring that $m \geq \max_u(n_u)$, where the maximisation is across all candidate thresholds.

3.5 Parameter and threshold uncertainty

Even if the true threshold u is known, relying on point estimates for the GPD parameters results in misleading inference (Coles and Pericchi, 2003). CIs are needed, but as standard errors and profile likelihoods rely on asymptotic arguments, they are not ideal due to the sparsity of threshold exceedances. We prefer parametric bootstrap methods which, as discussed in Section 3.2, perform similarly to the profile likelihood for large samples. Algorithm 1 details the bootstrapping procedure to account for GPD parameter uncertainty when u is known. A GPD is fitted to the n_u data excesses of u from a sample $\boldsymbol{x}$ of size n ($n \geq n_u$). Using the fitted parameters, B_1 parametric bootstrap samples above u are simulated, each of size n_u, and the GPD is re-estimated for each sample. A summary statistic, e.g., a return level, $s(u, \lambda_u, \sigma_u, \xi)$, may be computed for each of the B_1 bootstrapped values for (σ_u, ξ). This enables the construction of CIs for the GPD parameters and return levels.

Algorithm 1 focuses on the uncertainty of the estimates of (σ_u, ξ) and is the typical approach for uncertainty quantification in the frequentist literature (along with the use of asymptotic likelihood methods), ignoring the uncertainty in the threshold exceedance rate parameter. We incorporate the additional uncertainty in the estimation of λ_u by replacing the fixed n_u in the loop over b with a random variate from a $\text{Bin}(n, \hat{\lambda}_u)$

Algorithm 1 Parameter uncertainty for known threshold

```
Require: $(\boldsymbol{x}, u, B_1)$
  Find $n_u = \#\{i : x_i > u\}$, set $\hat{\lambda}_u = n_u/n$, and fit a GPD to $\boldsymbol{x}$ data above $u$ to give $(\hat{\sigma}_u, \hat{\xi}_u)$.
  for $b = 1, \ldots, B_1$ do
      Simulate sample $\boldsymbol{y}_u^b$ consisting of $n_u$ excesses of $u$ from GPD$(\hat{\sigma}_u, \hat{\xi}_u)$.
      Obtain parameter estimates $(\hat{\sigma}_b, \hat{\xi}_b)$ for $\boldsymbol{y}_u^b$ and summary of interest $s(u, \hat{\lambda}_u, \hat{\sigma}_b, \hat{\xi}_b)$.
  end for
  return A set of $B_1$ bootstrapped estimates for the summary statistic of interest.
```

distribution for each bootstrap sample, with this extension then referred to as Algorithm 1b.

GPD inferences are sensitive to the choice of threshold (Davison and Smith, 1990) but uncertainty about this choice is not represented in Algorithms 1 or 1b. This omission would be important when return levels inform the design of hazard protection mechanisms, where omitting this source of uncertainty could lead to over-confidence in the inference and have dangerous consequences. Algorithm 2 provides a novel method to propagate both threshold and parameter uncertainty through to return level estimation, using a double-bootstrap procedure. To focus on the threshold uncertainty and to forgo the need for a parametric model below the threshold, we employ a non-parametric bootstrap procedure on the original dataset. We resample with replacement n values from the observed data B_2 times, estimate a threshold for each such bootstrap sample using the automated selection method of Section 3.4, and fit a GPD to the excesses of this threshold. For each one of the B_2 samples, we employ Algorithm 1 to account for the subsequent uncertainty in the GPD parameters. Calculating a summary statistic for each of the $B_1 \times B_2$ samples leads to a distribution of bootstrapped estimates that accounts for uncertainty in the threshold selection as well as in the GPD and threshold exceedance rate parameters. We use $B_1 = B_2 = 200$. To run this algorithm using the EQD method for the threshold selection step (which itself has B bootstraps), it would require $B_2(B + B_1)$ bootstrap samples to be generated. Specifically, for the B_2 samples initially generated for Algorithm 2, we have $B_2 \times B$ in selecting the threshold values and

$B_2 \times B_1$ in capturing the GPD parameter uncertainty above these selected thresholds. In Section 3.6, we illustrate how using Algorithm 2 improves the coverage probability of CIs, and in Section 3.7 how it widens CIs for return levels of the River Nidd.

Algorithm 2 Parameter uncertainty for unknown threshold

```
Require: $(\boldsymbol{x}, n, B_2, B_1)$
  for $b = 1, \ldots, B_2$ do
      Obtain sample $\boldsymbol{x}_b$ of size $n$ by sampling $n$ times with replacement from $\boldsymbol{x}$.
      Estimate threshold $\hat{u}_b$ for $\boldsymbol{x}_b$ and record number of excesses as $n_{\hat{u}_b}$.
      Employ Algorithm 1 with inputs: $(\boldsymbol{x}_b, \hat{u}_b, B_1)$.
  end for
  return A set of $B_1 \times B_2$ bootstrapped estimates for the summary statistic of interest.
```

3.6 Simulation study

3.6.1 Overview

We illustrate the performance of the EQD method against the Wadsworth (2016) and Northrop et al. (2017) approaches, which we term the *Wadsworth* and *Northrop* methods respectively. Danielsson et al. (2001, 2019) approaches perform considerably worse than all others in threshold selection and quantile estimation; so results for these methods are only given in supplementary material A.4.2. We utilised the following R code for Wadsworth, Northrop and EQD methods respectively: code given in the supplementary materials of Wadsworth (2016), *threshr* (Northrop and Attalides, 2020), and `https://github.com/conor-murphy4/automated_threshold_selection` (Murphy et al., 2023).

The performance of all of the methods depends somewhat on the choice of the set of candidate thresholds which we denote by:

$$C_u = \{u_i, i = 1, \ldots, k : u_1 < \ldots < u_k\}, \tag{3.6.1}$$

where we restrict the u_i to be sample quantiles evaluated at equally-spaced probabilities. The range $[u_1, u_k]$, the number of candidates k and the inter-threshold spacing are all potentially important in terms of how they affect the performance of the methods. As emphasised in Section 3.1, we are aiming for an automated threshold selection method which can achieve accurate results without any user inputs, so a key element of our study is to investigate how these features of the set C_u impact on the methods' relative performance. When fitting a GPD with decreasing density (i.e., for $\xi > -1$), it would be inadvisable to use a threshold which clearly lies below the mode of the distribution. As we want to avoid the requirement of user estimation of the mode, our standard choice for the range of the candidate grid is $[u_1, u_k]$: $(u_1, u_k) = (0\%, 95\%)$ sample quantiles of all the data. However, we also explore several cases where only the data lying above the mode are used with $[u_1, u_k]$: $(u_1, u_k) = (0\%, 95\%)$ now sample quantiles of the remaining data. To remove the uncertainty arising from the choice of estimator of the mode, we use the true mode which has a unique value in our simulated cases. Results in supplementary material A.4.4 indicate that our original choice for the candidate threshold set does not unfairly favour the EQD method in any way.

We consider two scenarios: Scenario 1 and Scenario 2 where the true threshold is known and unknown respectively. We present the results using a candidate threshold grid across the whole distribution for Scenario 1 and above the sample median for Scenario 2, with the latter chosen as the Wadsworth method fails when applied across the default range in that setting. The Wadsworth method relies on asymptotic arguments, which limits how large k can be relative to the sample size, n', above the mode, with $n' \leq n$, where n is the total sample size. To assess how the Wadsworth method performs as a fully automated method, we apply the method despite the value of k not always aligning with the guidance in Wadsworth (2016) about its size relative to n'.

We assess the methods' ability to estimate the true threshold (when it exists) and the true return levels, using the RMSE to measure performance. The true quantiles and all bias-variance components of RMSE, discussed in this section, are given in supplementary materials A.2 and A.4.2 respectively. We also investigate the merits of including the uncertainty in threshold selection in our inference, as discussed in Section 3.5, in terms of how they improve the coverage levels of CIs relative to their nominal values.

3.6.2 Scenario 1: True threshold for GPD tail

We consider Cases 0-8, with different properties above and below the true threshold of $u = 1.0$ and various sample sizes. Case 0, where all of the data are from a GPD, is reported in supplementary material A.4.3, with the EQD performing notably better than the existing methods. Here, we present detailed results for Cases 1-4, with Table 3.6.1 providing outline model and sample size properties, with full details and density plots given in supplementary material A.2. Cases 5-8 are considered briefly after discussing Cases 1-4 below.

Cases 1-3 all have a distinct changepoint in the density and density mode both at the true threshold which should make all methods of threshold selection perform better than in situations without these features. Cases 1 and 2 have the same distribution, with $\xi > 0$, with Case 2 having a smaller sample size. We find that the Wadsworth method fails to estimate a threshold in samples with $\xi < -0.05$ irrespective of sample size, so Case 3 is selected near that boundary where the method works and has double the sample size of Case 1. Case 4 is a more difficult example with a continuous density and a small number of exceedances of the true threshold. The data are derived from a partially observed GPD, denoted GPD_p, with data drawn from a GPD above 0 and rejected if less than an independent realisation from a Beta(1,2) distribution.

For each case, the results are based on 500 replicated samples, for which we test the candidate thresholds C_u, with $k = 20$, as given in (3.6.1), with the true threshold being the 16.67% quantile for Cases 1-3 and the 72.10% quantile for Case 4.

Models	Below threshold	Sample size	Above threshold	Sample size
Case 1	$\mathrm{U}(0.5, 1.0)$	200	$\mathrm{GPD}(0.5, 0.1)$	1000
Case 2	$\mathrm{U}(0.5, 1.0)$	80	$\mathrm{GPD}(0.5, 0.1)$	400
Case 3	$\mathrm{U}(0.5, 1.0)$	400	$\mathrm{GPD}(0.5, -0.05)$	2000
Case 4	$\mathrm{GPD}_\mathrm{p}(0.5, 0.1)$	721	$\mathrm{GPD}(0.6, 0.1)$	279

Table 3.6.1: Model specifications for Cases 1-4.

Cases 1-4, Threshold recovery: Table 3.6.2 shows the RMSE of the chosen thresholds for each method in Cases 1-4, with the EQD achieving RMSEs 1.2-7.7 (1-11.2) times smaller than the Wadsworth (Northrop) method. The EQD has the lowest bias by a considerable margin in Cases 1-3 and shows the lowest variance in threshold estimation in all cases. In fact, the variance is reduced by a factor of at least 20 relative to both the Wadsworth and Northrop methods (see Table A.4.1). The very strong performance of the EQD relative to both the Wadsworth and Northrop methods is particularly noteworthy in Cases 1-3, and is also seen for Case 0 and later for Cases 5-7. We believe that the key reason for this is the discontinuity in the density, a feature common to all of these cases, as that appears to lead to a very small bias for the EQD method relative to the other methods. Specifically, the variance penalty of the EQD metric seems to push the threshold as low as possible, but its complementary goodness-of-fit measure almost entirely stops the threshold being selected below the clear discontinuity in the density. For Case 4, which has a continuous density, the EQD achieves the smallest RMSE almost entirely due to it having the smallest variance but with a bias component broadly comparable with the other methods.

	EQD	*Wadsworth*[1]	*Northrop*
Case 1	**0.048**	0.349	0.536
Case 2	**0.060**	0.461	0.507
Case 3	**0.060**	0.221	0.463
Case 4	**0.526**	0.628	0.543

Table 3.6.2: RMSE of the threshold choices for each method-case combination. The smallest values for each case are highlighted in bold.

[1] Results for Wadsworth are calculated only on the samples where a threshold was estimated. It

Cases 1-4, Quantile recovery: Table 3.6.3 presents the RMSEs for the $(1-p_{j,n})$-quantiles where $p_{j,n} = 1/(10^j n)$ for $j = 0, 1, 2$ for sample size n, which ensures that extrapolation is equally difficult over n for a given j. When $j = 0$, no extrapolation is required so the choice of u should not be too important; the similar RMSEs across methods reflect this. As j increases, all RMSEs increase and the differences between methods become clear. The EQD method is best uniformly, followed by the Wadsworth and then the Northrop methods. This pattern reflects the findings in Table 3.6.2, although here with differential performances sensitive to j. However, in terms of quantile estimation, the EQD method does not retain the large differential relative to the other methods which was seen for threshold selection in Cases 1-3. In contrast, the differences between methods in Case 4 are now more apparent, as controlling the variance is more important than any small differences in bias when we are concerned with a RMSE assessment of quantiles which lie far into the tail. The EQD achieves the lowest bias in the majority of cases and leads to quantile estimates with considerably less variance in all cases, particularly as j increases.

	EQD	*Wadsworth*[1]	*Northrop*	*EQD*	*Wadsworth*[1]	*Northrop*
j	**Case 1**			**Case 2**		
0	**0.563**	0.594	0.755	**0.599**	0.631	0.736
1	**1.258**	1.391	2.376	**1.488**	1.644	3.513
2	**2.447**	2.717	7.097	**3.119**	3.484	22.916
	Case 3			**Case 4**		
0	**0.190**	0.195	0.230	**0.677**	0.800	0.791
1	**0.323**	0.344	0.450	**1.563**	2.059	2.217
2	**0.483**	0.516	0.744	**3.043**	4.485	5.568

Table 3.6.3: RMSEs in the estimated quantiles in Cases 1-4 based on fitted GPD above chosen threshold. The smallest RMSE for each quantile are highlighted in bold.

Summary for Cases 5-8: Cases 5-8 are very similar in form to Cases 1-3 but with different shape parameters and sample sizes. The results for these cases are presented in supplementary material A.4.3, with a brief summary given here. Specifically, for Cases

failed estimate a threshold for 2.4%, 26.4%, 0%, 4.4% of the simulated samples in Cases 1-4, respectively.

5-7, we find that the EQD exhibits the strongest performance and the Wadsworth method consistently fails due to the small sample sizes or computational issues with numerical integration when $\xi < -0.05$. Case 8 is parameterised similarly to Case 1 but with an unrealistic sample size of $n = 20000$. Although the data in Case 8 are more suited to a method reliant on asymptotic theory, the EQD performs comparably with the Wadsworth method, with both performing better than the Northrop method.

Case 4, True quantile coverage: We apply Algorithms 1, 1b and 2 to data from Case 4, the hardest case for threshold selection. Table 3.6.4 presents the coverage probabilities of the nominal 80% and 95% CIs of the estimated $(1 - p_{j,n})$-quantiles as well as the average ratio of the CI widths (based on *Alg 2* relative to *Alg 1*) over the 500 samples, termed CI ratio. Results for extra quantile levels, as well as coverage for the 50% CI, are given in supplementary material A.5. Overall, incorporating only parameter uncertainty (*Alg 1* and *Alg 1b*) leads to underestimation of interval widths and inadequate coverage of the true quantiles, especially as we extrapolate further. The additional uncertainty, given in *Alg 1b*, by also accounting for uncertainty in the rate of threshold exceedance, typically makes a very small improvement in coverage, and for some quantiles, this actually leads to a reduction in coverage due to Monte Carlo variation in the simulations. In contrast, the inclusion of the additional threshold uncertainty (*Alg 2*) leads to much more accurate coverage of the true quantiles across all exceedance probabilities. The CI ratios show that this highly desirable coverage is achieved with only 43-62% increase in the CI widths on average.

	80% confidence			95% confidence		
j	0	1	2	0	1	2
Alg 1	0.646	0.618	0.606	0.834	0.804	0.794
Alg 1b	0.656	0.638	0.612	0.830	0.814	0.794
Alg 2	0.798	0.772	0.758	0.954	0.948	0.944
CI ratio	1.430	1.452	1.475	1.484	1.546	1.621

Table 3.6.4: Coverage probabilities for estimated quantiles using Algorithms 1, 1b and 2 for 500 replicated samples from Case 4 with sample size of 1000. CI ratio gives the average ratio of the CIs for Algorithm 2 relative to Algorithm 1 over the 500 samples.

3.6.3 Scenario 2: Gaussian data

In applications, there is no true value for the threshold above which excesses follow a GPD, so we explore this case here. We select the standard Gaussian distribution as it has very slow convergence towards an extreme value limit (Gomes, 1994), so threshold selection is likely to be difficult. We assess threshold selection methods based on estimation of the true quantiles $\Phi^{-1}(1-p_{j,n})$ where $p_{j,n}=1/(10^j n)$, for $j=0,1,2$. We simulate 500 samples, for $n=2000$ and 20000, with C_u, given in (3.6.1), now having range $[u_1,u_k]:(u_1,u_k)=(50\%,95\%)$ sample quantiles of the data and $k=10$ and 91 (i.e., steps of 5% and 0.5%) for the two choices of n respectively. As with Case 8 in Section 3.6.2, $n=20000$ is unrealistic, but we include it to show the slow convergence.

Quantile recovery: Table 3.6.5 shows the RMSEs of the estimated quantiles. For $n=2000$, the EQD method achieves the smallest RMSE with the Northrop method a close second, with the reverse when $n=20000$. The median and 95% CI of the chosen thresholds are given in supplementary material A.4.2. The Northrop method tends to choose slightly higher thresholds than the EQD method, leading to a small reduction in bias, but for only the smaller n is the additional variability relative to the EQD a disadvantage. The Wadsworth method performs the worst, selecting lower thresholds and so incurring the most bias.

	$n=2000$			$n=20000$		
j	*EQD*	*Wadsworth*[2]	*Northrop*	*EQD*	*Wadsworth*	*Northrop*
0	**0.214**	0.239	0.225	0.187	0.214	**0.172**
1	**0.430**	0.529	0.461	0.368	0.422	**0.331**
2	**0.703**	0.890	0.765	0.594	0.672	**0.533**

Table 3.6.5: RMSEs of estimated $(1-p_{j,n})$-quantiles for 500 replicated samples from a Gaussian distribution for samples of size n. The smallest RMSE are highlighted in bold.

True quantile coverage: For assessing the coverage of true quantiles using Algorithms 1, 1b and 2 for Gaussian data, Table 3.6.6 presents the coverage probabilities of

[2] Results for the Wadsworth method, which failed on 0.4% of the samples here, are calculated only for samples where a threshold estimate was obtained.

the nominal 80% and 95% CIs of the estimated quantiles, when $n = 2000$, as well as the average ratio of the CI widths (again, of *Alg 2* relative to *Alg 1*) over the 500 samples, with more results given in supplementary material A.5. Across the $p_{j,n}$, both *Alg 1 and 1b* give very low coverage probabilities in both cases, with performance deteriorating as j increases. The added threshold uncertainty from *Alg 2* results in large increases in coverage though still somewhat less than required, with this achieved through increases in CI widths by 45-66% on average. This weaker performance than we find in Section 3.6.2 suggests that no sample threshold (for realistic sample sizes) is large enough to overcome bias in making extreme value approximations for Gaussian data, but the improvement in coverage using *Alg 2* demonstrates the importance of including the additional threshold uncertainty.

	80% confidence			95% confidence		
j	0	1	2	0	1	2
Alg 1	0.588	0.450	0.366	0.750	0.618	0.510
Alg 1b	0.592	0.442	0.364	0.746	0.620	0.508
Alg 2	0.718	0.598	0.492	0.866	0.814	0.756
CI ratio	1.457	1.480	1.509	1.495	1.576	1.665

Table 3.6.6: Coverage probabilities for estimated quantiles using Algorithms 1, 1b and 2 for 500 replicated samples from a Gaussian distribution with sample size of 2000. CI ratio gives the average ratio of the CIs for Algorithm 2 relative to Algorithm 1 over the 500 samples.

3.7 Application to river flow data

The River Nidd dataset consists of 154 storm event peak daily river flow rates that exceeded 65 m^3/s in the period 1934-1969, i.e., an average exceedance rate of 4.4 events per year. Each observation can be deemed “extreme” and IID, though not necessarily well-described by a GPD. Davison and Smith (1990) identify the difficulties these data present for threshold selection and parameter uncertainty, which we reiterated in discussion of Figure 3.2.1. Given the small sample size for the River Nidd, any increase in the threshold value is more significant in terms of parameter uncertainty, than for larger datasets studied in Section 3.6.

Table 3.7.1 shows the selected thresholds of each of the methods for a range of candidate grids[3]. The remarkable robustness of the EQD (evaluated with $B = 200$ bootstrap samples) across grids stems from the method's novel incorporation of data uncertainty. The Wadsworth method fails to estimate a threshold unless the grid is made very coarse, and even then exhibits considerable sensitivity (varying between 0% and 90% sample quantiles) over grids of equal size but different endpoints and increments. This is problematic as a coarse grid is likely to remove the most appropriate threshold from consideration. The Northrop method critically depends on the validation threshold, and we find that increasing this level above the 90%-quantile leads to failure or convergence warnings. The thresholds selected by this method are quite variable (between 0% and 80% sample quantiles) over the grids.

Estimated thresholds for the River Nidd dataset			
Grid (% quantile)	*EQD*	*Wadsworth*	*Northrop*
0 (1) 93	67.10 (3%)	NA	68.45[3] (6%)
0 (1) 90	67.10 (3%)	NA	65.08 (0%)
0 (1) 80	67.10 (3%)	NA	100.28 (75%)
0 (20) 80	65.08 (0%)	NA	109.08 (80%)
0 (30) 90	65.08 (0%)	149.10 (90%)	65.08 (0%)
0 (25) 75	65.08 (0%)	100.28 (75%)	81.53 (50%)
0, 10, 40, 70	65.08 (0%)	65.08 (0%)	69.74 (10%)

Table 3.7.1: River Nidd dataset selected thresholds (and quantile %) for each method for different grids of candidate thresholds. The Grid column gives *start (increment) end* for each grid.

Comparing thresholds selected between the methods is complicated due to the sensitivity of the Wadsworth and Northrop methods to the grid choice. For the EQD, it is natural to use the densest and widest grid, giving $\hat{u} = 67.10$. This threshold, which is lower than previously found, gives far more data for the extreme value

[3] In marked cases, the Northrop method outputted a chosen threshold with some convergence warnings.

analysis. As all the River Nidd data are "extreme", we believe taking u so close to the lower endpoint of the data is not problematic, and it may indicate that the preprocessing level used to produce these data was too high. The first estimated threshold from the Wadsworth (Northrop) methods, without convergence or warning issues, is $\hat{u} = 149.10$ ($\hat{u} = 65.08$). For these three thresholds, the GPD parameter estimates (and 95% CIs) are: $\hat{\sigma}_{u:EQD} = 23.74$ $(17.78, 29.70)$ and $\hat{\xi} = 0.26$ $(0.06, 0.46)$ for the EQD; $\hat{\xi} = -0.15$ $(-1.00, 0.70)$ for the Wadsworth method; and for the Northrop method, $\hat{\xi} = 0.20$ $(0.02, 0.38)$, where we omit the latter two scale parameter estimates as they are estimating different quantities which depend on the threshold, see Section 3.2. Provided all estimated thresholds are high enough for the GPD to be appropriate, the values of $\hat{\xi}$ should be similar across methods, due to the threshold stability property (see Section 2.3.1). The Wadsworth method leads to an extremely wide CI, which results in meaningless inference. However, the EQD and Northrop findings about ξ are similar, but the sensitivity to the candidate grid is a problem for the Northrop method.

Figure 3.7.1 shows a QQ-plot for the GPD model using the EQD estimate $\hat{u} = 67.10$. The tolerance bounds show a reasonable agreement between model and data. For $\hat{u}$, Figure 3.7.1 also shows the T-year return level estimates, with $1 \leq T \leq 1000$. The 95% CIs incorporate parameter uncertainty alone and both parameter and threshold uncertainty via Algorithms 1 and 2 respectively, with an increase in uncertainty from the latter for larger T; e.g., for the 100- and 1000-year return levels, the CI width increases by a factor of 1.38 and 1.52 respectively. This reiterates how vital it is to incorporate threshold uncertainty into inference.

3.8 Conclusion and discussion

We proposed two substantial developments to univariate extreme value analysis. Firstly, we addressed the widely-studied problem of how to automatically select/estimate a

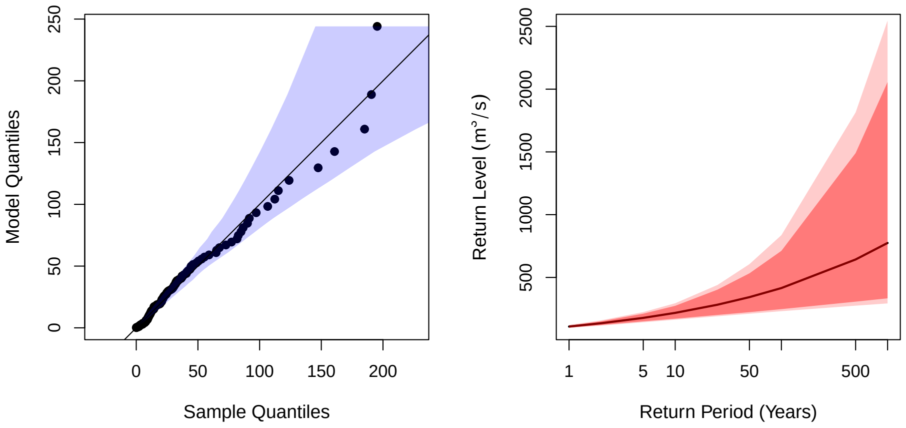

Figure 3.7.1: River Nidd analysis: QQ-plot [left] showing model fit with 95% tolerance bounds [shaded] and return level plot [right] based on EQD threshold choice with 95% CIs incorporating parameter uncertainty [dark-shaded] and additional threshold uncertainty [light-shaded].

threshold above which an extreme value, generalised Pareto, model can be fitted. We presented a novel and simple approach, which we termed the EQD method, that minimises an approximation to the IAE of the model quantiles and quantiles of the data generating process. Secondly, we proposed a new approach to improve the calibration of confidence intervals for high quantile inference, addressing an important but under-studied problem. We achieve this through an intuitively simple, but computationally intensive, double-bootstrapping technique which propagates the uncertainty in the threshold estimation through to quantile inference.

Regarding the threshold selection component of the work, we compared the EQD method to the leading existing threshold selection methods in terms of both threshold selection and consequent high quantile estimation. This was conducted using data from IID continuous univariate random variables and the superiority of the EQD method was illustrated across a range of examples using various metrics. Relative to existing approaches, we showed that the EQD exhibits greater robustness to changes in the set of candidate thresholds, to tuning parameters, and avoids a reliance on asymptotic theory in existing likelihood methods. The EQD method is applicable for all data set sizes and for any set of candidate thresholds.

So why does the EQD method perform much better than the existing approaches? Our analysis has identified two core reasons: the choice of a robust measure of goodness of fit for a given (bootstrapped) sample, which controls bias; and the use of bootstrapped replicates, which leads to reduced variance and also appears to reduce bias. Specifically, in comparison to existing methods, the use of our goodness-of-fit measure, over simply exploiting the GPD threshold stability property, ensures better model fits and hence better threshold selection, and the bootstrapping removes the variation that arises if only the observed sample is used, as that may not be a typical realisation from the underlying data generating process.

In assessing our suggested improvement for the calibration of confidence intervals, we compared the coverage of true quantiles using our proposed approach and the widely-adopted approach of incorporating the GPD parameter uncertainty alone in quantile inference once a threshold has been selected. We showed that the coverage of the existing approach was substantially less than the nominal confidence levels and our proposed approach led to much more reliable confidence intervals without an undue increase in their width.

While this paper has demonstrated the effectiveness of both the EQD method and our proposed approach for confidence interval construction in the univariate IID setting, we believe that the findings suggest that these approaches could have much wider utility. For example, the Varty et al. (2021) method, which motivated the structure of the EQD method, was originally developed for non-identically distributed data, with the transformation of excesses of a time-varying threshold to a common marginal Exponential(1) distribution. As such cases typically find that excesses have a common shape parameter ξ (Chavez-Demoulin and Davison, 2005), we could use the EQD variant of Varty et al. (2021) by transforming instead to a common GPD with parameters

$(1, \xi)$ given we have seen here that by retaining the scale of the original data, the EQD out-performs the Varty et al. (2021) approach. We also believe that the strategy of our new methods could be used to improve threshold estimation in multivariate extremes, in cases of multivariate regular variation assumptions (Wan and Davis, 2019) or for asymptotically independent variables (Heffernan and Tawn, 2004), and allow for the uncertainty in this threshold estimation to be incorporated in the subsequent joint tail inferences. Such developments would naturally have similar implications for spatial extreme value modelling as the threshold selection in this context currently comes down to a multivariate (at the data sites) threshold selection process.

Chapter 4

Automated tail-informed threshold selection for extreme coastal sea levels

4.1 Introduction

Natural hazards such as flooding, earthquakes and wildfires devastate communities and livelihoods around the world. Extreme value analysis (EVA) applied to the historical records of such events provides a useful tool for describing the frequency and intensity of these processes, and can be used by practitioners, community leaders, and engineers to prepare in advance for catastrophic events. Example applications include flood risk assessment (D'Arcy et al., 2023), nuclear regulation (Murphy-Barltrop and Wadsworth, 2024), ocean engineering (Jonathan et al., 2014), and structural design analysis (Coles and Tawn, 1994). Furthermore, stakeholders with assets spread across large geographical regions also utilise these tools to understand the hazard across regional, continental, and global scales; see Keef et al. (2013b), Quinn et al. (2019), and Wing et al. (2020).

Coastal flood events, driven by high tides, surges, or waves, are commonly recorded at tide gauge stations, which cover large proportions of the populated global coastline. When characterising extreme sea level events, these tide gauge records are a primary source of information available to coastal managers. Due to the large number of sites involved, automated techniques for the characterisation of extreme events are preferable.

The earliest EVA techniques used the annual maximum approach, whereby a theoretically motivated distribution is fitted to the observed yearly maxima. However, this approach suffers from the drawback that only one observation is recorded for each year, resulting in inefficient use of the data. In practice, this can lead to an incomplete picture of the upper tail and less accurate estimates of tail quantities, such as return levels. Consequently, recent consensus has been to move away from the annual maximum approach (Davison and Smith, 1990; Coles, 2001; Scarrott and MacDonald, 2012; Pan and Rahman, 2022).

As a result, the POT approach has become the most popular technique for EVA modelling; see Section 4.3 and Coles (2001) for further details. This approach involves fitting a statistical model to data above some high threshold. However, the choice of this threshold is not arbitrary, and inappropriate choices can result in poor model fits and extrapolation into the tail. Traditional approaches rely on visual assessments of parameter stability above the appropriate threshold. Such approaches suffer from subjectivity (Caballero-Megido et al., 2018) and the time input required to apply such techniques to global tide gauge records is not feasible. Consequently, many efforts have been made to reduce the time burden incurred by manual threshold selection. These include simplifications that allow large amounts of data to be processed, but at the cost of accuracy, e.g., using a static threshold, such as the 0.98 quantile or a fixed number of exceedances per year (Hiles et al., 2019; Collings et al., 2024). We refer to the approach of selecting a static 0.98 quantile across all sites (or variables) as the Q98 approach henceforth. Other approaches aim to automate much of the subjective decision-making process while retaining a flexible method that can capture the underlying behaviour of the physical processes (Solari et al., 2017; Curceac et al., 2020; Murphy et al., 2025).

In this study, our aim is to build upon existing techniques to provide a novel approach to automating threshold selection, which is applicable to a wide range of datasets whereby the extremes are characterised by different drivers. As a motivating example, we apply our method to a global dataset of 417 tide gauge records, demonstrating the performance of our approach over a variety of locations and benchmarking against other commonly used techniques.

The layout of this chapter is as follows; in Section 4.2 we introduce the dataset used in this study and in Section 4.3 we discuss the common difficulties in using the POT approach across such a large, varied dataset, as well as some of the methods used to simplify the process. In Section 4.4, we describe our novel approach to automating threshold selection and explain the subjective choices we have made in the method. In Section 4.5, we present the results of applying our method to the global tide gauge dataset described in Section 4.2. In Section 4.6, we discuss our results in the context of uncertainty, bias, and the underlying physical processes and finally, in Section 4.7 we provide a conclusion to our study.

4.2 Data

The locations of the considered tide gauge stations are illustrated in Figure 4.2.1. These data are obtained from the Global Extreme Sea Level Analysis (GESLA) database (Haigh et al., 2023), version 3.1. The GESLA database was collated from many organisations that collect and publish tide gauge data. The water level records are prepared using the quality control flags published by the authors alongside the data set, and duplicate timestamps in the records are also removed. The water level records that contain over 40 years of good data (defined as at least 75% complete) are retained. This results in a total of 417 water level records from around the world, which have an average record length of 66 years. The raw time series data are provided on a range of time steps (10, 15, and 60 minutes), and so are averaged to hourly resolution. A linear trend is calculated and removed to account for mean sea level rise. Generally, mean sea level rise can be approximated as linear in most locations globally. Some areas, especially areas where glacial isostatic rebound is important, can have non-linear signals, but any differences between non-linear and linear estimates of sea level change are generally small, especially in comparison to the magnitudes of the extreme sea levels, and so, in our analysis, we find a linear trend to be adequate. Daily maxima data are obtained from the hourly records, and the data is subsequently declustered (see Section 2.5 for review of declustering techniques) using a 4-day storm window to ensure event independence (Haigh et al., 2016; Sweet et al., 2020). Given the range of oceans and coastlines covered, one would generally expect to observe a wide variety of tail behaviours across the records.

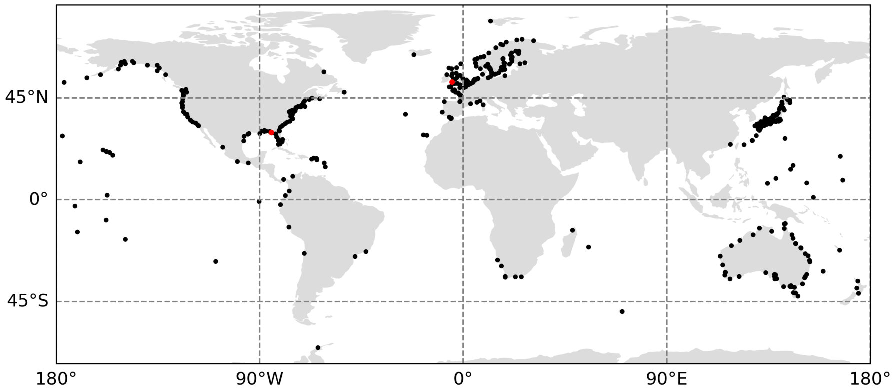

Figure 4.2.1: Map of GESLA record locations with record lengths greater than 40 years. The two locations highlighted in red are Apalachicola, US and Fishguard, UK, which are discussed in more detail in Section 4.5.4.

4.3 POT modelling

The POT approach, whereby a theoretically motivated distribution is fitted to the excesses of some high threshold (see, e.g., Coles, 2001), is the most common technique for assessing tail behaviour in environmental settings. Given any independent and identically distributed (IID) random variable X and a threshold u, the results of Balkema and de Haan (1974) and Pickands (1975) demonstrate that under weak conditions, the excess variable $Y := (X - u \mid X > u)$ can be approximated by a *generalised Pareto distribution* (GPD) – so long as the threshold u is 'sufficiently large'. The GPD has the form

$$H(y; \sigma, \xi) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi}, \quad y > 0, \tag{4.3.1}$$

where $z_+ = \max(0, z)$, $\sigma > 0$, and $\xi \in \mathbb{R}$. We refer to σ and ξ as the scale and shape parameters, respectively, and we remark that the latter parameter quantifies important information about the form of tail phenomena; see Davison and Smith (1990) for further discussion. A wide range of statistical techniques have been proposed, including both Bayesian and frequentist frameworks, to fit the model in equation (4.3.1) (Dupuis, 1999; Behrens et al., 2004; Scarrott and MacDonald, 2012; Northrop et al., 2017), although we note that maximum likelihood estimation (MLE) remains the most common technique (e.g., Gomes and Guillou, 2015). Consequently, we restrict attention to MLE techniques throughout this chapter.

In many practical contexts, distribution (4.3.1) is used to obtain estimates of *return levels* for some *return period* N of interest. Such values offer a straightforward interpretation: the N-year return level is the value x_N that one would expect to exceed once, on average, every N years. Return levels are easily obtained by inverting equation (4.3.1) (see Coles, 2001), and their estimates are often used to inform decision making. For example, in the contexts of flood risk analysis and nuclear infrastructure design, regulators specify design levels corresponding to return periods of $N = 100$ years (D'Arcy et al., 2023) and $N = 10000$ years (Murphy-Barltrop, 2024), respectively.

The ambiguity of the statement 'a sufficiently large threshold u' requires careful consideration. This is a problem that is commonly overlooked in many applications, and selecting a threshold u is entirely non-trivial. In particular, this selection represents a bias-variance trade-off: selecting a threshold too low will induce bias by including observations that do not represent tail behaviour, while extremely high thresholds will result in more variability due to lower sample sizes. Furthermore, the estimates of return levels are very sensitive to the choice of threshold, and biased estimates can significantly impact the cost and effectiveness of certain infrastructures, such as flood defences (Zhao et al., 2024).

Owing to the importance of this choice, a plethora of methods have been proposed which aim to balance the aforementioned trade-off; see Belzile et al. (2023) for a recent review of the literature. The standard and most-widely used approach for threshold selection involves a visual assessment of the stability of the GPD shape parameter across a range of increasing thresholds (Coles, 2001). This approach suffers from subjectivity in the choice of stable region. Furthermore, visual assessments for individual sites are simply not feasible (within a reasonable time scale) for large scale applications.

Automatic approaches seek to remove this subjectivity by selecting a threshold based on some criterion or goodness-of-fit metric; Wadsworth and Tawn (2012) and Northrop and Coleman (2014) utilise penultimate models and hypothesis testing; Bader et al. (2018) and Danielsson et al. (2019) use goodness-of-fit diagnostics; Wadsworth (2016) utilise a sequential assessment of a changepoint model; and Northrop et al. (2017) create a measure of predictive performance in a Bayesian framework. Tancredi et al. (2006) avoid the prior selection of the threshold by employing a Bayesian mixture model where the threshold is incorporated into the parameter estimation, allowing for straight-forward estimation of threshold uncertainty. In the applied literature, Durocher et al. (2018) and Curceac et al. (2020) compare several automated goodness-of-fit approaches for selecting an appropriate threshold in the hydrological setting. Furthermore, Choulakian and Stephens (2001), Li et al. (2005) and Solari et al. (2017) automate goodness-of-fit procedures and apply these techniques to a range of precipitation and river flow data sets.

Recently, Murphy et al. (2025) proposed a novel threshold selection technique, corresponding to Chapter 3 of this thesis, building on the work of Varty et al. (2021). This method, termed the *expected quantile discrepancy* (EQD), aims to select a threshold u for which the sample excesses are most consistent with a GPD model. We briefly outline this method below. Let $\boldsymbol{x}_u = (x_1, \ldots, x_{n_u})$ be the sample of excesses of some candidate threshold u, i.e., a sample from Y. For each candidate threshold, the EQD method assesses the expected deviation between sample and theoretical quantiles at a set of fixed probabilities $\mathcal{P}_m := \{j/(m+1) : j = 1, \ldots, m\}$, where m denotes some large whole number. This assessment is done across a number of bootstrapped samples, say B, to incorporate sampling variability and stablise the threshold choice. More specifically, letting $\boldsymbol{x}_u^b$ denote the b^{th} bootstrapped sample of $\boldsymbol{x}_u$, with $b = 1, \ldots, B$, Murphy et al. (2025) propose the metric

$$d_b(u) := \frac{1}{m} \sum_{j=1}^{m} \left| \frac{\hat{\sigma}_u^b}{\hat{\xi}_u^b} \left[\left(1 - \frac{j}{m+1} \right)^{-\hat{\xi}_u^b} - 1 \right] - Q\left(\frac{j}{m+1}; \boldsymbol{x}_u^b \right) \right|, \tag{4.3.2}$$

where $(\hat{\sigma}_u^b, \hat{\xi}_u^b)$ denote the GPD parameter estimates for $\boldsymbol{x}_u^b$, obtained using MLE, and $Q(j/(m+1); \boldsymbol{x}_u^b)$ denotes the $j/(m+1)$ empirical quantile of $\boldsymbol{x}_u^b$. Considering equation (4.3.2) over each bootstrapped sample, an overall measure of fit for u is given by $d(u) = \sum_{b=1}^{B} d_b(u)/B$. Finally, the selected threshold, u^*, is the value that minimises d, i.e., $u^* := \arg\min d(u)$. Through an extensive simulation study, Murphy et al. (2025) show that their approach convincingly outperforms the core existing approaches for threshold selection. Therefore, at the time of writing, the EQD technique is a leading approach for automating threshold selection.

In this chapter, we argue and demonstrate that while the EQD approach appears to work well in a wide variety of cases, it can suffer from drawbacks in certain applications that result in less than ideal threshold choices. Specifically, the chosen thresholds can result in model fits that do not match up well at the most extreme observations. We briefly explore the reasons for why this may occur below.

To begin, consider two candidate thresholds $u_1 < u_2$ satisfying $\Pr(X > u_1) = 0.5$ (i.e., the median) and $\Pr(X > u_2) = 0.99$. Taking each threshold in turn, the EQD computes quantiles from the (bootstrapped) conditional variables $(X - u_1) \mid (X > u_1)$ and $(X - u_2) \mid (X > u_2)$ that correspond with the probability set $\mathcal{P}_m$. When considered on the scale of the data, however, this results in very different quantile probabilities. Letting $x_{u_1,j}$ denote the (true) $j/(m+1)$ quantile of $(X - u_1) \mid (X > u_1)$ for any $j = 1, \ldots, m$, we have

$$\begin{aligned}\Pr(X \leq x_{u_1,j} + u_1) &= 1 - \Pr(X - u_1 > x_{u_1,j} \mid X > u_1)\Pr(X > u_1) \\ &= 1 - [1 - j/(m+1)]0.5 =: q_{u_1,j},\end{aligned}$$

with an analogous formula following for u_2, i.e., $q_{u_2,j} := 1 - [1 - j/(m+1)]0.99$. The resulting probability sets $\{q_{u_1,j}\}_{j=1}^{m}$ and $\{q_{u_2,j}\}_{j=1}^{m}$, with $m = 100$, are illustrated in Figure 4.3.1. This demonstrates that the lower the threshold level u, the lower the quantile probabilities evaluated by the EQD. Thus, quantiles lying far in the tail of the data will carry significantly less weight for lower thresholds than for higher thresholds.

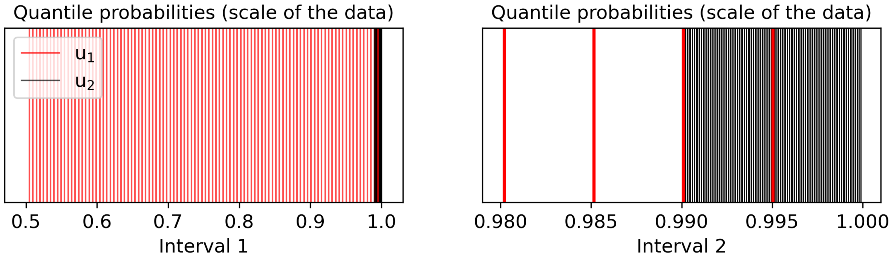

Figure 4.3.1: The probability sets $\{q_{u_1,j}\}_{j=1}^{m}$ and $\{q_{u_2,j}\}_{j=1}^{m}$ illustrated in red and black vertical lines, respectively. The left and right plots are given on different intervals to illustrate the fact the quantile probabilities exist in entirely different subregions of $[0, 1]$.

On a similar note, we remark that the metric described in equation (4.3.2) is equally weighted across all probability levels. We argue that this somewhat disagrees with intuition in the sense that many practitioners mainly care about a models' ability to estimate very extreme return levels, and one only wants observations in the tail to be driving this estimation. Including non-extreme observations will bias the estimation procedure and therefore assessing quantile discrepancies mainly for lower quantile levels, as will occur for lower candidate thresholds, provides little to no intuition as to how the fitted model will perform at the most extreme levels.

Taking these points into account, we propose an extension of the EQD procedure to improve the model fit to the most extreme observations. Our proposed extension results in models fits which more accurately capture the upper tail of the data in contexts where the EQD method struggles. Specifically, in the context of coastal modelling, we demonstrate that the EQD approach selects thresholds that do not appear appropriate for capturing the most extreme observations across many coastal sites; such issues do not arise for our extended approach.

Consider the example illustrated in Figure 4.3.2 for a tide gauge record located in Penascola Bay, US, which is in the Gulf of Mexico. This record was selected as it is located in a region impacted by tropical cyclones, where the uncertainty in the model fits using the historical records is typically large. This particular dataset may be better modelled by a two component mixture distribution to describe the tropical cyclones and other extremes. We choose to ignore this here as threshold estimation is the priority of this chapter and in any case, the heavier tail of such a mixture will dominate for high quantiles. As demonstrated in the left panel of this figure, the threshold chosen by the EQD method was the 84% quantile and the resulting model fit performs poorly within the upper tail. For this particular example, this indicates that the overall model fit is being driven mainly by lower observations, biasing the fit in the upper tail. Such findings were replicated across many coastal sites, indicating that this is not an unusual phenomenon. We also illustrate the model fit that arises from our proposed method (see Section 4.4), which selected a threshold at the 97% quantile, in the right panel of Figure 4.3.2. One can observe that even though the updated model fit has a higher discrepency value $d(u)$, the model quantiles appear better able to capture the upper tail in the data, at the cost of additional parameter uncertainty from the fewer exceedances included in the fit.

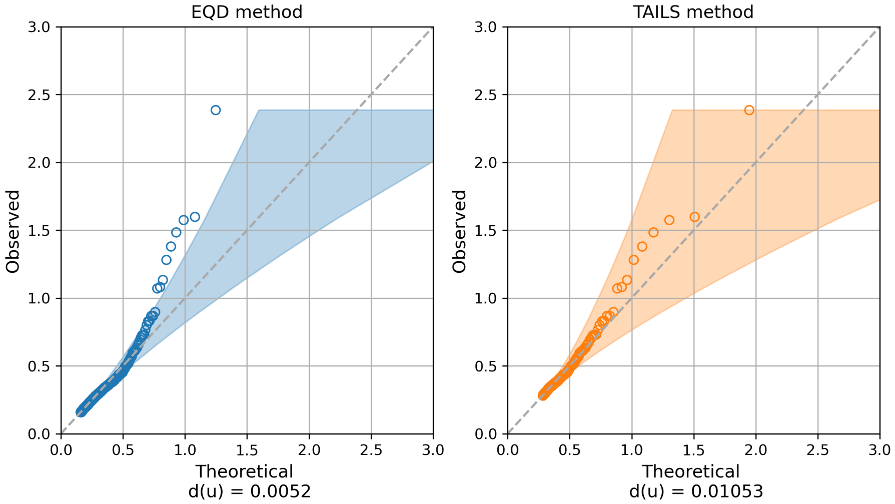

Figure 4.3.2: QQ plots for the thresholds selected using the EQD (left) and TAILS (right) approaches; see Section 4.4 for more details of the TAILS method. 95% tolerance bounds are shown as shaded regions. The sub captions, in both cases, give the EQD score $d(u)$ of the threshold chosen by each method.

These findings indicate that whilst the EQD approach outperforms many existing techniques, it can, in some cases, result in model fits that fail to capture the most extreme observations. This drawback motivates novel developments, and in this work we propose an adaptation of the EQD technique, which we term the ***Ta****il-****i****nformed thresho****l****d* ***s****election* (TAILS) approach. Unlike the EQD approach, our technique focuses exclusively on quantiles within a pre-defined upper tail of the data, independent of the choice of threshold. Furthermore, we demonstrate in Section 4.5 that TAILS results in improved model fits across a wide range of tide gauge records. Code for implementing the TAILS approach is available at `https://github.com/callumbarltrop/TAILS`.

4.4 The TAILS approach

In this section, we introduce the TAILS approach for GPD threshold selection. To begin, let $\mathscr{P} := \{p_i : i = 1, \ldots, m\}$ denote a set of increasing quantile levels close to 1: the selection of $\mathscr{P}$ is subsequently discussed. Given a candidate threshold u, let $\boldsymbol{x}_u^b$, $b = 1, \ldots, B$, be defined as in Section 4.3 and let $\pi_u = \Pr(X \leq u)$. We propose the following metric

$$\tilde{d}_b(u) := \frac{\sum_{i=1}^{m} \mathbb{1}(\pi_u < p_i) \left| \frac{\hat{\sigma}_u^b}{\hat{\xi}_u^b} \left[\left(\frac{1-p_i}{1-\pi_u} \right)^{-\hat{\xi}_u^b} - 1 \right] - Q\left(1 - \frac{1-p_i}{1-\pi_u}; \boldsymbol{x}_u^b\right) \right|}{\sum_{i=1}^{m} \mathbb{1}(\pi_u < p_i)}, \tag{4.4.1}$$

with $Q(\cdot\ ;\cdot)$ and $(\hat{\sigma}_u^b, \hat{\xi}_u^b)$ defined as before. For each threshold u, this metric ensures that the same quantile probabilities are evaluated, when considered on the scale of the data. Furthermore, observe that equation (4.4.1) accounts for cases when the threshold probability, π_u, exceeds a subset of $\mathscr{P}$; in such instances, the metric is only evaluated on probabilities greater than the threshold non-exceedance probability, corresponding to the region where the given GPD model is valid. Analogous to the original approach, an overall measure of fit for a candidate threshold u is given by $\tilde{d}(u) = \sum_{b=1}^{B} \tilde{d}_b(u)/B$, and the selected threshold, u^*, is the value that minimises $\tilde{d}$, i.e., $u^* := \arg\min \tilde{d}(u)$.

The motivation behind (4.4.1) is to only evaluate quantile differences within the tail of the data, independent of the threshold candidate. This ensures that the threshold choice is driven entirely by the model fit within the most extreme observations. However, prior to applying the method, one must select a probability set $\mathscr{P}$. This choice is non-trivial, and is crucial for ensuring the proposed method selects a sensible threshold. For instance, selecting probabilities very close to one is meaningless in a practical setting, since the corresponding quantiles cannot be estimated empirically from data of a finite sample size. On the other hand, selecting probabilities too low will defeat the objective of our proposed technique.

With this in mind, we term p_1 the *baseline probability*, i.e., the smallest probability in $\mathscr{P}$. This corresponds to the 'baseline' observation frequency below which one treats any events to be extreme relative to the sample size. Naturally, this represents a subjective choice, and the best choice of baseline probability is likely to be context dependent. In practice, we recommend selecting p_1 based on expert or domain-specific knowledge; for example, what magnitude of return period normally results in a relatively low-impact, but significant event within a given context? Take coastal flood risk mitigation and the occurrence of 'nuisance' flooding as an example. Nuisance flooding is defined as '*low levels of inundation that do not pose significant threats to public safety or cause major property damage, but can disrupt routine day-to-day activities, put added strain on infrastructure systems such as roadways and sewers, and cause minor property damage*' (Moftakhari et al., 2018). Although the exact return period of these events varies by location, a study carried out in the US demonstrated that these events generally occur at sub-annual frequencies, and that the median across their study sites was 0.5 years (Sweet et al., 2018). In this study, we chose to use a return period of 0.25 years for p_1, to include events below the median obtained in the study above. This choice was further supported by a sensitivity analysis, the results of which are presented in Appendix B. Note that this does not imply that the optimum threshold choice will lie close to the baseline event, since this choice is driven exclusively by the asymptotic rate of convergence to the underlying tail distribution.

Alongside the baseline probability, we also set p_m (the largest probability in $\mathscr{P}$), such that we ensure that we observe 10 exceedances above the corresponding quantile, on average, over the observation period. Extrapolating beyond this level is unlikely to be meaningful, since we cannot estimate empirical quantiles outside of the range of data. Furthermore, we impose that all candidate thresholds are less than the 1 year return level. This upper threshold is used in similar automated threshold selection studies, such as Durocher et al. (2018).

Finally, for the remaining probabilities in $\mathscr{P}$, we set $p_j := p_1 + (j-1)(p_m - p_1)/(m-1)$, $j = 2, \ldots, m-1$, corresponding to equally spaced values from the p_1 to p_m. For the number of quantile levels m, we follow Chapter 3 and set $m = 500$; such a value ensures a wide range of probabilities are evaluated without too much linear interpolation between observed quantile levels. Similar to Chapter 3, for this setting, we found that the choice of m made very little difference to the thresholds selected by the approach. See Appendix B for more details.

4.5 Results

We now assess the performance of the TAILS approach using the dataset introduced in Section 4.2. In Section 4.5.1, we apply both the EQD and TAILS approaches over all locations with $m = 500$ and $B = 100$. The same values for m and B were used by Murphy et al. (2025). In Section 4.5.2, we provide spatial plots of the results to determine if there are any patterns present in the thresholds selected by the TAILS method or in the differences between the selected thresholds of the TAILS and EQD approaches. In Section 4.5.3, we assess, with a right-sided Anderson-Darling (ADr) test, the GPD model fits obtained using the selected thresholds from each approach, as well as the model fits using the static threshold of the Q98. Lastly, in Section 4.5.4, we show the distance metrics from the EQD and TAILS approaches for two tide gauge records, and present the resulting return levels from the two automated methods and the Q98 approach.

4.5.1 Selected thresholds

Since the scales of data at different locations vary, we present the quantile probabilities of the selected thresholds rather than the threshold values themselves; these are illustrated in Figure 4.5.1. The TAILS approach clearly selects higher thresholds compared to the EQD approach, as expected. The minimum and maximum quantiles selected by the TAILS and EQD methods are (0.903, 0.993) and (0.501, 0.991), respectively. The lowest threshold selected by the EQD approach is very close to the lower limit of candidate thresholds provided (i.e., the median).

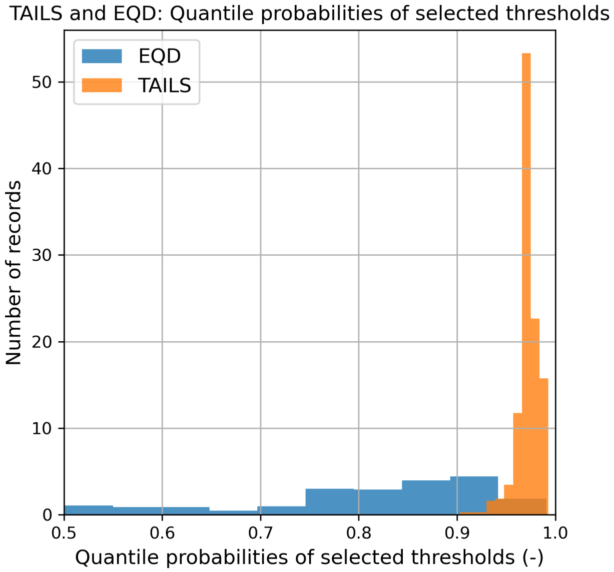

Figure 4.5.1: The results from applying the EQD and TAILS methods to every GESLA record used in this study, showing the distributions of quantile probabilities of the selected thresholds.

4.5.2 Spatial analysis

Figure 4.5.2 (a) shows the quantile probabilities of the original data for the TAILS selected thresholds plotted spatially. There are no obvious large-scale spatial patterns in the selected threshold probabilities. However, with the exception of a few outliers, the changes across space are generally small. In Australia, the quantile probabilities of selected thresholds appear to be marginally lower in the tide gauge records located in the south, compared with records located on the east and west coasts of the country. In contrast, the quantile probabilities of the selected thresholds around Japan and Hawaii look comparatively uniform, with very little spatial change.

Figure 4.5.2 (b) illustrates the differences between the quantile probabilities of the selected thresholds from the TAILS and EQD approaches. All thresholds selected using the TAILS method are greater than those of the EQD. Strong spatial patterns are present, particularly at tide gauge locations in north-eastern Europe. The tide gauge records with the largest increases are located in the Baltics and show increases of nearly 0.5 (in quantile probability). Spatial trends are also visible around Australia, with the TAILS approach selecting much higher thresholds around the south of the country, compared with the north.

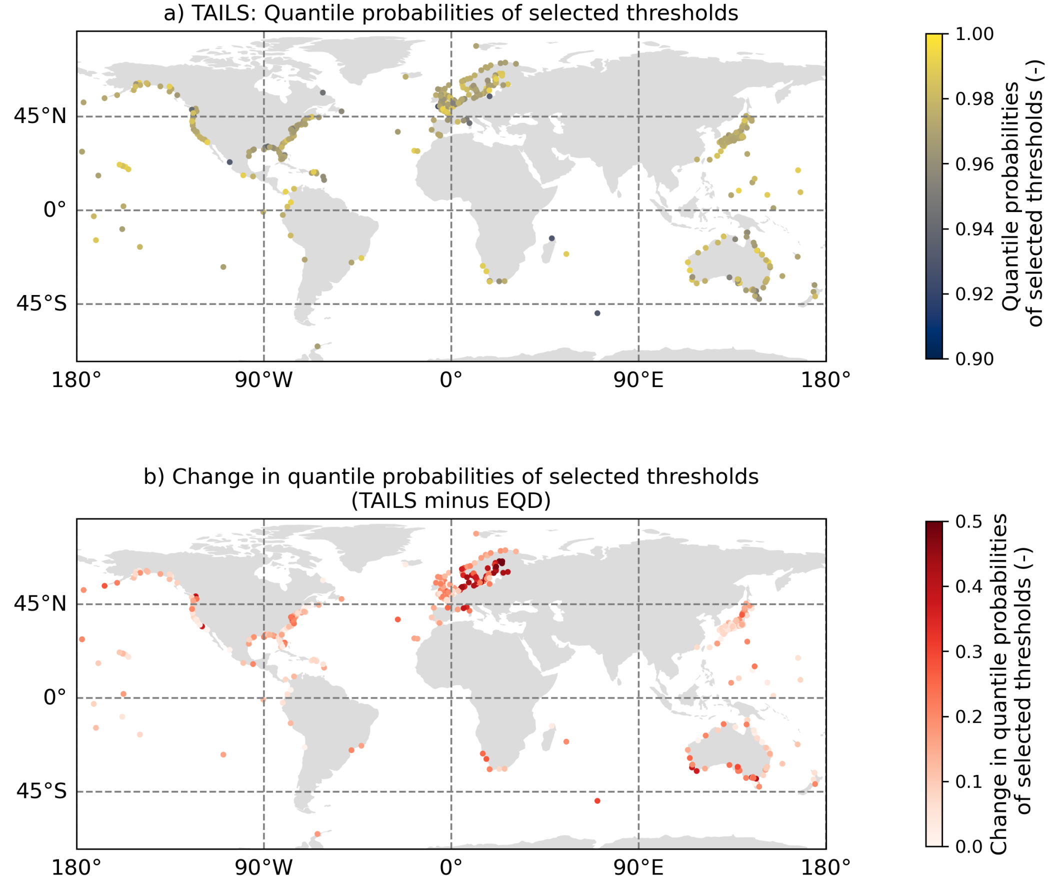
a) TAILS: Quantile probabilities of selected thresholds
45°N
0°
45°S
180°
90°W
0°
90°E
180°
1.00
0.98
0.96
0.94
0.92
0.90
Quantile probabilities of selected thresholds (-)
b) Change in quantile probabilities of selected thresholds (TAILS minus EQD)
45°N
0°
45°S
180°
90°W
0°
90°E
180°
0.5
0.4
0.3
0.2
0.1
0.0
Change in quantile probabilities of selected thresholds (-)

Figure 4.5.2: Spatial plots of a) the quantile probabilities of selected thresholds using the TAILS methods, and b) the difference in the quantile probabilities of the selected thresholds between the TAILS and EQD approaches.

4.5.3 Right-sided Anderson-Darling test

The ADr test statistic (Sinclair et al., 1990; Solari et al., 2017) is used to measure the goodness-of-fit of the exceedances over the thresholds selected using both the EQD and TAILS methods, as well as the model fits computed using the Q98 approach. The test compares the theoretical and empirical distributions, with more weight placed on the tails of the distribution. The statistic quantifies the deviation of the data from the specified distribution. A p-value is obtained by bootstrapping the test statistic, and indicates the probability of observing deviation seen between the threshold exceedances and the GPD model assuming, under the null hypothesis, that the GPD model is appropriate. The null hypothesis is typically rejected for p-values below 0.05, corresponding to a 5% significance level.

A larger test statistic (or equivalently, a lower p-value) indicates more deviation from the model distribution being tested, which in this case, is a GPD. As shown in Figure 4.5.3 (a), the EQD approach yields larger ADr test statistics than the TAILS method. The range of test statistics computed using the TAILS method are all less than 1, whereas the EQD approach has many values exceeding 1. This indicates that the EQD method could be selecting a threshold in this context over which the exceedances are not well characterised by a GPD. This is further corroborated by the p-values obtained for each method, plotted in Figure 4.5.3 (b). The median p-value across all model fits obtained using the TAILS method is 0.615, compared with 0.312 for the EQD approach. The TAILS method also outperforms the Q98 approach, with a smaller test statistic average and greater average p-value. While all the methods achieve adequate fits for most of the dataset, in some of the cases where the EQD and Q98 method lead to poor model fits (p-value less than 0.05), the TAILS method can significantly improve results. Of the 417 tide gauge records that were assessed, 89 records had an ADr p-value of less than 0.05 when using the EQD method. By comparison, using the TAILS approach, we obtain only 17 model fits with ADr p-values less than 0.05.

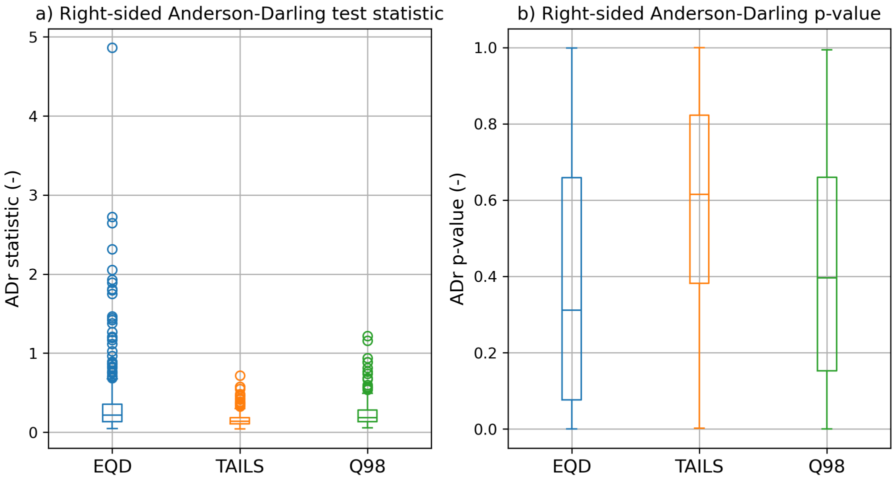

Figure 4.5.3: Box and whisker plots showing the results from applying an ADr test to all the exceedances over the thresholds selected using the EQD and TAILS approaches, as well as using a static Q98 threshold.

4.5.4 Distance metrics and return levels

As a further illustration, consider the two sites; Apalachicola in the US and Fishguard in the UK. The two sites have been selected based on the differences in geographic location and the associated extreme water level drivers, which lead to contrasting return level estimates. Apalachicola, located on the western coast of Florida in the Gulf of Mexico, is subjected to violent tropical cyclones which drive huge storm surges due to the large and shallow continental shelf (Chen et al., 2008; Zachry et al., 2015). The GPD model fit that characterises the return levels of the water level record therefore has a large positive shape parameter, which leads to a power-law growth in the return level curve. In contrast, Fishguard is located on the southern side of Cardigan Bay, near the inlet of the Irish Sea. The events driving extreme sea levels in this location are a combination of strong extratropical storms and astronomical tidal variation, which are characterised by a wholly different return level curve (Amin, 1982; Olbert and Hartnett, 2010). The GPD model fit for this record has a negative shape parameter, which leads to a plateau in the return levels as the return period increases. Figure 4.5.4 shows stability plots for the shape parameter for each site with the selected thresholds from the EQD and TAILS methods plotted as vertical lines. For both sites, the shape parameter looks approximately stable above both threshold choices. There is some deviation for Apalachicola at the largest quantiles due to the increased variability in parameter estimates and the effect of the largest extreme events which are driven by tropical cyclones. The return level plots and distance metrics for each site are shown in Figure 4.5.5.

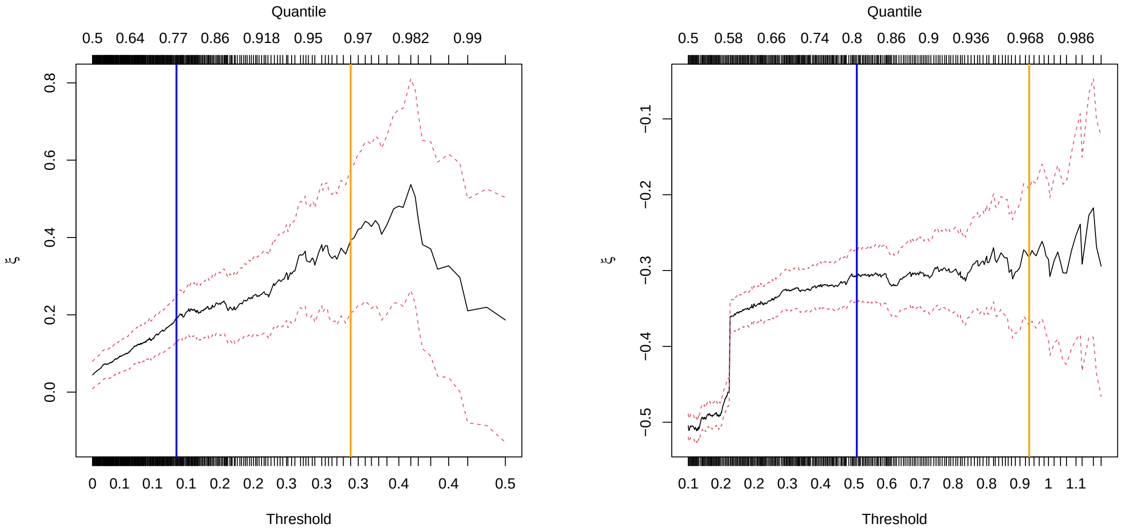

Figure 4.5.4: Parameter stability plots for two locations: Apalachicola, US [left] and Fishguard, UK [right]. The selected thresholds from the EQD and TAILS approaches are shown as blue and orange lines, respectively.

In the top row of Figure 4.5.5 (panels a and b), one can observe the EQD and TAILS distance metrics (i.e., expressions (4.3.2) and (4.4.1)) plotted as a function of the threshold probability for both tide gauge records. The global minima obtained from the two approaches are starkly different, as a result of the different quantile levels on which each metric is evaluated. Panels c and d of Figure 4.5.5 show the estimated return levels and 95% bootstrapped confidence intervals (incorporating GPD parameter uncertainty) from each of the TAILS, EQD and Q98 approaches, at Apalachicola and Fishguard, respectively.

In the case of Apalachicola, the minimum distance (panel a) obtained using the

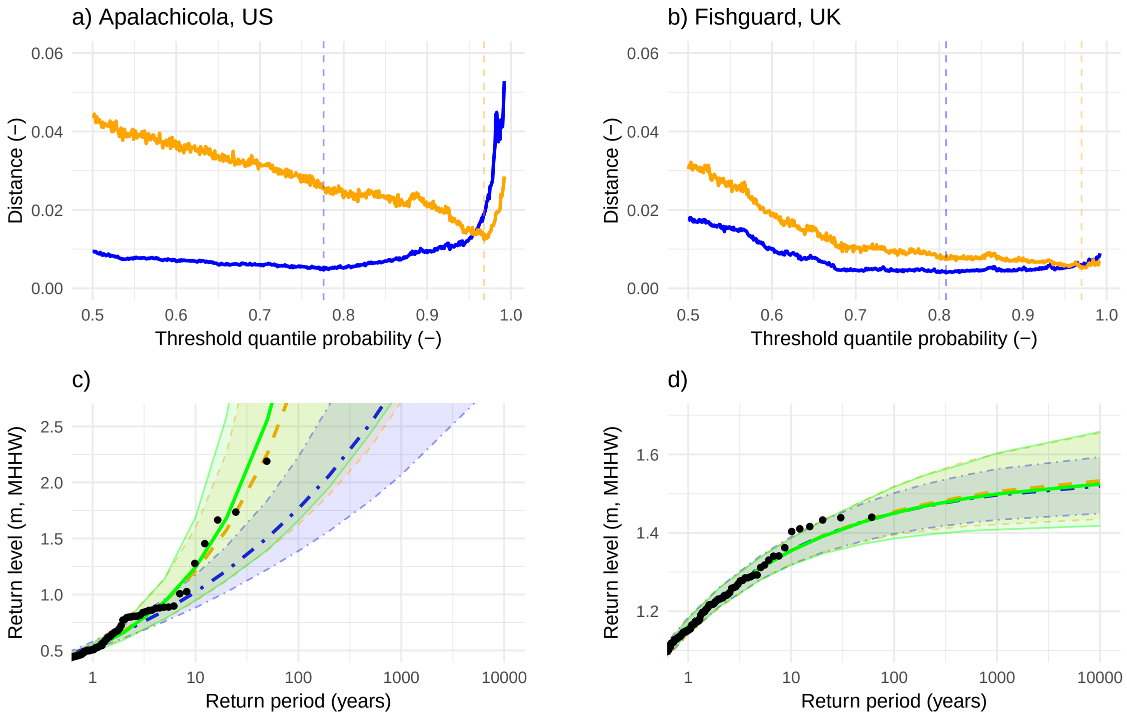

Figure 4.5.5: Model fits for two locations. Left column: Apalachicola, US (a and c). Right column: Fishguard, UK (b and d). The top row (a and b) shows the TAILS (orange) and EQD (blue) distance metrics, plotted as a function of the threshold probability. The vertical dashed lines indicate the distance minima, and therefore the selected threshold quantile probability. The bottom row (c and d) displays the unconditional return level estimates resulting from the EQD method (blue dot-dash), TAILS method (orange dashed) and the Q98 approach (green), with observations shown as black points. The shaded areas indicate the 95% bootstrapped confidence intervals incorporating GPD parameter uncertainty.

TAILS method (0.012) is more than double the minimum distance obtained using the EQD approach (0.005). Despite having a larger minimum distance, the TAILS approach captures the largest empirical observations much better than the EQD method. In fact, five of the historical events even lie outside of the 95% confidence interval for the EQD method. Contrast this with the results from Fishguard (panel b), where the minimum distances obtained using each approach are much more comparable; 0.005 for TAILS and 0.004 for the EQD approach. The resulting return level estimates (panel d) are also similar, with very small differences in the mean return levels between each of the three methods. The key difference observed in panel d is the uncertainty bounds, with the

EQD method having narrower uncertainty bounds at the higher return periods than the other two methods. However, all methods lead to a poor fit for some of the most extreme observations here. With the inclusion of threshold uncertainty, as directed in Chapter 3, the confidence intervals would likely cover all observations.

4.6 Discussion

In this work, we have introduced an automated threshold selection technique that addresses certain limitations of a leading existing approach in this particular context. This extension of Chapter 3 was driven largely by domain expectations where concern lies more closely on fitting the most extreme observations. This goal differs from Chapter 3 where we were concerned with providing an adequate fit to the data generating process, incorporating both the GPD model fit and sampling variability into the threshold choice. In this chapter, we have utilised those key aspects of the EQD metric but focussed the metric evaluation on the upper tail to ensure adequate fit to the most extreme observations. Using a global tide gauge dataset, both methods have been rigorously compared alongside a commonly used static threshold. We have examined the spatial patterns present in the differences between the TAILS and EQD approaches, and tested the resulting GPD model fits using an ADr test. Two tide gauge records have been investigated in more detail, to highlight the differences in the EQD and TAILS distance metrics, and to demonstrate how the parameter uncertainty changes between the different approaches.

4.6.1 Comparisons to existing approaches

At all locations, the TAILS method selects higher thresholds than the EQD approach. Particularly large increases are observed in Europe and the Baltic regions, as well as South Australia. The processes driving these increases are likely multifactorial. In the Baltic Sea, for example, the tidal range is very small (less than 10 cm in some locations). This makes any non-tidal variability in sea level much larger relative to the daily oscillation of the sea level due to tide. This could have an impact on the EQD approach, although it is unlikely to explain all the differences. Other regions in the world also have small tidal ranges, such as the Mediterranean and Gulf of Mexico, and yet these areas do not show such large increases in the quantile probabilities selected by the TAILS method relative to the EQD. Factors that could affect the selected thresholds include the meteorological forcing type (i.e. tropical cyclone vs extra-tropical storm) and the dominant driver of extreme water levels in a particular location (e.g. storm surge, waves or tides), but determining the impacts of each of these remains beyond the scope of this study.

Regardless of why individual differences occur, we demonstrate that in most cases in this coastal flood context, the methodological differences of the TAILS approach lead to more accurate GPD fits for the most extreme observations, compared to the EQD technique, and improvement over the commonly-used static Q98 threshold (when assessed using an ADr test). Furthermore, when applying methods to a large number of sites, employing an automated procedure, like the TAILS or EQD methods, avoids the need for manual checks on individual threshold choices. The TAILS method guarantees that the resulting model fits will be driven by data observed in the upper tail, which is desirable for practical applications where estimation of extreme quantities (e.g., return levels) is required. We also believe that calibrating threshold selection to focus on the upper tail will encourage more practitioners to adopt our approach, since we are more likely to obtain a model fit that accurately captures the observed tail behaviour.

4.6.2 Sensitivity to extreme observations and uncertainty

Focusing the model fit to the upper tail comes at the cost of additional uncertainty, since by definition, less data is available for inference. Since uncertainty quantification is a key focus of the approach proposed by Murphy et al. (2025), the EQD technique will generally offer lower model uncertainty than TAILS. In other applications, this may be more desirable than capturing the most extreme observations. Thus, when deciding whether to use the EQD or TAILS methods, one must consider the following question: is it more important that the model is more certain and robust, or that the model better captures the most extreme observations? We recommend that practitioners consider this question within the context of their application before selecting a technique.

For the application demonstrated in this paper, acknowledging and embracing uncertainty is key for any practitioner, despite the focus here on fitting the largest observations. Take the example of Apalachicola, US given in Section 4.5.4. This region is impacted by tropical cyclones, making the return level estimates made from the historical record very uncertain. To illustrate this point, two major Category 4 hurricanes (Helene and Milton) made landfall on the west coast of Florida in September and October 2024, after the GESLA 3.1 update was collated. Preliminary data recorded during the event suggest that Hurricane Helene broke the highest recorded water levels at three tide gauges located in Florida, and Hurricane Milton set the second highest water level ever recorded at the tide gauge located in Fort Myers, US (Powell, 2024a,b). Fitting distributions to these records pre and post these events results in different mean return levels being estimated, especially when considering the most extreme return periods (e.g., the 1 in 500 year event). We tested this and found that, when using the TAILS approach, the mean return level for the 1 in 500 year event increased by 55 cm if the tide gauge record is extended beyond the GESLA 3.1 update, to include these events. By recognising the uncertainty in the underlying processes and the uncertainty inherent in the estimates made from observations, we can be more confident that our models will be able to capture extreme events which are yet to occur. It must be noted that when estimating confidence intervals for the return level estimates shown Figure 4.5.5, we did not account for the uncertainty in the threshold estimation, as recommended by

Murphy et al. (2025). Accounting for this aspect of uncertainty using the Q98 threshold is not possible and so to allow comparison between the three approaches, we omitted this additional feature.

4.6.3 Incorporating more complex characteristics within the TAILS approach

Throughout this work, we make the implicit assumption that data are IID, even though we acknowledge that this is unrealistic for environmental processes such as sea levels in an ever-changing climate. This choice was motivated by practical implications; IID models are simpler to implement and best practices (i.e., using a POT model) are well established. However, there is no reason why the TAILS framework could not be expanded to incorporate more complex models. When applying simple IID models to such contexts, the TAILS approach may be favoured as the generally higher threshold choices should remove some of the complexity, leading to more accurate fits. However, there is well-established methods for incorporating covariate dependence into the threshold and parameters of a GPD (Davison and Smith, 1990; Chavez-Demoulin and Davison, 2005; Youngman, 2019). This aspect could be a reason for the poor fit in the upper tail for the EQD in Figure 4.3.2. Accounting for this aspect in the EQD or TAILS approach could allow for the use of lower thresholds without the loss of accuracy for the more extreme observations, providing a way to balance between the two goals mentioned above, i.e., uncertainty and accuracy in the upper-tail.

For example, one could extend the method to include non-stationary data by allowing the GPD parameters to be functions of time or covariates. A wide range of modelling approaches have been proposed for this purpose (Eastoe and Tawn, 2009; Sigauke and Bere, 2017; Youngman, 2019; Mackay and Jonathan, 2020). Relevant covariates are those that impact the number of extreme events that occur within a given year; for example, indices related to the ENSO and NAO phenomena, which affect the likelihood of temperature and precipitation extremes (Dong et al., 2019), could be incorporated. Only minor modification would be needed to apply the TAILS approach here; specifically, we would assess quantile discrepancies on a transformed scale, rather than the observed scale (see Varty et al. (2021) for related discussion). However, we note that standard practices for applying non-stationary POT models are not well established, and it is not clear how one should select which covariates to include or how much flexibility to include within the model (for example). Therefore, the development of automated threshold selection approaches for non-stationary data structures represents an important line of future research.

We also remark that we assume a constant baseline event for our approach. Future work could incorporate a variable baseline event, which is linked to the underlying forcing mechanisms in an area. As discussed in Section 4.5.4, tide gauges around the world are characterised by different patterns of extreme water levels. It might be possible to link a dominant forcing type to the baseline event, which could improve the ability of TAILS to capture the tail behaviour in the estimated return levels.

We also expect our automated selection technique to be useful for improved threshold estimation in the wider context of multivariate and spatial extremes. The method could be applicable, with suitable adjustment, to cases relying on multivariate regular variation assumptions such as Wan and Davis (2019) or for variables exhibiting asymptotic independence (Heffernan and Tawn, 2004). The data-driven approach would allow for the threshold estimation uncertainty to be propagated through to the joint tail inferences. Possibly of even more importance for this context, is the natural extension from multivariate to spatial modelling (Shaby and Reich, 2012; Richards et al., 2022).

4.6.4 Incorporating threshold uncertainty

While results may indicate in certain examples that the Q98 approach outperforms the EQD, the benefits of a data-driven approach can not be understated. When relying on TAILS or the EQD, not only is the threshold justified by a goodness-of-fit measure but sampling variability has also been taken into account. This leads to a well-justified threshold choice and an easier characterisation of the uncertainty in resulting estimates. It also allows for the uncertainty in the threshold choice to be incorporated when making inference; see Chapter 3. As mentioned in Section 4.6.2, we omitted this aspect of uncertainty in our confidence interval estimation for this work. As shown in Chapter 3, including this additional uncertainty results in well-calibrated confidence intervals. In our context, including this uncertainty may result in better capture of extreme events yet to be observed (see Section 4.6.2), and should provide a better understanding of the uncertainty in return level estimates beyond the observed data, which can contribute to improved decision-making for future hazard mitigation.

4.6.5 Selecting tuning parameters

TAILS requires the selection of several non-trivial tuning parameters; this includes probability set $\mathcal{P}$, m, B, and the limit on candidate thresholds, which we define as the 1-year return level in Section 4.4. Our choices were motivated by the specific application at hand, and we consequently recommend that practitioners experiment with these parameters to assess whether such values have a practical effect on the resulting model, using diagnostics such as QQ and return level plots to guide this procedure. The code has been written in such a way as to make it easily parallelised, allowing for fast testing of multiple baseline and maximal probabilities across a variety of datasets. We encourage and invite fellow researchers to utilise this method on other applications, such as rainfall or river flow measurements. Exploring data-driven techniques (e.g., cross validation) for selecting tuning parameters of automated threshold selection approaches represents an important line of future research.

4.7 Conclusions

Accurately estimating the extreme tail behaviour of historical observations is of great importance to researchers and practitioners working in natural hazards. POT methods are regularly used in these fields for this purpose, but selecting the threshold above which to consider an exceedance requires careful consideration. In this paper, we present TAILS, a new method for automating the threshold selection process building upon the recently published EQD method (Murphy et al., 2025).

We apply two key innovations to improve upon the EQD method in the context of extreme coastal sea levels. Firstly, we fix the quantiles that we consider when computing the distance metrics. This avoids oversampling the most extreme quantiles when assessing higher thresholds. Secondly, we limit the quantiles considered for our distance metric to be only above a predetermined baseline probability. This means that when optimising the distance metric to select a threshold, we are only considering quantiles that we deem to be extreme, and hence worth considering when selecting a threshold. Here, the baseline probability was decided using the literature and a sensitivity test.

We show that the TAILS approach selects, on average, higher thresholds than the EQD method. When the resulting model fits are evaluated using an ADr test against the EQD and Q98 approaches, the TAILS method outperforms both with respect to the ADr test statistic and the p-value. We also illustrate that the TAILS method typically results in larger uncertainty bounds, but argue that when considering water level records located in regions that experience tropical cyclones, this can be useful in adequately representing the uncertainty of extremes that are more appropriately modelled by a mixture distribution.

We applied the methods to a large number of tide gauge records. We expect that the TAILS method may also be useful to better estimate the intensities and frequencies of other natural hazards. The code has been written in such a way as to make it easily accessible and easily parallelised so as to encourage uptake from fellow researchers.

Chapter 5

Spatio-temporal modelling for extreme induced seismicity in the presence of an evolving measurement network

5.1 Introduction

Extraction or injection of gases from/into underground reservoirs of porous rock cause poroelastic deformations in the subsurface, which can lead to seismic activity, known as *induced seismicity* (Majer et al., 2007; Suckale, 2009). Where these underground reservoirs are located in populated areas, there are significant risks of public safety and damage to infrastructure (Ellsworth, 2013), so accurately estimating the distribution of induced seismicity magnitudes under future extraction or injection scenarios is of paramount importance. Such estimates can be used to inform the construction or reinforcement of infrastructure and keep the hazard associated with induced seismic events at an acceptable level. A key example of a region where this induced seismicity is prevalent is the Groningen gas field in the Netherlands, one of the largest gas fields globally. Here, extraction has now stopped (despite substantial remaining gas reserves) but earthquakes continue to occur and there is still debate on how best to determine the funds needed to mitigate against future earthquake damage.

To aid decision making on the future risks linked to the Groningen gas field, we must accurately estimate the distribution of the magnitudes of seismic events, with particular interest in the values beyond those that have already been observed. Two specific quantities are typically given focus. Firstly, the magnitude with a 90% probability of occurrence over a 50-year span is widely-used in the design of earthquake-resistant infrastructure (Code, 2005), which if earthquakes were identically-distributed over time, corresponds to a 475-year return level. The second is the largest possible earthquake within the region, denoted M_{max}, used to address concerns of worst case scenarios. For estimating M_{max}, there are purely statistical models (Beirlant et al., 2019) as well as substantial geophysical literature (McGarr, 2014; Galis et al., 2017; Weng et al., 2021). Whichever approach is used, high-quality earthquake data catalogues are required.

Unlike tectonically-driven earthquakes, the largest induced earthquakes are small in magnitude, but they occur at shallow depths and so, can still cause significant damage locally in relation to their epicentres. Catalogues of observed events typically have small sample sizes and cover a limited time-window $\mathcal{T}$ and spatial region $\mathcal{X}$. Thus, it is vital to make use of all of the reliable available information.

The key challenge with induced seismicity is that small magnitude events often go undetected. Induced seismic events are located and measured by a network of geophones spread across the region of interest. Earthquakes are only detected if their magnitude is sufficiently large such that its location may be identified by the geophone network. Investment in this network over time has improved the detection ability which, in turn, sheds light on the occurrence rate of earthquakes which were undetected during periods when the network was too sparse or insensitive to detect such events. In seismicity studies, a key quantity is the *magnitude of completion*, denoted by $m_c(\boldsymbol{x}, t)$, which is the smallest earthquake magnitude which can be detected with certainty if it occurs at a location $\boldsymbol{x} \in \mathcal{X}$ and time $t \in \mathcal{T}$. As the value of $m_c(\boldsymbol{x}, t)$ relates to the density and sensitivity of the geophones around location $\boldsymbol{x}$ at time t, it can only be estimated using the observed earthquake catalogue.

Only detected earthquakes with magnitudes that exceed the estimated magnitude of completion, $\hat{m}_c(\boldsymbol{x}, t)$, for that earthquake's estimated location $\boldsymbol{x}$ and time t should be used in the subsequent analysis of seismic rate changes, static and dynamic triggering, mapping of seismicity parameters, earthquake forecasting, and probabilistic seismic hazard assessment (Mignan et al., 2011). Therefore, efficient estimation of the magnitude of completion function m_c is vital. In estimating m_c, there is a bias-variance trade-off which affects design level and $M_{\max}$ inferences; too low an estimate incorporates levels impacted by undetected earthquake values (which biases inferences); while over-estimation excludes valid data (leading to unnecessary variance). The simplest inference methods for the function m_c assume that the function is constant over $\mathcal{X} \times \mathcal{T}$ but this is potentially highly inefficient for regions with a geophone network that has evolved substantially across $\mathcal{T}$, as it makes the investment redundant. Therefore, it is essential to estimate the function m_c as well as possible, which means estimating its spatial-temporal changes.

Almost all existing methods for estimation of $m_c(\boldsymbol{x}, t)$ follow Ogata (1988) by assuming that the true values of earthquake magnitudes (detected or not) are realisations of an independent and identically distributed (IID) variable, with distribution function denoted by F. For a given choice of statistical model for F, $m_c(\boldsymbol{x}, t)$ is estimated as the largest level where there is not a statistically significant departure of the fitted distribution F from the empirical distribution (using the catalogue values that fall within a selected neighbourhood of space and time of $(\boldsymbol{x}, t)$). This whole process is complicated as (i) F is unknown and even an appropriate parametric family for F is unclear, (ii) the earthquakes missing from the catalogue make it difficult to fit F accurately, and (iii) the results will be sensitive to the choice of the neighbourhood.

The typical choice for F is an exponential distribution, in line with the Gutenberg-Richter law for seismicity (Gutenberg and Richter, 1956). Various simplifying assumptions for the form of the function m_c have been used: Mignan and Woessner (2012) take it to be constant over $\mathcal{X} \times \mathcal{T}$; Hutton et al. (2010) and Das et al. (2012) both take a piecewise-constant function, with changepoints at pre-determined timepoints and spatial regions respectively; whilst Woessner and Wiemer (2005) allow for short-term increases following large magnitude earthquakes. These methods are likely to be inefficient as (i) knowledge of the geophone network is not exploited, (ii) the exponential distribution is known to result in inadequate fits to the data in the upper tail, (iii) the assumption that the earthquakes are identically distributed is likely too simplistic for induced earthquakes with evidence of space-time variations in the distribution due to changes in the incremental stress fields (Bourne and Oates, 2020; Richter et al., 2020).

Mignan et al. (2011) focussed on estimating $m_c(\boldsymbol{x})$ as constant over time, but overcome aspects of the first two of the above inefficiencies by employing a two-step Bayesian mapping procedure. They use a non-parametric method to develop provisional estimates and merge this with prior information about $m_c(\boldsymbol{x})$ based on a regression model $\phi_1[V_i(\boldsymbol{x})]^{\phi_2} + \phi_3$, where $V_i(\boldsymbol{x})$ is the surface distance from $\boldsymbol{x}$ to the i^{th} nearest geophone, and (ϕ_1, ϕ_2, ϕ_3) are parameters. They explore $i = 3-5$, to ensure accurate measurement of earthquakes for their monitors, and choose $i = 5$ as best for their analysis. Mignan et al. (2011) assume an exponential distribution for the excesses of their spatially varying m_c function. Other work has addressed concerns about the exponential distribution being unbounded in its upper tail, which does not align with physical understanding that the seismic energy which can be released in any region must have a finite upper bound, and hence $M_{\max} < \infty$. Potential approaches to address this include the truncated exponential distribution (Raschke, 2015) and the tapered Gutenburg-Ritcher distribution (Vere-Jones et al., 2001). Yue et al. (2025b) discuss other distributions that have been proposed which exhibit truncation and tapering relative to the exponential distribution. None of these distributions have a mathematical or physical justification, so even if they fit well in certain applications, there is no basis to believe that they are suitable in other regions.

For the Groningen gas field data, Varty et al. (2021) propose a range of new developments for estimating m_c as a solely temporally-varying function and for the estimation of its associated excess distribution. This research builds upon that work as we set out below.

Varty et al. (2021) model F utilising extreme value methods, using the generalised Pareto distribution (GPD) to be precise. This choice of distribution seems a well-justified assumption because (i) the Gutenberg-Ritcher law (i.e., the exponential distribution) is known to be a good approximation for seismic processes and it is a special case of the GPD; (ii) Varty et al. (2021) found strong empirical evidence for the GPD, (iii) the exponential distribution tends to overestimate the upper tail of earthquake magnitudes due to the infinite upper-endpoint, in contrast to the physical upper limit of induced earthquakes, and (iv) the GPD is known to provide a flexible, parsimonious model with a finite upper-endpoint (for $\xi < 0$) allowing accurate estimation of the upper tail including levels beyond what has previously been observed (Coles, 2001).

Varty et al. (2021) propose an automated method for estimating $m_c(t)$ as the threshold for the GPD. A variant of this method was shown to outperform the leading threshold selection techniques for fitting a GPD in the IID context in Chapter 3. Finally, they assume that $m_c(t)$ follows a parametric sigmoid function over time, with four parameters. This choice of parametric form was somewhat arbitrary but accounted for empirical evidence of a smooth transition between time periods with constant values. Although information about changes in the Groningen geophone network were not used in the Varty et al. (2021) analysis, the estimate found changes which were consistent with periods of known investment into the network and almost doubled the number of excesses over $m_c(t)$ relative to the established best constant estimate of this function.

Here, we focus on developing the first automated method for making inference on the spatial-temporal function $m_c(\boldsymbol{x}, t)$, which explicitly accounts for the known details of the evolving geophone network for the Groningen gas field, with the resulting estimate exhibiting clear spatio-temporal variations which reflect the geophone network. We propose to estimate $m_c(\boldsymbol{x}, t)$, using the threshold of a covariate-dependent GPD, as motivated in Varty et al. (2021). To simplify notation and align with standard terminology for the GPD threshold, we subsequently use $u(\cdot)$ in place of $m_c(\cdot)$ for our inference, but refer to existing methods as estimating $m_c(\cdot)$.

Our work builds upon that of Varty et al. (2021) and Murphy et al. (2025) to provide an automated threshold selection procedure that accounts for different functional forms linking $u(\boldsymbol{x}, t)$ to a spatio-temporal extension of the covariate $V_i(\boldsymbol{x})$ of Mignan et al. (2011) and to decide on the best choice over i. Following Varty et al. (2021), we assume that F is a GPD, but we are the first to exploit, during threshold selection, the knowledge of Bourne and Oates (2020) that the parameters of F vary with the changing stresses due to gas extraction. Using a key geophysical stress covariate, we can estimate the rate of occurrence of earthquakes exceeding different magnitudes into the future under various extraction scenarios.

Critically, there is novelty from the perspectives of both earthquake modelling and extreme value methods, as we are the first to explore the impact of the uncertainty in the estimation of the magnitude of completion or equivalently the threshold function, $u(\boldsymbol{x}, t)$, on the subsequent tail inferences of induced earthquake magnitudes. In particular, we evaluate the additional uncertainty as a result of the unknown threshold function as well as the unknown formulation of the threshold function with covariates.

The chapter is structured as follows. In Section 5.2, we present details of the Groningen seismic data and the geophone network. Section 5.3 provides background of the extreme value methods that underpin the use of the GPD and the associated threshold selection approaches that we build on. In Section 5.4, we present our statistical methods for exploiting the knowledge of the geophone network to derive estimates of both $u(\boldsymbol{x}, t)$, the distribution of excesses over $u(\boldsymbol{x}, t)$ which varies with both $\boldsymbol{x}$ and t, and the underlying intensity function of true earthquakes over $\mathcal{X} \times \mathcal{T}$. Within our methodology, we allow for a range of functional forms that link V_i and the choice of i to $u(\boldsymbol{x}, t)$. In Section 5.5, we propose methods to account for the uncertainty in the estimation of $u(\boldsymbol{x}, t)$ in the subsequent inferences. For the Groningen data, we illustrate our methods for model inference and selection, and demonstrate the improved performance relative to the widely-adopted conservative estimate of the magnitude of completion in Section 5.6. We finish with a discussion in Section 5.7.

5.2 Groningen data

5.2.1 Earthquake data and existing estimates of $m_c(t)$

The Groningen earthquake catalogue covers the period, $\mathcal{T}$, from April 1995 to January 2024 and consists of $n = 1565$ seismic occurrences within the region of interest, $\mathcal{X}$, which has been determined by practitioners. Events in this region pose a significant hazard and should be included in any analysis of the risks associated with the gas field, $\mathcal{G}$, where $\mathcal{G} \subset \mathcal{X}$, see Figure 5.2.1. The catalogue includes the event time, hypocentre (a three-dimensional location of surface position and depth), and magnitude, denoted respectively by $(t_k, \boldsymbol{x}_k, y_k)$ for $k = 1, \ldots, n$. The hypocentres $\boldsymbol{x} = (x_1, x_2, x_3)$ are given as two-dimensional RD coordinates (a grid-based planar projection of locations across the Netherlands) and a corresponding depth with the majority of depths of 3-4 km. The magnitudes are recorded on the local magnitude ($\mathrm{M_L}$) scale, a logarithmic scale used to measure the energy released by an earthquake. Typically, catalogues report magnitudes to one decimal place, resulting in a dataset of rounded observations, an added modelling challenge considered by Varty et al. (2021). Like Yue et al. (2025b), we use a catalogue with magnitudes reported to at least two decimal places.

Figure 5.2.1 shows the event magnitudes over time and their spatial locations separately. The lines on the temporal panel of this figure represent two previously-used formulations of $m_c(t)$: a constant level of $m_c = 1.45\mathrm{M_L}$ (Dost and Kraaijpoel, 2013) and a piece-wise constant function with a single changepoint at 2015-12-25 (Yue et al., 2025b). The former is accepted as a conservative estimate, after accounting for rounding, whereas the latter is a special case of the sigmoid function of Varty et al. (2021).

The temporal panel shows that the rate of occurrences of recorded earthquake magnitudes in the catalogue has increased over $\mathcal{T}$. In fact, the rate per year has almost doubled, from 46.33 to 81.25 before and after the changepoint in the piece-wise constant $m_c(t)$ function. Above the conservative estimate of $m_c(t) = 1.45\mathrm{M_L}$, the rates per year are very similar, suggesting the primary changes are seen in the occurrence rates of smaller earthquakes. To help us understand the nature of the change, we calculate empirical estimates of the probabilities (and standard errors based on asymptotic normality) of recording an earthquake below $0.76\mathrm{M_L}$ (i.e., the value of $m_c(t)$ used by Varty et al. (2021) after 2015-12-25) for t before and after this changepoint; these are 0.187 (0.013) and 0.448 (0.020) respectively. There are two potential reasons for this type of change: improvements to the geophone network and the impact of gas extraction. We explore each of these aspects in Sections 5.2.2 and 5.2.3 respectively. As the changing rate is specifically related to small magnitude events, it would appear that the former aspect is likely the dominant factor influencing this feature of the catalogue. The spatial panel of Figure 5.2.1 shows that most of the detected earthquakes occurring within $\mathcal{X}$ are located within $\mathcal{G}$, and furthermore, there are clear sub-regions of $\mathcal{G}$ where the earthquake activity is focussed.

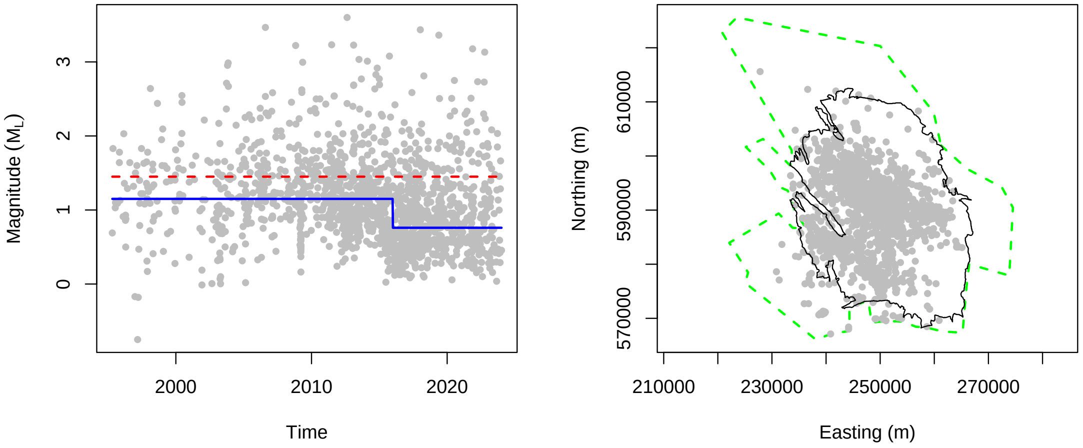

Figure 5.2.1: Temporal and spatial features of the induced earthquakes in the Groningen catalogue: [left] magnitudes (in $\mathrm{M_L}$) against date of occurrence with two previously used $m_c(t)$: a conservative level of $m_c = 1.45$ (red-dashed line) and a changepoint (solid blue line). [right] occurrence locations in the region of interest $\mathcal{X}$ (green-dashed line) with gas field $\mathcal{G}$ (solid black line).

5.2.2 The geophone network and the magnitude of completion

The Royal Netherlands Meteorological Institute (KNMI) measure seismic activity across the Netherlands through an extensive network of geophones (KNMI, 2020). For this study, a geophone dataset is available containing their locations (in RD coordinates), depths and dates of operation over the Netherlands, a subset of which, denoted $\mathcal{R}$, is plotted in Figure 5.2.2, with $\mathcal{X} \subset \mathcal{R}$. This dataset was not available to Varty et al. (2021), they only had access to knowledge of the time-window of the major developments across the network between 2014-17.

Figure 5.2.2 illustrates the drastic change in the network over $\mathcal{T}$ in terms of the number of geophones and their spatial coverage in regions $\mathcal{G}, \mathcal{X}$ and $\mathcal{R}$. Temporally, we see a slow growth in geophones prior to 2014, then a massive growth in the period 2014-17. There have also been some smaller changes post 2017 for $\mathcal{G}$ and $\mathcal{X}$, and ongoing evolution in the network in $\mathcal{R}\backslash\mathcal{X}$. For assessing the spatial evolution of the network, we show the locations of the geophones which were in operation in 2010 and

2020 separately. These snapshots of the geophone network show that geophones are not located uniformly across space, and the network expands at different rates over different regions. Specifically, geophones are placed to achieve adequate spatial coverage with a focus on areas which have seen a high intensity of earthquakes and extraction rates, see Figures 5.2.1 and 5.2.3 respectively. There are even geophones outside of the region $\mathcal{X}$, with their locations selected to improve detection of events occurring in $\mathcal{X}$.

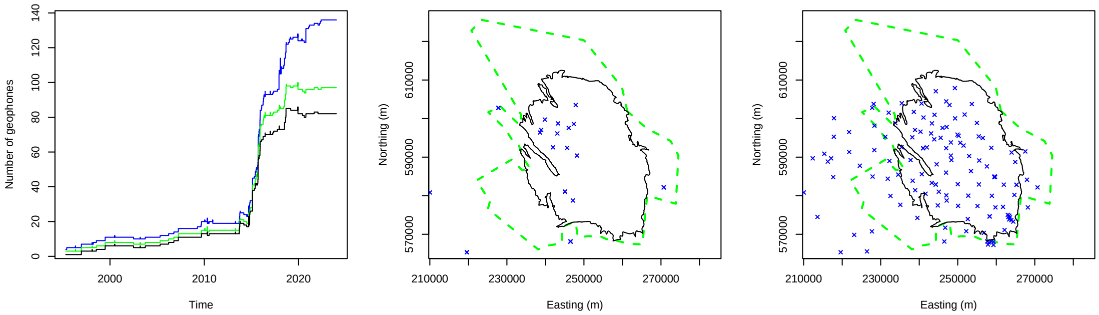

Figure 5.2.2: Features of the Groningen region geophone network present over time and space: [left] number of geophones in operation daily throughout $\mathcal{T}$ within the regions $\mathcal{R}$ (blue), $\mathcal{X}$ (green) and $\mathcal{G}$ (black); [centre] and [right] the respective locations (blue crosses) of geophones in operation in the years 2010 and 2020.

We exploit more information about the Groningen geophone network than Mignan et al. (2011) did in their study for Taiwan, which only considered the surface distance from the i^{th} nearest geophone by $V_i(\boldsymbol{x})$. We improve understanding of the spatio-temporal variation in detection capability by measuring such distances in three-dimensional space and incorporating operation times of individual geophones, with an updated measure given by $V_i(\boldsymbol{x}, t)$. The inclusion of the depth data is particularly helpful for capturing differences in distances when V_i is small due to the vertical resolution not available when measuring surface distances.

We use $V_i(\boldsymbol{x}, t)$, over $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}$ for a range of i, to provide our covariate information for estimating $u(\boldsymbol{x}, t)$. Unlike Mignan et al. (2011), we tie this information into our inference for a parametric model for the distribution of exceedances of $u(\boldsymbol{x}, t)$ and we also consider values of $i < 3$ as these may provide a better reflection of the magnitude of completion. Using $V_i(\boldsymbol{x}, t)$ in this way opens up the first possibility for spatial and temporal inference for $u(\boldsymbol{x}, t)$, and it enables a lowering of $u(\boldsymbol{x}, t)$ for some t and $\boldsymbol{x}$ relative to previous studies at Groningen. This should improve efficiency of all subsequent inferences based on the exceedances of the estimated magnitude of completion function. Although the total number of geophones has increased in $\mathcal{R}$ over $\mathcal{T}$, it doesn't necessarily mean that $V_i(\boldsymbol{x}, t)$ has decreased similarly over time, for any specific location $\boldsymbol{x}$, as this depends on the local configuration of the geophones near $\boldsymbol{x}$.

5.2.3 Extraction stress covariate and the intensity inference

We have access to other physical covariates from the Groningen field. These covariates, which are on a grid over $\mathcal{X} \times (\mathcal{T} \cup \mathcal{T}_F)$, where $\mathcal{T}_F$ is the period from February 2024 - January 2055, with the covariates for this future time period, relative to the catalogue, derived under the assumption of no further extraction from the gas field. The covariates arise from physics-based reservoir models calibrated using measurements taken at boreholes and seismic imaging. Seismic activity is not observed at a particular location until the previous maximum stress level is exceeded since that stress will have already caused all feasible earthquakes at the location (Tang and Hudson, 2010; Zang et al., 2014). Bourne and Oates (2020), Smith et al. (2022) and Kaveh et al. (2024) all use the Kaiser stress (KS) which, for each $\boldsymbol{x} \in \mathcal{X}$, is the maximum of the difference up to time t, of the vertically averaged maximum stress from the initial stress state at the start of gas extraction in $\mathcal{G}$.

We denote the KS field for the two time periods of interest by $\mathcal{S} = \{s(\boldsymbol{x}, t) : s(\boldsymbol{x}, t) \geq 0, \boldsymbol{x} \in \mathcal{X}, t \in \mathcal{T}\}$ and $\mathcal{S}_F = \{s(\boldsymbol{x}, t) : s(\boldsymbol{x}, t) \geq 0, \boldsymbol{x} \in \mathcal{X}, t \in \mathcal{T}_F\}$. KS is a monthly covariate, presented here in units of MPa, and we take it to be constant throughout each respective month. For each event included in the catalogue, i.e., for $k = 1, \ldots, n$, we define $s_k = s(\boldsymbol{x}_k, t_k)$ as the value of the KS field $s(\boldsymbol{x}, t)$ at the grid point nearest to $\boldsymbol{x}_k$ at time t_k. Figure 5.2.3 [left] provides KS averaged over the year 2020 for a fixed depth of 3km, presented for all $\mathcal{G}$, with some values in $\mathcal{X}\backslash\mathcal{G}$ not shown as these values are zero. A comparison of the regions of highest KS with the spatial locations for the catalogued events shown in Figure 5.2.2 indicates that KS is likely to be a useful covariate for describing spatial variation in earthquake locations. For each $t \in \mathcal{T} \cup \mathcal{T}_F$, Figure 5.2.3 [centre] shows KS at a fixed depth of 3km averaged over the spatial region shown in Figure 5.2.3 [left] along with KS values at three individual locations on a north-south transect through $\mathcal{G}$. The values for $t \in \mathcal{T}_F$ are obtained under the assumption that no further extraction takes place in $\mathcal{G}$. This plot shows that, on average, the KS grows over $t \in \mathcal{T}$, and continues to grow at a slower rate for $t \in \mathcal{T}_F$ as the stresses stabilise across the region, with the temporal development varying over locations, with KS constant in $\mathcal{T}_F$ for some locations.

Figure 5.2.3 [right] gives a crude impression of the effect of KS on the magnitudes of the observed events. Specifically, we use (s_k, y_k) pairs such that $y_k > 1.45\mathrm{M_L}$, to ensure that no issues with magnitude of completion affect the relationship. Of the pairs which satisfy this constraint, we find the median KS and present the boxplots for the magnitudes with corresponding KS values below and above the median. The boxplots reveal that the typical size of the magnitudes increases with KS. Hence, in Section 5.4, we incorporate KS as a covariate into our model for the distribution of true magnitudes, a departure from the identical distribution assumption (Ogata, 1988).

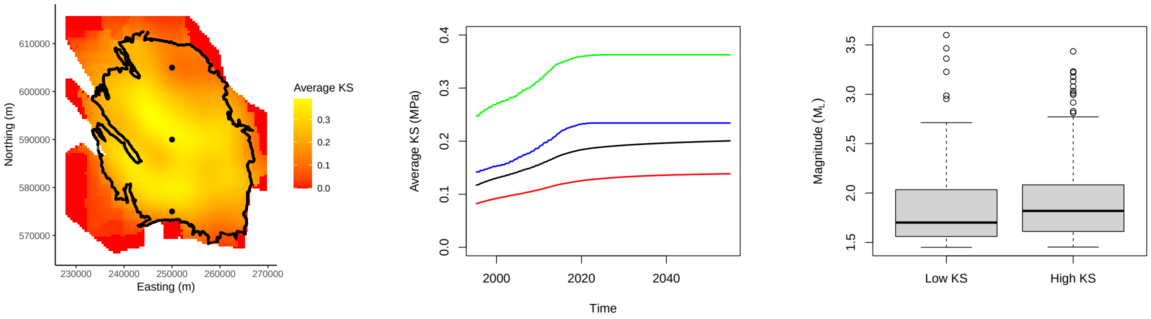

Figure 5.2.3: Features of the Kaiser stress covariate field: [left] Average Kaiser stress (KS) for 2020, for a fixed depth of 3km, across all of $\mathcal{G}$ (solid black line) and earthquake active region $\mathcal{X}\backslash\mathcal{G}$. [centre] temporally varying spatial average of KS over region shown in the left panel (black) and location specific $s(\boldsymbol{x}, t)$ series for three sites shown at dots on left panel: top (red), middle (green) and bottom (blue). [right] Boxplots of magnitudes greater than $m_c = 1.45\mathrm{M_L}$ corresponding to KS above and below the median KS value associated with magnitudes above $m_c = 1.45\mathrm{M_L}$.

To avoid issues with missing observations of earthquakes, Bourne et al. (2018) take the conservative estimate of the magnitude of completion, i.e., $m_c = 1.45\mathrm{M_L}$, and estimate the rate of occurrence of earthquakes above this level over $\mathcal{X} \times \mathcal{T}$. They estimate this intensity, which we denote by $\lambda_{1.45}(\boldsymbol{x}, t)$, using a parametric geophysics-based model. When the gas reservoir is of constant thickness over $\mathcal{X}$, this model simplifies to

$$\lambda_{1.45}(\boldsymbol{x}, t; \gamma_0, \gamma_1) = \frac{\partial s(\boldsymbol{x}, t)}{\partial t} \exp[\gamma_0 + \gamma_1 s(\boldsymbol{x}, t)], \text{ for } (\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}, \tag{5.2.1}$$

with parameters $(\gamma_0, \gamma_1) \in \mathbb{R}^2$, where the temporal partial derivative term for $s(\boldsymbol{x}, t)$ is evaluated using finite differencing. Bourne and Oates (2017) estimate the parameters (γ_0, γ_1) of $\lambda_{1.45}(\boldsymbol{x}, t)$ using the subset of catalogue values $\{(t_k, \boldsymbol{x}_k) : y_k > 1.45, k = 1, \ldots, n\}$ by making the assumption that these are realisations of a Poisson process with intensity $\lambda_{1.45}(\boldsymbol{x}, t)$ over $\mathcal{X} \times \mathcal{T}$. The intensity model (5.2.1) increases with both $s(\boldsymbol{x}, t)$ and its temporal derivative (when this is positive), with Figure 5.2.3 [centre] providing insight into how KS varies over three selected locations, e.g., for the southerly and centre locations KS has been constant since 2021 which would imply, under model (5.2.1), that earthquakes could not occur at these locations after 2021.

One limitation of the KS covariate is that it does not account for changes in the stresses in either $\mathcal{S}$ or $\mathcal{S}_F$ that arise locally in time and space as a consequence of earthquakes, i.e., the fact that earthquakes can increase the intensity for further aftershocks near the original location. As these additional local stresses are not incorporated into the available KS covariate, this leads to the possibility of observing induced after-shock events in regions where the available covariates suggest that an earthquake should be impossible. This is not of major concern for Groningen, as when the classification method of Zaliapin et al. (2008) was applied to the Groningen catalogue, only 20% of all earthquakes were classified as after-shocks. If interest lies in modelling occurrence rates of main- and after-shock events, then the current approach is to use epidemic-type aftershock sequence (ETAS) models (Ogata, 1988), however, this is not needed for our purposes.

5.3 Underpinning extreme value methods

5.3.1 Distributional model

Consider a univariate random variable Y with continuous distribution function F, and upper endpoint $y^F := \sup\{y : F(y) < 1\}$ and threshold $u < y^F$. Under weak assumptions on F, an asymptotic argument justifies the use of the generalised Pareto distribution (GPD) as a model for the conditional distribution function, $F_u(y)$, of excesses of a high threshold u, where $F_u(y) = [F(u + y) - F(u)]/[1 - F(u)]$ for $y > 0$. Specifically, (Pickands, 1975) shows that as $u \to y^F$, if there exists a function $a(u) > 0$ such that $F_u(a(u)y)$ is non-degenerate in the limit, then $F_u(a(u)y) \to G(y)$, where

$$G(y; \sigma, \xi) = 1 - (1 + \xi y/\sigma)_+^{-1/\xi}, \tag{5.3.1}$$

with $y > 0$, $w_+ = \max(w, 0)$, the shape parameter $\xi \in \mathbb{R}$, which is determined by F, and the scale parameter $\sigma \in \mathbb{R}_+$. This distribution is denoted by $\text{GPD}(\sigma, \xi)$.

Applying limit distribution G as an approximation for the excesses over a threshold u, with $u < y^F$ being a high quantile of F, leads to a statistical model for the tail of F, popularised by Davison and Smith (1990) and Coles (2001), given by:

$$F(y) = 1 - \lambda_u[1 - G_u(y - u; \sigma_u, \xi)], \tag{5.3.2}$$

for $y > u$ with unknown threshold exceedance rate, scale and shape parameters $\boldsymbol{\theta}_u := (\lambda_u, \sigma_u, \xi) \in [0, 1] \times \mathbb{R}_+ \times \mathbb{R}$. The upper tail behaviour of the GPD is determined by the value of ξ: $\xi = 0$ (taken as the limit as $\xi \to 0$) gives the exponential distribution, $\xi > 0$ gives an unbounded distribution with power law decay, and for $\xi < 0$, the distribution has a finite upper bound $u - \sigma_u/\xi$. The $(1 - p)^{\text{th}}$-quantile of Y, denoted by $y_p(\boldsymbol{\theta}_u)$, where Y has a GPD tail, satisfies $\Pr(Y \leq y_p; \boldsymbol{\theta}_u) = 1 - p$. So, if $p < \lambda_u$, i.e., $y_p > u$, then

$$y_p(\boldsymbol{\theta}_u) = \begin{cases} u - \sigma_u \left[1 - (p/\lambda_u)^{-\xi}\right]/\xi & \text{for } \xi \neq 0, \\ u - \sigma_u \log(p/\lambda_u) & \text{for } \xi = 0. \end{cases}$$

Inference for the GPD above threshold u is well established in using likelihood and Bayesian approaches (Davison and Smith, 1990; Coles and Tawn, 1996), with comparisons recently made by Yue et al. (2025b) for incorporating a penalty function to account for experts' knowledge on the distribution of $M_{\max}$, i.e., an upper bound for the GPD upper endpoint y^F. Specifically, Yue et al. (2025b) found that likelihood and Bayesian uncertainty analyses gave very similar results provided prior knowledge was weak and that the likelihood uncertainty in the GPD parameters was handled via a parametric bootstrap. We adopt the latter approach for inference and uncertainty evaluation in this paper.

A key property of the GPD for our modelling is its distributional stability under a time-varying threshold. Let $(Y_1, \ldots, Y_n)$ be IID random variables, distributed as Y, where $(Y-u)|(Y>u) \sim \mathrm{GPD}(\sigma_u, \xi)$. Then, for any set of thresholds $(v_1 \ldots, v_n)$, with $u \leq v_i < y^F$ for each $i = 1, \ldots, n$, it follows that

$$(Y_i - v_i)|(Y_i > v_i) \sim \mathrm{GPD}(\sigma_{v_i}, \xi), \text{ with } \sigma_{v_i} = \sigma_u + \xi(v_i - u), \tag{5.3.3}$$

i.e., the GPD form and the shape parameter are stable with respect to the varying threshold. So, the exceedances of the thresholds $(v_1 \ldots, v_n)$ are not identically distributed unless $v_i = v$ for all $i = 1, \ldots, n$ for $u \leq v < y^F$ or $\xi = 0$, despite the $\{Y_i\}$ being identically distributed.

There are also well-established extreme values methods for non-identically distributed variables, and in particular when covariates $\boldsymbol{Z}$, observed as $\boldsymbol{z}$, affect the parameters of the GPD above a threshold function, $u(\boldsymbol{z})$, which can vary with $\boldsymbol{z}$. Specifically, it is assumed that, for $y > u(\boldsymbol{z})$,

$$\Pr(Y > y \mid \boldsymbol{Z} = \boldsymbol{z}) = 1 - \lambda_u(\boldsymbol{z}) \left(1 + \xi(\boldsymbol{z}) \frac{y - u(\boldsymbol{z})}{\sigma_u(\boldsymbol{z})}\right)_+^{-1/\xi(\boldsymbol{z})},$$

with $\lambda_u(\cdot), \sigma_u(\cdot)$ and $\xi(\cdot)$ the respective covariate-dependent threshold exceedance rate, scale and shape parameter functions such that

$$(Y - u(\boldsymbol{z}))|(Y > u(\boldsymbol{z}), \boldsymbol{Z} = \boldsymbol{z}) \sim \mathrm{GPD}(\sigma_u(\boldsymbol{z}), \xi(\boldsymbol{z})).$$

For a given threshold function, the typical approaches for modelling the functional forms of the parameters are using linear models (Davison and Smith, 1990) or variations of generalised additive models (Chavez-Demoulin and Davison, 2005; Youngman, 2019), each with suitable link functions. Usually, a log-link function for $\sigma_u(\boldsymbol{z})$ is used, though Eastoe and Tawn (2009) show that for the threshold stability property (5.3.3) to hold across covariate values, the identity link is required. It is relatively standard to assume that $\xi(\boldsymbol{z})$ is constant, i.e., $\xi(\boldsymbol{z}) = \xi$ for some unknown ξ for all $\boldsymbol{z}$. There are two core reasons for this choice (i) simplicity, the shape parameter is difficult to estimate accurately, so typically, there is insufficient evidence to choose a shape parameter function that is more complex than a constant, and (ii) across a range of application areas, there appears to be evidence for a different constant shape parameter being suitable for each different hazard (Healy et al., 2025).

5.3.2 Threshold choice

The first challenge when employing a GPD is the selection of an appropriate threshold, which requires a trade-off between bias and variance: too low a threshold is likely to violate the asymptotic basis of the GPD, leading to bias, whilst too high a threshold results in very few threshold excesses to estimate parameters (Coles, 2001). Even in the context of IID data, this is not an easy task, and has been the focus of much research, with a recent review of the variety of methods provided by Belzile et al. (2023). Extensions of this framework enable covariate-dependence in the threshold choice and in the GPD parameters (Kyselý et al., 2010; Northrop and Jonathan, 2011; Yue et al., 2025a).

Our approach stems from the work of Varty et al. (2021), developed specifically for the Groningen catalogue, in which the underlying variables were assumed to be IID GPD but with a time-varying threshold, as in the setup of formulation (5.3.3). For this set up, but with a constant threshold, Murphy et al. (2025) proposed a novel automated threshold selection procedure, termed the *expected quantile discrepancy* (EQD), which selected the threshold value as the level above which the sample excesses are most consistent with a GPD model out of all possible choices of level. Through an extensive simulation study, they found that the EQD approach convincingly outperforms the leading existing automated methods for threshold selection.

Specifically, the EQD method, detailed in Chapter 3, minimises a metric which approximates the integrated absolute error (IAE) between GPD model quantiles and the empirical quantiles for a set of m probabilities $\{j/(m+1) : j = 1, \ldots, m\}$. A brief outline of the approach is given here. For a choice of threshold u, let $\boldsymbol{y}_u = (y_1, \ldots, y_{n_u})$ be the sample of excesses of u. The EQD value $d(u)$ is an average calculated across B bootstrapped samples which incorporates sampling variability into the selection and stabilises the threshold choice. For $\boldsymbol{y}_u^b$, the b^{th} bootstrapped sample of $\boldsymbol{y}_u$, with $b = 1, \ldots, B$, we obtain maximum likelihood estimates $(\hat{\sigma}_u^b, \hat{\xi}_u^b)$ and evaluate the approximate IAE as:

$$d_b(u) := \frac{1}{m} \sum_{j=1}^{m} \left| \frac{\hat{\sigma}_u^b}{\hat{\xi}_u^b} \left[\left(1 - \frac{j}{m+1}\right)^{-\hat{\xi}_u^b} - 1 \right] - Q\left(\frac{j}{m+1}; \boldsymbol{y}_u^b\right) \right|, \tag{5.3.4}$$

where $Q(j/(m+1); \boldsymbol{y}_u^b)$ denotes the $j/(m+1)$ empirical quantile of $\boldsymbol{y}_u^b$. The EQD value for a choice of threshold u, is then given by $d(u) = \sum_{b=1}^{B} d_b(u)/B$. Given formulation (5.3.3), Varty et al. (2021) address the fact that excesses of different candidate thresholds are not identically distributed by using the fitted model to transform the sample excesses to standard exponential margins by the probability integral transform (see Section 5.4.3) and measure the IAE on that scale. For model-fit assessment of non-identically distributed GPD variables, transforming to a common standard exponential distribution is an established approach (Coles, 2001).

5.4 Statistical modelling and inference

5.4.1 Threshold model

To properly account for the changing data quality over time and space, we must acknowledge the main factor in this, namely, the evolving geophone network. The key starting point is to consider geophysical evidence to suggest a structure for the model of $u(\boldsymbol{x}, t)$. Consider how the magnitude of completion varies as a function of the Euclidean distance $r \geq 0$ from the hypocentre of an earthquake to a geophone, which we denote by $m_c(r)$ with a slight abuse of notation. In idealised conditions, such as uniform rock types, and with a single geophone detector, geophysicists have identified that, for $r > 0$,

$$m_c(r) = \phi_0 + \phi_1 \log r + \phi_2 r \tag{5.4.1}$$

for constants (ϕ_0, ϕ_1, ϕ_2) determined by the rock properties and the shock wave amplitude (Stange, 2006; Freudenreich et al., 2012; Gaucher, 2016). The logarithmic and linear terms are linked to shock-wave attenuation and geometric spreading respectively. Depending on the conditions, simplifications of relationship (5.4.1) have been proposed, with Goertz et al. (2012) and Demuth et al. (2016) setting $\phi_0 = 0$ and $\phi_1 = 0$ respectively.

Relative to the idealised conditions that relationship (5.4.1) is based on, for Groningen, we have two additional aspects to consider. Firstly, the observed distances from hypocentres to the nearest geophone, r, have a narrow range due to the small size of the gas field and the density of geophones. This feature makes inference using function (5.4.1) poor due to co-linearity when both modes of variation, corresponding to $\log r$ and r, are included in a single model. For our observed range of r, we anticipate that one of the two modes of variation with r will dominate, or possibly the actual variation will be better represented by some intermediate form. So, in addition to $\log r$ and r, we consider the function $r^{1/2}$, as from the Box-Cox transformation, $(r^\lambda - 1))/\lambda$ with $\lambda = 1/2$, with the other two variations in relationship (5.4.1) corresponding to $\lambda = 0$ and $\lambda = 1$. Secondly, the network as a whole is used to obtain the most accurate earthquake measurements possible, so using the nearest geophone distance is possibly overly-simplistic, and hence like Mignan et al. (2011), we consider a range of possible distance choices, i.e., $V_i(\boldsymbol{x}, t)$ for a range of $i \geq 1$. However, the range of i is not directly related to the number of geophones that were used to record events in the catalogue. Exploring a large range of i would appear prudent, but in early stages of the network development, there were insufficient geophones to calculate the relevant distances if i

is too large, so we restrict our range to $i = 1 - 4$.

Consequently, we propose three distinct threshold function model types A, B and C for the functional relationship to distance, and four covariates of distance, $i = 1 - 4$, as input to the covariate model for the GPD threshold $u(\boldsymbol{x}, t)$. Specifically, we consider the threshold functions:

$$\begin{aligned} A_i &: u(\boldsymbol{x}, t) = \alpha_0 + \alpha_1 V_i(\boldsymbol{x}, t); \\ B_i &: u(\boldsymbol{x}, t) = \alpha_0 + \alpha_1 \log(V_i(\boldsymbol{x}, t)); \\ C_i &: u(\boldsymbol{x}, t) = \alpha_0 + \alpha_1 \sqrt{V_i(\boldsymbol{x}, t)}, \end{aligned} \tag{5.4.2}$$

for $i = 1 - 4$, where $(\alpha_0, \alpha_1) \in \mathbb{R} \times \mathbb{R}_+$ and $V_i(\boldsymbol{x}, t)$ is the spatio-temporal covariate, introduced in Section 5.2.2, corresponding to the three-dimensional Euclidean distance from $\boldsymbol{x}$ to the i^{th} nearest geophone in the network at time t. The restriction $\alpha_1 \geq 0$ reflects that the threshold u decreases as the network becomes denser.

Each of the resulting 12 (types $A - C$ and $i = 1 - 4$) threshold models (5.4.2) capture the changing magnitude of completion over time, in a broadly similar way to the sigmoid model of Varty et al. (2021). Here, the formulation need not change smoothly over time, and critically it is allowed to vary spatially, with a physical basis. It is more parsimonious with two parameters to estimate rather than four, despite the richer spatial-temporal variation. For our subsequent inference, we face the dual challenges of fitting the first ever spatio-temporal parametric threshold while also accounting for the model choice uncertainty over the 12 different possible formulations.

5.4.2 Distributional model within the observed period $\mathcal{T}$

The natural and most parsimonious way to approach the model specification for a single distributional model of the magnitudes of the true earthquakes is to model only the distribution of earthquakes above a given level. Traditionally, this has been achieved using a constant conservative magnitude of completion estimate, e.g., 1.45M$_\mathrm{L}$ for Groningen. However, with time improvements in the measurement process, a constant magnitude of completion estimate is clearly not optimal. What is required is to specify a model for the distribution of the magnitudes of true earthquakes above a level which we are certain lies below $u(\boldsymbol{x}, t)$ for all $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}$. As $u(\boldsymbol{x}, t)$ is unknown, we choose to model the distribution above 0 M$_\mathrm{L}$ as the Groningen catalogue contains almost entirely positive values (see Figure 5.2.1) and this is mathematically convenient. Unlike when using $m_c = 1.45\mathrm{M_L}$, we cannot fit the model using all the observations above this level, but instead fit using the data above the estimate of $u(\boldsymbol{x}, t)$, see Section 5.4.3.

We assume that the distribution of true earthquakes F has the GPD tail form of (5.3.2), for $u = 0$, which gives a parametric model for all earthquakes above 0 but does not specify the form of $F(y)$ for $y < 0$. It follows that F_0, the conditional distribution of excesses of 0, has a lower endpoint at 0 and follows a GPD of the form (5.3.1). We use $Y_0(\boldsymbol{x}, t)$ to denote an earthquake excess of 0 at hypocentre and time $(\boldsymbol{x}, t)$.

Unlike Ogata (1988), and the majority of the literature referenced in Section 5.1, we do not assume that $Y_0(\boldsymbol{x}, t)$ is identically distributed over $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}$. Following the discussion in Section 5.2.3 of the recent use of the spatio-temporal covariate KS and our findings from the exploratory analysis presented in Figure 5.2.3, we propose a simple formulation for the inclusion of $s(\boldsymbol{x}, t)$ in the GPD parameterisation. As discussed in Section 5.3, we choose to keep the shape parameter constant and incorporate the covariate-dependence into the GPD scale parameter. In line with the view of Eastoe and Tawn (2009) on how to optimally incorporate covariates in the GPD scale parameter to ensure the threshold stability property holds, we make the distributional assumption for the true earthquakes above 0 to be

$$Y_0(\boldsymbol{x}, t) \sim \mathrm{GPD}(\sigma_0(\boldsymbol{x}, t), \xi), \text{ with } \sigma_0(\boldsymbol{x}, t) = \beta_0 + \beta_1 s(\boldsymbol{x}, t), \text{ for } (\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}, \quad (5.4.3)$$

where $(\beta_0, \beta_1) \in \mathbb{R}_+ \times \mathbb{R}_+$, where $\beta_1 \geq 0$ as increased stress from extraction cannot lead to a stochastic decrease in earthquake magnitudes.

Model (5.4.3) cannot be fitted directly due to the biasing effects of data below the unknown magnitude of completion function missing from the earthquake catalogue. However, any recorded earthquake $(t, \boldsymbol{x}, y)$, with $y > u(\boldsymbol{x}, t)$ can be considered to be a realisation of a variable linked directly to $Y_0(\boldsymbol{x}, t)$. With our setup, we have the threshold function, with the model for $u(\boldsymbol{x}, t)$ given by one of the set of formulations (5.4.2), and so we focus on the excesses of this function. Exploiting the threshold stability property (5.3.3) results in the following conditional distribution of the excesses of the threshold $u(\boldsymbol{x}, t)$:

$$[Y(\boldsymbol{x}, t) - u(\boldsymbol{x}, t)] \mid [Y(\boldsymbol{x}, t) > u(\boldsymbol{x}, t)] \sim \text{GPD}(\sigma_u(\boldsymbol{x}, t), \xi), \text{ for } (\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}, \quad (5.4.4)$$

where $\sigma_u(\boldsymbol{x}, t) = \sigma_0(\boldsymbol{x}, t) + \xi u(\boldsymbol{x}, t)$, with the scale and threshold given by models (5.4.3) and (5.4.2) respectively, and $\sigma_u(\boldsymbol{x}, t)$ is a function of parameters $(\alpha_0, \alpha_1, \beta_0, \beta_1) \in \mathbb{R}_+^4$.

5.4.3 Joint threshold and excess model inference

Here, we detail our inferences for the parameters $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \beta_0, \beta_1, \xi, \gamma_0, \gamma_1)$ of the combined threshold-distributional model for magnitudes and for the baseline intensity of true earthquake occurrences. For the former, our inference procedure optimises the fit over three different elements of our model; the GPD parameters (β_0, β_1, ξ); the threshold parameters (α_0, α_1); and the threshold formulation $(A - C, i = 1 - 4)$. For the latter, we optimise over parameters (γ_0, γ_1) of the baseline intensity model formulation (5.2.1). Given the threshold function choice, inferences on (β_0, β_1, ξ) and (γ_0, γ_1) are orthogonal and so we can make inference on these separately.

For the combined threshold-distributional model, the strategy is as follows. Firstly, we select one of the 12 specific functional threshold formulations (5.4.2) and minimise our extension of the EQD metric, introduced in Section 5.3.2, to optimise values for

(α_0, α_1); for each choice of (α_0, α_1) that is considered, the GPD parameters are optimised by maximum likelihood estimation. This procedure is then repeated for each threshold formulation $(A - C, i = 1 - 4)$. Finally, once estimates for $(\alpha_0, \alpha_1, \beta_0, \beta_1, \xi)$ have been obtained for all 12 threshold formulations, we use the EQD metric to compare between these formulations to select the most appropriate combined distributional model and threshold function fit overall. We expand on the details of these steps below.

For a given choice of the threshold function, its functional form and values of (α_0, α_1), the set of observed magnitude exceedance indices are defined as $K_u = \{k \in \{1, \dots, n\} : y_k > u_k\}$, where $u_k := u(\boldsymbol{x}_k, t_k)$, and the vector of the associated exceedances $\boldsymbol{y}_u = \{y_k : k \in K_u\}$. Then the GPD likelihood is

$$L((\beta_0, \beta_1, \xi); \boldsymbol{y}_u, (\alpha_0, \alpha_1)) = \prod_{k \in K_u} \left\{ \frac{1}{\sigma_{u,k}} \left(1 + \xi(y_k - u_k)/\sigma_{u,k}\right)_+^{-1-1/\xi} \right\} \tag{5.4.5}$$

where $\sigma_{u,k} := \sigma_u(\boldsymbol{x}_k, t_k)$. We denote the maximum likelihood estimates, given the threshold function, by $\hat{\sigma}_{u,k} = \hat{\sigma}_u(\boldsymbol{x}_k, t_k)$ for $k \in K_u$ and $\hat{\xi}_u$.

For a given choice of functional form for $u(\boldsymbol{x}, t)$, we need to estimate the threshold parameters (α_0, α_1). We adapt the methods of Varty et al. (2021) and Murphy et al. (2025) to assess the fit of the GPD over possible values of (α_0, α_1). For a given (α_0, α_1), we define the vectors of hypocentres, times and stresses of the earthquakes that exceed $u(\boldsymbol{x}, t)$, namely $(X_u, \boldsymbol{t}_u, \boldsymbol{s}_u) := \{(\boldsymbol{x}_k, t_k, s(\boldsymbol{x}_k, t_k) : k \in K_u\}$. The evaluation of our EQD metric broadly follows the approach set out in Section 5.3.2, with details given here when the model detailed in Sections 5.4.1 and 5.4.2 requires additional information for EQD evaluation. We resample with replacement the rows of the array $(\boldsymbol{y}_u, X_u, \boldsymbol{t}_u, \boldsymbol{s}_u)$ to obtain bootstrapped samples, where the b^{th} bootstrap is $(\boldsymbol{y}_u^b, X_u^b, \boldsymbol{t}_u^b, \boldsymbol{s}_u^b)$, and the k^{th} row of this array is denoted by $(y_{u,k}^b, \boldsymbol{x}_{u,k}^b, t_{u,k}^b, s_{u,k}^b)$ for $k \in K_u$. Using this bootstrapped sample, we maximise likelihood (5.4.5) to obtain parameter estimates $(\hat{\beta}_0^b, \hat{\beta}_1^b, \hat{\xi}^b)$ and hence obtain $\hat{\sigma}_{u,k}^b = \hat{\sigma}_u^b(\boldsymbol{x}_k^b, t_k^b)$, the function $\hat{\sigma}_u^b$ is the maximum likelihood estimate of the function σ_u, defined by expression (5.4.4), for the bth bootstrap sample. We transform the vector of bootstrapped magnitudes $\boldsymbol{y}_u^b$, via the probability integral transform, to the vector $\boldsymbol{y}_u^{E,b}$, with its k^{th} component $y_{u,k}^{E,b} = F_{\text{Exp}}^{-1}\{G(y_{u,k}^b - u_k; \hat{\sigma}_{u,k}^b, \hat{\xi}^b)\}$, where F_{Exp}^{-1} is the inverse distribution function of a standard exponential and G is the GPD distribution function (5.3.1). If the threshold was a good choice for the model then $\boldsymbol{y}_u^{E,b}$ would resemble a sample from a standard exponential distribution. Hence, we use the metric

$$d(\alpha_0, \alpha_1) = \frac{1}{B}\sum_{b=1}^{B} d_b(\alpha_0, \alpha_1)$$

where

$$d_b(\alpha_0, \alpha_1) = \frac{1}{m}\sum_{j=1}^{m}\left|F_{\text{Exp}}^{-1}\left(\frac{j}{m+1}\right) - Q\left(\frac{j}{m+1}; \boldsymbol{y}_u^{E,b}\right)\right|, \qquad (5.4.6)$$

in the EQD method, with notation as in expression (5.3.4). This metric provides a comparable measure of fit across different values of (α_0, α_1), which does not require threshold functions to be ordered in value across $\mathcal{X} \times \mathcal{T}$, i.e., the elements of the set K_u need not be nested across different choices of (α_0, α_1).

To enable the best choice of model over the different functional forms for the threshold covariates, we separately minimise the metric $d(\alpha_0, \alpha_1)$ for each threshold function formulation. This provides 12 metric values ($A - C$ and $i = 1 - 4$) with the best model formulation simply selected as the one achieving the minimum EQD. In Section 5.5, we detail our procedure for accounting for the various uncertainties in this inference procedure, i.e., the uncertainty in the GPD parameter estimates, the uncertainty in threshold parameter estimation, and the uncertainty due to the selection of the functional formulation of the threshold.

Our combined threshold-distributional model may consider excesses of a threshold as low 0M_L, and so we need a model for the intensity of true earthquakes above this level, which we denote by $\lambda_0(\boldsymbol{x}, t)$ and term the baseline intensity. We cannot directly estimate this intensity function because the magnitude of completion is above 0M_L and below this level, there are earthquakes missing not-at-random from the catalogue. To overcome this limitation of the data, we estimate the parametric intensity model for $\lambda_0(\boldsymbol{x}, t)$ using only earthquake data above $\hat{u}(\boldsymbol{x}, t)$, which, for convenience, we sometimes denote by $\hat{u}$. Specifically, for any $\boldsymbol{x} \in \mathcal{X}$ and $t \in \mathcal{T}$, our model for $\lambda_0(\boldsymbol{x}, t; \gamma_0, \gamma_1)$ uses an identical parametric model formulation to that used for $\lambda_{1.45}$ in model (5.2.1). Under this parametric model for λ_0, earthquakes with magnitudes exceeding our spatio-temporal threshold $u(\boldsymbol{x}, t)$ have hypocentres and occurrence times described by the intensity function

$$\lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}_\gamma) = \lambda_0(\boldsymbol{x}, t; \gamma_0, \gamma_1) \left[1 + \xi \frac{\hat{u}(\boldsymbol{x}, t)}{\hat{\sigma}_0(\boldsymbol{x}, t)}\right]_+^{-1/\xi}, \tag{5.4.7}$$

where $\hat{\boldsymbol{\theta}}_\gamma = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}_0, \hat{\beta}_1, \hat{\xi}, \gamma_0, \gamma_1)$. Intuitively, $\lambda_{\hat{u}}$ is the intensity of earthquake occurrences with magnitudes above $0\mathrm{M_L}$ scaled by the estimated GPD-based probability of such events exceeding the estimated threshold $\hat{u}(\boldsymbol{x}, t)$. The estimated functions $\hat{u}$ and $\hat{\sigma}_0$ are functions of $(\hat{\alpha}_0, \hat{\alpha}_1)$ and $(\hat{\beta}_0, \hat{\beta}_1, \hat{\xi})$ respectively, although this is not explicit in the notation. To estimate (γ_0, γ_1), we follow the approach of Bourne et al. (2018) in using the Poisson process-based likelihood including only the exceedances of $\hat{u}$, i.e.,

$$L(\gamma_0, \gamma_1) \propto \left(\prod_{k=1}^{n} \lambda_u(\boldsymbol{x}_k, t_k; \hat{\boldsymbol{\theta}}_\gamma)^{I(y_k > u_k)}\right) \exp\left(-\int_{\boldsymbol{x}\in\mathcal{X}} \int_{t\in\mathcal{T}} \lambda_u(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}_\gamma)\, \mathrm{d}t\mathrm{d}\boldsymbol{x}\right), \tag{5.4.8}$$

with $\lambda_u(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}_\gamma)$ given by expression (5.4.7), $u_k = u(\boldsymbol{x}_k, t_k)$ and $I(A)$ is the indicator function of event A. Then, $(\hat{\gamma}_0, \hat{\gamma}_1)$ are obtained by maximising likelihood (5.4.8). The Poisson process is not a full description of the occurrences, due to aftershocks which result in local clustering of events in space and time (Ross, 2016), but, as discussed in Section 5.2.3, we do not incorporate known after-shocks into this likelihood.

5.4.4 Model diagnostics

The estimated magnitude of completion function $\hat{u}(\boldsymbol{x}, t)$ defines a dataset of its exceedances with catalogue values indexed by K_u, as defined in Section 5.4.3. To assess $\hat{u}$ and the resulting model fit we explore the model performance for both the distribution of the magnitudes of threshold excesses and the spatial-temporal density of earthquakes as these are of key importance. We use only the subset indexed by K_u of the catalogue as they are deemed the complete and reliable data based on $\hat{u}(\boldsymbol{x}, t)$.

Given that each excess of $\hat{u}(\boldsymbol{x}, t)$ follows a different distribution for each $(\boldsymbol{x}, t)$, we transform the excesses into a common unit exponential distribution, following Varty et al. (2021). Under the fitted model, the excesses are assumed to be realisations of the distribution (5.4.4). We define the transformed values by $\{y_k^E : k \in K_u\}$ where

$$y_k^E = -\log\{[1 + \xi(y_k - u_k)/\sigma_{u,k}]_+^{-1/\xi}\} \text{ for } k \in K_u,$$

which can be assessed as a unit exponential sample through standard techniques.

To assess the performance of $\lambda_{\hat{u}}(\boldsymbol{x}, t; \boldsymbol{\theta})$, our estimated spatial-temporal occurrence rate of the excesses of the estimated threshold function $\hat{u}$, we use two summaries of the intensity estimates: $\Lambda_{\hat{u}}^{\mathcal{X}}(T; \hat{\boldsymbol{\theta}})$, the estimated expected yearly aggregated intensity (in $\mathcal{X}$) of events over $\hat{u}$ in year T, and $\Lambda_{\hat{u}}(\boldsymbol{x}, T; \hat{\boldsymbol{\theta}})$, the spatially dis-aggregated version of that summary, where

$$\Lambda_{\hat{u}}^{\mathcal{X}}(T; \hat{\boldsymbol{\theta}}) = \int_{\boldsymbol{x} \in \mathcal{X}} \int_{t \in T} \lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}) \, \mathrm{d}t \mathrm{d}\boldsymbol{x} \text{ and } \Lambda_{\hat{u}}(\boldsymbol{x}, T; \hat{\boldsymbol{\theta}}) = \int_{t \in T} \lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}) \, \mathrm{d}t. \quad (5.4.9)$$

We compare these expected values with the observed numbers and locations of earthquakes in year T as a diagnostic assessment for the intensity element of the model.

5.4.5 Inference for future extreme magnitude events

Section 5.4.3 provides estimates, based on earthquakes exceeding the threshold function $\hat{u}(\boldsymbol{x}, t)$, of how the GPD F_0 and the intensity $\lambda_0(\boldsymbol{x}, t)$ of true earthquakes above $0\mathrm{M_L}$ vary with $s(\boldsymbol{x}, t)$ for $\boldsymbol{x} \in \mathcal{X}, t \in \mathcal{T}$. We follow Beirlant et al. (2019) and Varty et al. (2021) by focusing exclusively on future extreme magnitude events, though geophysicists typically take the inference a step further to develop estimates of the ground motion across the entire region $\mathcal{X}$ (Bommer et al., 2017). We focus on two-types of future extreme event summary: the largest possible earthquake magnitude and events that exceed magnitude v in some sub-region of $\mathcal{X}$ under the scenario of no further extraction from the Groningen gas field over the period of 30 years from January 2025 until April 2055, i.e., the period $\mathcal{T}_F$. For this scenario, we have access to the geophysical model-based predictions for the Kaiser stress covariate over the period $\mathcal{T}_F$, i.e., $\mathcal{S}_F$, although as noted in Section 5.2.3, this covariate ignores local time and space changes that arise from earthquakes that induce after-shock events. We extend our notation for $\sigma_0(\boldsymbol{x}, t)$ and $\lambda_0(\boldsymbol{x}, t)$ to incorporate the future covariate estimates, so that conditioning on $\mathcal{S}_F$, we have $\sigma_0(\boldsymbol{x}, t \mid \mathcal{S}_F)$ and $\lambda_0(\boldsymbol{x}, t \mid \mathcal{S}_F)$ respectively.

Under an assumption of temporal stationarity and ignoring any spatial variation, Beirlant et al. (2019) estimate the upper-endpoint of the magnitude distribution. Under the GPD distributional assumption of Section 5.3.1, the endpoint corresponds to $u - \sigma_u/\xi$. Under our GPD covariate model of Section 5.4.2, the upper endpoint varies temporally and spatially into the future with form, for all $(\boldsymbol{x}, t), \in \mathcal{X} \times \mathcal{T}_F$, of

$$e(\boldsymbol{x}, t \mid \mathcal{S}_F) := u(\boldsymbol{x}, t) - \sigma_u(\boldsymbol{x}, t \mid \mathcal{S}_F)/\xi = -\sigma_0(\boldsymbol{x}, t \mid \mathcal{S}_F)/\xi = -[\beta_0 + \beta_1 s(\boldsymbol{x}, t)]/\xi. \tag{5.4.10}$$

To provide endpoint values which are practically useful for planning of infrastructure maintenance and reinforcement, we consider two summaries of the endpoint function (5.4.10) through its maximum $e_{\max}(\mathcal{S}_F)$ and a weighted average $e_{\mathrm{wm}}(\mathcal{S}_F)$, with the weights given by a probability density function $g_0(\boldsymbol{x}, t \mid \mathcal{S}_F)$ which accounts for the occurrence rates of true earthquakes (in terms of exceedances of $0\mathrm{M_L}$), i.e.,

$$e_{\max}(\mathcal{S}_F) = \max_{(\boldsymbol{x},t)\in\mathcal{X}\times\mathcal{T}_F} e(\boldsymbol{x}, t \mid \mathcal{S}_F) \text{ and } e_{\mathrm{wm}}(\mathcal{S}_F) = \sum_{T\in\mathcal{T}_F} e_{\mathrm{wm}}(T \mid \mathcal{S}_F)\frac{\Gamma^{\mathcal{X}}(T)}{\Gamma^{\mathcal{X}}(\mathcal{T}_F)}$$

where

$$e_{\mathrm{wm}}(T \mid \mathcal{S}_F) = \int_{\boldsymbol{x}\in\mathcal{X}}\int_{t\in T} e(\boldsymbol{x}, t \mid \mathcal{S}_F) g_0(\boldsymbol{x}, t \mid \mathcal{S}_F)\,\mathrm{d}\boldsymbol{x}, \text{ with } g_0(\boldsymbol{x}, t \mid \mathcal{S}_F) := \frac{\lambda_0(\boldsymbol{x}, t \mid \mathcal{S}_F)}{\Gamma^{\mathcal{X}}(T)}.$$

where $\Gamma^{\mathcal{X}}(T) = \int_{\tau\in T}\int_{\boldsymbol{z}\in\mathcal{X}} \lambda_0(\boldsymbol{z}, \tau \mid \mathcal{S}_F)\,\mathrm{d}\boldsymbol{z}\mathrm{d}\tau$ and $\Gamma^{\mathcal{X}}(\mathcal{T}_F) = \int_{\tau\in\mathcal{T}_F}\int_{\boldsymbol{z}\in\mathcal{X}} \lambda_0(\boldsymbol{z}, \tau \mid \mathcal{S}_F)\,\mathrm{d}\boldsymbol{z}\mathrm{d}\tau$. Here, $e_{\mathrm{wm}}(T \mid \mathcal{S}_F)$ gives a measure of the endpoints which is related to the likely earthquake locations in year T.

It is reasonable to assume that the geophone network is designed to a sufficient level that it will be certain to record any future extreme events in the region. So we consider earthquakes with magnitudes exceeding level v in future time period, with $v > u(\boldsymbol{x}, t)$ for all $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}_F$. As noted in Section 5.1, design standards require structures to withstand all earthquakes with a 90% probability of occurrence over a 50-year span (Code, 2005), which if the process was stationary corresponds to the 475-year return level. Here, we have $\mathcal{S}_F$ for 30 years into the future, so focus on estimating an equivalent level of design risk such that the maximum earthquake magnitude over $\mathcal{T}_F$ must be less than v with a 93.87% probability.

Hence, we focus on the extreme event of the form $R_v(\mathcal{W}, \mathcal{T}_F) = \{\boldsymbol{x} \in \mathcal{W} \subseteq \mathcal{X}, t \in \mathcal{T}_F, y \in \mathbb{R}_+ : y > v\}$, with possible choices for $\mathcal{W}$ being $\mathcal{G}$ or a region of dense housing, e.g., the city of Groningen. When $\mathcal{W} = \mathcal{X}$, this enables comparisons of our modelling approach with previous studies which ignore the spatial context of the data. Under the scenario of no future extraction, the expected number of future occurrences of extreme event $R_v(\mathcal{W}, \mathcal{T}_F)$ is given by

$$\Lambda_v(\mathcal{W}, \mathcal{T}_F \mid \mathcal{S}_F) := \int_{\boldsymbol{x}\in\mathcal{W}} \int_{t\in\mathcal{T}_F} \lambda_0(\boldsymbol{x}, t \mid \mathcal{S}_F) \left[1 + \xi \frac{v}{\sigma_0(\boldsymbol{x}, t \mid \mathcal{S}_F)}\right]_+^{-1/\xi} \mathrm{d}t \, \mathrm{d}\boldsymbol{x},$$

using the same logic as for expression (5.4.7). Letting $N_v(\mathcal{W}, \mathcal{T}_F \mid \mathcal{S}_F)$ be the number of future v-level extreme events, then, under the assumption of a Poisson process of earthquakes, we have that the probability that no earthquakes with magnitude in excess of v occur in $\mathcal{W} \times \mathcal{T}_F$ is given by

$$\Pr(N_v(\mathcal{W}, \mathcal{T}_F \mid \mathcal{S}_F) = 0) = \exp[-\Lambda_v(\mathcal{W}, \mathcal{T}_F \mid \mathcal{S}_F)].$$

Hence, for a level of risk specified by Code (2005), we require that $\mathcal{W} = \mathcal{X}$ and v is such that $\Lambda_v(\mathcal{X}, \mathcal{T}_F \mid \mathcal{S}_F) = -\log(0.9387)$. We estimate the level v by solving this equation with the parameters of the statistical models replaced by estimated values using the methods of Section 5.4.3.

5.5 Uncertainty quantification

Similarly to Murphy et al. (2025), we use bootstrapping methods to quantify uncertainty in our modelling procedure. Methods to generate confidence intervals (CIs) using standard errors or profile likelihoods cannot account for threshold uncertainty and relying on asymptotic arguments would not work well in our setting due to the sparsity of exceedances of the threshold function $u(\boldsymbol{x}, t)$. Below, we propose three algorithms which capture uncertainty in (i) the GPD parameter estimation and the parameters of the rate of exceedance of the threshold function, (ii) the parameters of threshold estimation and aspect (i), and (iii) the selection of the threshold functional formulation and aspect (ii). The latter two cover aleatoric and epistemic uncertainty about the magnitude of completion, while Murphy et al. (2025) considered aspects (i) and (ii) in the simplified setting of IID variables. Accounting for uncertainty in both the threshold model function formulation in aspect (iii) and the inclusion of covariates are entirely novel for extreme value analyses.

Current seismic studies do not account for uncertainty in m_c. Accounting for only the uncertainty in the excess distribution of the estimated magnitude of completion is not sufficient, as the inference for $u(\boldsymbol{x}, t)$ relies on observed earthquakes so its value is both unknown and inferences are sensitive to its choice. Hence, we incorporate this additional source of uncertainty to ensure CIs for seismic hazards are not too narrow. The algorithms detailed below provide methods for uncertainty quantification when estimating a summary of interest $w(\boldsymbol{\theta})$, e.g., the quantities discussed in Section 5.4.5.

Algorithm 1 details the parametric bootstrapping procedure for the uncertainty of both the GPD parameters (β_0, β_1, ξ) and the threshold exceedance rate through the parameters (γ_0, γ_1) of the baseline intensity. Algorithm 1 treats the threshold function as known, with the corresponding estimates $(\hat{\alpha}_0, \hat{\alpha}_1)$ provided as input and fixed and the remaining parameters of $\hat{\boldsymbol{\theta}}$ estimated as explained in Section 5.4.3. For each b of the B_{par} bootstraps, the number of exceedances $n^b_{\hat{u}}$ of the threshold function $\hat{u}$ is generated along with the corresponding hypocentre for each exceedance and the respective magnitude excess values. Here, $n^b_{\hat{u}}$ is a realisation of a Poisson$(\Lambda_{\hat{u}}(\hat{\boldsymbol{\theta}}))$ variable, where

$$\Lambda_{\hat{u}}(\hat{\boldsymbol{\theta}}) = \int_{\boldsymbol{x}\in\mathcal{X}} \int_{t\in\mathcal{T}} \lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}) \, \mathrm{d}t\mathrm{d}\boldsymbol{x}$$

and $\lambda_u(\boldsymbol{x}, t; \boldsymbol{\theta})$ is given by expression (5.4.7). We sample $n^b_{\hat{u}}$ hypocentres independently according to the density $g_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}) = \lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}})/\Lambda_{\hat{u}}(\hat{\boldsymbol{\theta}})$, for $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}$. For these simulated exceedance hypocentres, we use the corresponding stress covariate values and the GPD model (5.4.4) parameter estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\xi})$, to generate the parametric bootstrap sample of $n^b_{\hat{u}}$ magnitude excesses. For each of the B_{par} bootstrapped samples, we keep $(\hat{\alpha}_0, \hat{\alpha}_1)$ fixed and re-estimate all other parameters in $\boldsymbol{\theta}$, obtaining $\hat{\boldsymbol{\theta}}^b_\alpha = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}^b_0, \hat{\beta}^b_1, \hat{\xi}^b, \hat{\gamma}^b_0, \hat{\gamma}^b_1)$ with which to estimate any summary of interest $w(\hat{\boldsymbol{\theta}}^b_\alpha)$, e.g., design levels, and construct CIs as quantiles of the sample of bootstrap estimates.

Algorithm 1 Parameter uncertainty for GPD with known threshold function and covariate formulation

```
Require: $(\hat{\alpha}_0, \hat{\alpha}_1, B_{\text{par}}, \{y_k, \boldsymbol{x}_k, t_k, s(\boldsymbol{x}_k, t_k) : k = 1, \dots, n\})$
  Estimate the remaining parameters of $\hat{\boldsymbol{\theta}}$ by fitting a GPD to the magnitude excesses of the estimated threshold function, defined using $(\hat{\alpha}_0, \hat{\alpha}_1)$, and obtaining GPD estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\xi})$ and estimates $(\hat{\gamma}_0, \hat{\gamma}_1)$ of the parameters of the Poisson baseline intensity function $\lambda_0$.
  for $b = 1, \dots, B_{\text{par}}$ do
    • Simulate the number of exceedances $n_{\hat{u}}^b$ above the threshold function $\hat{u}$, as a Poisson$(\Lambda_{\hat{u}}(\hat{\boldsymbol{\theta}}))$ variable, generate indepndently the $n_{\hat{u}}^b$ hypocentres using $g_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}})$ and extract corresponding stress and threshold values $s_k^b$ and $u_k$, $k = 1, \dots, n_{\hat{u}}^b$.
    • Simulate sample $\boldsymbol{y}_u^b$ independently from a GPD with parameters $(\hat{\sigma}_{u,k} = \hat{\beta}_0 + \hat{\beta}_1 s_k^b + \hat{\xi}\hat{u}_k, \hat{\xi})$ for $k = 1, \dots, n_u^b$.
    • Refit covariate GPD model to $\boldsymbol{y}_u^b$ and Poisson baseline intensity to the the hypocentre bootstrap data of exceedances to obtain parameter estimates $(\hat{\beta}_0^b, \hat{\beta}_1^b, \hat{\xi}^b, \hat{\gamma}_0^b, \hat{\gamma}_1^b)$ and evaluate $w(\hat{\boldsymbol{\theta}}_\alpha^b)$.
  end for
  return A set of $B_{\text{par}}$ bootstrapped estimates for $w(\boldsymbol{\theta})$.
```

We next incorporate the uncertainty in the estimation of the parameters (α_0, α_1) of the threshold function for a given functional form of $u(\boldsymbol{x}, t)$ from the options (5.4.2). Algorithm 2 builds on Algorithm 1 by using a double bootstrap procedure to account for the uncertainty in the estimation of the threshold function parameters, (α_0, α_1), for a particular formulation of the threshold function. Firstly, we resample with replacement the rows of the array $\{y_k, \boldsymbol{x}_k, t_k, s(\boldsymbol{x}_k, t_k) : k = 1, \dots, n\}$, to generate B_{nonpar} bootstrapped samples of the array, each with n rows. Secondly, for each of the B_{nonpar} bootstrapped arrays, we obtain point estimates of the threshold function parameters (α_0, α_1) by minimising $d(\alpha_0, \alpha_1)$ as defined in metric (5.4.6) and employ Algorithm 1 to account for the uncertainty in the estimation of $(\beta_0, \beta_1, \xi, \gamma_0, \gamma_1)$. For the b^{th} bootstrapped sample this gives $\hat{\boldsymbol{\theta}}^b = (\hat{\alpha}_0^b, \hat{\alpha}_1^b, \hat{\beta}_0^b, \hat{\beta}_1^b, \hat{\xi}^b, \hat{\gamma}_0^b, \hat{\gamma}_1^b)$. Finally, as above, we calculate a summary of interest $w(\hat{\boldsymbol{\theta}}^b)$ for each of the $B_{\text{par}} \times B_{\text{nonpar}}$ samples to construct CIs that incorporate uncertainty in the estimation of the entire parameter vector $\boldsymbol{\theta}$, for the chosen formulation of the threshold function.

Algorithm 2 Parameter uncertainty for GPD and threshold parameters with known threshold covariate formulation

```
Require: $(B_{\text{nonpar}}, B_{\text{par}}, \{y_k, \boldsymbol{x}_k, t_k, s(\boldsymbol{x}_k, t_k) : k = 1, \dots, n\})$
  for $b = 1, \dots, B_{\text{nonpar}}$ do
    • Sample $n$ rows with replacement from array $\{y_k, \boldsymbol{x}_k, t_k, s(\boldsymbol{x}_k, t_k) : k = 1, \dots, n\}$ to generate new array $\{y_k^b, \boldsymbol{x}_k^b, t_k^b, s(\boldsymbol{x}_k^b, t_k^b) : k = 1, \dots, n\}$.
    • Estimate values $(\hat{\alpha}_0^b, \hat{\alpha}_1^b)$ for the particular covariate formulation for the threshold function $u$.
    • Employ Algorithm 1 with inputs: $(\hat{\alpha}_0^b, \hat{\alpha}_1^b, B_{\text{par}}, \{y_k^b, \boldsymbol{x}_k^b, t_k^b, s(\boldsymbol{x}_k^b, t_k^b) : k = 1, \dots, n\})$.
  end for
  return A set of $B_{\text{par}} \times B_{\text{nonpar}}$ bootstrapped estimates of $w(\boldsymbol{\theta})$.
```

The spatial-temporal formulation of the true threshold function is unknown but, as motivated in Section 5.3.2, we consider 12 possible threshold formulations as there is no clear geophysical basis to select between them. Accounting for the uncertainty over these 12 options is needed to provide reliable CIs for design of hazard-resistant infrastructure. Algorithm 3 details a procedure to propagate this uncertainty through to tail inference along with the uncertainties described in Algorithms 1 and 2. It follows a similar procedure to Algorithm 2 but additionally, it allows the formulation of the threshold function (as given by expressions (5.4.2)) to vary for each bootstrapped resample of the observed data.

5.6 Application to Groningen earthquakes

5.6.1 Threshold model selection and GPD inference

We apply our developed model and inference methods to the Groningen earthquakes and covariates data described in Section 5.2. We use each of the 12 threshold function model formulations identified in Section 5.4.1, which cover all combinations of covariate $(V_i : i = 1, \dots, 4)$ and transformation $A - C$. As there were only three geophones active

Algorithm 3 Parameter uncertainty for GPD and threshold parameters with unknown threshold covariate formulation

```
Require: $(B_{\text{nonpar}}, B_{\text{par}}, \{y_k, \boldsymbol{x}_k, t_k, s(\boldsymbol{x}_k, t_k) : k = 1, \dots, n\})$
  for $b = 1, \dots, B_{\text{nonpar}}$ do
    • Sample $n$ rows with replacement from array $\{y_k, \boldsymbol{x}_k, t_k, s(\boldsymbol{x}_k, t_k) : k = 1, \dots, n\}$ to generate new array $\{y_k^b, \boldsymbol{x}_k^b, t_k^b, s(\boldsymbol{x}_k^b, t_k^b) : k = 1, \dots, n\}$.
    • Apply each of the 12 formulations of the threshold function $u$, $(A - C, i = 1 - 4)$, select the most appropriate threshold formulation by minimisation of the EQD values, and record values $(\hat{\alpha}_0^b, \hat{\alpha}_1^b)$ for the selected covariate formulation of $u(\mathcal{X}, \mathcal{T})$.
    • Employ Algorithm 1 with inputs: $(\hat{\alpha}_0^b, \hat{\alpha}_1^b, B_{\text{par}}, \{y_k^b, \boldsymbol{x}_k^b, t_k^b, s(\boldsymbol{x}_k^b, t_k^b) : k = 1, \dots, n\})$.
  end for
  return A set of $B_{\text{par}} \times B_{\text{nonpar}}$ bootstrapped estimates of $w(\boldsymbol{\theta})$.
```

for the first earthquake in the catalogue, we exclude that earthquake from the analyses to allow comparisons of V_i with $i = 1 - 4$ on the same data. We choose the number of evaluation points and number of bootstraps within the evaluation of the EQD to be $(m, B) = (500, 200)$, see Section 5.3.2, based on the sensitivity analysis of Murphy et al. (2025), with our value for B larger than the default in that paper due to the added complexity in our models. Re-running the threshold selection procedure with $B = 1000$ did not change the selected threshold formulation.

The EQD metric of fit for each of these 12 threshold function formulations is given in Table 5.6.1. The best fitting model is A_2, a linear function of the distance to the second nearest geophone, with the models B_1 and C_2 showing the best fits for the other two forms B and C respectively. The EQD values are very close suggesting that there is not much difference between the threshold function models in terms of fit across the distribution, with no systematic better performance for a given form $A - C$ over covariate V_i or vice versa. However, there may be important differences between these models when inferences are made far into the tail of the distribution, so accounting for the uncertainty in the threshold functional form may be needed for valid evaluation of the inference uncertainty.

Model	1	2	3	4
A	0.0333	**0.0321**	0.0348	0.0327
B	*0.0327*	0.0330	0.0358	0.0338
C	0.0332	*0.0329*	0.0362	0.0342

Table 5.6.1: EQD value for each threshold model formulation, A–C and V_i: $i = 1, \dots, 4$. Smallest EQD value for each formulation $A-C$ is given in italics with overall minimum in bold.

To help understand the form of the best fitted threshold function model A_2, Figures 5.6.1 and 5.6.2 provide temporal and spatial summaries respectively. Figure 5.6.1 [left] shows the estimated threshold function for the observed earthquakes $\{\hat{u}(\boldsymbol{x}_k, t_k) : k = 1, \dots n\}$, plotted as a continuous function over time (to aid visibility), and the spatial average of the A_2 threshold across $\mathcal{G}$ for each time, i.e., $\int_{\boldsymbol{x} \in \mathcal{G}} \hat{u}(\boldsymbol{x}, t)\, \mathrm{d}\boldsymbol{x}$ over $t \in \mathcal{T}$. As reference points, the two previously studied conservative and changepoint thresholds, given in Figure 5.2.1, are also included. The A_2 estimate, both for the observed earthquake locations and in its average form, has a broadly similar temporal behaviour to the changepoint threshold but differs in that at the start of the catalogue we estimate a higher magnitude of completion for the observed earthquakes than the conservative threshold. Following this, our estimate varies around the first constant value of the changepoint model, before reducing in variability near the changepoint and lying close to the subsequent level of the changepoint threshold in the later period. The spatially averaged A_2 estimate shows good agreement with the sigmoidal threshold model of Varty et al. (2021). The key difference is that the A_2 threshold function incorporates the spatial evolution of the geophone network, which is particularly important in early periods when the network was sparse in sub-regions of $\mathcal{X}$ where earthquakes occurred.

Figure 5.6.1 [centre] shows the A_2 threshold estimate, over time, for the three locations on a north-south transect through $\mathcal{G}$, as identified in Figure 5.2.3. These curves reveal that the key differences are early in $\mathcal{T}$, with the threshold being largest for the northern and central locations and the southern location having values which are close to the changepoint threshold. From 2016, the three site's thresholds are closely aligned.

Figure 5.6.1 [right] identifies that the fitted threshold for A_2 is mostly above both B_1 and C_2. There is variability in the both threshold differences throughout most of the period with the differences diminishing after 2016, particularly for the C_2 threshold. Despite these differences in the thresholds, the numbers of exceedances are rather similar, with A_2, B_1 and C_2 having 849, 890 and 851 exceedances respectively, in comparison to the conservative threshold $1.45\mathrm{M_L}$ and changepoint thresholds having 364 and 817 exceedances respectively. Threshold model A_2, which provides the best fit to the excesses according to the EQD metric, captures similar behaviour in the magnitude of completion over time to the changepoint threshold, and incorporates more exceedances than previous threshold choices, which suggests that our estimator for the tail of the distribution of earthquakes is preferable over previous analyses.

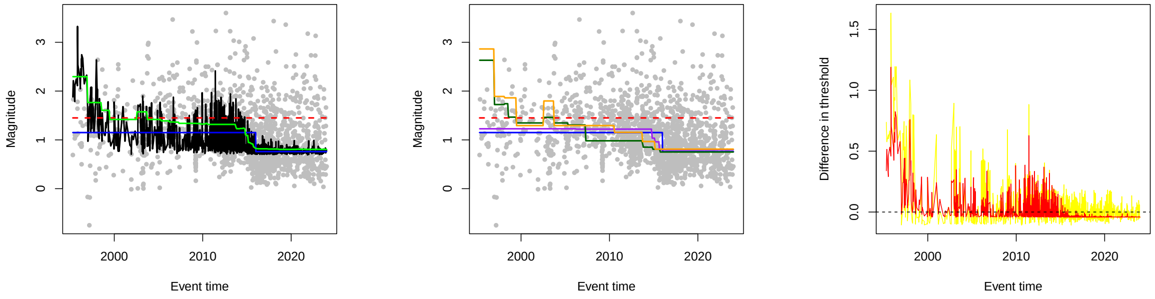

Figure 5.6.1: Comparisons of fitted threshold functions. [Left] Best-performing threshold A_2 over time: (black) is $\hat{u}(\boldsymbol{x}_k, t_k)$ for the kth earthquake in the catalogue; (green) is the spatial average of the A_2 threshold across $\mathcal{G}$. Also on this plot is the conservative level of $m_c = 1.45$ (red-dashed line) and the changepoint threshold (solid blue line). [Centre] Threshold A_2 over time for the three locations shown in Figure 5.2.3: north (orange), middle (dark green), south (purple), with conservative and changepoint thresholds for reference. [right] Difference between thresholds: $A_2 - B_1$ (yellow) and $A_2 - C_2$ (red).

Figure 5.6.2 illustrates how the A_2 threshold function $\hat{u}(\boldsymbol{x}, t)$ varies over $\boldsymbol{x} \in \mathcal{X}$ relative to geophone locations and time. The estimated function is plotted for the 1st January in 2010 and 2020, which span the major change in geophone density over $\mathcal{T}$. The figures show clearly how the threshold function is lowered in the vicinity of the geophones, as they are added, and that the wide coverage of geophones across $\mathcal{X}$ in 2020 has reduced considerably the presence of sub-regions of $\mathcal{X}$ where the estimated threshold function exceeds 1 $\mathrm{M_L}$.

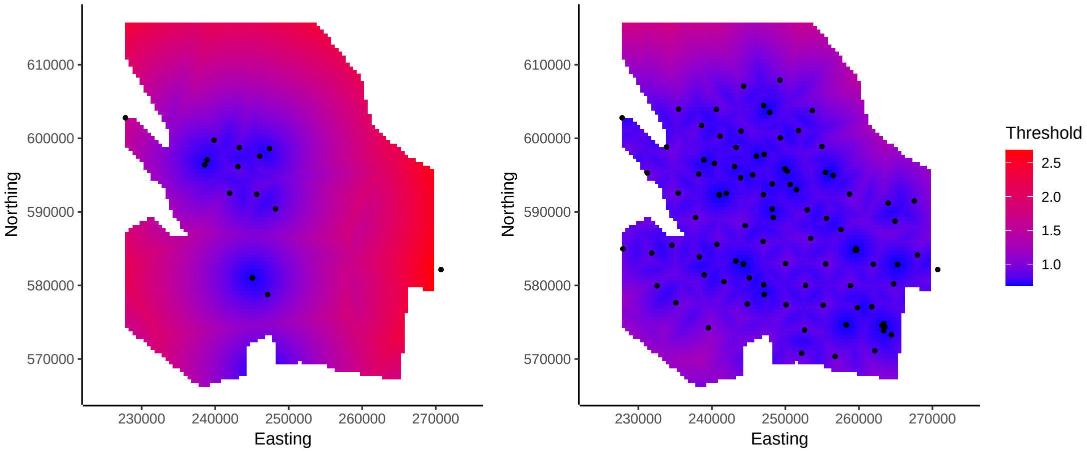

Figure 5.6.2: Spatial plots of model A_2, threshold function $\hat{u}(\boldsymbol{x}, t)$ for $\boldsymbol{x} \in \mathcal{X}$ for the dates 2010-01-01 (left) and 2020-01-01 (right). Active geophones are shown as black dots.

The upper tail features of the fitted GPD are sensitive to the choice of threshold function form. To illustrate this, we fit our covariate GPD model to the excesses of A_2, B_1 and C_2, treating the threshold as known in each case. First consider ξ, which is viewed as the key parameter of extreme value inference. The three corresponding maximum likelihood estimates (and bootstrapped standard errors obtained using Algorithm 1) are $\hat{\xi}^{A_2} = -0.154$ (0.030), $\hat{\xi}^{B_1} = -0.158$ (0.031) and $\hat{\xi}^{C_2} = -0.141$ (0.024). For the conservative threshold of $u = 1.45$ and our covariate model structure for σ_0, these values are $\hat{\xi} = -0.069$ (0.057). It is reassuring that the inferences for ξ are so similar over the three selected threshold function models in terms of both point estimates and uncertainties. Furthermore, our models almost halve the standard error relative to the conservative threshold predominantly due to the increased sample size. The reduced standard errors show strong evidence that $\xi < 0$ and hence, that there exists a finite upper endpoint for the distribution of magnitudes.

The discrepancies between these three estimates of ξ are emphasised in the inferences for the upper endpoint at the time and hypocentre location for the largest observed earthquake of $3.6\mathrm{M_L}$. The point estimates and 95% CIs, all in units of $\mathrm{M_L}$, for the upper endpoints, for models A_2, B_1 and C_2, are 4.89 $(3.875, 6.809)$, 4.77 $(3.757, 6.234)$ and 5.07 $(4.106, 6.318)$ respectively. The C_2 model's larger ξ estimate and smaller standard error are reflected in the largest endpoint estimate with the smallest CI width. The estimated uncertainty in these three endpoint estimates using just Algorithm 1 is much larger than the difference in the three point estimates, indicating that the uncertainty arising from the selection of the covariate model formulation should be less important than the distributional uncertainty when estimating far into the tails.

As most previous analyses of earthquakes treat excesses of the threshold to be identically distributed, we also assess the significance of the KS stress covariate in the GPD scale parameter. Using the same methods and summaries as for ξ, we obtain that $\hat{\beta}_1^A = 0.984$ (0.375), $\hat{\beta}_1^B = 1.023$ (0.339) and $\hat{\beta}_1^C = 0.951$ (0.247). In each case, β_1 differs from zero by approximately three standard errors and hence, the GPD scale parameters vary statistically significantly with the KS covariate, confirming the exploratory analysis presented in Figure 5.2.3.

To assess the global fit of the GPD of a given threshold function, we use the QQ and PP diagnostics discussed in Section 5.4.4. Figure 5.6.3 compares these diagnostics for the GPD fitted above the conservative threshold of $m_c = 1.45\mathrm{M_L}$ with KS covariate and the GPD excess model (5.4.4) fitted above the estimated threshold function A_2. For the conservative threshold, the fit appears good based on the pointwise 95% tolerance intervals, though there is some evidence of under-estimation by this model in the body of the distribution. The conservative threshold fit is based on a much smaller sample than for threshold A_2, and so its tolerance intervals are wider. In comparison, collectively the QQ and PP plots for threshold A_2 suggest the fit across the whole distribution is excellent even after accounting for the much tighter tolerance intervals.

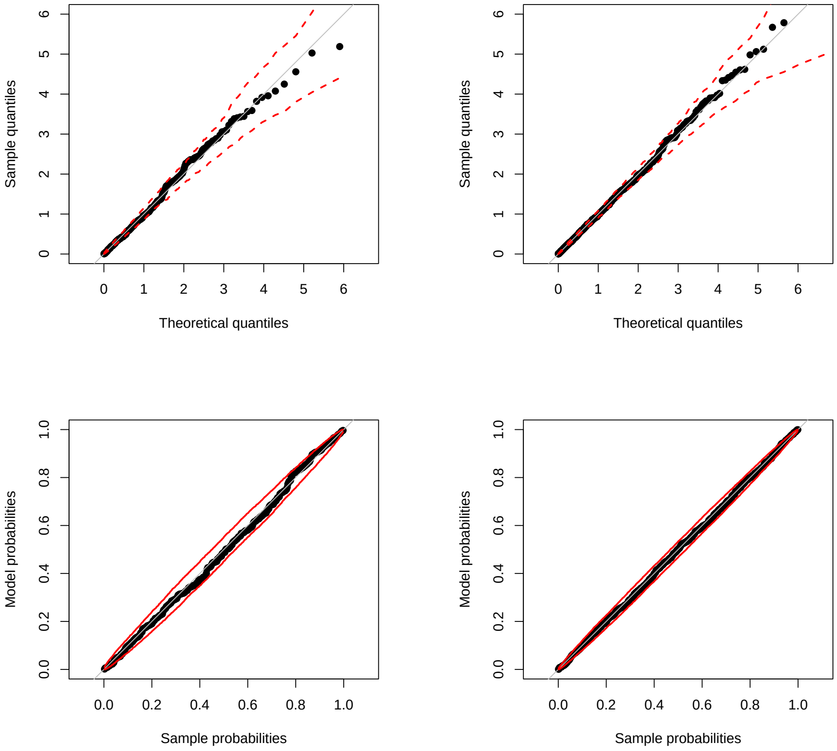

Figure 5.6.3: Assessment of GPD fit [left] above a conservative threshold $1.45\mathrm{M_L}$ and [right] above the threshold A_2. [top] QQ-plots of excesses transformed onto standard exponential margins, [bottom] corresponding PP-plots. Pointwise 95% tolerance intervals are in red.

5.6.2 Intensity inference

We now consider the fit of $\lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}})$, the intensity model (5.4.7) for exceedances of the A_2 estimated threshold function. In Section 5.4.3, we outlined how we fit model (5.4.7) using estimates from the GPD fit above $\hat{u}(\boldsymbol{x}, t)$ together with a Poisson likelihood fit of the parameters in the model $\lambda_0(\boldsymbol{x}, t, \gamma_0, \gamma_1)$. As discussed in Section 5.2.3, when $s(\boldsymbol{x}, t)$ is constant between consecutive monthly values, model (5.2.1) gives a zero value, and we choose to interpret earthquakes at those times as after-shocks. This led to 26 such earthquakes being removed from our inference and diagnostics for the intensity model. With the remaining exceedances, we obtained estimates (and standard errors) $\hat{\gamma}_0 = -0.4$ (0.2) and $\hat{\gamma}_1 = 15.6$ (0.7). We can compare our method to that of Bourne et al. (2018), which uses threshold $u = 1.45$, in terms of efficiency for estimating features of the intensity function, by comparing estimates of γ_1, as this parameter captures the important geophysical effect of KS on the intensity. With our spatio-temporal threshold, we use 485 extra data values from the catalogue that are omitted from the Bourne et al. (2018) analysis, namely $\{y_k : \hat{u}(\boldsymbol{x}_k, t_k) < y_k < 1.45 \text{ for } k = 1, \ldots, n\}$. In particular, omitting 14 events which resulted in zero intensity (i.e., deemed aftershocks), with $u = 1.45$ (with our covariate formulation for σ_0), the conservative threshold gives $\hat{\gamma}_1 = 13.2$ (1.8), which is consistent with, but 2.5 times more uncertain than, our estimate of γ_1.

Figures 5.6.4 [left] and 5.6.5 respectively show temporal and spatial intensity summaries $\lambda_{\hat{u}}^{\mathcal{X}}(T; \hat{\boldsymbol{\theta}})$ and $\lambda_{\hat{u}}(\boldsymbol{x}, T; \hat{\boldsymbol{\theta}})$ of Section 5.4.4 and the earthquakes that exceed the estimated threshold, excluding those deemed to be aftershocks. Figure 5.6.4 [left] compares observed and expected annual earthquake counts above the threshold A_2. The fit closely captures the rapid growth in earthquakes above this threshold caused by a combination of overall gas extraction stresses increasing and expansion in the geophone network lowering the threshold. The estimates closely follow the observed decline in excesses despite the lowering of the threshold in this period. This reduction comes from the KS becoming constant due to the cessation of extraction. The estimated integrated intensity over a year for each grid box in $\mathcal{X}$, shown in Figure 5.6.5 for the years 2010 and 2020, matches closely with the locations of earthquakes exceeding threshold A_2 in these years: observations are clustered around the two clear peaks in the estimated intensity, and in 2020, a few events in the south-east of the region are centred on local intensity maxima.

The close agreement between the observed and expected annual earthquake counts suggests that A_2 is at least as large as the magnitude of completion over time and space.

In Figure 5.6.4 [centre], we compare the observed and expected numbers of earthquakes above $0\mathrm{M_L}$, both excluding earthquakes with zero KS gradient as mentioned above. The expected numbers shown are estimated under our model given by $\Lambda_0^{\mathcal{X}}(T;\hat{\boldsymbol{\theta}}_{\hat{u}})$ (see expression (5.4.9)). The expected and observed counts per year T show a marked difference, with the observed count less in all years because the probability of detecting an earthquake magnitude below the magnitude of completion is less than one. To quantify the probability of recording an earthquake above $0\mathrm{M_L}$ in each year, Figure 5.6.4 [right] shows the ratio of the observed and expected annual earthquake counts above $0\mathrm{M_L}$ over T. This estimated probability rises steadily from 20% to 80% over $\mathcal{T}$ due to the expansion of the geophone network in this period, with the most rapid change occurring around 2015-2016. This is consistent with the step change in the number of active geophones in $\mathcal{X}$, shown in Figure 5.2.2, and it provides a novel estimator for the probability of recording of an earthquake, which is of interest to geophysicists. In Section 5.7, we discuss the inference for a related, but more specific, measure of missingness which accounts for the values of the observed magnitudes, rather than solely their relation to the threshold.

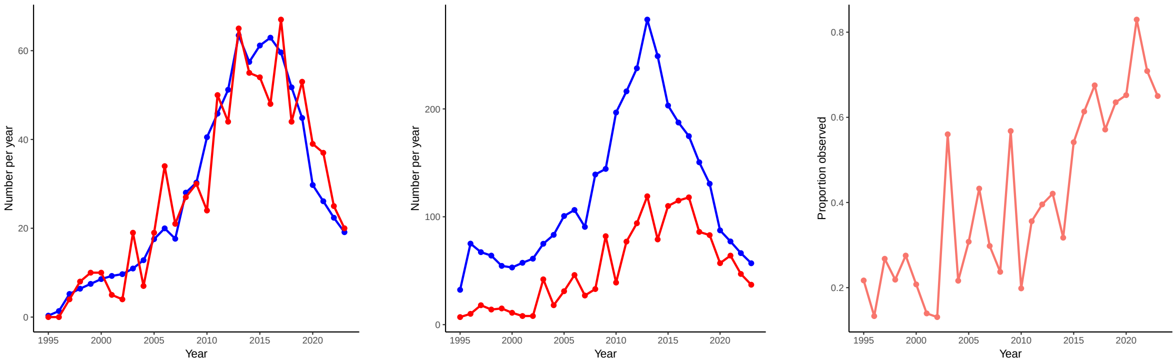

Figure 5.6.4: Estimates of features of the occurrence properties of earthquakes for years $T = 1995, \ldots, 2023$: [left] observed and expected numbers of exceedances of threshold $\hat{u}(\boldsymbol{x}, t)$ per year based on estimated aggregated intensity $\Lambda_{\hat{u}}^{\mathcal{X}}(T)$ using model formulation A_2 (blue) and counts of events in the catalogue (red); [centre] as for the left panel but for exceedances of 0 $\mathrm{M_L}$; and [right] annual estimate of the probability an earthquake above 0 $\mathrm{M_L}$ is recorded, i.e., the ratio of observed and expected annual earthquake counts above 0 $\mathrm{M_L}$ for each $T \in \mathcal{T}$.

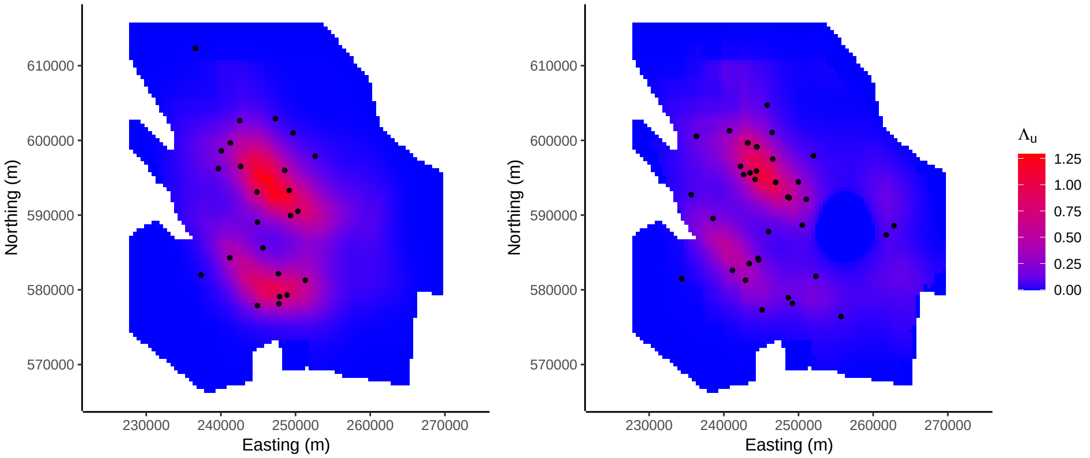

Figure 5.6.5: Spatial plots of aggregated intensity $\Lambda_u(\boldsymbol{x}, T)$ per km^2 for years $T = 2010$ and $T = 2020$ with exceedances of A_2 threshold occurring throughout each year shown as black dots.

5.6.3 Threshold uncertainty

We now explore the impact of threshold uncertainty as captured by Algorithm 3. For non-parametric samples of the original earthquake data, the algorithm selects from the 12 different threshold function model formulations, i.e., $A - C$ and $i = 1, \dots, 4$. From $B_{\text{nonpar}} \times B_{\text{par}} = 200 \times 200 = 40000$ bootstrap replicates using Algorithm 3, the model forms A, B and C (i.e., model form A corresponds to models $A_1 - A_4$ collectively) make up the respective percentages of $55.5\%, 24.5\%$ and 20% of the selected models. The corresponding percentages for $i = 1 - 4$ (i.e., $i = 1$ corresponds to models A_1, B_1, C_1, collectively) are respectively $30.5\%, 24\%, 5.5\%$ and 40%, whereas for the best model for each of $A - C$, we have A_2, B_1 and C_2 occurring $10\%, 9.5\%$ and 5% respectively. These results show that across the two aspects of covariate inclusion in the threshold function model, there is no overwhelming best choice and so it is important to assess the effect of that element of uncertainty in subsequent inference. It is somewhat surprising that the formulations with $i = 3$ are selected so infrequently given the physical motivation that $i = 3$ is the minimum number of geophones required for providing adequate location accuracy for observed earthquakes - the reasoning used by Mignan et al. (2011) for their choices of i.

To illustrate the uncertainty in the A_2 threshold function, which was selected as the best threshold formulation, Figure 5.6.6 shows the point estimate for A_2 and 200 bootstrapped summaries of the threshold function using Algorithms 2 and 3 separately. Specifically, we show the estimated spatial average of the threshold function across $\mathcal{G}$ over time. Both plots also show pointwise 95% CIs for this quantity. The patterns and spread of these bootstrapped spatial average threshold function estimates (and the associated CIs) are very similar using each of these algorithms. Both show much greater variability for the start of $\mathcal{T}$ than after 2016, and after 2016 almost all replicated spatial average functions are approximately constant over time.

To provide further insight into the variation over the bootstrapped threshold function estimates, we also investigated how the number of exceedances $|K_{\hat{u}^b}|$ of the estimated threshold functions $\hat{u}^b$ varied across the $b = 1, \dots B_{\text{nonpar}} = 200$ replicates. For Algorithms 2 and 3, the minimum, mean, maximum, and standard deviation of the number of exceedances were respectively $(546, 838, 1127, 93)$ and $(545, 897, 1212, 143)$. The mean values here show that using Algorithm 3 leads to 59 more exceedances on average than Algorithm 2, and the maximums and standard deviations for both these algorithms show there to be some samples with much larger numbers of exceedances arising from Algorithm 3. So, allowing the formulation of the threshold function to vary over the 12 different forms, leads to generally more exceedances being used and as a result, a better quality of fit for each bootstrap sample.

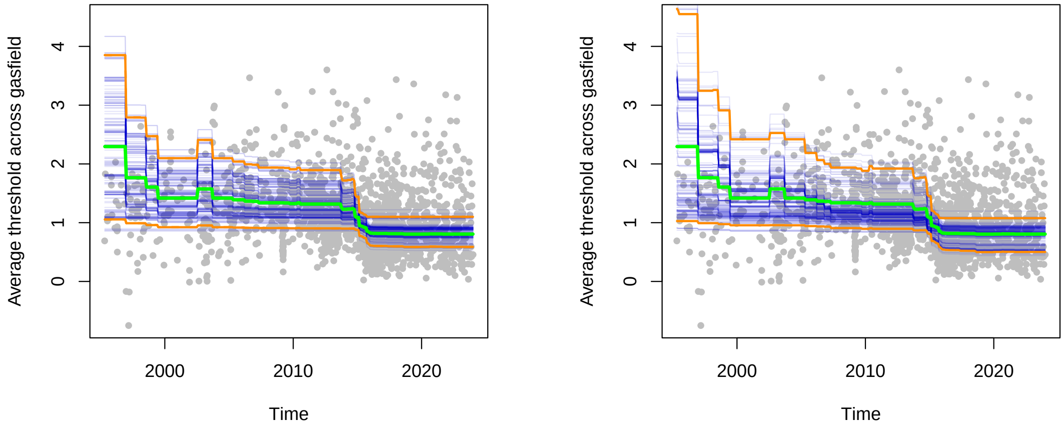

Figure 5.6.6: Spatial average over $\mathcal{G}$ threshold function uncertainty. The $B_{\text{nonpar}} = 200$ bootstrapped best threshold function estimates for Groningen according to EQD metric averaged over $\mathcal{G}$ (blue) across time $\mathcal{T}$: [left] Algorithm 2 and [right] Algorithm 3. Spatial average over $\mathcal{G}$ for the fitted A_2 threshold function is shown in green. Pointwise 95% confidence intervals for the spatial average are shown in orange.

5.6.4 Inference and uncertainty for design parameters

We first look at the inference uncertainty for shape parameter ξ, with ξ assumed common over time, space and covariates, it underpins all aspects of our extrapolations. For the selected best threshold function model A_2, we obtained $\hat{\xi}^A = -0.154$. Now we explore the assessment of uncertainty of ξ in terms of estimated 95% CIs using the bootstrapping Algorithms 1-3, with the intervals given in Table 5.6.2, where we have used $B_{\text{par}} = B_{\text{nonpar}} = 200$, giving 40000 bootstrapped estimates of ξ for Algorithms 2 & 3. When the threshold function is only taken to have the structure of A_2 with (α_0, α_1) unknown (i.e., Algorithm 2), the CI for ξ increases in width by 43% relative to when the threshold function structure of A_2 and the resulting estimates $(\hat{\alpha}_0, \hat{\alpha}_1)$ are treated as known (i.e., Algorithm 1). In comparison, the CI allowing for the uncertainty in the threshold functional form, in Algorithm 3, slightly reduces the width of the CI relative to Algorithm 2. At first sight, this is a surprising finding, as it should be expected that incorporating additional uncertainty would widen the interval. However, by allowing threshold function form to vary across bootstrapped samples, we obtain less variable shape parameter estimates.

The 95% CIs do not reveal the occurrence rate, over the bootstraps, of $\hat{\xi}^b > 0$, i.e., estimated distributions with no finite upper endpoint. For Algorithms 1-3, we observed $0\%, 0.26\%$ and 0.21% respectively, which is promising as it shows our inference is strongly consistent with geophysical knowledge about the existence of an upper bound to the distribution. By comparison, for the conservative threshold, even when applying Algorithm 1, we find the percentage of samples which obtain $\hat{\xi}^b > 0$ to be 7.5%.

Parameter	Estimate	Alg 1	Alg 2	Alg 3
ξ	-0.154	(-0.221,-0.099)	(-0.240,-0.066)	(-0.236,-0.064)
$e_{\text{max}}(\mathcal{S}_F)$	5.746	(4.398, 8.280)	(4.064,13.832)	(4.048,13.629)
$e_{\text{wm}}(\mathcal{S}_F)$	5.037	(4.017, 6.901)	(3.718, 11.663)	(3.721,11.370)
design-level v	3.943	(3.524, 4.328)	(3.406,4.619)	(3.392,4.611)

Table 5.6.2: Inference for key measures of earthquake hazard for Groningen: maximum likelihood estimates and associated 95% confidence intervals derived using Algorithms 1-3.All values reported in rows 2-4 of the table are in units of $\mathrm{M_L}$.

Now consider the inference for the endpoints of magnitudes into the future. Unlike the analyses of Beirlant et al. (2019), Varty et al. (2021) and Yue et al. (2025b) which assume that the magnitudes are identically distributed, we account for distributional changes across space and time under a scenario that there will be no future extraction. Previous analyses focussed on a single endpoint. We utilise the forecasted KS values $\mathcal{S}_F$ over $\mathcal{X} \times \mathcal{T}_F$ to estimate a spatio-temporal endpoint field. We summarise this field through its maximum $e_{\text{max}}(\mathcal{S}_F)$ and a weighted mean, $e_{\text{wm}}(\mathcal{S}_F)$. The weights are given by the intensity of earthquakes over $\mathcal{X} \times \mathcal{T}_F$, see Section 5.4.5. The maximum $e_{\text{max}}(\mathcal{S}_F)$ is comparable to the endpoint estimate in previous studies which ignored the spatial and temporal variations in the endpoint.

Table 5.6.2 presents point estimates and CIs for these two summaries using Algorithms 1-3. Consider our point estimate and 95% CIs for $\hat{e}_{\text{max}}(\mathcal{S}_F)$ using Algorithm 1. For background, Beirlant et al. (2019) use a constant threshold of $1.5\mathrm{M_L}$ with rounded earthquake data and present point estimates for $\hat{e}_{\max}(\mathcal{S}_F)$ in the range $3.61-3.80$, with 90% upper confidence bounds varying from $3.85-4.50$. These estimates are substantially lower in value and uncertainty than what we obtain using a lower threshold and the point estimates lie exceptionally close to the largest observed earthquake. In contrast, using the conservative threshold of $1.45\mathrm{M_L}$ with unrounded data (equivalent to the threshold of $1.5\mathrm{M_L}$ used by Beirlant et al. (2019)), and including the KS covariate, we find a point estimate $\hat{e}_{\max}(\mathcal{S}_F) = 8.139$ and 95% CI $(5.072, \infty)$, obtained using Algorithm 1). The difference in findings relative to Beirlant et al. (2019) are substantial given that the same threshold is used and the quality of fit exhibited by our model with this threshold is adequate, as seen in Figure 5.6.3. Relative to the inferences for $e_{\max}(\mathcal{S}_F)$ using our estimated threshold function, we see that the conservative threshold produces a much larger point estimate and with an unbounded confidence interval. This would suggest very strong assumptions are used in the inferences of Beirlant et al. (2019) to gain this level of extra precision.

Neither Varty et al. (2021) nor Yue et al. (2025b) report endpoint estimates, but the latter provides the distribution of a quantity of interest to geophysicists, namely the maximum possible earthquake in this region, denoted $M_{\max}$. This distribution is drawn from the report NAM (2022), which provides purely geophysical evidence for such values based on the fault structure and other geophysical aspects of the gas reservoir. The distribution has a median of $4.488\mathrm{M_L}$ with lower and upper bounds of $3.75\mathrm{M_L}$ and $6.75\mathrm{M_L}$. Our point estimate and 95% CI lower bound from Algorithm 1 are consistent with this $M_{\max}$ distribution, although the upper limit of the CI appears too large. This is not surprising as unlike the geophysical approach, there is limited statistical information to constrain this upper limit.

For uncertainty in $e_{\max}(\mathcal{S}_F)$, Table 5.6.2 shows that both Algorithms 2 and 3 provide much wider CIs than Algorithm 1, with the increased uncertainty reflected in massive increases (small reductions) in the upper (lower) limits respectively. As we found for

ξ, Algorithms 2 and 3 give almost identical intervals, with the latter slightly narrowing the interval, despite allowing for an additional source of uncertainty. Similar findings are obtained for the estimates of $e_{\text{wm}}(\mathcal{S}_F)$, with all values slightly less than the $e_{\max}(\mathcal{S}_F)$ quantities, as we would expect by its construction. The values for $\hat{e}_{\text{wm}}(\mathcal{S}_F)$ are likely to provide more practically useful information than $\hat{e}_{\max}(\mathcal{S}_F)$ as they better reflect the occurrence rates of earthquake hypocentres. See Appendix C.1 for the estimates of the annual behaviour of this endpoint summary.

Given the potential unbounded GPD model, inferences for endpoints are always problematic in terms of their interpretation, particularly when considering 95% CIs. We believe it is insightful to provide 50% CIs for the endpoints too. In particular, Algorithms 2 and 3 lead to 50% CIs for $e_{\max}(\mathcal{S}_F)$ of $(4.889, 6.736)$ and $(4.872, 6.756)$ respectively. For $e_{wm}(\mathcal{S}_F)$, the corresponding intervals are $(4.339, 5.778)$ and $(4.337, 5.811)$. Thus, we can see that these intervals align more closely with the $M_{\max}$ values reported by Yue et al. (2025b).

Finally, consider the inferences for the design level v, a quantity which meets the design criteria of (Code, 2005). Our point estimate and 95% CIs for v are given in Table 5.6.2. Of the previous analyses, only Yue et al. (2025b)[Figure 6] estimates this quantity, doing so as the 475-year return level under the assumption that earthquake magnitudes are identically distributed into the future. They find it to be approximately 4.5M_L. This appears to be an over-estimate as it does not take into account the cessation of extraction. Under the scenario of no further extraction, we obtain $\hat{v} = 3.943\text{M}_\text{L}$. Unlike for the endpoint summaries, the 95% CIs for v are all quite narrow and the upper limits do not exceed the geophysicist's upper bound estimate for $M_{\max}$, with Algorithms 2 and 3 again being very similar. It is interesting to note that when estimating v using the conservative threshold, the point estimate and 95% CI, under Algorithm 1, is 4.255 $(3.727, 4.804)$. This is a larger estimate and a larger upper limit for the CI relative to that for v obtained using our estimated $\hat{u}$, with uncertainty based on using

Algorithm 3, which accounts for the uncertainty in both the threshold parameters and form, in addition to that covered by Algorithm 1.

5.7 Discussion

We have developed spatial and temporal extensions of the methods of Varty et al. (2021) and Murphy et al. (2025) for extreme value threshold and excess modelling, which incorporate threshold function selection uncertainty into subsequent quantile inferences. Our methodological developments were motivated by the continuing need for accurate future hazard assessments in the Groningen gas field. To accomplish these goals, we needed to incorporate considerable contextual complexity into our statistical modelling framework. Key to our approach is the inclusion of geophysical covariates which capture the spatial and temporal changes of both the measurement network and of geophysical model-generated stress fields that describe the resultant effects of gas extraction. A range of diagnostic methods indicate that our model provides an excellent fit to the data. The fitted model provides improved scientific understanding of the form and sources of the spatio-temporal variability of the intensity of earthquake occurrences and the values of large magnitude earthquakes. Our analysis has led to improved estimators of the magnitude of completion function (the smallest magnitude which can be detected with certainty at a given time and location), which is lower than previously estimated. This reduction in level has led to more excess data being used for the analysis and hence less uncertainty in the parameter estimation. It has also achieved useful inferences for the tail behaviour of earthquake magnitudes into the future, both for design levels and upper limits, with both providing inferences that are much more consistent with geophysical knowledge than previous analyses. Even after accounting for the additional threshold function uncertainty, we have greater confidence in lower estimated design levels relative to the results for the conservative threshold.

We provided these estimates under the scenario of no further extraction from the gas field but our approach allows for estimates to be drawn under other future scenarios.

The societal importance of mitigation from earthquakes associated with gas extraction from the Groningen gas field provided a strong motivation for the extreme value methods that we have developed. However, the methodology is generic and so has potential for wide use in other gas extraction fields. Going forward, it will be likely be most impactful for use in the rapidly growing body of research on model development and hazard assessment for carbon capture (Bauer et al., 2019), which involves the injection of gas into underground storage. In these cases, earthquakes are expected but the number of geophones to be used per region is anticipated to be much lower than for Groningen, so efficient estimation of the magnitude of completion will be vital.

The methods developed here to quantify the effect of threshold uncertainty in subsequent tail inference have the potential for wide impact in core extreme value methodology. This paper substantially expands on previous work which focussed on realisations of IID variables (Murphy et al., 2025). Here, we have proposed effective methods to account for non-identically distributed data and the uncertainty in the functional choice of covariates in the threshold. The methodology also has the potential to be useful for a range of extreme value contexts where data are missing not-at-random due to limitations in measurement equipment. Furthermore, our model for earthquake magnitudes contributes to the recent and exciting evolution of work on extremes of marked point processes, with developments in this area having a particular focus on extreme wildfires and landslides, Turkman et al. (2010), Koh et al. (2023), Yadav et al. (2023).

Regarding the specifics of our data analysis, there are some practical further steps which may improve the inference. For the threshold covariate, $V_i := V_i(\boldsymbol{x}, t)$, we focused on three different transformations, V_i, $\log V_i$ and $V_i^{1/2}$ incorporated into linear functions for the threshold and we also considered the uncertainty of the choice between these forms. However, we could have looked more broadly at the covariate $(V_i^{\phi} - 1)/\phi$, for unknown $\phi \geq 0$, and estimated ϕ in the fitting. Here, there are direct parallels to the work of Wadsworth et al. (2010) on exploring the benefits of Box-Cox transforming data before applying standard extreme value methods. There is also the issue of the choice of i in the V_i covariate. It may not be best to pick a single i, but instead to allow i to change over space and/or time, possibly through a mixture of the V_i functions.

Our focus has been on accurately modelling the earthquake exceedances above the magnitude of completion. However, it is also valuable to assess the detection ability of the geophone network. An aspect of this was illustrated in Figure 5.6.4 [right]. However to do this more precisely, we would need to estimate the probability of detection function $\alpha(\boldsymbol{x}, t, y)$ for an earthquake of magnitude y with hypocentre at $\boldsymbol{x}$ and occurrence time t for all $(\boldsymbol{x}, t) \in (\mathcal{X}, \mathcal{T})$. For $y \geq m_c(\boldsymbol{x}, t)$, $\alpha(\boldsymbol{x}, t, y) = 1$, but what can be determined for $y < m_c(\boldsymbol{x}, t)$? Our paper provides the framework for the first such inference on $\alpha(\boldsymbol{x}, t, y)$. Specifically, if $\lambda_X(\boldsymbol{x}, t, y)$ is the intensity of recorded earthquakes of magnitude y at hypocentre and time $(\boldsymbol{x}, t)$, then

$$\lambda_X(\boldsymbol{x}, t, y) = \lambda_0(\boldsymbol{x}, t) \frac{1}{\sigma_0(\boldsymbol{x}, t)} \left[1 + \xi \frac{y}{\sigma_0(\boldsymbol{x}, t)}\right]_+^{-1-1/\xi} \alpha(\boldsymbol{x}, t, y), \text{ for } (\boldsymbol{x}, t, y) \in \mathcal{X} \times \mathcal{T} \times \mathbb{R}_+.$$

As $\lambda_X(\boldsymbol{x}, t, y)$ can be empirically estimated and all terms on the right hand side, other than $\alpha(\boldsymbol{x}, t, y)$ have been estimated in this paper, it is clearly possible to now estimate the detection probability function.

There are other aspects of the modelling that could be evolved in future work. As mentioned in Section 5.4.5, we could take the modelling to the next step of a full probabilistic seismic hazard analysis by incorporating a spatial spreading effect of an earthquake at a point through incorporating ground motion models. Although we have used details about the geophone network to an unprecedented level in our analysis, we have not attempted to incorporate information about the varying accuracy of different geophones in the region and how the accuracy has improved in time. In fact, addressing measurement error in terms of extreme value methods has hardly been addressed in this area, with an interesting approach having been proposed by Lin and Newberry (2023). Our analysis focuses on information in the catalogue and knowledge of the geophone network. In contrast, Yue et al. (2025b) incorporate expert knowledge of the physically-motivated worst possible earthquake in the region which provides an upper bound of the GPD upper endpoint, and our analysis could be extended similarly, which would alleviate the non-physical estimates of uncertainty we currently encounter. Finally, when modelling earthquake baseline rates of occurrence, i.e., in modelling $\lambda_0(\boldsymbol{x}, t)$, we have not accounted explicitly for the clustering of events due to the dependence between triggering main-shock earthquakes and after-shocks, so the intensity modelling could be adapted to cover this feature through the use of ETAS models of Ogata (1988).

Chapter 6

Extreme value methods for estimating rare events in Utopia

6.1 Introduction

This paper details an approach to the data challenge organised for the Extreme Value Analysis (EVA) 2023 Conference. The objective of the challenge was to estimate extremal probabilities, or their associated quantiles, for simulated environmental data sets for various locations in a fictitious country called Utopia. The data challenge is split into 4 challenges; challenges C1 and C2 focus on a setting where data is obtained from a single location while challenges C3 and C4 concern multivariate data sets, where data is obtained simultaneously from multiple locations.

Challenge C1 requires estimation of the 0.9999-quantile of the distribution of the environmental response variable Y conditional on a covariate vector $\boldsymbol{X}$, for 100 realisations of covariates. To do so, we model the tail of $Y \mid \boldsymbol{X} = \boldsymbol{x}$ using a generalised Pareto distribution (GPD; Pickands, 1975) and employ the extreme value generalised additive modelling (EVGAM) framework, first introduced by Youngman (2019), to account for the non-stationary data structure. We consider a variety of model formulations and select our final model using cross-validation. Furthermore, central 50% confidence intervals are estimated via a non-stationary bootstrapping technique, and the final model performance is assessed using the number of times the true conditional quantile lies in the confidence intervals (Rohrbeck et al., 2023). For Challenge C2, we are interested in estimating the value of q that satisfies $\Pr(Y > q) = 1/(300T)$, where $T = 200$.

Challenges C3 and C4 concern the estimation of probabilities for extreme multivariate regions, subsets of $\mathbb{R}^d$, where some or all of the components are so large that we seldom observe any data in them. Such estimates require techniques for modelling and extrapolating within the joint tail. For challenge C3, we want to estimate two joint tail probabilities for three unknown non-stationary environmental variables. To achieve this, we propose a non-stationary extension of the model introduced by Wadsworth and Tawn (2013). Lastly, for challenge C4, we wish to estimate the probability that 50 variables (locations) jointly exceed prespecified extreme thresholds. Based on an initial analysis, we separate the variables into five independent groups, and obtain distinct probability estimates for each group using the conditional extremes approach of Heffernan and Tawn (2004).

The remainder of the paper is structured as follows. A suitable background to EVA is provided in Section 6.2, introducing concepts required throughout our work. Section 6.3 covers our approach to the univariate challenges C1 and C2, and the multivariate challenges C3 and C4 are considered in Sections 6.4 and 6.5, respectively. The paper ends with a discussion of the results of all challenges in Section 4.6.

6.2 EVA background

6.2.1 Univariate modelling

Univariate EVA methods are concerned with capturing the behaviour of the tail of a distribution which allows for extreme quantities to be estimated. A common univariate approach is the peaks-over-threshold framework. Consider a continuous, independent and identically distributed (IID) random variable Y with distribution function F and upper endpoint $y^F := \sup\{y : F(y) < 1\}$. Pickands (1975) shows that, for some high threshold $v < y^F$, the excesses $(Y - v) \mid Y > v$, after suitable rescaling, converge in distribution to a GPD as $v \to y^F$. Davison and Smith (1990) provide an overview of the properties of the GPD, and also propose an extension of this framework to the non-stationary setting: given a non-stationary process Y with associated covariate(s) $\boldsymbol{X}$, the authors propose the following model

$$\Pr(Y > y + v \mid Y > v, \boldsymbol{X} = \boldsymbol{x}) = \left(1 + \frac{y\xi(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\right)_+^{-1/\xi(\boldsymbol{x})}, \tag{6.2.1}$$

for $y > 0$, where $\sigma(\cdot)$ and $\xi(\cdot)$ are the covariate-dependent scale and shape parameters, respectively. Recent extensions of the Davison and Smith (1990) framework include allowing the threshold to be covariate-dependent, i.e., $v(\boldsymbol{x})$ (Kyselý et al., 2010; Northrop and Jonathan, 2011), and using generalised additive models (GAMs; Chavez-Demoulin and Davison, 2005; Youngman, 2019) to capture the functions $\sigma(\cdot)$ and $\xi(\cdot)$ flexibly.

6.2.2 Extremal dependence measures

In addition to analysing marginal tail behaviours, multivariate EVA methods are concerned with quantifying the dependence between extremes of the individual components. An important classification of this dependence is obtained through the measure χ (Joe, 1997): given a d-dimensional random vector $\boldsymbol{Z}$, with $d \geq 2$ and $Z_i \sim F$ for all $i \in \{1, \ldots, d\}$,

$$\chi(u) := \left(\frac{1}{1-u}\right) \Pr(F(Z_1) > u, \ldots, F(Z_d) > u), \tag{6.2.2}$$

with $u \in [0, 1)$. Where the limit exists, we set $\chi := \lim_{u \to 1} \chi(u) \in [0, 1]$. When $\chi > 0$, we say that the variables in $\boldsymbol{Z}$ exhibit asymptotic dependence, i.e., can take their largest values simultaneously, with the strength of dependence increasing as χ approaches 1.

If $\chi = 0$, the variables cannot all take their largest values together. In particular, for $d = 2$, we refer to the case $\chi = 0$ as asymptotic independence.

We also consider the coefficient of tail dependence proposed by Ledford and Tawn (1996). Using the formulation given in Resnick (2002), let

$$\eta(u) := \frac{\log(1-u)}{\log \Pr\left(F(Z_1) > u, \ldots, F(Z_d) > u\right)},$$

with $u \in [0, 1)$. When the limit exists, we set $\eta := \lim_{u \to 1} \eta(u) \in (0, 1]$. The cases $\eta = 1$ and $\eta < 1$, correspond to $\chi > 0$ and $\chi = 0$, respectively. For $\eta < 1$, this coefficient quantifies the form of dependence for random vectors that do not take their largest values simultaneously.

As χ and η are limiting values, they must be approximated using numerical techniques in practice. Therefore, when quantifying extremal dependence, we approximate χ (η) using empirical estimates of $\chi(u)$ $\big(\eta(u)\big)$ for some high threshold u.

6.3 Challenges C1 and C2

Both challenges concern 70 years of daily data for the capital city of Amaurot. Each year has 12 months of 25 days and two seasons (season 1 for months 1-6, and season 2 for months 7-12). Suppose Y is an unknown response variable, and $\boldsymbol{X} = (V_1, \ldots, V_8)$ is a vector of covariates, (V_1, V_2, V_3, V_4) denoting unknown environmental variables and (V_5, V_6, V_7, V_8) denoting season, wind direction (radians), wind speed (unknown scale), and atmosphere (recorded monthly), respectively.

For C1, we build a model for $Y \mid \boldsymbol{X}$ and estimate the 0.9999-quantile, with associated 50% confidence intervals, for 100 different covariate combinations denoted $\tilde{\boldsymbol{x}}_i$ for $i \in \{1, \ldots, 100\}$. Note $\tilde{\boldsymbol{x}}_i$ are not covariates observed within the data set, but new observations provided by the challenge organisers.

For C2, we estimate the marginal quantile q corresponding to a once in 200-year event in the IID setting such that, with 300 observations per year, $\Pr(Y > q) = 1/300(200) = (6 \times 10^4)^{-1}$. Furthermore, q is obtained subject to a predefined loss function. We first estimate the marginal distribution $F_Y(y)$ using Monte-Carlo techniques; see for instance, Eastoe and Tawn (2009). Since we have a large sample size, $n = 21,000$, it is reasonable to assume that the observed covariate sample is representative of $\boldsymbol{X}$. Thus, we can approximate the marginal distribution $F_Y(y)$ as follows,

$$\hat{F}_Y(y) = \int_{\boldsymbol{X}} F_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \approx \frac{1}{n} \sum_{t=1}^{n} F_{Y_t|\boldsymbol{X}_t}(y_t \mid \boldsymbol{x}_t). \tag{6.3.1}$$

where $F_{Y|\boldsymbol{X}}(\cdot)$ is the conditional distribution function of $Y \mid \boldsymbol{X}$ and $f_{\boldsymbol{X}}(\cdot)$ denotes the joint probability density of the covariates $\boldsymbol{X}$.

We incorporate the following loss function provided by the challenge organisers,

$$\mathcal{L}(q, \hat{q}) = \begin{cases} 0.9(0.99q - \hat{q}) & \text{if } 0.99q > \hat{q}, \\ 0 & \text{if } |q - \hat{q}| \leq 0.01q, \\ 0.1(\hat{q} - 1.01q) & \text{if } 1.01q < \hat{q}, \end{cases} \tag{6.3.2}$$

where q and $\hat{q}$ are the true and estimated marginal quantiles, respectively. This loss function penalises under-estimation more heavily than an over-estimation.

We conduct the same exploratory data analysis for both challenges given the same covariates are used; this is outlined in Section 6.3.1. In Section 6.3.2 we introduce our techniques for modelling $Y \mid \boldsymbol{X}$, which is then used for modelling Y via (6.3.1). Our approach for uncertainty quantification is outlined in Section 6.3.3, and we give our results for both challenges in Section 6.3.4.

6.3.1 Exploratory data analysis

Given the covariate vector $\boldsymbol{X_t} = \{V_{1,t}, \dots, V_{8,t}\}$, the environmental response variable Y_t, $t \in \{1, \dots, n\}$, is temporally independent (Rohrbeck et al., 2023). However, it is not clear which covariates affect Y, and what form these covariate-response relationships take. In what follows, we aim to explore these relationships so we can account for them in our modelling framework.

We explore the dependence between all variables to understand the relationships between covariates, as well as the relationships between individual covariates and the response variable. We investigate dependence in the main body of the data using Kendall's τ measure, while for the joint tails, we use the pairwise extremal dependence coefficients χ and η defined in Section 6.2; values for all pairs are shown in Figure 6.3.1, with the threshold u set at the empirical 0.95-quantile for the extremal measures.

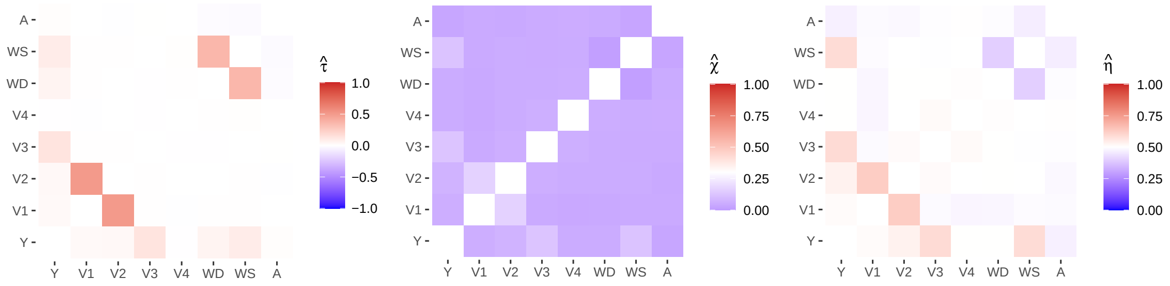

Figure 6.3.1: Heat maps for dependence measures for each pair of variables: Kendall's τ (left), χ (middle) and η (right). Note the scale in each plot varies, depending on the support of the measure, and the diagonals are left blank, where each variable is compared against itself.

The response variable Y has the strongest dependence with V_3 in the body of the distribution (see $\hat{\tau}$ in Figure 6.3.1), followed by V_6 (wind speed) then V_7 (wind direction), . For the tail of the distribution, Y has strongest dependence with V_2, V_3 and V_6 (see $\hat{\chi}$ and $\hat{\eta}$ in Figure 6.3.1). We also find strong dependence between V_6 and V_7 in the body, but evidence of weak dependence in the tail (dark blue for $\hat{\chi}$ and $\hat{\eta}$). There is also strong dependence between V_1 and V_2 in both the body and tail (see dark red for $\hat{\eta}$). We find very similar dependence relationships when the data are split into seasons. In the supplementary material (Appendix D.1), we show scatter plots of each covariate against the response variable. The scatter plots show no clear relationships between Y and V_1, V_2, V_4 or V_8. The remaining covariates show some evidence of relationships with the response, in particular V_3 appears to have a strong non-linear relationship. V_6 (wind speed) and V_7 (wind direction) show weaker relationships with the response, see Appendix D.1 for further discussion.

We also explore temporal relationships for the response variable Y. We first find temporal non-stationarity as the distribution of Y varies significantly with V_5 (season); see Appendix D.1 for more detail. The mean and range of Y is higher in season 1 than season 2, with greater extreme values observed in season 1. However, within each season, across months, there is little temporal variation in the distribution of Y. We also find that while Y exhibits statistically significant temporal dependence for a large number of lags, the auto-correlation function (acf) values are very near zero and thus, we choose to treat Y as independent at all lags; see Appendix D.1.

As noted in Rohrbeck et al. (2023), 11.7% of the observations have at least one value missing completely at random (MCAR). A detailed breakdown of the pattern of missing predictor observations is provided in Appendix D.1. Since we can assume the data are MCAR, ignoring the observations that have a missing predictor covariate will not bias our inference, however, a complete case analysis is undesirable due to the amount of data loss. To mitigate against this, we attempted to impute the observations where predictors are missing but ultimately could not find an imputation method that satisfactorily retained the dependence structure between the response and covariates, particularly in the tails of the distribution. Therefore, we use a case analysis approach, whereby an observation is only removed if a predictor covariate of interest is missing. This results in only 4% of observations being removed for our final model.

6.3.2 Methods

Due to the complex nature of the data, we consider various non-stationary GPD models, as in equation (6.2.1), that are formulated as GAMs to fit $Y \mid \boldsymbol{X}$. For threshold selection, we extend the method proposed by Murphy et al. (2025) to select a threshold for non-stationary, covariate-dependent GPD models; the details are provided in Section 6.3.2. Our inference and model selection procedures are then provided in Sections 6.3.2 and 6.3.2, respectively. We note that the same model formulation is used for both C1 and C2 with a small adjustment to the parameter estimation procedure for C2 to incorporate the provided loss function given in (6.3.2). We utilise equation (6.3.1) to obtain the marginal distribution of Y.

General model formulation

Let $\tilde{\boldsymbol{X}}_t$ denote the set of predictor covariates with $t \in \{1, \dots, n\}$. Then y_t and $\tilde{\boldsymbol{x}}_t$ denote the observations of the response variable and predictive covariates, respectively. We consider models with the following form,

$$F_{Y_t|\tilde{\boldsymbol{X}}_t}(y_t|\tilde{\boldsymbol{X}}_t = \tilde{\boldsymbol{x}}_t) = 1 - \zeta(\tilde{\boldsymbol{x}}_t)\left[1 + \xi(\tilde{\boldsymbol{x}}_t)\left(\frac{y_t - v(\tilde{\boldsymbol{x}}_t)}{\sigma(\tilde{\boldsymbol{x}}_t)}\right)\right]_+^{-1/\xi(\tilde{\boldsymbol{x}}_t)}, \tag{6.3.3}$$

where $v(\tilde{\boldsymbol{x}}_t)$ and $\zeta(\tilde{\boldsymbol{x}}_t)$ are covariate-dependent threshold and rate parameters, respectively. The rate parameter corresponds to the probability of exceeding the threshold.

Our analysis in Section 6.3.1 indicates that V_3, V_5 (season), and V_6 (wind speed) exhibit non-trivial dependence relationships with the response variable. Therefore we assume these variables can be used as predictor variables for modelling Y, and set $\tilde{\boldsymbol{x}} := (\boldsymbol{V}_j)_{j\in\{3,5,6\}}$. Although V_7 (wind direction) also exhibits strong dependence with Y, we do not consider it here since it is highly correlated with wind speed so would involve adding complex interaction terms to the model formulation, and V_6 has a stronger relationship with Y compared to V_7, as measured by each of Kendall's τ, χ and η (see

Figure 6.3.1).

Owing to the complex covariate structure observed in the data, as described in Section 6.3.1, we employ the flexible EVGAM framework proposed in Youngman (2019) for modelling tail behaviour. Under this framework, GAM formulations are used to capture non-stationarity in the threshold, scale and shape functions given in equation (6.3.3). Without loss of generality, consider the scale function $\sigma(\cdot)$. We assume that

$$h(\sigma(\tilde{\boldsymbol{x}})) = \psi_\sigma(\tilde{\boldsymbol{x}}), \quad \text{with} \quad \psi_\sigma(\tilde{\boldsymbol{x}}) = \beta_0 + \sum_{\kappa=1}^{K} \sum_{p=1}^{P_\kappa} \beta_{\kappa p} b_{\kappa p}(\tilde{\boldsymbol{x}}), \tag{6.3.4}$$

where $h(x) := \log(x)$ denotes the link function which ensures the correct support, with coefficients $\beta_0, \beta_{\kappa p} \in \mathbb{R}$ and basis functions $b_{\kappa p}$ for $p \in \{1, \ldots, P_\kappa\}, \kappa \in \{1, \ldots, K\}$, where K is the number of splines in the GAM formulation and P_κ is the basis dimension relating to spline κ. The basis functions can be in terms of individual covariates, i.e., $b_{\kappa p} : \mathbb{R} \mapsto \mathbb{R}$, or multiple covariates, i.e., $b_{\kappa p} : \mathbb{R}^m \mapsto \mathbb{R}$, $1 < m \leq 8$. Analogous forms can be taken for $v(\cdot)$ and $\xi(\cdot)$, adjusting the link function $h(\cdot)$ as appropriate, although these are not considered here for reasons detailed below.

To select an appropriate threshold, we employ the threshold selection method of Murphy et al. (2025), corresponding to Chapter 3, and extend this approach to select a threshold for non-stationary, covariate-dependent GPD models. The method selects a threshold based on minimising the expected quantile discrepancy (EQD) between the sample quantiles and fitted GPD model quantiles. When fitting a non-stationary model, the excesses will not be identically distributed across covariates. Thus, to utilise the EQD method in this case, we use the fitted non-stationary GPD parameter estimates to transform the excesses to common standard exponential margins and compare sample quantiles against theoretical quantiles from the standard exponential distribution, similar to the Varty et al. (2021) approach which the EQD built upon. This transformation is a common approach for checking the model fit of a non-stationary GPD (Coles, 2001).

We follow Murphy et al. (2025) and set the tuning parameters at the default values of $m = 500$ and $B = 100$. Increasing B in this non-IID setting would be beneficial, however due to the fine grid of thresholds, the incorporation of the `evgam` package, and the number of models we assessed (only a subset are shown in the chapter), we chose not to increase the computational intensity any further.

We use a bi-seasonal piecewise-constant threshold as there is clear variation in the distribution, and thereby the extremes, of Y between seasons; see Appendix D.1 for more details. Specifically, we set $v(\tilde{\boldsymbol{x}}_t) := \mathbb{1}(\tilde{x}_{2,t} = 1)v_1 + \mathbb{1}(\tilde{x}_{2,t} = 2)v_2$, $v_1, v_2 \in \mathbb{R}$, with corresponding rate parameter $\zeta(\tilde{\boldsymbol{x}}_t) := \mathbb{1}(\tilde{x}_{2,t} = 1)\zeta_1 + \mathbb{1}(\tilde{x}_{2,t} = 2)\zeta_2$, where $\zeta_1, \zeta_2 \in [0, 1]$ denote the probabilities of exceeding the threshold for seasons 1 and 2, respectively, and $\tilde{x}_{r,t}$ are realisations of the r^{th} component of $\tilde{\boldsymbol{x}}$ for $r \in \{1, 2, 3\}$. This seasonal threshold significantly improves model fits; see Appendix D.1 for further details. GAM forms for the threshold were also explored, but did not offer significant improvement. Furthermore, the smooth GAM formulation of the GPD scale parameter adequately captures any residual variation in the response arising due to covariate dependence.

Inference

For all GAM formulations, we only consider basis functions of singular covariates, since specifying basis functions of multiple variables requires a detailed understanding of covariate interactions and can significantly increase the computational complexity of the modelling procedure (Wood, 2017). We keep the shape function $\xi(\boldsymbol{x}) := \xi \in \mathbb{R}$ constant across covariates; this is common in non-stationary analyses, since this parameter is difficult to estimate (Chavez-Demoulin and Davison, 2005). Within the GAM formulation, we consider several parametric forms to account for the predictive covariates in the scale parameter using linear models, indicator functions and thin-plate regression splines.

When using splines, we must select a basis dimension $P_\kappa \in \mathbb{N}$; this determines the number of coefficients to be estimated. Basis dimension is the most important choice within spline modelling procedures and directly corresponds with the flexibility of the framework (Wood, 2017). We only consider splines for V_3 and V_6. For each $\tilde{X}_r$, $r \in \{1, 3\}$, we determine the basis dimension P_1 and P_2, respectively, by building a model for $Y_t \mid \tilde{X}_{r,t}$, to consider the effect of this predictor on the response directly. We vary the basis dimension and compare the resulting models using cross validation (CV), detailed in the following section. We set $P_1 = 4$ and $P_2 = 3$ for V_3 and V_6, respectively.

For C2, we incorporate the loss function of equation (6.3.2) into the estimation procedure. Let $\mathcal{I}_v := \{t \in \{1, \ldots, n\} \mid y_t > v(\tilde{\boldsymbol{x}}_t)\}$ denote the set of temporal indices corresponding to threshold exceedances and $n_v := |\mathcal{I}_v|$. We consider the objective function

$$S(\boldsymbol{\theta}) := -l_R(\boldsymbol{\theta}) + \sum_{i \in \mathcal{I}_v} \mathcal{L}(q_i^*, \hat{q}_i)/n_v, \tag{6.3.5}$$

where $l_R(\boldsymbol{\theta})$ denotes the penalised log-likelihood function of the restricted maximum likelihood estimation (REML) approach (Wood, 2017), $\boldsymbol{\theta}$ denotes the parameter vector associated with the GPD formulation of equation (6.3.4), and $\sum_{i \in \mathcal{I}_v} \mathcal{L}(q_i^*, \hat{q}_i)/n_v$ denotes the average loss between the sample quantiles of the transformed excesses and the theoretical standard exponential quantiles. Specifically, we transform the excesses, $(y_t - v(\tilde{\boldsymbol{x}}_t))_{t \in \mathcal{I}_v}$, to standard exponential margins using the fitted non-stationary GPD parameter estimates and compare the ordered excesses, $\boldsymbol{q}^*$, to the theoretical quantiles, $\hat{\boldsymbol{q}}$, from a standard exponential distribution evaluated at probabilities $\{p_i = i/(n_v + 1), i = 1, \ldots, n_v\}$. Minimising the objective function $S(\boldsymbol{\theta})$ ensures that the parameter estimates also account for and minimise the loss function, $\mathcal{L}$. We use this formulation to adjust the GPD parameters for challenge C2 once a threshold is selected.

Model selection

To determine the best-fitting model, we use a forward selection process and aim to minimise the model's CV score. For each model, we apply k-fold CV (Hastie et al., 2001, Ch 7.) utilising the continuous ranked probability score (CRPS, Gneiting and Katzfuss, 2014) as our goodness-of-fit metric. CRPS describes the discrepancy between the predicted distribution function and observed values without the specification of empirical quantiles. We explore model ranking by taking both $k = 10$ and 50, and find that both give an equivalent ranking; we present results for the latter. We also provide the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values to aid in model selection. A subset of models used in the forward selection process are detailed in Table 6.3.1 where, for each model, we provide the change in the CRPS, AIC and BIC relative to model 1. The parameterisation of model 7 achieves the largest reduction for all three metrics relative to the baseline model.

Table 6.3.1: Table of selected models considered for challenge C1. $\mathbb{1}(\cdot)$ denotes an indicator function, $s_i(\cdot)$ for $i \in \{1, 2\}$ denote thin-plate regression splines, β_0, β_1 are coefficients to be estimated, and $\tilde{x}_{r,t}$ is defined as in the text. All values have been given to one decimal place.

Model	$\sigma(\tilde{\boldsymbol{x}}_t)$	ΔCRPS	ΔAIC	ΔBIC
1	β_0	0	0	0
2	$\beta_0 + \beta_1 \mathbb{1}(\tilde{x}_{2,t} = 1)$	-0.5	-33.4	-26.1
3	$\beta_0 + s_1(\tilde{x}_{1,t})$	-0.9	-408.5	-379.2
4	$\beta_0 + s_2(\tilde{x}_{3,t})$	-0.5	-284.3	-276.8
5	$\beta_0 + \beta_1 \mathbb{1}(\tilde{x}_{2,t} = 1) + s_1(\tilde{x}_{1,t})$	-0.9	-425.8	-388.1
6	$\beta_0 + s_1(\tilde{x}_{1,t}) + s_2(\tilde{x}_{3,t})$	-1.0	-752.7	-717.2
7	$\beta_0 + \beta_1 \mathbb{1}(\tilde{x}_{2,t} = 1) + s_1(\tilde{x}_{1,t}) + s_2(\tilde{x}_{3,t})$	**-1.1**	**-780.0**	**-735.3**

6.3.3 Uncertainty

For each of the 100 different covariate combinations, $\tilde{\boldsymbol{x}}_i$ for $i \in \{1, \ldots, 100\}$, we need to construct central 50% confidence intervals. We use a bootstrapping procedure to avoid making potentially inaccurate assumptions such as the asymptotic normality approximation of maximum likelihood estimates, for example. Traditional bootstrap approaches are non-parametric and randomly resample the data with replacement. However, in Section 6.3.1 we find that the response variable is dependent on covariates, and these covariates exhibit temporal dependence. A standard bootstrap procedure would therefore not retain this dependence. Instead, we preserve the temporal dependence structure of covariates and their relationship with the response variable by approximating our confidence intervals using the stationary, semi-parametric bootstrapping procedure adopted by D'Arcy et al. (2023).

First, the response variable Y_t is transformed to Uniform(0,1) margins to preserve its non-stationary behaviour; denote this sequence $U_t^Y = F_{Y_t|\tilde{\boldsymbol{X}}_t}(Y_t|\tilde{\boldsymbol{X}}_t = \tilde{x}_t)$ where $F_{Y_t|\tilde{\boldsymbol{X}}_t}$ is the estimated model given in equation (6.3.3). We then adopt the stationary bootstrap procedure of Politis and Romano (1994) to retain the temporal dependence in the response and explanatory variables by sampling blocks of consecutive observations. The block length L is random and simulated from a Geometric($1/l$) distribution, where the mean block length $l \in \mathbb{N}$ is carefully selected based on the autocorrelation function. This was selected at 50 days, the maximum lag for which the autocorrelation was significant across all variables; see Appendix D.1. Denote this bootstrapped sequence on Uniform margins by U_t^B. We transform U_t^B back to the original scale using our fitted model, preserving the original structure of Y_t; we denote this series Y_t^B. Then we fit our model to Y_t^B to re-estimate all of the parameters and thus the quantile of interest. We repeat this procedure to obtain 200 bootstrap samples.

6.3.4 Results

For C1, we use our final model of Section 6.3.2 to estimate the 0.9999-quantile of $Y \mid \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}_i$, $i \in \{1, \ldots, 100\}$, for the set of 100 covariate combinations. The left panel of Figure 6.3.2 shows the QQ-plot for our model. There is general alignment between the model and empirical quantiles; however, there is some over-estimation in the upper tail, and our 95% tolerance bounds do not contain some of the most extreme response values. The right panel of Figure 6.3.2 shows our predicted quantiles, and their associated confidence intervals, compared to their true quantiles. As expected,

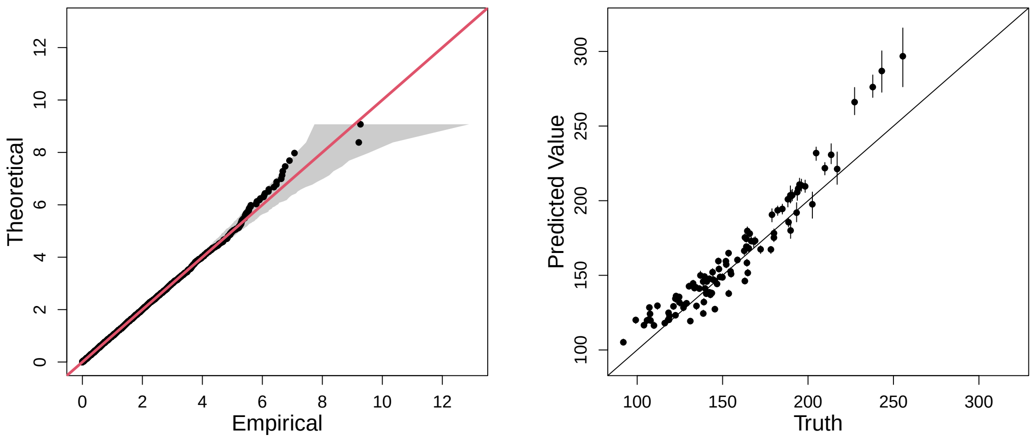

Figure 6.3.2: QQ plot for our final model (model 7 in Table 6.3.1) on standard exponential margins. The $y = x$ line is given in red and the grey region represents the 95% tolerance bounds (left). Predicted $0.9999-$quantiles against true quantiles for the 100 covariate combinations. The points are the median predicted quantile over 200 bootstrapped samples and the vertical error bars are the corresponding 50% confidence intervals. The $y = x$ line is also shown (right).

our predictions tend to over-estimate the true quantiles. We note this figure is different from the one presented by Rohrbeck et al. (2023) due to an error in our code being fixed after submission. In this scenario, our estimated confidence intervals lead to a 14% coverage of the true quantiles, which does not alter our ranking for this challenge. Our performance and model improvements are discussed in Section 4.6.

For challenge C2, we estimate the quantile of interest as $\hat{q} = 213.1\ (209.3, 242.1)$. A 95% confidence interval for the estimate is given in parentheses based on the bootstrapping procedure outlined in Section 6.3.2. Due to a coding error, this value differs from the original estimate submitted for the EVA (2023) Conference Data Challenge. The updated value over-estimates compared to the truth ($q = 196.6$).

6.4 Challenge C3

6.4.1 Exploratory data analysis

For challenge C3, we are provided with 70 years of daily data of an environmental variable for three towns on the island of Coputopia. These data are denoted by $Y_{i,t}$, $i \in \{1, 2, 3\}$, $t \in \{1, \ldots, n\}$, where i is the index of each town and t is the point in time. Each year consists of 12 months, each lasting 25 days, resulting in $n = 21,000$ observations for each location.

We are also provided with daily covariate observations $\boldsymbol{X}_t = (S_t, A_t)$, where S_t and A_t denote seasonal and atmospheric conditions, respectively. Season is a binary variable, taking values in the set $\{1, 2\}$, with each year of observations exhibiting both seasons for exactly 150 consecutive days. Atmospheric conditions are piecewise constant over months, with large variation in the observed values between months. A descriptive figure of both covariates is given in Appendix D.1.1.

In Rohrbeck et al. (2023), we are informed that $Y_{i,t}$ are distributed identically across all sites and over time, with standard Gumbel margins. However, it is not known whether the covariates $\boldsymbol{X}_t$ influence the dependence structure of $\boldsymbol{Y}_t := (Y_{1,t}, Y_{2,t}, Y_{3,t})$. We are also informed that, conditioned on covariates, the process is independent over time, i.e., $(\boldsymbol{Y}_t \mid \boldsymbol{X}_t) \perp\!\!\!\perp (\boldsymbol{Y}_{t'} \mid \boldsymbol{X}_{t'})$ for any $t \neq t'$. In this section, we examine what influence, if any, the covariate process $\boldsymbol{X}_t$ may have on the dependence structure of $\boldsymbol{Y}_t$.

We begin by transforming the time series $Y_{i,t}$ to standard exponential margins, denoted by $\boldsymbol{Z}_{i,t}$, via the probability integral transform. This transformation is common in the study of multivariate extremes and can simplify the description of extremal dependence (Keef et al., 2013b). To explore the extremal dependence in the Coputopia time series, we consider all 2- and 3-dimensional subvectors of the process, i.e., $\{Z_{i,t}, i \in I, t \in \{1, \ldots, n\}\}$, $I \in \mathcal{I} := \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$. This separation is important to ensure the overall dependence structure is fully understood, since intermediate scenarios can exist where a random vector exhibits $\chi = 0$, but $\chi > 0$ for some 2-dimensional subvector(s) (Simpson et al., 2020).

Furthermore, to explore the impact of covariates on the dependence structure, we partition the time series into subsets using the covariates. For the seasonal covariate, let $G^S_{I,j} := \{Z_{i,t}, i \in I, S_t = j\}$ for $j = 1, 2$, and for the atmospheric covariate, let $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$ denote the permutation associated with the order statistics of A_t, defined so that ties in the data are accounted for. We then split the data into 10 equally sized subsets corresponding to the atmospheric order statistics, i.e., $G^A_{I,k} := \left\{Z_{i,t}, i \in I, t \in \Sigma^k\right\}$ for $k = 1, 2, \dots, 10$, where $\Sigma^k := \{t \mid (k-1)n/10 + 1 \leq \pi(t) \leq kn/10\}$. Thus, the atmospheric values associated with each subset $G^A_{I,k}$ will increase over k.

The idea behind these subsets is to examine whether altering the values of either covariate impacts the extremal dependence structure. Consequently, we set $u = 0.9$ and estimate $\chi(u)$ using the techniques outlined in Section 6.2, with uncertainty quantified through bootstrapping with 200 samples. The bootstrapped χ estimates for $G^A_{I,k}$ with $I = \{1, 2, 3\}$ are given in Figure 6.4.1. The plots for the remaining index sets in $\mathcal{I}$, along with the subsets associated with the seasonal covariate, are given in Appendix D.1.1. The estimates of χ appear to vary, in the majority of cases, across both subset types (seasonal and atmospheric), suggesting both covariates have an impact on the dependence structure. For the atmospheric process in particular, the values of χ tend to decrease for higher atmospheric values, suggesting a negative association between the strength of positive extremal dependence and the atmospheric covariate. We also observe that across all subsets, χ appears consistently low in magnitude, suggesting the extremes of some, if not all, of the sub-vectors are unlikely to occur simultaneously. As such, for modelling the Coputopia time series, we require a framework that can capture such forms of dependence. We also consider pointwise estimates of the function $\lambda(\cdot)$, as defined later in equation (6.4.1), over $G^S_{I,j}$ and $G^A_{I,k}$ for fixed simplex points; these results are given in Appendix D.1.1. Similar to χ, estimates of $\lambda(\cdot)$ vary significantly across subsets, providing additional evidence of non-stationarity within extremal dependence structure.

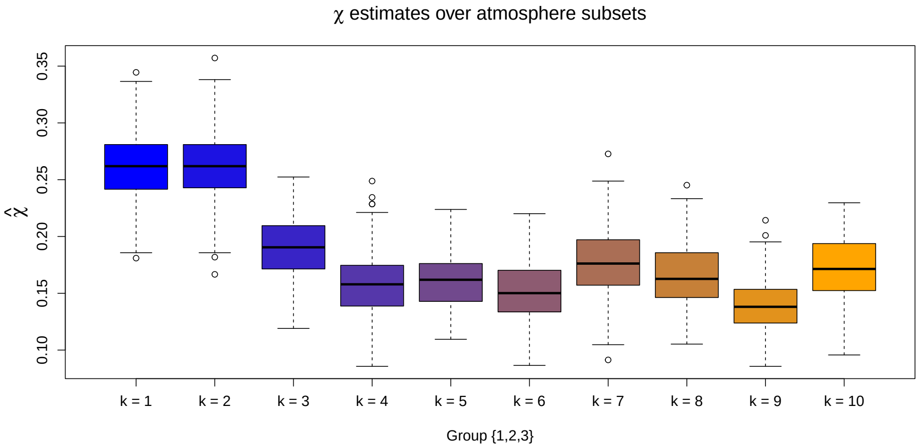

Figure 6.4.1: Boxplots of empirical χ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in χ estimates as the atmospheric values are increased.

6.4.2 Joint tail probabilities under asymptotic independence

For challenge C3, we are required to estimate $p_1 := \Pr(Y_1 > y, Y_2 > y, Y_3 > y)$ and $p_2 := \Pr(Y_1 > v, Y_2 > v, Y_3 < m)$, with $y = 6$, $v = 7$ and $m = -\log(\log(2))$. Note that p_1 and p_2 are independent of the covariate process and correspond to different extremal regions in $\mathbb{R}^3$; we refer to p_1 and p_2 as parts 1 and 2 of the challenge, respectively. For the remainder of this section we will consider the transformed exponential variables (Z_1, Z_2, Z_3), omitting the subscript t for ease of notation. Observe that $F_{(-Z_3)}(z) = e^z$, for $z < 0$; setting $\tilde{Z}_3 := -\log(1 - \exp(-Z_3))$, we have

$$p_2 = \Pr(Z_1 > \tilde{v}, Z_2 > \tilde{v}, Z_3 < \tilde{m}) = \Pr\left(Z_1 > \tilde{v}, Z_2 > \tilde{v}, \tilde{Z}_3 > \tilde{m}\right),$$

where $\tilde{v}$ and $\tilde{m}$ denote the values v and m transformed to the standard exponential scale, e.g., $\tilde{v} := -\log(1 - \exp(-\exp(-v)))$. Similarly, $p_1 = \Pr(Z_1 > \tilde{y}, Z_2 > \tilde{y}, Z_3 > \tilde{y})$.

Consequently, both p_1 and p_2 can be considered as joint survivor probabilities.

Since not all extremes of Z_1, Z_2 and Z_3 are observed simultaneously, we employ the framework by Wadsworth and Tawn (2013), which is a generalisation of the approach proposed in Ledford and Tawn (1996). The model of Wadsworth and Tawn (2013) assumes that for any ray $\boldsymbol{\omega} \in \boldsymbol{S}^2 := \{(\omega_1, \omega_2, \omega_3) \in [0,1]^3 : \omega_1 + \omega_2 + \omega_3 = 1\}$, where $\boldsymbol{S}^2$ denotes the standard 2-dimensional simplex,

$$\begin{aligned}\Pr\left(Z_1 > \omega_1 r,\ Z_2 > \omega_2 r,\ Z_3 > \omega_3 r\right) =& \Pr\left(\min\{Z_1/\omega_1,\ Z_2/\omega_2,\ Z_3/\omega_3\} > r\right)\\ =& \mathcal{L}(e^r; \boldsymbol{\omega}) e^{-r\lambda(\boldsymbol{\omega})}, \qquad (6.4.1)\end{aligned}$$

as $r \to \infty$, where $\lambda(\boldsymbol{\omega}) \geq \max(\boldsymbol{\omega})$ is known as the angular dependence function (ADF). Asymptotic dependence occurs at the lower bound, i.e., $\lambda(\boldsymbol{\omega}) = \max(\boldsymbol{\omega})$ for all $\boldsymbol{\omega} \in \boldsymbol{S}^2$, and the coefficient of tail dependence is related to the ADF via $\eta = 1/\{3\lambda(1/3, 1/3, 1/3)\}$. In practice, equation (6.4.1) can be used to evaluate extreme joint survivor probabilities; in particular, probabilities p_1 and p_2 can be identified with the rays $\boldsymbol{\omega}^{(1)} := (\tilde{y}, \tilde{y}, \tilde{y})/r^{(1)}$ and $\boldsymbol{\omega}^{(2)} := (\tilde{v}, \tilde{v}, \tilde{m})/r^{(2)}$ in $\boldsymbol{S}^2$, respectively, where $r^{(1)} := \tilde{y} + \tilde{y} + \tilde{y}$ and $r^{(2)} := \tilde{v} + \tilde{v} + \tilde{m}$. See Section 6.4.4 for further details.

6.4.3 Accounting for non-stationary dependence

In the stationary setting, pointwise estimates of $\lambda(\cdot)$ can be obtained via the Hill estimator (Hill, 1975), from which tail probabilities can be approximated. However, alternative procedures are required for data exhibiting trends in dependence, such as the Coputopia data set. Existing approaches for capturing non-stationary dependence structures are sparse in the extremes literature, and most approaches are limited to asymptotically dependent data structures. For the case when data are not asymptotically dependent, Mhalla et al. (2019) and Murphy-Barltrop and Wadsworth (2024) propose non-stationary extensions of the Wadsworth and Tawn (2013) framework, while

Jonathan et al. (2014) and Guerrero et al. (2023) propose non-stationary extensions of the Heffernan and Tawn (2004) model (see Murphy-Barltrop and Wadsworth (2024) for a detailed review).

To account for non-stationary dependence in C3, we propose an extension of the Wadsworth and Tawn (2013) framework. With $\boldsymbol{Z}_t = (Z_{1,t}, Z_{2,t}, Z_{3,t})$ and $\boldsymbol{X}_t$, defined as in Section 6.4.1, we define the structure variable $T_{\boldsymbol{\omega},t} := \min\{Z_{1,t}/\omega_1, Z_{2,t}/\omega_2, Z_{3,t}/\omega_3\}$, for any $\boldsymbol{\omega} \in \boldsymbol{S}^2$; we refer to $T_{\boldsymbol{\omega},t}$ as the min-projection variable at time t. From Section 6.4.1, we know that the joint distribution of $\boldsymbol{Z}_t$ is not identically distributed over t; which implies non-stationarity in the distribution of $T_{\boldsymbol{\omega},t}$. To account for this, Mhalla et al. (2019) and Murphy-Barltrop and Wadsworth (2024) assume the following model given the vector of covariates $\boldsymbol{x}_t$:

$$\Pr\left(T_{\boldsymbol{\omega},t} > u \mid \boldsymbol{X}_t = \boldsymbol{x}_t\right) = \mathcal{L}\left(e^u \mid \boldsymbol{\omega}, \boldsymbol{x}_t\right) e^{-\lambda(\boldsymbol{\omega};\boldsymbol{x}_t)u} \text{ as } u \to \infty, \tag{6.4.2}$$

for all t, where $\lambda\left(\cdot;\boldsymbol{x}_t\right)$ denotes the non-stationary ADF. Note that this assumption is very similar in form to equation (6.4.1), with the primary difference being the function $\lambda(\cdot;\boldsymbol{x}_t)$ is non-stationary over t. From equation (6.4.2), we have

$$\Pr\left(T_{\boldsymbol{\omega},t} - u > z \mid T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\right) = e^{-\lambda(\boldsymbol{\omega};\boldsymbol{x}_t)z} \text{ as } u \to \infty, \tag{6.4.3}$$

for $z > 0$. Consequently, equation (6.4.2) is equivalent to assuming $(T_{\boldsymbol{\omega},t} - u) \mid \{T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\} \sim \text{Exp}(\lambda\left(\boldsymbol{\omega};\boldsymbol{x}_t\right))$ as $u \to \infty$.

We found that equation (6.4.2) was not flexible enough to capture the tail of $T_{\boldsymbol{\omega},t}$ for the Coputopia data; see Section 6.4.3 for further discussion. Thus, we propose the following model: given any $z > 0$ and a fixed $\boldsymbol{\omega} \in \boldsymbol{S}^2$, we assume

$$\Pr\left(T_{\boldsymbol{\omega},t} - u > z \mid T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\right) = \left(1 + \frac{\xi\left(\boldsymbol{\omega};\boldsymbol{x}_t\right) z}{\sigma\left(\boldsymbol{\omega};\boldsymbol{x}_t\right)}\right)^{-1/\xi(\boldsymbol{\omega};\boldsymbol{x}_t)} \text{ as } u \to \infty, \tag{6.4.4}$$

where $\sigma(\cdot;\boldsymbol{x}_t), \xi(\cdot;\boldsymbol{x}_t)$ are non-stationary scale and shape parameter functions, respectively. This is equivalent to assuming $(T_{\boldsymbol{\omega},t}-u) \mid \{T_{\boldsymbol{\omega},t} > u, \boldsymbol{X}_t = \boldsymbol{x}_t\} \sim \text{GPD}(\sigma\left(\boldsymbol{\omega};\boldsymbol{x}_t\right), \xi\left(\boldsymbol{\omega};\boldsymbol{x}_t\right))$ as $u \to \infty$, and equation (6.4.3) is recovered by taking the limit as $\xi\left(\boldsymbol{\omega};\boldsymbol{x}_t\right) \to 0$ for all t.

Our proposed formulation in equation (6.4.4) allows for additional flexibility within the modelling framework by including a GPD shape parameter $\xi\left(\boldsymbol{\omega};\boldsymbol{x}_t\right)$, which quantifies the tail behaviour of $T_{\boldsymbol{\omega},t}$. Given the wide range of distributions in the domain of attraction of a GPD (Pickands, 1975), it is reasonable to assume that the tail of $T_{\boldsymbol{\omega},t}$ can be approximated by equation (6.4.4). For the Coputopia time series, this assumption appears valid, as demonstrated by the diagnostics in Section 6.4.3.

Model fitting

To apply equation (6.4.4), we first fix $\boldsymbol{\omega} \in \boldsymbol{S}^2$ and assume that the formulation holds approximately for some sufficiently high threshold level from the distribution of $T_{\boldsymbol{\omega},t}$; we denote the corresponding quantile level by $\tau \in (0,1)$. For simplicity, the same quantile level is considered across all t. Further, let $v_\tau(\boldsymbol{\omega},\boldsymbol{x}_t)$ denote the corresponding threshold function, i.e., $\Pr(T_{\boldsymbol{\omega},t} \leq v_\tau(\boldsymbol{\omega},\boldsymbol{x}_t) \mid \boldsymbol{X}_t = \boldsymbol{x}_t) = \tau$ for all t. Under our assumption, we have $(T_{\boldsymbol{\omega},t} - v_\tau(\boldsymbol{\omega},\boldsymbol{x}_t)) \mid \{T_{\boldsymbol{\omega},t} > v_\tau(\boldsymbol{\omega},\boldsymbol{x}_t), \boldsymbol{X}_t = \boldsymbol{x}_t\} \sim \text{GPD}(\sigma\left(\boldsymbol{\omega};\boldsymbol{x}_t\right), \xi\left(\boldsymbol{\omega};\boldsymbol{x}_t\right))$. We emphasise that $v_\tau(\boldsymbol{\omega},\boldsymbol{x}_t)$ is not constant in t, and we would generally expect $v_\tau(\boldsymbol{\omega},\boldsymbol{x}_t) \neq v_\tau(\boldsymbol{\omega},\boldsymbol{x}_{t'})$ for $t \neq t'$.

As detailed in Section 6.4.2, both p_1 and p_2 can be associated with points on the simplex $\boldsymbol{S}^2$, denoted by $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$, respectively. Letting $\boldsymbol{\omega} \in \{\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}\}$, our estimation procedure consists of two stages: estimation of the threshold function $v_\tau(\boldsymbol{\omega},\boldsymbol{z}_t)$ for a fixed $\tau \in (0,1)$, followed by estimation of GPD parameter functions $\sigma\left(\boldsymbol{\omega};\boldsymbol{x}_t\right), \xi\left(\boldsymbol{\omega};\boldsymbol{x}_t\right)$. For both steps, we take a similar approach to Section 6.3.2 and use GAMs to capture these covariate relationships. To simplify our approach, we falsely assume that the atmospheric covariate A_t is continuous over t; this step allows us to utilise GAM formulations containing smooth basis functions. Given the significant variability in A_t between months, discrete formulations for this covariate would significantly increase the number of model parameters and result in higher variability.

Let $\log(v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)) = \psi_v(\boldsymbol{x}_t)$, $\log(\sigma(\boldsymbol{\omega}; \boldsymbol{x}_t)) = \psi_\sigma(\boldsymbol{x}_t)$ and $\xi(\boldsymbol{\omega}; \boldsymbol{x}_t) = \psi_\xi(\boldsymbol{x}_t)$ denote the GAM formulations of each function, where $\psi_\cdot$ denotes the basis representation of equation (6.3.4). Exact forms of basis functions are specified in Section 6.4.3. As in Section 6.3.2, model fitting is carried out using the `evgam` software package (Youngman, 2022). For the first stage, $v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)$ is estimated by exploiting a link between the loss function typically used for quantile regression and the asymmetric Laplace distribution (Yu and Moyeed, 2001). The spline coefficients associated with ψ_σ and ψ_ξ are estimated subsequently using the obtained threshold exceedances.

Selection of GAM formulations and diagnostics

Prior to estimation of the threshold and parameter functions, we specify a quantile level τ and formulations for each of the GAMs. To begin, we fix $\tau = 0.9$ and consider formulations for each ψ_v, ψ_σ and ψ_ξ. By comparing metrics for model selection, namely AIC, BIC and CRPS, we found the following formulations to be sufficient

$$\psi_v(\boldsymbol{x}_t) = \beta_u + s_v(a_t) + \beta_s \mathbb{1}(s_t = 2), \quad \psi_\sigma(\boldsymbol{x}_t) = \beta_\sigma + s_\sigma(a_t) \quad \text{and} \quad \psi_\xi(\boldsymbol{x}_t) = \beta_\xi, \tag{6.4.5}$$

for parts 1 and 2, where $\beta_u, \beta_\sigma, \beta_\xi \in \mathbb{R}$ denote constant intercept terms, $\mathbb{1}$ denotes the indicator function with corresponding coefficient $\beta_s \in \mathbb{R}$, and s_u, s_σ denote cubic regression splines of dimension 10. The shape parameter is set to constant for the reasons outlined in Section 6.3.2. Cubic basis functions are used for ψ_v and ψ_σ since they have several desirable properties, including continuity and smoothness (Wood, 2017). A dimension of size 10 appears more than sufficient to capture the trends relating to the atmosphere variable. Alternative formulations were tested for both parts, but this made little difference to the resulting model fits.

We remark that the seasonal covariate is only present with the formulation for ψ_v. Once accounted for in the non-stationary threshold, the seasonal covariate appeared to have little influence on the fitted GPD parameters. More complex GAM formulations were tested involving interaction terms between the seasonal and atmospheric covariates, which showed little to no improvement in model fits. Thus, we prefer the simpler formulations on the basis of parsimony.

We now consider the quantile level $\tau \in (0,1)$. To assess sensitivity in our formulation, we set $\mathrm{T} := \{0.8, 0.81, \dots, 0.99\}$ and fit the GAMs outlined in equation (6.4.5) for each $\tau \in \mathrm{T}$. Letting $\delta_{\boldsymbol{\omega},t}$ and $\mathcal{T}_\tau := \{t \in \{1,\dots,n\} \mid \delta_{\boldsymbol{\omega},t} > v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)\}$ denote the min-projection observations and indices of threshold-exceeding observations, respectively, we expect the set $\mathcal{E} := \{-\log\{1 - F_{GPD}(\delta_{\boldsymbol{\omega},t} - v_\tau(\boldsymbol{\omega}, \boldsymbol{x}_t)) \mid \sigma(\boldsymbol{\omega}; \boldsymbol{x}_t), \xi(\boldsymbol{\omega}; \boldsymbol{x}_t)\} \mid t \in \mathcal{T}_\tau\}$ to follow a standard exponential distribution.

With all exceedances transformed to a unified scale, we compare the empirical and model exponential quantiles using QQ plots, through which we assess the relative performance of each $\tau \in \mathrm{T}$. We selected τ values for which the empirical and theoretical quantiles appeared most similar in magnitude. From this analysis, we set $\tau = 0.83$ and $\tau = 0.85$ for parts 1 and 2, respectively. The corresponding QQ plots are given in Figure 6.4.2, where we observe reasonable agreement between the empirical and theoretical quantiles. However, whilst these values appeared acceptable within T, we stress that adequate model fits were also obtained for other quantile levels, suggesting our modelling procedure is not particularly sensitive to the exact choice of quantile. Furthermore, we also tested a range of quantile levels below the 0.8-level, but were unable to improve the quality of model fits.

Plots illustrating the estimated GPD scale parameter functions are given in Appendix D.1.1, with the resulting dependence trends in agreement with the observed trends from Section 6.4.1. We also remark that the estimated GPD shape parameters obtained for parts 1 and 2 were $0.042\,(0.01, 0.075)$ and $0.094\,(0.059, 0.128)$, respectively,

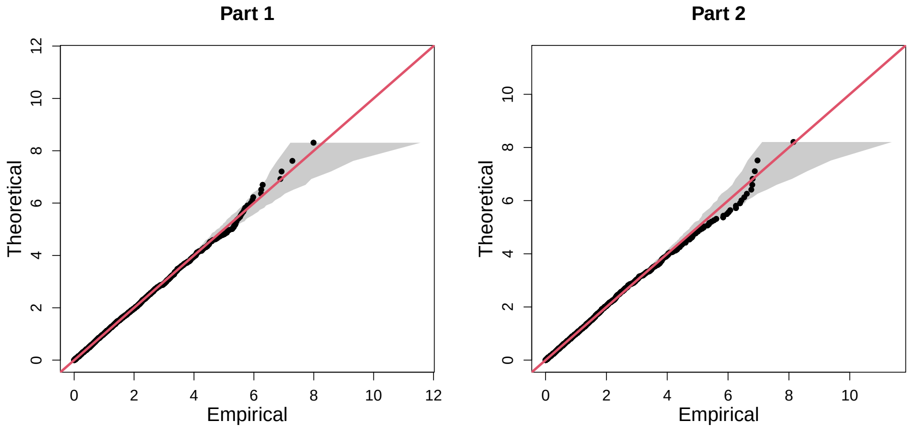

Figure 6.4.2: Final QQ plots for parts 1 (left) and 2 (right) of C3, with the $y = x$ line given in red. In both cases, the grey regions represent the 95% bootstrapped tolerance bounds.

where the brackets denote 95% credible intervals obtained using posterior sampling (Wood, 2017). These estimates, which indicate slightly heavy-tailed behaviour within the min-projection variable, provide insight into why the original exponential modelling framework is not appropriate for C3.

Overall, these results suggest different extremal dependence trends exist for the two simplex points $\boldsymbol{\omega}^{(1)}$ and $\boldsymbol{\omega}^{(2)}$, illustrating the importance of the flexibility in our model. These findings are also in agreement with empirical trends observed in Section 6.4.1, suggesting our modelling framework is successfully capturing the underlying extremal dependence structures.

6.4.4 Results

Given estimates of threshold and parameter functions, probability estimates can be obtained via Monte Carlo techniques. Taking p_1, for instance, we have

$$\begin{aligned}
p_1 =& \Pr(Z_1 > \tilde{y}, Z_2 > \tilde{y}, Z_3 > \tilde{y}) \\
=& \Pr\left(\min\left(Z_1/\omega_1^{(1)}, Z_2/\omega_2^{(1)}, Z_3/\omega_3^{(1)}\right) > r^{(1)}\right) \\
=& \int_{\boldsymbol{X}_t} \Pr\left(T_{\boldsymbol{\omega}^{(1)},\, t} > r^{(1)} \mid \boldsymbol{X}_t = \boldsymbol{x}_t\right) f_{\boldsymbol{X}_t}(\boldsymbol{x}_t) \mathrm{d}\boldsymbol{x}_t \\
=& (1-\tau) \int_{\boldsymbol{X}_t} \Pr(T_{\boldsymbol{\omega}^{(1)},\, t} > r^{(1)} \mid T_{\boldsymbol{\omega}^{(1)},\, t} > v_\tau(\boldsymbol{\omega}^{(1)}, \boldsymbol{x}_t), \boldsymbol{X}_t = \boldsymbol{x}_t) f_{\boldsymbol{X}_t}(\boldsymbol{x}_t) \mathrm{d}\boldsymbol{x}_t \\
\approx& \frac{1-\tau}{n} \sum_{t=1}^{n} \left(1 + \frac{\xi(\boldsymbol{\omega}^{(1)}; \boldsymbol{x}_t)\left(r^{(1)} - v_\tau(\boldsymbol{\omega}^{(1)}, \boldsymbol{x}_t)\right)}{\sigma\left(\boldsymbol{\omega}^{(1)}; \boldsymbol{x}_t\right)}\right)^{-1/\xi\left(\boldsymbol{\omega}^{(1)}; \boldsymbol{x}_t\right)},
\end{aligned}$$

assuming $\{\boldsymbol{x}_t : t \in \{1, \ldots, n\}\}$ is a representative sample from $\boldsymbol{X}_t$. The procedure for p_2 is analogous. We note that this estimation procedure is only valid when $r^{(1)} > v_\tau(\boldsymbol{\omega}^{(1)}, \boldsymbol{x}_t)$, or $r^{(2)} > v_\tau(\boldsymbol{\omega}^{(2)}, \boldsymbol{x}_t)$, for all t: however, for each $\tau \in \mathrm{T}$, this inequality is always satisfied, owing to the very extreme nature of the probabilities in question. Through this approximation, we obtain $\hat{p}_1 = 1.480 \times 10^{-5}$ and $\hat{p}_2 = 2.461 \times 10^{-5}$.

6.5 Challenge C4

6.5.1 Exploratory data analysis

Challenge C4 entails estimating survival probabilities across 50 locations on the island of Utopula. As stated in Rohrbeck et al. (2023), the Utopula island is split in two administrative areas, for which the respective regional governments 1 and 2 have collected data concerning the variables $Y_{i,t}$, $i \in I = \{1, \ldots, 50\}$, $t \in \{1, \ldots, 10, 000\}$. Index i denotes the i^{th} location, with locations $i \in \{1, \ldots, 25\}$ and $i \in \{26, \ldots, 50\}$ belonging to the administrative areas of governments 1 and 2, respectively. Index t denotes the timepoint in days; however, since $Y_{i,t}$ are IID for all i, we drop the subscript t for the remainder of this section.

Since many multivariate extreme value models are only applicable in low-to-moderate dimensions, we consider dimension reduction based on an exploration of the extremal dependence structure of the data. In particular, we analyse pairwise estimates of the extremal dependence coefficient $\chi(u)$, introduced in equation (6.2.2), for all possible pairwise combinations of sites; the resulting estimates, using $u = 0.95$, are presented in the heat map of Figure 6.5.1. Identification of any dependence clusters is achieved through visual investigation, which seems appropriate for this data. We note, however, that should visual considerations not suffice, alternative more sophisticated clustering methods are available and can be applied; see for example Bernard et al. (2013).

Figure 6.5.1 suggests the existence of 5 distinct subgroups where all variables within each subgroup have similar extremal dependence characteristics, while variables in different subgroups appear to be approximately independent of each other in the extremes. It is worth mentioning that the same clusters are identified when we analyse pairwise estimates of the extremal dependence coefficient $\eta(u)$; the resulting estimates can be found in Appendix D.1.2. Moreover, examining the magnitudes of $\chi(\cdot)$ and $\eta(\cdot)$ estimates, it does not appear reasonable to assume asymptotic dependence between variables in the same group. We therefore consider models that can be applied to data structures that do not take their extreme values simultaneously. The indices of the five aforementioned subgroups are $G_1 = \{4, 14, 19, 28, 30, 38, 43, 44\}$, $G_2 = \{3, 10, 15, 18, 22, 29, 45, 47\}$, $G_3 = \{8, 21, 25, 26, 32, 33, 34, 40, 41, 42, 48, 49, 50\}$, $G_4 = \{1, 2, 5, 7, 9, 17, 20, 31, 46\}$ and $G_5 = \{6, 11, 12, 13, 16, 23, 24, 27, 35, 36, 37, 39\}$. Groups G_1 and G_2 include the most strongly dependent variables (shown by the darkest color blocks in Figure 6.5.1), followed by group G_3, while groups G_4 and G_5 contain the most weakly dependent variables. We henceforth assume independence between these groups of variables, i.e., $\Pr((Y_i)_{i\in G_k} \in A_k, (Y_i)_{i\in G_{k'}} \in A_{k'}) = \Pr((Y_i)_{i\in G_k} \in A_k) \Pr((Y_i)_{i\in G_{k'}} \in A_{k'})$, $A_k \subset$

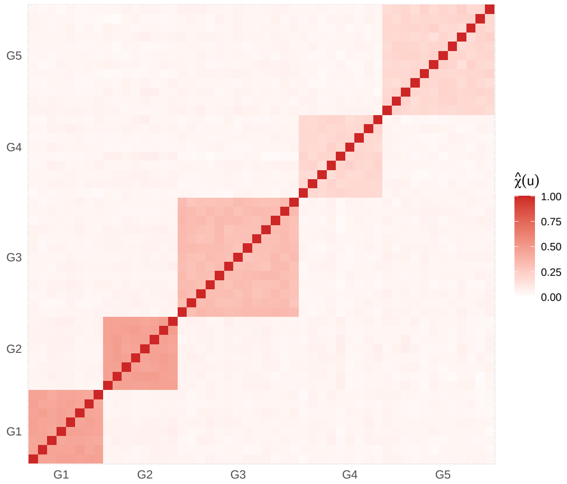

Figure 6.5.1: Heat map of estimated empirical pairwise $\chi(u)$ extremal dependence coefficients with $u = 0.95$.

$\mathbb{R}^{|G_k|}$, $A_{k'} \subset \mathbb{R}^{|G_{k'}|}$, for any $k \neq k' \in \{1, \ldots, 5\}$.

Challenge C4 requires us to estimate the probabilities $p_1 = \Pr(Y_i > s_i;\, i \in I)$ and $p_2 = \Pr(Y_i > s_1;\, i \in I)$, where $s_i := \mathbb{1}(i \in \{1, 2, \ldots, 25\})s_1 + \mathbb{1}(i \in \{26, 27, \ldots, 50\})s_2$ and s_1 (s_2) denotes the marginal level exceeded once every year (month) on average. Under the assumption of independence between groups, the challenge can be broken down to 5 lower-dimensional challenges involving the estimation of joint tail probabilities for each G_k, $k \in \{1, \ldots, 5\}$. These can then be multiplied together to obtain the required overall probabilities due to (assumed) between-group independence; specifically, we have $p_1 = \prod_{k=1}^{5} \Pr(Y_i > s_i;\ i \in G_k)$ and $p_2 = \prod_{k=1}^{5} \Pr(Y_i > s_1;\ i \in G_k)$.

6.5.2 Conditional extremes

The conditional multivariate extreme value model (CMEVM) of Heffernan and Tawn (2004) provides a flexible framework capable of capturing a range of extremal dependence forms without making assumptions about the specific form of joint dependence structure. Consider a d-dimensional random variable $\boldsymbol{W} = (W_1, \ldots, W_d)$ on standard Laplace margins. For $i \in \{1, \ldots, d\}$, the CMEVM approach assumes the existence of parameter vectors $\boldsymbol{\alpha}_{-|i} \in [-1,1]^{d-1}$ and $\boldsymbol{\beta}_{-|i} \in (-\infty,1]^{d-1}$ such that

$$\lim_{u_i \to \infty} \Pr\left\{\boldsymbol{W}_{-i} \leq \boldsymbol{\alpha}_{-|i} W_i + W_i^{\boldsymbol{\beta}_{-|i}} \boldsymbol{z}_{|i}, W_i - u_i > w \mid W_i > u_i\right\} = e^{-w} H_{|i}\left(\boldsymbol{z}_{|i}\right), \quad w > 0,$$

with non-degenerate distribution function $H_{|i}(\cdot)$, vector operations being applied componentwise, and conditional threshold u_i. The vector $\boldsymbol{W}_{-i}$ denotes $\boldsymbol{W}$ excluding its i^{th} component and $\boldsymbol{z}_{|i}$ is within the support of the residual random vector $\boldsymbol{Z}_{|i} = (\boldsymbol{W}_{-i} - \boldsymbol{\alpha}_{-|i} w_i)/w_i^{\boldsymbol{\beta}_{-|i}} \sim H_{|i}(\cdot)$. We apply this model to data where $W_i > u_i$, for some finite conditioning threshold u_i, to estimate the probabilities p_1 and p_2 defined in Section 6.5.1, using the inference procedure of Keef et al. (2013b).

6.5.3 Results

Let $\boldsymbol{W} := (W_1, \ldots, W_{50})$ denote the random vector after transformation to standard Laplace margins. This vector is divided into the five subgroups identified in Section 6.5.1, and the subgroup probabilities are estimated using predictions obtained from the sampling method of Heffernan and Tawn (2004). We condition on the first variable of each subgroup being extreme, and simulate 10^8 predictions from each of the resulting fitted conditional extremes models. To account for uncertainty in the estimates, we perform a parametric bootstrapping procedure with 100 samples.

Sensitivity analyses of the estimated probabilities to the choice of conditioning variable suggest no significant effect. Furthermore, we consider a range of conditioning thresholds; the corresponding estimates of subgroup probabilities defined in Section 6.5.1 appear relatively stable with respect to the conditioning threshold quantile. We ultimately select 0.85-quantiles for the conditioning thresholds of our final probability estimates. These are given by $\hat{p}_1 = 1.094 \times 10^{-26}$ $(2.150 \times 10^{-36}, 1.359 \times 10^{-24})$ and $\hat{p}_2 = 1.076 \times 10^{-31}$ $(1.596 \times 10^{-46}, 1.850 \times 10^{-29})$, with 95% confidence intervals obtained from parametric bootstrapping given in parentheses.

6.6 Discussion

We have proposed a range of statistical methods for estimating extreme quantities for challenges C1-C4. For the univariate challenge C1, we estimated the 0.9999-quantile, and the associated 50% confidence intervals, of $Y \mid \boldsymbol{X} = \boldsymbol{x_i}$, $i \in \{1, \ldots, n\}$. For challenge C2, we estimated a quantile, corresponding to a once in 200 year level, of the marginal distribution Y whilst incorporating the loss function in equation (6.3.2). Overall we ranked 6$^{\text{th}}$ and and 4$^{\text{th}}$ for challenges C1 and C2, respectively.

For challenge C1, our final model (model 7 in Table 6.3.1) was chosen to minimise the model selection criteria; however, QQ plots showed over-estimation of the most extreme values of the response (see Figure 6.3.2). As a result, the conditional quantiles calculated for C1 are generally over-estimated when compared with the true quantiles. If we ignored the model selection criteria and chose the model based on a visual assessment of QQ plots, we would have chosen model 5 in Table 6.3.1 and this would have covered the true quantile on fewer occasions than our chosen model. Therefore, the main issue with our results concerns the width of the confidence intervals.

Narrow confidence intervals are an indication of over-fitting and this could have arisen in several places. For instance, Rohrbeck et al. (2023) suggested all the seasonality is captured in the threshold, while our model includes a seasonal threshold and a covariate for seasonality in the scale parameter of the GPD model. As well as over-fitting, the model may not have been flexible enough; this could be, in part, due to our model missing covariates. For instance, the true model contained V_2 as a covariate (Rohrbeck et al., 2023) whilst our model did not. In addition, the basis dimensions for our splines are low. In practice, a higher dimension should be considered and, although we chose the dimension using a model-based approach, it may have resulted in the splines not being flexible enough to capture all of the trends in the data.

Narrow confidence intervals may have also resulted from the choice of uncertainty quantification procedure. Changing the average block length l in our stationary bootstrap procedure would alter the confidence interval widths, although this was carefully chosen to reflect the temporal dependence in the data. Alternative methods, such as the standard bootstrap procedure or the delta method, could be implemented to investigate how this affects the confidence interval widths. We expect that such confidence intervals will be wider than those presented here since the dependence in the data is not accounted for, but assuming temporal independence would be inaccurate. Therefore, whilst adopting an alternative procedure may widen confidence intervals, thus improving our performance, such intervals may not be well calibrated for this data set.

The over-fitting and over-estimation issues encountered in C1 are carried through to C2 since the same model is used for both challenges. However, one aspect specific to C2 is the choice of quantile evaluation within the loss function. Many methods exist for evaluating the non-stationary quantiles which feed into the loss function term of the objective function $S(\boldsymbol{\theta})$ in equation (6.3.5). As the loss function will be dominated by the log-likelihood in $S(\boldsymbol{\theta})$, we choose to transform to standard exponential margins when evaluating the quantiles in order to give more importance to the loss function. Since the data is light tailed ($\xi < 0$) this transformation elongates the tail and therefore inflates any deviations between the model and theoretical quantiles which in turn, inflates the contribution of the average loss function to $S(\boldsymbol{\theta})$. However, this approach means that the objective function will have a preference to minimise the deviations in the upper-tail of the distribution, leading to potential over-fitting to the upper-tail and possibly, a poor fit in the rest of the tail. This may not necessarily be undesirable since the loss function penalises under-estimation more than over-estimation, however, since the model in C1 already over-fits, this method may only exacerbate the problem for C2.

For the first multivariate challenge C3, we employed an extension of the method of Wadsworth and Tawn (2013) to estimate probabilities of three variables lying in extremal sets. Our extension accounts for non-stationarity in the extremal dependence structure, with GAMs used to represent covariate relationships. The QQ-plots for the resulting model suggested reasonable fits. For this challenge, we ranked 5th and our estimates are on the same order of magnitude as the truth (Rohrbeck et al., 2023).

We note similarities in the methodologies presented for the challenges C1, C2, and C3. Specifically, each of the proposed methods used the EVGAM framework for capturing non-stationary tail behaviour via a GPD. We acknowledge that the model selection tool proposed for C1 and C2 could also be applied for C3. However, we opted not to use this tool for several reasons. Firstly, unlike the univariate setting, there is no guarantee of convergence to a GPD in the limit, and the GPD tail assumption thereby needs to be tested. Moreover, in exploratory analysis, we tested the model selection tool for C3 but found the selected models and quantiles to not be satisfactory, particularly in the upper tail of the min-projection variable. Therefore, we selected a model manually, using QQ plots to evaluate performance, despite the potential for over-fitting. Exploring threshold and model selection techniques for multivariate extremes represents an important area of research.

In the final multivariate challenge C4, we estimated very high-dimensional joint survival probabilities. To do so, we split the probability into 5 lower-dimensional components which are assumed independent of each other, then estimated each using the CMEVM of Heffernan and Tawn (2004). In the final rankings of Rohrbeck et al. (2023), we ranked 3rd for this challenge. A more prudent method could have been implemented, as groups of variables were never truly independent. Alternatively, although we achieve relatively stable probability estimates with respect to threshold in Section 6.5.2 (see Appendix D.1.2), our approach may have been improved by estimating individual group probabilities across varying thresholds and taking an average value as our final result. We also do not report the effect of the choice of the conditioning variable on our estimates. Preliminary analysis suggested this to be negligible. However, conditioning on each site in a given subgroup and then taking a weighted sum of the resulting probabilities (e.g., Keef et al., 2013a) may have resulted in more robust estimates.

Chapter 7

Conclusions and further work

In this final chapter, we summarise the contributions that this thesis makes to the area of extreme value theory and methods and to induced earthquake modelling. Section 7.1 provides summaries of the content and contributions of the individual chapters. In Section 7.2, we outline potential developments and avenues of further work of the proposed methods and analyses.

7.1 Summary of contributions

In Chapter 3, we proposed novel methods to improve two particular aspects for univariate IID extreme value analysis: (i) the fundamental problem of selecting a threshold for identifying values consistent with extreme value theory in the peaks-over-threshold framework with a generalised Pareto distribution (GPD) used to model the threshold excesses and; (ii) the propagation of the uncertainty in this threshold selection through to subsequent tail inference. For the first aspect, we developed a simple but effective approach, called the expected quantile discrepancy (EQD) method. The EQD is a robust measure of goodness of fit, which accounts for uncertainty by bootstrapping. We automate the selection of the threshold by minimising the EQD metric across candidate thresholds. This selected threshold minimies the approximate integrated absolute error

(IAE) of the model quantiles and quantiles of the data generating process. Through an extensive simulation study that compares the EQD and the leading existing approaches, we demonstrated that the EQD performs better in terms of threshold selection and in subsequent quantile estimation. The EQD also shows less sensitivity to changes in the candidate threshold set and to values of its tuning parameters. Furthermore, the EQD does not rely on asymptotic theory and so, is applicable for all data set sizes with candidate threshold sets as fine as every observation if necessary. For the second aspect, we proposed a double-bootstrap procedure to incorporate the uncertainty in the selection of the threshold along with the GPD parameter uncertainty through to tail inference. Our proposed method led to major improvements in the calibration of confidence intervals for high quantile estimation, as shown by a coverage assessment across a range of confidence levels.

In Chapter 4, we adapted the threshold selection procedure of Chapter 3 for IID variables used in coastal flood risk assessments. Here, the focus was on capturing accurately the GPD fit to the most extreme observations. We developed the extension of the EQD method, known as the TAILS approach, which relies on the same principles of the EQD with two key innovations; the quantile levels considered in the IAE approximation are fixed across candidate thresholds to avoid oversampling of the upper tail for higher candidate thresholds, and these quantiles are set above a predefined level to ensure accurate fitting to the observed upper tail. Using tide gauge records, we demonstrated that the selected thresholds using this adaptation are generally higher across the global tide gauge records and lead to improved fits for the most extreme observations at selected sites explored in more detail. This improvement in upper tail fit comes at the cost of additional uncertainty in the GPD parameter estimates.

Chapter 5 presented a further extension of the work in Chapter 3, moving away from the IID context, to improve the modelling of induced seismicity (i.e., magnitudes of earthquakes caused from human activity) in the Groningen gas field in the Netherlands. Here, the observed data are not identically distributed due to spatio-temporal changes in stresses over the gas field due to the extraction of gas and further spatio-temporal variability in the measurement process. In this work, we adapted the work of Chapter 5 in a number of ways. Firstly, we built upon the approach of Varty et al. (2021) which considered the time variation in the development of the data measurement. We additionally considered the combined spatial-temporal variability in the geophone network characteristics (this information was not available to Varty et al. (2021)) and proposed to use the geophone network more directly as a covariate. We developed a specialised extension of the EQD here to select a physically-motivated spatio-temporal threshold function for Groningen as an improved estimator for the magnitude of completion (i.e., the smallest magnitude which can be detected with certainty at a given time and location). Secondly, we utilised a key physical covariate, the Kaiser stress, in the modelling of the spatio-temporal rate of occurrence of earthquakes and for the magnitude excess distribution. Thirdly, we developed adaptations of the uncertainty algorithms, proposed in Chapter 3, to account for the uncertainty in the inference for the spatio-temporal threshold function, the rate of occurrence of earthquakes model and the non-identical GPD model. We incorporated these uncertainties in our endpoint and quantile inference, providing estimates of key future hazard quantities. Novelly, we also proposed an approach to incorporate the uncertainty in the functional form of the threshold function with covariates in such future hazard assessments.

Our proposed approaches led to (i) a larger catalogue of exceedances of the estimated magnitude of completion with which to fit the models, (ii) excellent fit diagnostics, (iii) improved understanding of the form and sources of the spatio-temporal variability in the magnitude of completion for Groningen, and (iv) useful future hazard assessments with reduced uncertainty relative to existing methods. Our new estimator for the magnitude of completion has a physical basis and captured the same changes in the measurement process highlighted by the Varty et al. (2021)'s smooth parametric temporal function but also, it revealed times and spatial regions where the magnitude of completion departs from this function due to the impact of the spatial variability in the changing measurement process. This estimator allows the use of more exceedances to fit the model and in comparison to the previous conservative approach leads to a more concrete reasoning, from a statistical perspective, for a finite upper endpoint due to the major reduction in the uncertainty of the GPD parameter estimates. The uncertainty algorithms presented in this work allowed for future estimates of endpoint summaries and typical hazard defence design quantities, with a more useful quantification of the uncertainty of such quantities.

Chapter 6 detailed the contributions, as part of a wider team, to the EVA data challenge 2023. In this chapter, we proposed a range of statistical methods for estimating extreme quantities. For the univariate challenges, C1 and C2, we proposed an extreme value model with a covariate-dependent threshold, rate of exceedance and GPD scale parameter. In this work, we adapted the EQD of Chapter 3 to allow for the selection of a seasonal threshold, and corresponding exceedance rate parameter, and combined this with generalised additive models (GAMs) for the GPD scale parameter incorporating a variety of covariates. To do this, we combined the EQD threshold selection method with the flexible estimation procedure of the `evgam` package (Youngman, 2022). For challenge C1, we estimated an extremal quantile, and the associated 50% confidence intervals. For challenge C2, we estimated the marginal 200-year return level and adjusted the estimation procedure to include an asymmetric loss function. For the first multivariate challenge C3, we employed an extension of the method proposed by Wadsworth and Tawn (2013) to estimate probabilities of three variables lying in extremal sets and accounted for the non-stationarity in the extremal dependence structure with GAMs used to represent covariate relationships. In challenge C4, we estimated joint survival probabilities for 50-dimensional variables. We first identified a clustering structure with independence between clusters, which led to splitting the 50-dimensional event probability into five lower-dimensional components and using the conditional extremes approach of Heffernan and Tawn (2004) to estimate each of these probabilities.

7.2 Further work

Chapter 3 demonstrated the effectiveness of both our proposed methodologies; the EQD method for automated threshold selection and the double-bootstrap procedure for confidence interval construction in the univariate IID setting. The flexibility of EQD method to adaptation for different settings has been shown in the subsequent chapters of this thesis where a selection of the potential avenues of extension within the area of extreme value modelling have been explored. We also believe there to be wider applicability of the proposed methods beyond the specific extensions of the EQD explored in this thesis.

In Chapter 4, we introduced an extension of the EQD which focusses on the GPD fit to the most extreme observations. We addressed concerns of practitioners for the specific context of coastal flood risk assessment. Within this work, we provided an uncertainty assessment of return levels for particular sites calculated using three proposed thresholds, selected by (i) the EQD method, (ii) the TAILS extension of the EQD method and (iii) the arbitrarily-chosen 98% quantile, typically used in coastal flood risk contexts. Within this comparison, we ignored the important aspect of threshold uncertainty which had been highlighted in Chapter 3, to allow fair comparison with this static 98% quantile threshold.

The missing component of threshold uncertainty is a key avenue of further research. Utilising the TAILS approach led to a more accurate fit to the most extreme observations but led to larger GPD parameter uncertainty. The inclusion of threshold uncertainty with this extension could lead to an assessment of coastal flood hazard which combines the positives of the TAILS method and the uncertainty quantification of Chapter 3. The higher threshold choices and accurate fitting to the most extreme observations would alleviate any concerns of practitioners while the well-calibrated uncertainty assessment would aid decision-making for future coastal flood defences. From a methodological point of view, it would also be interesting to compare the levels of threshold uncertainty when using the EQD versus the TAILS approach. The generally higher thresholds selected by the TAILS approach lead to an increase in the GPD parameter uncertainty but the focus on the smaller region of observations may lead to more certain threshold choice and it would be interesting to see how the resulting confidence intervals compare when both aspects of uncertainty are taken into account.

Similar to the EQD approach, in Chapter 4, we make the assumption that data are identically distributed. Environmental processes such as sea levels are unlikely to be identically-distributed particularly, given the effects of the changing climate. Furthermore, such datasets can exhibit short-range temporal dependence both in the stochastic component and due to the effect of known tidal component. The tidal component could be accounted for via a tidal covariate while the dependence in the stochastic component of sea levels (i.e., the surge) could be modelled by estimating the subasymptotic extremal index (D'Arcy et al., 2023). Similar approaches to the extensions of the EQD in Chapters 5 & 6 for more complex, non-IID settings, could be applied to the TAILS method. This may involve utilising physically-motivated parameterisations of thresholds and GPD parameters, as performed in Chapter 5, or making use of the well-established methods for incorporating covariate dependence into the threshold and parameters of a GPD through smooth GAM formulations (Chavez-Demoulin and Davison, 2005; Youngman, 2019). Such extensions of the TAILS approach could prove useful in coastal flood risk assessments and in the wider modelling of environmental processes.

For the coastal flood context, one could extend the TAILS method to incorporate inter-annual non-stationarity by utilising relevant covariates that impact the number of extreme events that occur within a given year e.g., indices related to the El Niño-

Southern Oscillation or North Atlantic oscillation phenomena, which describe fluctuations in weather patterns in their respective regions. Similarly to the extensions of the EQD and the work of Varty et al. (2021), minor modification of the TAILS approach to assess quantile discrepancies on a common transformed scale, e.g., exponential margins, would enable its use in these more complex settings. However, when fitting a non-stationary GPD model, without the clear physical motivation of the covariates as in Chapter 5, there are not well-established methods for selecting which covariates to include with different thresholds. Furthermore, if using smooth GAM formulations, there is the additional challenge of selecting the level of flexibility within the smooth functions that is most appropriate. Therefore, the development of selection techniques for model and threshold formulation with covariates for non-stationary data structures is an important line of future research.

In Chapter 5, we utilised the EQD approach to select the most appropriate threshold function formulation using a selection of physically-motivated formulations with covariates. While the need for mitigation of the induced seismic hazard of the Groningen gas field provided a strong motivation for the methods used in Chapter 5, we expect that the methodology will be useful for other gas extraction or injection contexts. In particular, we see the insights of the magnitude of completion formulation, the stress-dependent magnitude distribution and the quantification of the uncertainties in these aspects to likely be especially useful in the model development and hazard assessment for underground carbon dioxide storage sites (Bauer et al., 2019), an important and growing area of greenhouse emissions mitigation. At such underground storage sites, similar characteristics of induced seismicity have been identified. However, gas injection sites come with their own modelling challenges due to differences in the number of geophones and structure of the networks. Thus, making use of the insights we gleamed here from the modelling of gas extraction sites is of paramount importance.

Within the specifics of our proposed methodology for induced earthquake modelling, there are key avenues of future development for the framework. In the selection of our geophone-based threshold formulation, we considered three transformations (linear, square-root and logarithm) of four distance covariates in a simple linear relationship with the threshold. The methodology could be extended to use a Box-Cox transformation of the covariate (which incorporates all these three forms as special cases) and estimate the relevant additional Box-Cox parameter as part of the threshold function selection. Secondly, to aid hazard assessments, we estimated quantities of interest relating to the future behaviour of earthquake magnitudes in this region under an assumption of no further extraction. Geo-physicists typically take the inference a step further by using the future magnitude estimates to develop ground motion assessments across the entire region in a full probabilistic seismic hazard analysis (Baker et al., 2021). This involves geo-physical spreading models which incorporate information on the nature of ruptures and the latent fault structure between the earthquake locations and the rest of the field. The ground motion field for an earthquake is then modelled as a realisation of a spatial log-Gaussian process, with mean and covariance functions given by the spreading model (Bommer and Stafford, 2016; Bommer et al., 2017). However, we did not take this additional step in the absence of an appropriate spatial ground motion model. This type of assessment is what's used to estimate the hazard to infrastructure and public safety and thus, using our improved models for earthquake magnitudes in such an assessment is a vital future avenue.

In our direct use of the geophone network, we uncovered characteristics of the spatial variability of the detection ability through time. An avenue which could potentially uncover even more subtle characteristics of the relationship between the magnitude of completion and the geophone network would be to incorporate information about the sensitivity or detection ability of individual geophones of different types in the region.

We provided endpoint summaries as part of our future inference. These summaries, and tail inferences near the endpoints, could possibly be improved by incorporating expert knowledge for the physical upper bounds of possible magnitudes, similar to the approach of Yue et al. (2025b). Lastly, in our model for the baseline rate of occurrence of earthquakes above a magnitude of 0 $\mathrm{M_L}$, we did not account for the dependence between main-shock and after-shock earthquake occurrences. Utilising approaches which address this, such as the epidemic-type aftershock sequence models (Ogata, 1988), could provide a better description of the clustering of earthquake occurrences. Combining a model for this clustering of earthquakes with our methodology for modelling the excesses of a spatio-temporal threshold could lead to a further improvement in the accuracy of hazard assessments for potential future occurrences of large magnitude events.

Beyond the gas extraction/injection contexts, we also believe our methods which account for non-identically distributed data have the potential to be useful in extreme value contexts where data are missing not-at-random due to measurement equipment quality. Furthermore, the methodology developed in Chapter 5 is generic in its structure, although our presentation is specific to induced seismicity modelling in terms of the choice of covariates and model formulation. With adjustment to the EQD and the parameterisations of the GPD model, this methodology of combined threshold function and model selection inferences could be widely applicable to any extreme value context. We see particular utility in applying these techniques when there is known seasonal or directional behaviour in environmental applications, and/or long-term trends (Jonathan and Ewans, 2007a; D'Arcy et al., 2023).

In environmental contexts, where we may have a set of potentially important covariates, the usual approach would be to use a method to select a threshold, whether it be constant or a function of covariates, by assessing the fit of a particular formulation of the GPD. For a simple example, suppose we want to estimate threshold $u(t)$ (which could be constant such that $u(t) = u$ for all days t) and we have identified within-year seasonality in the data Y_t on day t through exploratory analysis which may be modelled well by a sinusoidal function. Thus, we have three possible models we might want to consider:

1. $(Y_t - u)|(Y_t > u) \sim \text{GPD}(\sigma_u, \xi)$,

2. $(Y_t - u)|(Y_t > u) \sim \text{GPD}(\sigma_u(t), \xi)$ with $\sigma_u(t) = a + b\sin(2\pi t/365)$,

3. $(Y_t - u(t))|(Y_t > u(t)) \sim \text{GPD}(\sigma_u, \xi)$ with $u(t) = \alpha + \beta\sin(2\pi t/365)$.

with $(a, \alpha) \in \mathbb{R}^2$ and $(b, \beta) \in \mathbb{R}_+^2$ and $\xi \in \mathbb{R}$.

First, suppose we were only interested in selecting between models 1 and 2. Typically, the threshold u would be estimated using standard approaches assuming a model 1, which we denote by $\hat{u}_1$. Then, following the selection of $\hat{u}_1$, comparisons of the two models would be made either using information criteria, or measures of goodness-of-fit, or by simply checking whether parameter b was statistically significant. The reason for this being a typical approach is the fact that information criterions require models to be compared on the same data. However, this poses a problem; $\hat{u}_1$ was selected assuming model 1 and so is only the appropriate threshold choice for model 1. Using model 2 fitted above $\hat{u}_1$ is suboptimal for that model form as a different constant threshold value is likely more appropriate. However, estimating a threshold for model 2, say $\hat{u}_2$, results in two nested sets of exceedances and violates the use of standard information criteria for deciding between these models.

Now suppose, following our initial comparison and exploratory analysis, we conclude that this sinusoidal behaviour is significant and must be included in the model in some way. We now must decide where to incorporate this variation, in the threshold or scale parameter, i.e., we want to select between models 2 and 3. Both models require the estimation of thresholds separately due to the differing parameterisations. Following this step, we now cannot select between these models via information criterions due to the differing exceedances involved in their estimation. One may consider simply making a visual assessment of the fit of each through standard QQ-plot approaches, however this incorporates a degree of subjectivity, especially if the differences are subtle. Based on our novel developments in Chapter 5 in assessing non-nested covariate models, we propose that all three models above could be compared using the EQD. We have shown the adaptability of the EQD approach to different model formulations in the extensions developed in this thesis. For other model selection scenarios, such as the example above, the EQD could be easily adapted to the different covariate formulations and used to select optimal thresholds for all candidate models. Once the EQD metric is calculated on the same scale, using a transformation to standard margins, the minimised EQD values can be compared as a method of model selection. For the example above, this would mean additionally estimating $\hat{u}_3(t)$ for model 3. Each threshold estimate would have a corresponding EQD value, $\hat{d}_1$, $\hat{d}_2$, $\hat{d}_3$ approximating the integrated absolute error between the model and data in each case. We can then compare $\hat{d}_1$, $\hat{d}_2$, and $\hat{d}_3$ and choose the minimum, leading us to the optimal combined model formulation and threshold choice. We see this avenue of model selection as a key aspect of the EQD method's wider use.

Another avenue which has not been explored as part of this thesis is the use of the EQD as a method for threshold estimation in multivariate extreme value contexts. The method could be applicable, with suitable adjustment, to cases relying on multivariate regular variation assumptions such as Wan and Davis (2019) or for variables exhibiting asymptotic independence (Heffernan and Tawn, 2004). As discussed in Chapters 3 & 5, a key aspect of proposed methods is accounting for the uncertainty in the threshold estimation, so applying these approaches in a multivariate context could allow improved calibration of uncertainty in joint tail inference. This idea also naturally extends to spatial extreme value modelling.

Finally, the methods we have developed in Chapters 3 and 5 for quantifying the different aspects of uncertainty in the threshold selection, model formulation and estimation of GPD parameters in the subsequent tail inference has potential for wider utility in core extreme value methodology and in a variety of applications where improved uncertainty assessments are needed. The algorithms developed are intuitively simple and easy to implement, although can be computationally intensive. A potential further avenue to explore is to optimise the efficiency of the coding of these algorithms to allow for a wider applicability. Here, we used $B_{\text{nonpar}} = B_{\text{par}} = 200$, but using the same number of 40000 boostrap samples with $B_{\text{nonpar}} \neq B_{\text{par}}$ may be more effective.

This thesis has made contributions to the core methodology of extreme value analysis and in particular, the fundamental problem of threshold selection. Our proposed methodology is highly adaptable and we see the potential for its use in a wide variety of extreme value contexts. Within this thesis, we have provided new insights for the specific context of Groningen induced seismicity and developed improved physically-motivated methodology for tackling the challenges of modelling induced earthquakes. This methodology also contributes to the wider area of extreme value modelling of non-identically distributed environmental processes. As evidenced by the multiple avenues of further work proposed in this chapter, many new questions, challenges and further extensions within the modelling of induced earthquakes and the wider modelling of environmental extremes have been identified.

Appendix A

Supplementary materials to Chapter 3

A.1 Introduction

This document provides further information to accompany Chapter 3. Section A.2 presents further details of the distributions in Cases 1-4 of Section 3.6, including quantile derivations. Section A.3 covers the reasoning for the omission of the Varty et al. (2021) approach from the simulation study in Section 3.6, justifies the choice of the default tuning parameters for the EQD method and discusses the choice of calibration data within the definition of $d_b(u)$ in Section 3.4 of the main text. Section A.4 presents additional simulation experiments, a detailed breakdown of the results outlined in Section 3.6, and an exploration into the effect of taking different candidate threshold grids. Finally, Section A.5 provides a more extensive analysis of the coverage of true quantiles achieved by both Algorithm 1 and 2, as outlined in Section 3.5.

A.2 The distributions of Cases 1-4

This section provides full details of the true quantile and density functions used for the simulation study for Cases 1-4 in Section 3.6.

A.2.1 Quantile calculations

Case 1-3: We simulate X from a mixture of a Uniform$(0.5, 1.0)$ distribution and a GPD(σ_u, ξ) distribution above the threshold $u = 1.0$. Consequently, X has distribution function:

$$F_X(x) = \begin{cases} \frac{x-0.5}{3}, & 0.5 \le x \le 1 \\ \frac{1}{6} + \frac{5}{6}\left[H(x-1;\sigma_u,\xi)\right], & x > 1. \end{cases} \tag{A.2.1}$$

where $H(x-1;\sigma_u,\xi)$ is the distribution function of a GPD with parameters (σ_u, ξ). Therefore, the true quantile x_p with exceedance probability p (for $p < 5/6$) is

$$x_p = 1 + \frac{\sigma_u}{\xi}\left[\left(\frac{6p}{5}\right)^{-\xi} - 1\right].$$

This formulation is identical across Cases 1-3, but the model parameters and simulation sample sizes differ over these cases, with these values being given in Table 3.6.1 of Section 3.6.

Case 4: Here X has distribution function, for $x > 0$, of

$$F_X(x) = \int_0^x \frac{h(s;\sigma,\xi)\mathbb{P}(\mathcal{B} < s)}{q + \bar{H}(1;\sigma,\xi)}\,\mathrm{d}s$$

where $\bar{H}(x;\sigma,\xi)$ and $h(x;\sigma,\xi)$ are the survivor and density functions of a GPD with parameters (σ, ξ) and a threshold 0, $q = \int_0^1 h(s; 0.5, 0.1)\mathbb{P}(\mathcal{B} < s)\mathrm{d}s$ with $\mathcal{B} \sim \text{Beta}(\alpha, \beta)$, so $0 \le \mathcal{B} \le 1$. This unusual distribution function transitions from a non-GPD distribution to an exact GPD for excesses of 1 (the upper bound of $\mathcal{B}$), so that the true threshold for X is $u = 1$ and the excess distribution, for $x > 1$, has survivor function

$\bar{H}(x-1; \sigma_1 = \sigma+\xi, \xi)$ following from the threshold stability property (see Section 2.3.1). Here $\tau := \mathbb{P}(X \leq 1) = q/(q + \bar{H}(1; \sigma, \xi))$ so the true quantile x_p for X with exceedance probability p, where $p \leq 1 - \tau$, is:

$$x_p = 1 + \frac{\sigma_1}{\xi}\left[\left(\frac{p}{1-\tau}\right)^{-\xi} - 1\right].$$

Although F_X appears complex, it is straightforward to simulate samples from X. Specifically, X is generated using a rejection method by first simulating a proposal variable $Y \sim \text{GPD}(\sigma, \xi)$ above the threshold of 0. The rejection step involves rejecting Y as a candidate for X only if $Y < \mathcal{B}$, where $\mathcal{B}$ is a random variate generated from a Beta(α, β) distribution. Since $\mathcal{B} \leq 1$, for $Y > 1$ it follows that $X = Y$, but only a proportion of the values of $Y < 1$ are retained in X, with the rate of retention dependent on the parameters of the Beta(α, β) distribution.

In our simulations, we selected $(\sigma, \xi) = (0.5, 0.1)$ and $(\alpha, \beta) = (1, 2)$, with the latter chosen such that the density of X has a mode below 1. For these parameters we obtain $\hat{\tau} = 0.721$, to 3 decimal places, where we evaluated τ using Monte Carlo integration methods. To ensure that we have the same numbers of exceedances of the true threshold across samples, we simulate until we have a sample proportion of threshold exceedances matching the true value of $1 - \tau$ and an overall sample size of 1000.

A.2.2 Density and quantile functions

The density functions of the random variable X are given in Figure A.2.1, for Cases 1-4. All four density plots show that there is a large probability of exceeding the true threshold (i.e., $X > 1$) so a large proportion of each sample is from a GPD tail. This is unusual in practice (where often thresholds correspond to 90-99% sample quantiles) and so threshold selection should be easier in these examples than typically. This is especially true for Cases 1-3 with the density having a large step change at the threshold and the density having a completely different shape below the threshold. In contrast, Case 4 has a much more subtle transition for the density across the threshold. The density is continuous and first-differentiable at the threshold, but with a clearly non-GPD distribution below the threshold, shown by the density's mode lying away from its lower endpoint. Thus, we see this case as much more challenging for threshold selection.

To further emphasise the differences between Cases 1-4, Figure A.2.2 shows return level plots for the simulated distributions of Cases 1-4 for a range of return periods (per observation). In particular, this plot emphasises the key difference between Case 3 and the other cases; Case 3 has a finite upper-endpoint due to $\xi < 0$ whereas the other cases have unbounded distributions. This difference is not apparent from the density plots in Figure A.2.1.

The idea behind our choice of distributions is that if methods struggle in cases where we have a clear true threshold then there will be significant problems when it comes to real datasets. So, collectively the four cases provide a natural testing ground for separating between threshold methods.

A.3 Supporting details for Section 3.4

A.3.1 Overview

In this section, we provide evidence, based on a range of simulation studies, to support several decisions we made in Section 3.4. Specifically, in Section A.3.2, we provide evidence to demonstrate the advantages of using the EQD over the Varty et al. (2021) method; Section A.3.3 outlines results indicating that our suggested default choices for the tuning parameters (B, m) of the EQD method are widely suitable, and that there is very little sensitivity to these choices; in Section A.3.4, we show that the choice of bootstrap data as the calibration data in the metric $d_b(u)$ works comparably relative to using the observed data; and in Section A.3.5, we demonstrate the benefits of the

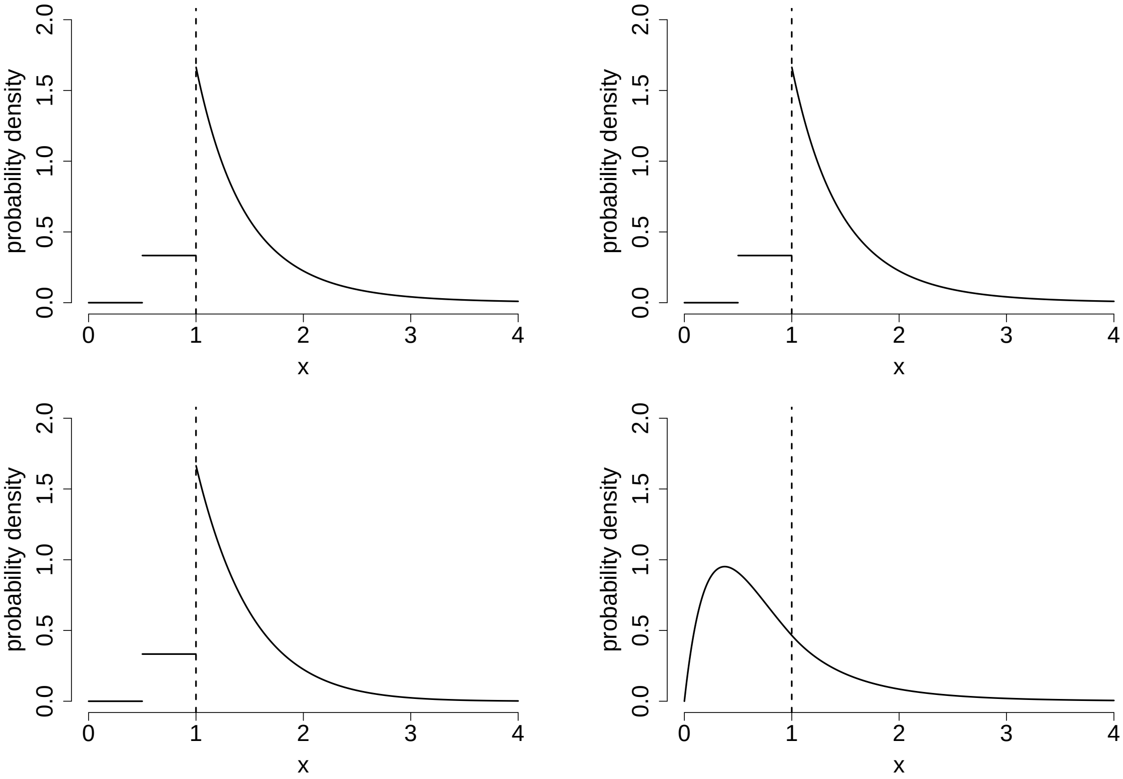

Figure A.2.1: True densities of simulated datasets from Cases 1-4 with numbering corresponding to left-right and then top to bottom.

bootstrapping component of the EQD method.

A.3.2 Comparison of EQD and Varty et al. (2021) methods

This section provides the full evidence basis for the decision, outlined in Section 3.4.1, to prefer the EQD method and omit the results of the Varty et al. (2021) method from the main text. Here, we compare both methods using the data simulated from Cases 1-4, as described in detail in Section A.2, and outlined overview in Section 3.6. As seen from Section 3.4.1, the EQD and the Varty et al. (2021) methods differ simply in the scale on which the metric of goodness-of-fit is compared, with the former evaluated on the observed data scale and the latter making the same comparison after transformation onto Exponential(1) margins.

We assess the performance of the two methods based on both the selection of

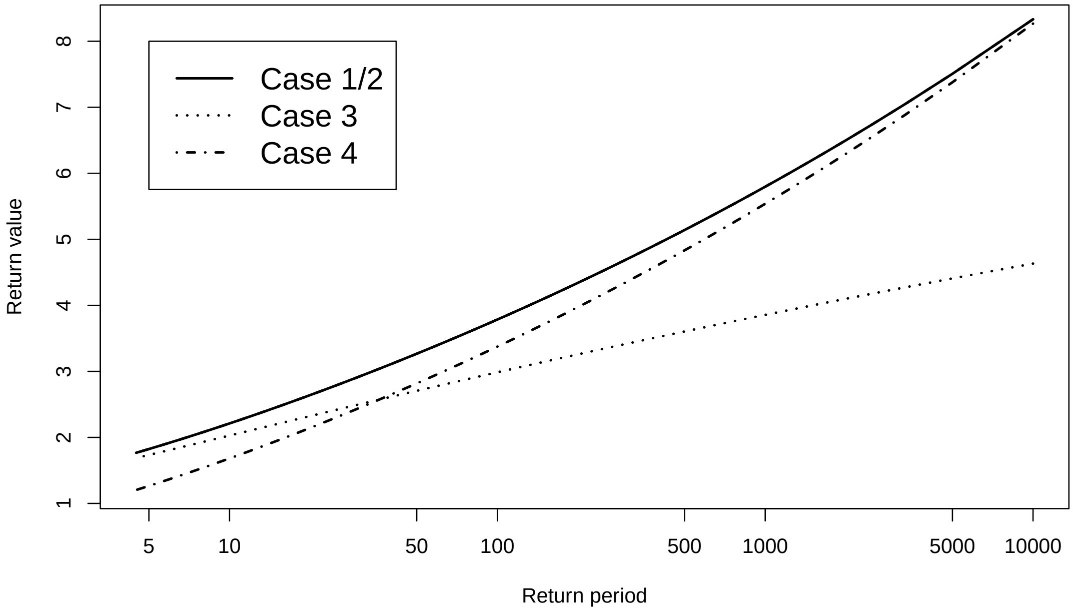

Figure A.2.2: True return values of simulated datasets from Cases 1-4.

thresholds (as the truth is known in each case) and on the subsequent estimation of quantiles for a range of exceedance probabilities, namely the $(1-p_{j,n})$-quantiles where $p_{j,n} = 1/(10^j n)$ for $j = 0, 1, 2$ with n denoting the length of the simulated dataset. For each case and each of the measures of fit, all comparisons between the methods are based on estimates obtained using the same set of 500 replicated samples, so any difference in the methods found is simply due to the two method's performance.

Table A.3.1 shows the RMSE, bias and variance of the thresholds chosen by the two methods. Based solely on threshold choice, it is difficult to distinguish between the approaches with each method narrowly outperforming the other in two of the four cases based on each of RMSE and bias. Cases 1-3 exhibit positive bias for both methods, as they are much less likely to pick a threshold too low in these cases given the sudden change in the density shown in Figure A.2.1. In contrast, for Case 4, both methods incur a negative bias, essentially due to the smooth transition from GPD in the density shown in Figure A.2.1. Additionally, in every case, there is no method that gives threshold estimates of lower variance than the EQD method.

The primary goal of an extreme value analysis is usually quantile inference, rather than threshold selection. Tables A.3.2 and A.3.3 compare the EQD and Varty et al. (2021) methods in terms of quantile inference. These tables show, for the three target quantiles, the RMSE, bias and variance of quantile estimates that are based on MLE fits of a GPD above the thresholds selected using each method. The differences between the two methods are evident. The EQD either matches or achieves the lower RMSE in all cases and all quantiles and the differential in performance becomes more evident for long-range extrapolation, i.e., as j increases, see Table A.3.2. This difference in RMSE seems to stem mainly from the smaller variance of the EQD estimates, see Table A.3.3 [right]. As with the results for threshold selection, the bias results for quantile estimation in Table A.3.3 [left] show the two methods perform similarly (with each method slightly better on a number of occasions across the cases and the three quantile levels of interest).

Given that the goal of a threshold selection method is to improve high quantile inference, we conclude from this study that, while results are similar, the EQD appears to perform better for quantile estimation. Furthermore, the Varty et al. (2021) method requires an additional, and non-intuitive (for IID data), transformation to Exponential(1) margins. Thus, we choose to omit the Varty et al. (2021) approach from subsequent studies in the supplementary material and from the simulation study discussed in Section 3.6 of the main text.

	EQD			*Varty method*		
Case	RMSE	Bias	Variance	RMSE	Bias	Variance
Case 1	**0.048**	**0.034**	**0.001**	0.059	0.041	0.002
Case 2	**0.060**	**0.031**	**0.003**	0.073	0.039	0.004
Case 3	0.060	0.042	0.002	**0.055**	**0.039**	0.002
Case 4	0.526	−0.515	**0.012**	**0.508**	**−0.492**	0.016

Table A.3.1: Measures of performance (RMSE, bias and variance) for threshold choices for the EQD and Varty et al. (2021) methods, for Cases 1-4. The smallest magnitude for each measure of performance are highlighted in bold for each case.

	EQD	*Varty*	*EQD*	*Varty*
j	**Case 1**		**Case 2**	
0	**0.563**	0.605	**0.599**	0.611
1	**1.258**	1.335	**1.488**	1.542
2	**2.447**	2.612	**3.119**	3.305
	Case 3		**Case 4**	
0	0.190	0.190	**0.677**	0.705
1	**0.323**	0.324	**1.563**	1.673
2	0.483	0.483	**3.043**	3.378

Table A.3.2: RMSE of the estimated $(1 - p_{j,n})$-quantiles in Cases 1-4 based on fitted GPD above chosen threshold for the EQD and Varty et al. (2021) methods. The smallest value for each quantile are highlighted in bold.

	EQD	*Varty*	*EQD*	*Varty*
j	**Case 1**		**Case 2**	
0	−0.021	**−0.005**	−0.049	**−0.018**
1	**−0.015**	0.030	**−0.046**	0.055
2	**0.044**	0.145	**0.069**	0.312
	Case 3		**Case 4**	
0	−0.008	**−0.007**	−0.283	**−0.233**
1	−0.007	**−0.005**	−0.722	**−0.571**
2	−0.002	0.002	−1.410	**−1.064**

	EQD	*Varty*	*EQD*	*Varty*
j	**Case 1**		**Case 2**	
0	**0.316**	0.335	**0.357**	0.373
1	**1.582**	1.716	**2.211**	2.376
2	**5.988**	6.638	**9.723**	10.824
	Case 3		**Case 4**	
0	0.036	0.036	**0.379**	0.444
1	0.105	0.105	**1.926**	2.479
2	0.233	0.233	**7.287**	10.299

Table A.3.3: Bias [left] and variance [right] of the estimated $(1-p_{j,n})$-quantiles in Cases 1-4 based on fitted GPD above chosen threshold for the EQD and Varty et al. (2021) methods. The smallest variance and absolute bias for each quantile are highlighted in bold.

A.3.3 Selection of default tuning parameters

This section provides the sources of evidence presented in Sections 3.4.2 and 3.4.3 in terms of the effect of m on the interpolation of quantiles in the metric and the suitability of our default values of the tuning parameters (B, m).

First, consider an analysis of the sensitivity of the EQD method to different choices of its two tuning parameters (B, m), where B denotes the number of the bootstraps for which $d_b(u)$ is evaluated in order to calculate the overall metric $d_E(u)$, and m is the number of quantiles used in the evaluation of the metric $d_b(u)$ for each bootstrap. In the main text, the values of $(B, m) = (100, 500)$ are proposed as the default values for the simulation studies of the performance of the EQD method. These values are used for the tuning parameters in all simulation studies of the EQD method in the main text and the supplementary material.

We focus our sensitivity analysis on the 500 replicated samples of Case 1 where $n = 1000$, detailed in Section 3.6. Tables A.3.4 & A.3.5 provide the RMSEs of the threshold estimates along with the computation time relative to that of the default value when using the EQD with different values of B and m respectively.

For the choice of B, the number of the bootstraps for which $d_b(u)$ is evaluated when calculating $d_E(u)$, Table A.3.4 shows that the computation time of the EQD method increases linearly with B. As B is simply the number of bootstrap samples in an average, then in principle we want to take B as large as possible to remove Monte Carlo noise in the average approximation of the associated expectation. Thus, in selecting B, we require it to be sufficiently large so that any residual Monte Carlo noise is not important (for threshold selection to be stable) whilst recognising the linear increase in computation time from this choice. Thus, for one-off analyses, as computation time is not of particular concern, it is ideal to take B as large as possible. However, for simulation studies, a more careful choice of B is required as accuracy needs to be balanced with computation time. Table A.3.4 provides evidence on how the Monte

Carlo noise is diminishing as B increases, and shows the RMSE value stabilising as B increases. The decreases in RMSE are only very slight, especially when compared to the differences we see between the EQD and other existing approaches in Section 3.6.

B	200	400	1000	100 (default)
RMSE	0.043	0.039	0.040	0.048
Relative time	2	4	10	1

Table A.3.4: RMSEs for threshold estimates and for the relative computation time compared to the default choice of $B = 100$ obtained using the EQD method for different values of B for Case 1. Each result uses $m = 500$ and 500 replicated samples.

For the choice of m, the number of quantiles used in the evaluation of the metric $d_b(u)$ for each bootstrap, Table A.3.5 shows results for two different strategies for selecting m: the first allowing m to be proportional to the data sample size n irrespective of threshold value, i.e. $m = cn$ for $c = 0.5, 1, 2, 10$, and the second allowing m to vary according to the number, n_u, of exceedances of each candidate threshold u, i.e. $m = cn_u$ for $c = 0.5, 1, 2, 10$. Our reason for exploring the second strategy is to ensure that across candidate thresholds we are using the same level of interpolation/extrapolation to non-sample quantiles. For each strategy, we examine the effect of different degrees of proportionality c. The RMSE of the threshold estimates obtained show little sensitivity to the value of m across the two strategies and all levels of proportionality.

	$m = cn$				$m = cn_u$				$m = 500$
c	0.5	1	2	10	0.5	1	2	10	(default)
RMSE	0.045	0.046	0.045	0.046	0.049	0.048	0.047	0.047	0.048
Relative time	1.1	1.4	2.0	7.2	0.9	1.1	1.4	4.6	1

Table A.3.5: RMSEs for threshold estimates and for the relative computation time compared to the default of $m = 500$ of the EQD method for different values of m for Case 1. Each result uses $B = 100$ and 500 replicated samples.

Table A.3.5 also shows that increasing m in either strategy is essentially wasting the increased computation time. When estimating $d_b(u)$ for a particular bootstrap, we are aiming to approximate the integrated absolute error (IAE) between the model quantiles and sample quantiles for that sample. This $d_b(u)$ then feeds into the overall

$d_E(u)$ for the excesses of candidate threshold u which approximates the IAE between model quantiles and the data generating process. This sensitivity analysis shows that once we choose a suitably large value for m, any changes in this approximation error are quite small in comparison to the $d_b(u)$ value itself.

Now consider the effect of m on the interpolation of quantiles in the metric. While we have shown that changing m does not meaningfully affect the RMSE of threshold choice over repeated samples, it is still important to investigate the effect of this choice for the values of $d_b(u)$ and $d_E(u)$ for a range of candidate thresholds and the effect, if any, on the resulting threshold choice for a particular dataset. In particular, we are interested in the effect of the choice of interpolation grid between $m = 500$ and $m = n_u$.

For a particular bootstrap sample, the choice of $m = 500$ can lead to under- or over-sampling of the upper tail when approximating the IAE of the QQ-plot, depending on if $m < n_u$ or $m > n_u$. While this may not be ideal, it is only important if it has a significant effect on the overall metric value in a way that unfairly or adversely affects the resulting choice of thresholds.

To explore the effect of the interpolation grid on the sampling distribution of metric values for thresholds (i.e., $d_b(u)$ values), $d_E(u)$ and on the resulting threshold choice, we have considered the following additional investigations. Let $d_b^{500}(u)$ and $d_b^{n_u}(u)$ denote the value of the metric for the b^{th} bootstrap sample above threshold u using $m = 500$ and $m = n_u$ respectively. For the first simulated sample of Case 1 and the Gaussian case, we look at:

1. The distribution of $d_b^{500}(u)$ and $d_b^{n_u}(u)$ for each u over the candidate grid of thresholds.

2. The distribution of the relative difference between $d_b^{500}(u)$ and $d_b^{n_u}(u)$ to the overall metric $d_E(u)$ over the candidate grid.

The reasons that we selected these two features to investigate are that the former looks at the effect of the interpolation on the sampling distribution of the metric while the latter assesses if the interpolation method could significantly change the overall metric value for any particular thresholds on a particular dataset.

For the first simulated data sample of Case 1, we explore the distribution of $d_b^{500}(u)$ and $d_b^{n_u}(u)$ in Figure A.3.1. Specifically, this figure shows boxplots of the distribution of $d_b^{500}(u)$ and $d_b^{n_u}(u)$ values for each value of u. The mean of these values, i.e., $d_E(u)$, is also plotted as a black point in each boxplot. For particular thresholds, comparison of the sampling distributions of $d_b^{500}(u)$ and $d_b^{n_u}(u)$ values shows only very minor differences. While there are some larger differences between the plots, particularly at higher thresholds, the black points in each plot indicate that these differences in $d_b(u)$ do not lead to large differences in the overall metric value $d_E(u)$.

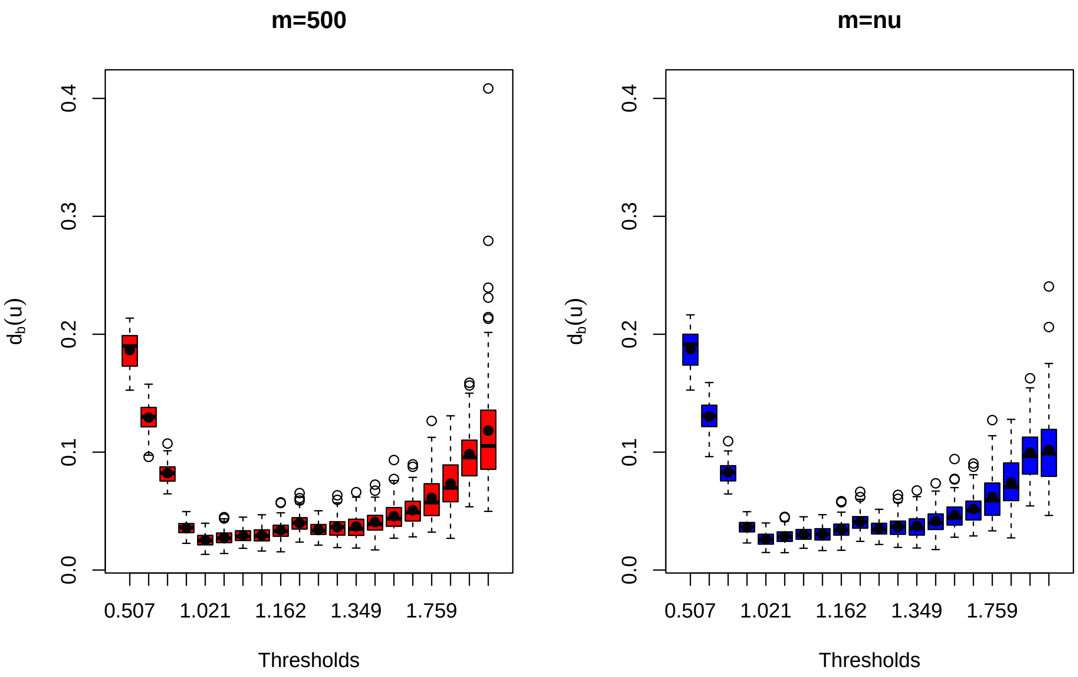

Figure A.3.1: Boxplots of $d_b^{500}(u)$ [left] and $d_b^{n_u}(u)$ [right] for each u over the candidate grid of thresholds for the first sample of Case 1. The mean for each threshold is shown as a black point.

To demonstrate our findings from Figure A.3.1 more concretely, Figure A.3.2 shows the sampling distribution of the difference $d_b^{500}(u) - d_b^{n_u}(u)$ relative to the metric value $d_E(u)$ for each threshold u. Across almost all thresholds, most of the values within the bootstrap sampling distribution lie very close to zero and more importantly, the mean of the sampling distribution lies very close to zero. These findings indicate that the choice of interpolation grid has no meaningful effect on the value of the metric for a particular threshold. However, for the very largest threshold shown in the plot, we see a much larger range for the bootstrapped distribution of these relative differences, showing the effect of the over-sampling of the upper tail as you mentioned in your original review. Importantly, referring back to Figure A.3.1, the $d_b(u)$ values for the highest threshold are all larger than the largest $d_b(u)$ value for the optimal threshold in both plots. Thus, while there is a clear effect on the values of $d^{500}(u)$ and $d^{n_u}(u)$ for particular bootstrap samples, both Figure A.3.1 and A.3.2 show that the effect on the mean is relatively small and certainly, would not alter the selected threshold in any way.

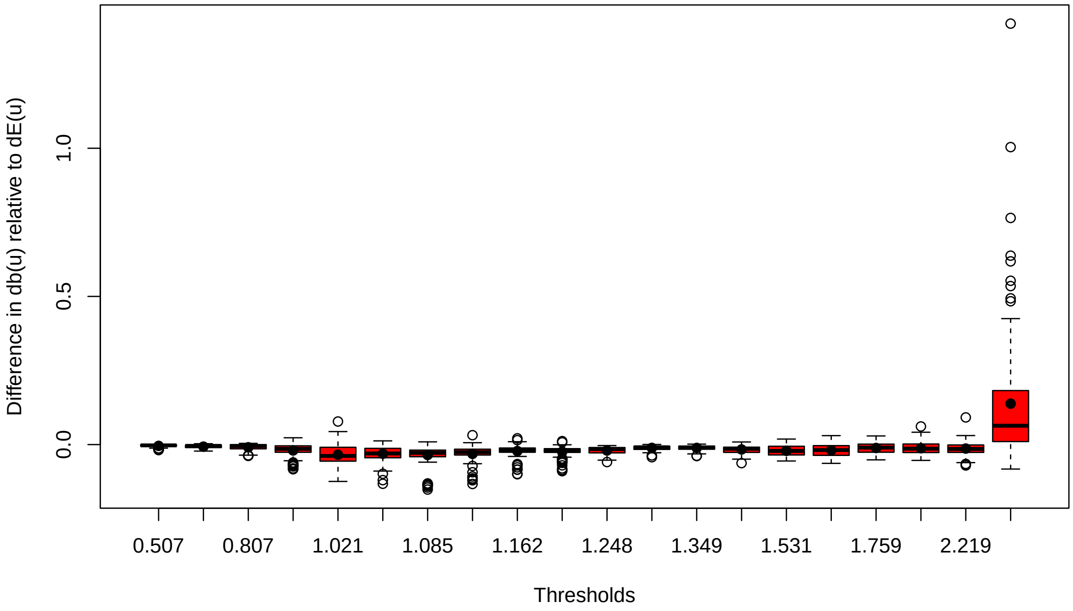

Figure A.3.2: Plot of $(d_b^{500}(u) - d_b^{n_u}(u))/d_E(u)$ for first sample of Case 1. The mean for each threshold is shown as black points.

For Case 1, the true threshold is at a low sample quantile (16.67%) and so, it is unlikely that the threshold choice would be affected by the under- or over-sampling of the tail as the only significant effect of this comes at very high thresholds. In a case where the optimal threshold lies at a higher sample quantile, based on the above analysis, we might expect the threshold choice to show greater sensitivity to choice of interpolation grid. To explore a case of this nature, we repeat the above analysis on the first sample from the replicated data of the Gaussian case. Here, we expect the optimal threshold to lie further into the tail due to the slow convergence of the Gaussian distribution to an extreme value limit. As a result, in theory, we expect that this case could show more sensitivity to the under- or over-sampling of the upper tail.

Figures A.3.3 and A.3.4 show the Gaussian results, in the same format as for Case 1. In Figure A.3.3, there is a clear threshold choice in both plots and there does not seem to be any effect from the choice of interpolation grid on the bootstrap sampling distribution of mean-absolute deviations, and certainly there is no effect on the mean values for any thresholds. In fact, surprisingly, any effect due to under- or over-sampling the upper tail is much smaller in this case than above. Figure A.3.4 reiterates this where the sampling distribution of relative differences lie close to zero for all thresholds. The choice of interpolation grid does not show any meaningful effect on the overall metric value and certainly, would not affect the choice of threshold for this dataset.

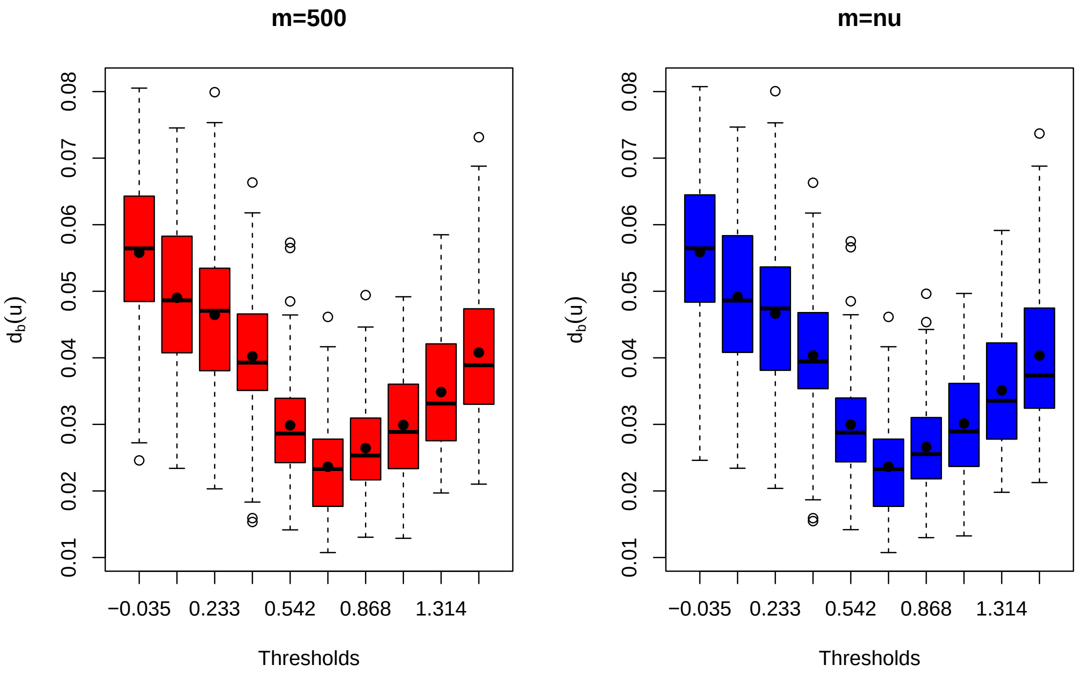

Figure A.3.3: Boxplots of $d_b(u)$ for the first sample of the Gaussian case with $m = 500$ (left) and $m = n_u$ (right). The mean for each threshold is shown as black points.

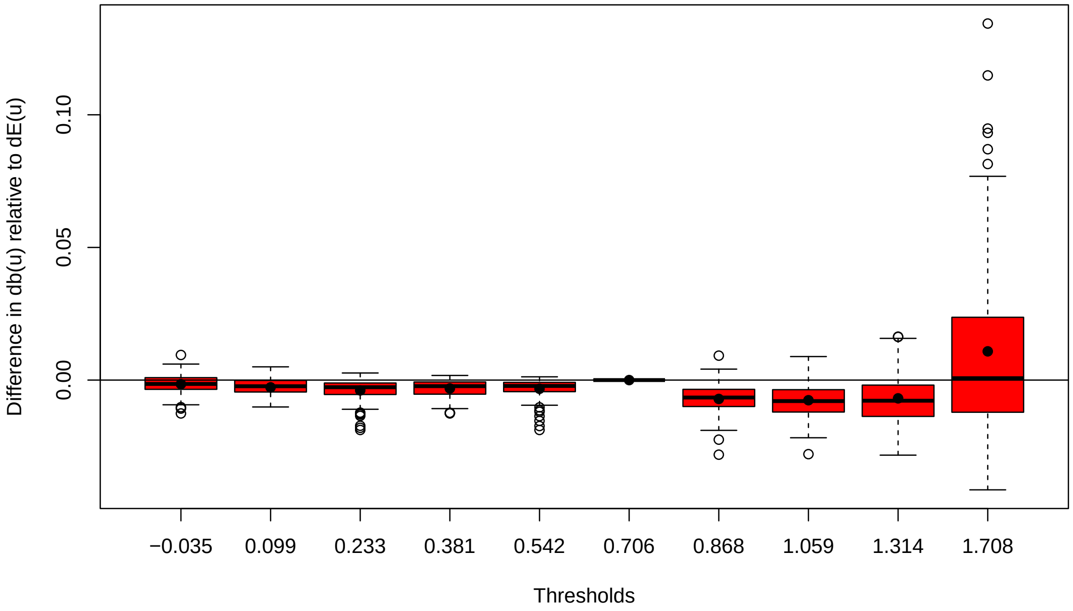

Figure A.3.4: Plot of $(d_b^{500}(u) - d_b^{n_u}(u))/d_E(u)$ for first sample of the Gaussian case. The mean for each threshold is shown as black points.

Given that our overall aim is to approximate the IAE between model quantiles and quantiles of the data generating process, we choose to keep the quality of the approximation of the IAE for particular samples fixed for a fair comparison across a range of candidate thresholds. Since m controls the quality of the approximation, we suggest $m = 500$ is sufficient based on the above results. Given we are confident in the quality of results for this choice, for convenience, we keep this fixed value across sample sizes and threshold levels in our simulation studies.

In conclusion, in order to balance accuracy and computation time for our large scale simulation studies, we find that the EQD method, with value $B = 100$ and $m = 500$, provides sufficiently accurate results in a timely manner, so we choose to use these values throughout, unless stated otherwise.

A.3.4 Selection of calibration data in the metric $d_b(u)$

Here, we provide results for an adjusted version of the EQD method (as suggested by a referee) with the findings being reported in Section 3.4.2. Here, the adjusted version takes the calibration data used for threshold estimation as the actual observed sample excesses $\boldsymbol{x}_u$ of candidate threshold u, as described below. This adjusted version differs from the EQD method as proposed in Section 3.4.1, where the calibration data for a particular bootstrapped sample are calculated based on the bootstrapped excesses $\boldsymbol{x}_u^b$. Thus, the only difference between the two approaches is that in $d_b(u)$, given by metric (3.4.1), the $Q(p_j; \boldsymbol{x}_u^b, \boldsymbol{q})$ term is replaced by $Q(p_j; \boldsymbol{x}_u, \boldsymbol{q})$ in the adjusted version.

Table A.3.6 shows the RMSE, bias and variance of threshold estimates using the proposed and the adjusted versions of the EQD method evaluated on 500 replicated samples from each of Cases 1-4. The relative performance of the two methods vary only slightly across all cases, and it is difficult to distinguish between the two methods, with each having the smaller RMSE an equal number of times. Furthermore, any difference between the two methods is very small relative to the differences between the EQD and the existing automated threshold selection approaches.

Using the adjusted method leads to larger variability in the distribution of $d_b(u)$ relative to the proposed method. This is because, unlike the proposed EQD method, the adjusted method does not compare like with like; one term in $d_b(u)$ is based on a bootstrap sample and the other on the actual data. Despite this increased variability, our analysis suggests that the adjustment does not necessarily lead to a notable change in either the $d_E(u)$ value for a particular candidate threshold or in the subsequently selected threshold. As a result, we choose to retain our proposed method and this is utilised throughout the analyses of the main text and the supplementary material.

	Proposed - $Q(p_j; \boldsymbol{x}_u^b, \boldsymbol{q})$			Adjusted - $Q(p_j; \boldsymbol{x}_u, \boldsymbol{q})$		
Case	RMSE	Bias	Variance	RMSE	Bias	Variance
Case 1	0.048	**0.034**	0.001	**0.047**	0.035	0.001
Case 2	0.060	0.031	0.003	**0.050**	**0.027**	**0.002**
Case 3	**0.060**	**0.042**	0.002	0.062	0.046	0.002
Case 4	**0.526**	**−0.515**	0.012	0.545	−0.535	**0.011**

Table A.3.6: RMSE, bias and variance of threshold estimates for Cases 1-4 for the EQD method found using the different calibration data in the metric $d_b(u)$: the proposed EQD with $Q(p_j; \boldsymbol{x}_u^b, \boldsymbol{q})$ and the adjusted version with $Q(p_j; \boldsymbol{x}_u, \boldsymbol{q})$.

A.3.5 Investigating the effect of bootstrapping

In this section, we provide further simulation experiments to investigate the effect of the bootstrapping component of the EQD method. We utilise a variant of the EQD method with no bootstrapping and compare the results against the original EQD method across Cases 0-4 and the Gaussian case (with $n = 2000$), with each case based on 500 replicated samples (Case 0 is outlined in Section A.4.3, while Cases 1-4 and the Gaussian case are outlined in Section 3.6).

Table A.3.7 provides results for the RMSE, bias and variance of threshold choices for the EQD method with and without bootstrapping, for Cases 0-4. For Cases 0-3, removing the bootstraps and only evaluating the metric on the original sample leads to

RMSEs of threshold choice almost twice as large than for the original EQD method. In each of these cases, there is an increase in positive bias and the variance increases by a factor of at least 4. Thus, removing the bootstrapping component leads to higher and more variable threshold choices for these cases as would be expected. In contrast, for Case 4, the removal of the bootstrapping component actually leads to a slight decrease in RMSE. This decrease stems from a reduction in the negative bias component as a result of threshold choices being slightly higher than the original method. This reduction in bias comes at a cost of greater variability in selected thresholds. This greater variability is to be expected by not averaging over bootstrap samples.

	Case 0			Case 1		
	RMSE	Bias	Variance	RMSE	Bias	Variance
Original	0.042	0.019	0.001	0.048	0.034	0.001
No bootstraps	0.095	0.039	0.008	0.090	0.054	0.005

	Case 2		
	RMSE	Bias	Variance
Original	0.060	0.031	0.003
No bootstraps	0.122	0.063	0.011

	Case 3			Case 4		
	RMSE	Bias	Variance	RMSE	Bias	Variance
Original	0.060	0.042	0.002	0.526	-0.515	0.012
No bootstraps	0.138	0.080	0.013	0.473	-0.441	0.030

Table A.3.7: RMSE, bias and variance of threshold choice for Cases 0-4 for EQD only evaluated on original sample (i.e., no bootstrapping).

Table A.3.8 shows the RMSE of quantile estimation following threshold selection using the original EQD method and the variant with no bootstrapping for the Gaussian case. There is a slight improvement in terms of RMSE from including the bootstrapping component across all quantiles but results are very similar across both methods.

Overall, across the studied cases, the addition of bootstrapping to the EQD method leads to a systematic reduction in the variance and lower threshold choices, typically resulting in reduced RMSE, for the selected thresholds and the subsequent quantile es-

j	Original	No bootstraps
0	0.214	0.217
1	0.430	0.435
2	0.703	0.715

Table A.3.8: RMSEs of estimated $(1 - p_{j,n})$-quantiles where $p_{j,n} = 1/(10^j n)$, for $j = 0, 1, 2$, for the Gaussian case.

timates. While the mean absolute deviation is a robust and effective metric for threshold selection, the bootstrapping component provides further stability in the threshold choices and allows us to account for the increasing uncertainty in parameter estimates as the threshold increases. These investigations support our choice to use bootstrapping as a key component of the EQD method.

A.4 Additional simulation study results

A.4.1 Overview

We provide results that expand on those in Section 3.6 of the main text. Section A.4.2 provides bias-variance decompositions for the RMSEs values given in Section 3.6 and presents additional results for the Danielsson et al. (2001, 2019) methods, evaluating threshold estimation for Cases 1-4 and quantile estimation for Gaussian data. In Section A.4.3, we provide further threshold estimation results for additional cases, which have alternative parameters and sample sizes to those of Cases 1-4. Section A.4.4 presents additional results to assess the sensitivity of the methods to the choice of candidate threshold grids using Cases 1 and 4 with candidate grids defined above the mode of the distribution and for the Gaussian case with candidate grids spanning the entire range of the sample.

For the results for the Danielsson et al. (2001) method given in Section A.4.2, we utilised the *tea* package (Ossberger, 2020), i.e., the package was not built by the authors of the paper; whereas for the Danielsson et al. (2019) method, as there did not seem to be code freely available, we constructed our own function.

A.4.2 Detailed results for the case studies

We provide the RMSE and bias-variance decomposition for the application of the EQD method and all methods described in Section 3.3 of the main text to the case studies detailed in Section 3.6 of the main text.

In what follows, we find here that the Danielsson et al. (2001, 2019) methods perform much worse that the EQD, Wadsworth (2016) and Northrop et al. (2017) methods both in terms of threshold and quantile estimation. We therefore omit results for these methods in Section 3.6 of the main text and do not to apply them beyond this section of the supplementary material.

Scenario 1: True GPD tail - Cases 1-4

Threshold recovery:

For Cases 1-4, Table A.4.1 shows the RMSE, bias and variance of the thresholds selected by the EQD, Wadsworth (2016) and Northrop et al. (2017) methods. The RMSE is also reported in Table 3.6.2, but is repeated here for completeness. The EQD method has the smallest RMSE and the least variable estimates in all cases, and is the least biased method Cases 1-3. Table A.4.1 also presents equivalent results for the Danielsson et al. (2001, 2019) methods. Both methods show considerably larger RMSEs than the EQD, Wadsworth (2016) and Northrop et al. (2017) methods, due to the large positive biases of these methods across all cases. In particular, for Cases 3 and 4, the Danielsson et al. (2019) method has the smallest variance of all the methods but its larger bias leads to RMSE values much larger than those of the EQD and other methods in Table A.4.1.

[1] Results for Wadsworth are calculated only on the samples where a threshold was estimated. It failed estimate a threshold for 2.4%, 26.4%, 0%, 3.6% of the simulated samples in Cases 1-4, respectively.

	EQD			*Wadsworth*[1]		
Case	RMSE	Bias	Variance	RMSE	Bias	Variance
Case 1	**0.048**	**0.034**	**0.001**	0.349	0.111	0.110
Case 2	**0.060**	**0.031**	**0.003**	0.461	0.204	0.172
Case 3	**0.060**	**0.042**	**0.002**	0.221	0.060	0.045
Case 4	**0.526**	−0.515	**0.012**	0.627	−0.407	0.230

	Northrop		
Case	RMSE	Bias	Variance
Case 1	0.536	0.276	0.212
Case 2	0.507	0.238	0.201
Case 3	0.463	0.256	0.149
Case 4	0.543	**−0.222**	0.246

	Danielsson et al. (2001)			*Danielsson et al. (2019)*		
Case	RMSE	Bias	Variance	RMSE	Bias	Variance
Case 1	2.767	2.416	1.825	1.635	1.633	0.007
Case 2	2.212	1.850	1.474	1.639	1.634	0.017
Case 3	2.528	2.441	0.435	1.314	1.314	0.001
Case 4	2.838	2.499	1.813	1.138	1.134	0.009

Table A.4.1: RMSE, bias and variance of the threshold estimates for Cases 1-4: for EQD, Wadsworth and Northrop methods (top) and Danielsson et al. (2001, 2019) (bottom). Results are based on 500 replicated samples.

Quantile recovery:

Tables A.4.2 & A.4.3 present the bias and variance of the quantile estimates for the EQD, Wadsworth (2016) and Northrop et al. (2017) methods applied to Cases 1-4. These are shown for the $(1 - p_{j,n})$-quantile estimates where $p_{j,n} = 1/(10^j n)$ for $j = 0, 1, 2$ and n denotes the length of the simulated dataset. As mentioned in Section 3.6, we use exceedance probabilities of this form because we have simulated samples of different sizes and want to make extrapolation equally difficult in each case. These bias and variance components correspond to the RMSE values presented in Table 3.6.3 in Section 3.6, where the EQD method achieves the lowest RMSEs in all cases and quantiles. Table A.4.3 shows that these lower RMSE values derive mainly from the variance component; for all j and in all cases the EQD method shows the least variability in quantile estimates, with the differences between the methods becoming more evident for higher j. The smallest absolute bias values in Table A.4.2 vary between each of the methods but the EQD incurs the least bias in the majority of cases and quantiles. In particular, the EQD achieves the smallest absolute bias for all j in Cases 2 and 3.

	EQD	*Wadsworth*[1]	*Northrop*	*EQD*	*Wadsworth*[1]	*Northrop*
j	**Case 1**			**Case 2**		
0	**−0.021**	−0.079	−0.075	**−0.049**	−0.118	−0.071
1	−0.015	−0.192	**−0.001**	**−0.046**	−0.316	0.245
2	**0.044**	−0.319	0.554	**0.069**	−0.532	2.568
	Case 3			**Case 4**		
0	**−0.008**	−0.026	−0.041	−0.283	−0.372	**−0.192**
1	**−0.007**	−0.047	−0.065	−0.722	−0.965	**−0.344**
2	**−0.002**	−0.066	−0.074	−1.410	−1.809	**−0.258**

Table A.4.2: Bias of the estimated quantiles in Cases 1-4 based on fitted GPD above chosen threshold. The smallest absolute bias for each quantile are highlighted in bold.

	EQD	*Wadsworth*[1]	*Northrop*	*EQD*	*Wadsworth*[1]	*Northrop*
j	**Case 1**			**Case 2**		
0	**0.317**	0.348	0.565	**0.358**	0.386	0.538
1	**1.585**	1.903	5.657	**2.215**	2.611	12.305
2	**6.000**	7.297	50.155	**9.743**	11.885	519.569
	Case 3			**Case 4**		
0	**0.036**	0.038	0.051	**0.379**	0.503	0.591
1	**0.105**	0.116	0.199	**1.926**	3.317	4.805
2	**0.233**	0.263	0.549	**7.287**	16.879	30.999

Table A.4.3: Variance of the estimated quantiles in Cases 1-4 based on fitted GPD above chosen threshold. The smallest variance for each quantile are highlighted in bold.

For completeness, Table A.4.4 presents the equivalent RMSE values for the Danielsson et al. (2001, 2019) methods. These can be compared with the RMSE results for the EQD, Wadsworth (2016) and Northrop et al. (2017) methods in Table 3.6.3. For $j = 0$ the threshold choice should not be too important because we are not extrapolating but Table A.4.4 shows that, even in this case, the Danielsson et al. (2001, 2019) approaches have RMSE values much greater than the other methods, by factors of between $1.5 - 3$ across the cases. This difference in performance is only exacerbated as we extrapolate further. For $j = 1, 2$, both approaches lead to RMSEs that are orders of magnitude larger than any of the other methods analysed. For example, when $j = 2$ the RMSEs of the Danielsson et al. (2001) and Danielsson et al. (2019) methods are respectively $4 - 70$ and $3 - 7$ times larger than those of the EQD method.

j	*Danielsson 2001*	*Danielsson 2019*	*Danielsson 2001*	*Danielsson 2019*
	Case 1		**Case 2**	
0	1.020	0.859	0.962	0.757
1	3.128	3.172	3.806	3.991
2	12.943	10.303	110.347	23.102
	Case 3		**Case 4**	
0	0.675	0.262	1.118	0.865
1	1.655	0.570	3.938	2.978
2	38.001	1.000	40.653	8.721

Table A.4.4: RMSEs in the estimated quantiles in Cases 1-4 based on fitted GPD above chosen threshold for the Danielsson et al. (2001) and Danielsson et al. (2019) methods.

Scenario 2: Gaussian data

Quantile recovery:

We next consider the case of Gaussian data, using 500 simulated datasets of size $n =$ 2000 and 20000. Tables A.4.5 and A.4.6 present the bias, variance and RMSE for the estimation of the $(1 - p_{j,n})$-quantiles (where $p_{j,n} = 1/(10^j n)$ and $j = 0, 1, 2$), based on the thresholds selected by the EQD, Wadsworth (2016) and Northrop et al. (2017) methods. These RMSE values are detailed in Table 3.6.5.

Table A.4.5 shows that for the smaller sample size of $n = 2000$, the EQD method achieves the smallest RMSEs and variance for all j for all methods. The Danielsson et al. (2019) and Northrop et al. (2017) methods incur the least absolute bias in quantile estimation due to their slightly higher threshold choices, see Table A.4.7, and have similar RMSE values smaller than that of the Wadsworth (2016) method in this aspect. The Danielsson et al. (2001) method incurs considerable bias with large variance in its quantile estimates, leading to the highest RMSEs of all analysed methods.

[2] Results for the Wadsworth method, which failed on 0.4% of the samples here, are calculated only for samples where a threshold estimate was obtained.

	EQD			*Wadsworth*[2]		
j	RMSE	Bias	Variance	RMSE	Bias	Variance
0	**0.214**	−0.086	**0.038**	0.239	−0.120	0.043
1	**0.430**	−0.275	**0.109**	0.529	−0.366	0.147
2	**0.703**	−0.521	**0.222**	0.890	−0.654	0.365

	Northrop		
j	RMSE	Bias	Variance
0	0.225	−0.076	0.045
1	0.461	−0.224	0.162
2	0.765	−0.414	0.414

	Danielsson et al. (2001)			*Danielsson et al. (2019)*		
j	RMSE	Bias	Variance	RMSE	Bias	Variance
0	0.758	−0.470	0.354	0.232	**−0.059**	0.050
1	1.550	−0.815	1.739	0.479	**−0.173**	0.200
2	33.183	−1.380	1099.182	0.790	**−0.321**	0.522

Table A.4.5: RMSE, bias and variance of estimated quantiles from a Gaussian distribution with sample size of 2000: for EQD, Wadsworth and Northrop methods (top) and Danielsson et al. (2001, 2019) (bottom). Results are based on 500 replicated samples.

As the sample size n increases, the relative importance of bias and variance terms within the RMSE shifts, with low bias becoming increasingly important. Table A.4.6 shows that when $n = 20000$, the Northrop et al. (2017) method achieves the lowest RMSE and bias over all j. The EQD method takes second place, again showing the least variability in its estimates. Note that all of the methods show decreased RMSEs as n increases from 2000 to 20000, even though the bias values do not all decrease. A possible reason for this lack of reduction in bias is the slow convergence of the Gaussian distribution to the extreme value limit, so an order of magnitude increase in sample sizes could be required for the bias to reduce. We attempted to apply the Danielsson et al. (2001, 2019) methods to the 500 Gaussian samples with the larger sample size of $n = 20000$ but the computation time was simply too large.

	EQD			*Wadsworth*		
j	RMSE	Bias	Variance	RMSE	Bias	Variance
0	0.187	−0.131	**0.018**	0.214	−0.165	0.019
1	0.368	−0.307	**0.042**	0.422	−0.366	0.044
2	0.594	−0.528	**0.074**	0.672	−0.611	0.078

	Northrop		
j	RMSE	Bias	Variance
0	**0.172**	**−0.104**	0.019
1	**0.331**	**−0.255**	0.045
2	**0.533**	**−0.450**	0.081

Table A.4.6: RMSE, bias and variance of estimated quantiles from a Gaussian distribution with sample size of 20000. Results are based on 500 replicated samples.

Threshold Recovery:

There is no true GPD threshold for the Gaussian scenario but the quantile estimates discussed in Tables A.4.5 and A.4.6 require a preceding step of selecting a suitable threshold above which the GPD approximation is adequate. In Table A.4.7, we provide information about the selected thresholds for the Gaussian data, presenting the 2.5%, 50%, 97.5% values of the sampling distribution of the threshold estimates (presented as a quantile of the Gaussian distribution for each method). The results show that the Wadsworth (2016) method tends to estimate the threshold lowest, followed by the EQD method, and then the Northrop et al. (2017) method which tends to estimate the highest threshold values, with this finding being consistent across the sampling distribution quantiles. It is interesting to see that, even with a very large sample size, almost always thresholds are estimated to be below the 95% quantile (the maximum of the candidate thresholds) of the Gaussian distribution. This is surprising given its widely known slow convergence issues.

A.4.3 Extra case studies

This section provides a description and the results of additional case studies, beyond Cases 1-4, which were omitted from Section 3.6. These are denoted by Cases 0, 5, 6, 7

	$n = 2000$			$n = 20000$		
	EQD	*Wadsworth*[2]	*Northrop*	*EQD*	*Wadsworth*	*Northrop*
50% Q	75	50	80	87.5	84	91.5
2.5%, 97.5% Q	55, 90	50, 95	60, 95	77.5, 94	69.5, 95	82.5, 95

Table A.4.7: Sampling distribution quantiles (2.5%, 50%, 97.5%) of quantile level Q (%) for the threshold estimates for each method derived from 500 replicated samples from a Gaussian distribution for two sample sizes n.

and 8. Specifically, these extra cases are:

Case 0: We simulate samples of size $n = 1000$ from a GPD(0.5,0.1) above a threshold $u = 1$ with no data below the threshold.

Case 5: We simulate samples from the distribution (A.2.1) with $(\sigma_u, \xi) = (0.5, 0.1)$, but with a reduced sample size of $n = 120$.

Case 6 & 7: We simulate samples from the distribution (A.2.1) with a sample size of $n = 1200$, but with shape parameters $\xi < -0.05$, i.e., $\xi = -0.2$ for Case 6 and $\xi = -0.3$ for Case 7.

Case 8: We simulate samples from the distribution (A.2.1) with $(\sigma_u, \xi) = (0.5, 0.1)$, but with an increased sample size of $n = 20000$.

We chose to omit Case 0 from the main text due to its simplicity, in that it should be the easiest case for threshold selection due to the lack of data below the threshold. Cases 5, 6 and 7 were also omitted due to poor performance of the Wadsworth (2016) method which failed to estimate a threshold in the majority of samples generated from each of these cases. This high failure rate occurs for two reasons; dependence between parameter estimates when using candidate thresholds which lie in close proximity for small sample sizes or in the cases where $\xi < -0.05$, where an error results from a divergent integral in the calculation of the inverse Fisher information matrix. Thus, for Cases 5, 6 and 7, here we only show comparisons of the EQD method against the Northrop et al. (2017) method. Finally, Case 8 was also omitted from the main text due to the large sample size being atypical of extreme value analyses, but we decided to include a large sample case to explore how the EQD compares with existing methods which base their theoretical justification on asymptotic arguments.

In all of the cases, results presented here, for each of the methods, are based on 500 replicated samples and we take the set of candidate thresholds as the sample quantiles at levels 0%, 5% . . . , 95%, as in Section 3.6. Specifically, for Case 8, we considered an additional finer grid of candidate thresholds as the sample quantiles at levels 0%, 0.5%, . . . , 95%. For Case 0, we initially used a set of candidate thresholds which contains values lower than the minimum of the sample, but Wadsworth (2016) and Northrop et al. (2017) methods had major problems, and so we omit those results. However, this restriction to consider only candidate thresholds which are sample quantiles automatically positively biases threshold estimates in Case 0. In any simulation study, it is reasonable to consider candidate thresholds above and below the true threshold, so the fact that the EQD method continues to work well for candidate thresholds below the sample quantiles is a particularly pleasing feature, even if not illustrated here. In Section A.4.4, we explore the effect of candidate thresholds which lie below the mode for Cases 1, 4 and the Gaussian case for the EQD, Wadsworth and Northrop methods.

Case 0 should be much easier to estimate than for even Cases 1-3, with the lowest candidate threshold being very close to the true threshold. Table A.4.8 provides the RMSE, bias and variance of threshold estimates for each of the methods for this case, with the EQD obtaining the lowest value for each of the three summary features by considerable margins. Hence, in the most ideal case for threshold estimation, the EQD method excels in its performance.

	EQD	*Wadsworth*[3]	*Northrop*
RMSE	**0.042**	0.564	0.566
Bias	**0.019**	0.228	0.326
Variance	**0.001**	0.266	0.214

Table A.4.8: RMSE, bias and variance of the threshold estimates for Case 0. Results are based on 500 replicated samples. The smallest in each case is given in bold.

[3] Results for Wadsworth are calculated only on the samples where a threshold was estimated, the method failed on 3.8% of the simulated samples in Case 0.

Table A.4.9 shows the RMSEs for the EQD and Northrop et al. (2017) methods for estimating the threshold in Case 5, 6 and 7, with the EQD performing the best on each occasion. Case 5 is particularly important as the small sample size is typical of many data applications, so it is especially pleasing to see that the EQD method outperforms the Northrop et al. (2017) method by the largest of the three margins in this case. Northrop et al. (2017) performs especially badly in Case 5 in terms of variance. Furthermore, the EQD method, relative to that of Northrop et al. (2017), has a smaller (equal) absolute error in threshold estimates in 70.0% (20.4%) of samples. Similarly, in Cases 6 and 7, the EQD achieves a smaller (equal) absolute error in 59.8% (18.0%) of samples and 50.2% (18.8%) respectively.

	Case 5	Case 6	Case 7
EQD	**0.078**	**0.107**	**0.185**
Northrop	0.602	0.373	0.341

Table A.4.9: RMSEs of the threshold estimates of the EQD and Northrop methods in Cases 5, 6 and 7. The smallest value in each case is given in bold.

Table A.4.10 shows the RMSEs of threshold choice for Case 8. The results are shown for each method using two different candidate grids of thresholds. In contrast to the previous results, the Wadsworth method slightly outperforms the EQD achieving the smallest RMSEs for these large samples. However, the sample size for this to be achieved significantly exceeds that for data in practice. This illustrates the potential benefits, but also serious limitations, of relying on asymptotic methods to guide threshold selection.

Grid (% quantile)	*EQD*	*Wadsworth*	*Northrop*
0 (5) 95	0.036	**0.021**	0.503
0 (0.5) 95	0.027	**0.003**	0.529

Table A.4.10: RMSE of the threshold choices for Case 8 for two different candidate grids given with notation *start (increment) end*. The smallest values are in bold.

Table A.4.11 shows the RMSEs of quantile estimation for Case 8, where for threshold selection the Wadsworth method has a slight benefit over the EQD method. Here, for quantile estimation, there are similar findings, the Wadsworth method achieves the lowest RMSEs, followed closely by the EQD and then, the Northrop method which obtains significantly higher RMSEs.

Grid	0 (5) 95			0 (0.5) 95		
j	*EQD*	*Wadsworth*	*Northrop*	*EQD*	*Wadsworth*	*Northrop*
0	0.370	**0.364**	0.546	0.369	**0.358**	0.573
1	0.694	**0.681**	1.157	0.692	**0.669**	1.200
2	1.199	**1.174**	2.189	1.196	**1.151**	2.242

Table A.4.11: RMSEs in the estimated quantiles in Case 8 samples based on fitted GPD above chosen threshold for two candidate grids given with notation *start (increment) end*. The smallest RMSE for each quantile are highlighted in bold.

A.4.4 Sensitivity to the choice of candidate threshold grids

In this section, we provide further simulation experiments where we choose the candidate threshold grids more in line with general extreme value analyses (i.e., above the mode) and explore the sensitivity of methods to candidate thresholds which lie below the mode of the data. Since the GPD density is monotonically decreasing for realistic values of ξ (i.e., when $\xi > -1$), when utilising a threshold selection procedure, it is unusual to consider thresholds which lie below an obvious mode. Many threshold selection procedures are simply not set up to handle data from non-monotonically decreasing densities. We explore a different choice for the candidate threshold grids to ensure that our choices made in the main text do not unfairly favour the EQD method.

Given that our aim is to provide a method which requires no user input, we consider candidate thresholds across the range of the sample data and allow the method to make the threshold selection whether the conditions are suitable or not. In the main text and the supplementary material, outside of Gaussian case where we specifically restricted our choice of candidate thresholds such that $u > q_{50}$ (with q_{50} being the sample median), and Case 0, where the optimal threshold is at the lowest sample quantile, we have only explored cases where candidate thresholds span the range of the dataset. Here, we reanalyse two cases; Case 1 and Case 4, now with candidate threshold grids which lie above the mode of the distributions.

To avoid sensitivities from different mode estimators clouding the results, we use the true mode to help define our range/set of candidate thresholds. In Case 1, the true mode is trivial to find and it is the optimal threshold. For Case 4, the mode needs to be found numerically, using the expression for the density function. Once the mode is known, we define the candidate threshold set as sample quantiles at levels $0\%, \ldots, 95\%$ of the data which lies above the mode.

First, we consider the effect of the range $[u_1, u_k]$ of the candidate thresholds on threshold selection performance. Table A.4.12 provides the RMSE, bias and variance of threshold estimation for samples from both of these cases, using the EQD, Wadsworth and Northrop methods each applied with candidate thresholds across the distribution, as given in Section 3.6.1, and only above the true mode. In terms of RMSE, the EQD method outperforms the other two methods regardless of how the candidate threshold grid is chosen. For the "Above mode" grid, we have rather different performances for the Wadsworth and Northrop methods relative to the "Original" candidate grid. In Case 1, these two methods perform very similarly, with RMSEs approximately 15 times larger than the EQD. In Case 4, the Wadsworth method fails completely with the "Above mode" grid due to the grid being too fine relative to the sample size. For the Northrop method, the RMSE of threshold estimation is increased with the "Above mode" grid and the differential between the Northrop and EQD becomes wider with a RMSE that is 1.15 times larger the EQD method.

For the EQD method, using only candidate thresholds above the true mode results in a smaller RMSE (and bias and variance) for both Cases 1 and 4, presumably as we are now using a finer grid of candidate thresholds so candidate thresholds lie closer to the optimal value. In contrast, the performance of the Wadsworth and Northrop methods is worsened when we restrict the candidate thresholds to be above the mode,

	RMSE	Bias	Variance	RMSE	Bias	Variance	RMSE	Bias	Variance
	Case 1								
	EQD			Wadsworth			Northrop		
Original	0.048	0.034	0.001	0.349	0.111	0.110	0.536	0.276	0.212
Above mode	0.038	0.018	0.001	0.577	0.236	0.277	0.597	0.348	0.235
	Case 4								
Original	0.526	−0.515	0.012	0.627	−0.407	0.230	0.543	−0.222	0.246
Above mode	0.514	−0.505	0.009	NA	NA	NA	0.589	−0.127	0.331

Table A.4.12: RMSE, bias and variance of threshold choice for Case 1 and Case 4 samples for EQD, Wadsworth and Northrop methods. The methods are evaluated for two sets of candidate threshold sets: Original, using candidate threshold across the whole sample; Above mode, using candidate thresholds which only lie above the mode.

primarily due to an increase in variance.

We consider the effect of the range of the candidate thresholds on quantile estimation for the Gaussian case. Previously, for the Gaussian case, we reported results only using candidate thresholds above the sample median, i.e., essentially the mode. Here, we assess the effect of allowing candidate thresholds to range across the whole sample, i.e., the sample quantiles at levels $0\%, 5\%, \ldots, 95\%$ of the whole data sample.

The results, in terms of RMSE for three different quantiles, are presented in Table A.4.13 for the two different candidate threshold grids. Firstly, we find that the Wadsworth method completely fails with this extended range of candidate thresholds, so this method is omitted from the table. For both the EQD and Northrop methods, we have almost identical values regardless of the set of candidate thresholds we use, with the RMSE always smallest for the EQD.

	EQD		Northrop	
j	Original	Whole data	Original	Whole data
0	0.214	0.214	0.225	0.225
1	0.430	0.431	0.461	0.460
2	0.703	0.707	0.765	0.763

Table A.4.13: RMSEs of estimated $(1 - p_{j,n})$-quantiles where $p_{j,n} = 1/(10^j n)$, for $j = 0, 1, 2$, for the Gaussian case when applying methods with candidate thresholds across the range of the data.

The results here indicate that the choice of candidate grid taken in the main text provides no unfair advantage to the EQD method, in fact it performs even better if the choice of candidate threshold grid is tuned by exploiting knowledge of the value of the mode of the distribution.

A.5 Further details on coverage probability results

This section provides more detailed results for the coverage probabilities for the 500 replicated samples of Case 4 and the Gaussian case outlined in Section 3.6 of the main text. Here, we also show results for the 50% confidence level and include an extended range of exceedance probabilities at which the coverage of the true quantiles is evaluated.

Tables A.5.1 and A.5.2 expanding on the results given in Tables 3.6.4 and 3.6.6, providing coverage probabilities and average CI width ratios using Algorithms 1 and 2 for 500 samples derived from Case 4 and the Gaussian distribution. These results allow us to conclude that the additional threshold uncertainty, captured by Algorithm 2, leads to a significant improvement in the calibration of CIs and that it is vital to include this uncertainty in any inference for extreme quantiles.

Scenario 1: True GPD tail - Case 4

Table A.5.1 shows that for all confidence levels and exceedance probabilities, Algorithm 1, which includes uncertainty only from the GPD parameter estimation, substantially under-estimates the uncertainty in quantile estimates. This leads to CIs which do not cover the true quantiles to the nominal levels of confidence. In contrast, the inclusion of the additional threshold uncertainty in Algorithm 2 leads to significant increases in the coverage of true quantiles with coverage probabilities at all quantile levels lying very close to the nominal confidence level, especially at the 95% level. From a practical perspective, it is reassuring that this improvement in coverage is achieved with only 40-68% average increases in the width of the CIs.

p	$1/n$	$1/3n$	$1/5n$	$1/10n$	$1/25n$	$1/50n$	$1/100n$	$1/200n$	$1/500n$
	50% confidence								
Alg 1	0.398	0.390	0.384	0.368	0.368	0.358	0.358	0.350	0.342
Alg 2	0.526	0.498	0.500	0.496	0.480	0.476	0.474	0.468	0.464
CI ratio	1.408	1.413	1.414	1.414	1.414	1.413	1.413	1.412	1.411
	80% confidence								
Alg 1	0.646	0.642	0.630	0.618	0.616	0.608	0.606	0.600	0.594
Alg 2	0.798	0.778	0.770	0.772	0.760	0.760	0.758	0.762	0.758
CI ratio	1.430	1.440	1.445	1.452	1.461	1.468	1.475	1.483	1.495
	95% confidence								
Alg 1	0.834	0.810	0.808	0.804	0.798	0.796	0.794	0.788	0.788
Alg 2	0.954	0.950	0.950	0.948	0.942	0.942	0.944	0.944	0.944
CI ratio	1.484	1.511	1.525	1.546	1.574	1.597	1.621	1.646	1.682

Table A.5.1: Coverage probabilities for estimated $(1-p)$-quantiles using Algorithms 1 and 2 for Case 4, with sample size of 1000. Values are based on 500 replicated samples.

Scenario 2: Gaussian data

Table A.5.2 shows that, for Gaussian variables across all quantiles, Algorithm 1 and 2 are less successful in the coverage of the true quantiles than in Case 4. Specifically, there is a significant under-estimation of the estimated uncertainty necessary to provide coverage probabilities near to the nominal confidence level, and the actual coverage is decreasing with the level of extrapolation required. This is not too surprising as it is well-established that Gaussian variables exhibit quite slow convergence to an extreme value limit. However, what Table A.5.2 shows is that the additional threshold uncertainty in Algorithm 2 leads to a substantial improvement in actual coverage across all quantiles and for all nominal confidence levels. This improvement is achieved with the CI widths on average being extended by 45-73%. In particular, for the $p = 1/n, 1/3n, 1/5n$, which are typical levels of extrapolation from a sample in practical contexts, Algorithm 2 achieves a workable performance, with coverage reasonably close to the nominal level.

p	$1/n$	$1/3n$	$1/5n$	$1/10n$	$1/25n$	$1/50n$	$1/100n$	$1/200n$	$1/500n$
	50% confidence								
Alg 1	0.358	0.300	0.292	0.278	0.240	0.218	0.204	0.190	0.168
Alg 2	0.462	0.398	0.354	0.322	0.278	0.262	0.232	0.212	0.200
CI ratio	1.461	1.463	1.462	1.465	1.465	1.466	1.465	1.466	1.465
	80% confidence								
Alg 1	0.588	0.522	0.498	0.450	0.402	0.388	0.366	0.348	0.316
Alg 2	0.718	0.656	0.630	0.598	0.542	0.516	0.492	0.476	0.446
CI ratio	1.457	1.468	1.473	1.480	1.493	1.501	1.509	1.517	1.526
	95% confidence								
Alg 1	0.750	0.674	0.650	0.618	0.580	0.550	0.510	0.490	0.466
Alg 2	0.866	0.838	0.828	0.814	0.794	0.772	0.756	0.742	0.722
CI ratio	1.495	1.531	1.549	1.576	1.611	1.638	1.665	1.692	1.729

Table A.5.2: Coverage probabilities for estimated $(1-p)$-quantiles using Algorithms 1 and 2 for a Gaussian distribution, with sample size of 2000. Values are based on 500 replicated samples.

Appendix B

Supplementary materials to Chapter 4

B.1 Sensitivity to the baseline probability, p_1

A range of baseline probabilities were tested across the whole dataset, and the resulting threshold and model fits were used to calculate a right-sided Anderson-Darling (ADr) test statistic and the p-value (Sinclair et al., 1990; Solari et al., 2017). For more details on the ADr test, see the main text. The return periods that were tested for the baseline probabilities were 0.083, 0.167, 0.25, 0.33, 0.5, 0.667, and 1.0 years. These equate to the 1 in 1, 2, 3, 4, 6, 8 and 12 month events.

The results of this sensitivity test are shown in Figure B.1.1. Panel (a) presents the ADr test statistic for the 7 return periods tested. When looking at the median and interquartile ranges of the ADr test statistics, the threshold selection looks relatively insensitive to the return period chosen, with very little differences between the 0.167, 0.25, 0.333, and 0.5 year return periods. When considering the ADr test p-value (panel (b)), there is only small differences between the 0.167, 0.25, 0.33 and 0.5 year return periods. We take this as evidence that any one of these values would suffice as the baseline probability, p_1.

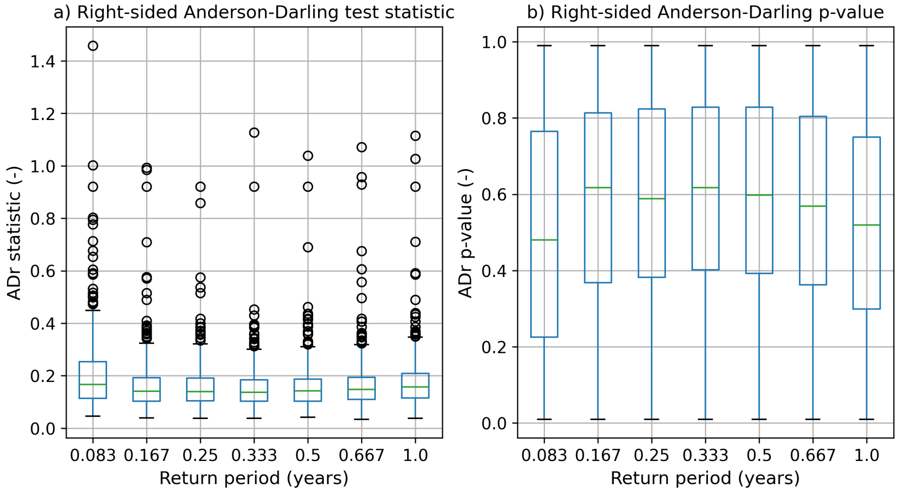

Figure B.1.1: The results from the sensitivity test of different baseline probabilities.

B.2 Sensitivity to the number of quantile levels, m

Following Murphy et al. (2025), a sensitivity test to the number of quantile levels, m was carried out. The values of m tested were 10, 50, 100, 200, 500, 1000 and ‘n_exceedances’, which denotes the number of exceedances over the baseline event for each tide gauge record. The range of m values for ‘n_exceedances’ are shown below in Figure B.2.1. The full range spreads between 161 to 811, and the median is centred on 231.

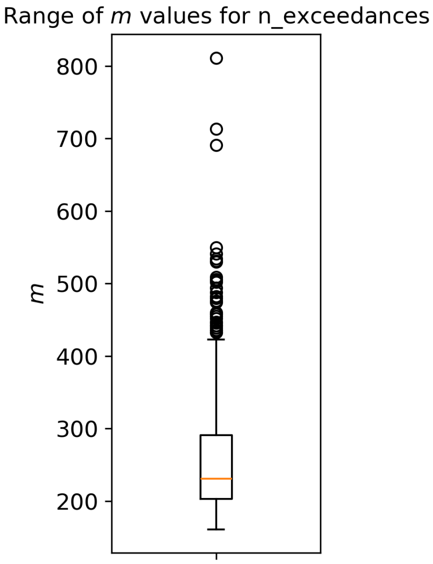

Figure B.2.1: The range of m values used by 'n_exceedances', i.e., the number of exceedances over the baseline probability for each tide gauge record.

The results of this sensitivity analysis are presented in Figure B.2.2, showing that the method is quite insensitive to the m value used. This is similar to the findings of Murphy et al. (2025). We recommend using any value over 10, and choose to use $m = 500$ in this study for consistency with Murphy et al. (2025).

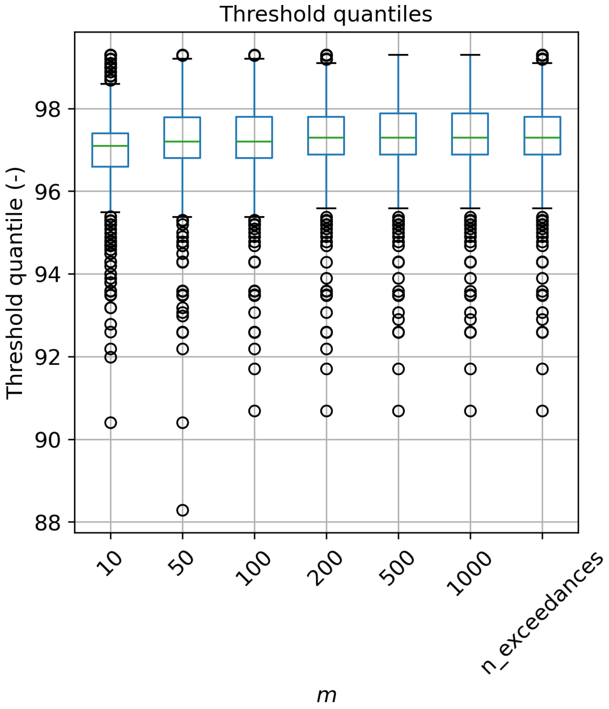

Figure B.2.2: The results of the sensitivity test using different m values. 'n_exceedances' refers to the number of exceedances over the baseline event, at each tide gauge record.

Appendix C

Supplementary materials to Chapter 5

C.1 Annual weighted mean endpoint estimates

In Section 5.4.5, we broke $e_{\mathrm{wm}}(\mathcal{S}_F)$ into a weighted sum of yearly weighted mean values, $e_{\mathrm{wm}}(T \mid \mathcal{S}_F)$ for year T. Figure C.1.1 provides point estimates of these aggregated weighting values and the annual endpoint summary for $T \in \mathcal{T}_F$. The weights show a progressive decrease over time over $\mathcal{T}_F$ representing that fewer earthquakes are predicted to occur as time progresses if the scenario of no future extraction holds. The values of $e_{\mathrm{wm}}(T \mid \mathcal{S}_F)$ also show a gradual reduction from $T = 2025$ as it nears $T = 2040$, after which the values of the weighted mean increase again. Figure C.1.2 provides clarification for this behaviour by showing spatial plots of the estimated endpoints across $\mathcal{X}$ for January 2025, the changes in the endpoint estimates when moving from January 2025 to January 2040, followed by the changes between Januray 2040 and January 2055. As the endpoint estimates are additive functions of the Kaiser stress, these plots also provide information on how the stresses change across $\mathcal{X}$ between these time points. Thus, insights can also be gleamed on how the intensity of earthquakes -

also dependent on the Kaiser stress and importantly, its temporal derivative - change through time. When moving from 2025-2040, there are large subregions of the gas field, particularly in the south-east, where the endpoint estimates change only very slightly (if at all), meaning that the Kaiser stress showed very little change in these areas. This would result in a near-zero temporal derivative, over the years 2025-2040, leading to reductions in the estimated intensity of earthquakes in these areas, and thus, reductions in the weights for these locations in $e_{\text{wm}}(T \mid \mathcal{S}_F)$. From 2040, the weights would be focussed mainly in the red regions in the centre panel in Figure C.1.2. This explains the gradual reduction in $e_{\text{wm}}(T \mid \mathcal{S}_F)$ over the years 2025-2040. Now, focussing on the changes from January 2040 - January 2055, we see changes in the endpoints occurring almost exclusively in these same areas. At this point, we have the weights resulting from the intensity estimates and the largest changes in the endpoint estimates occurring in the same subregions, which now leads to the increases we see in $e_{\text{wm}}(T \mid \mathcal{S}_F)$ from $T = 2040 - 2055$.

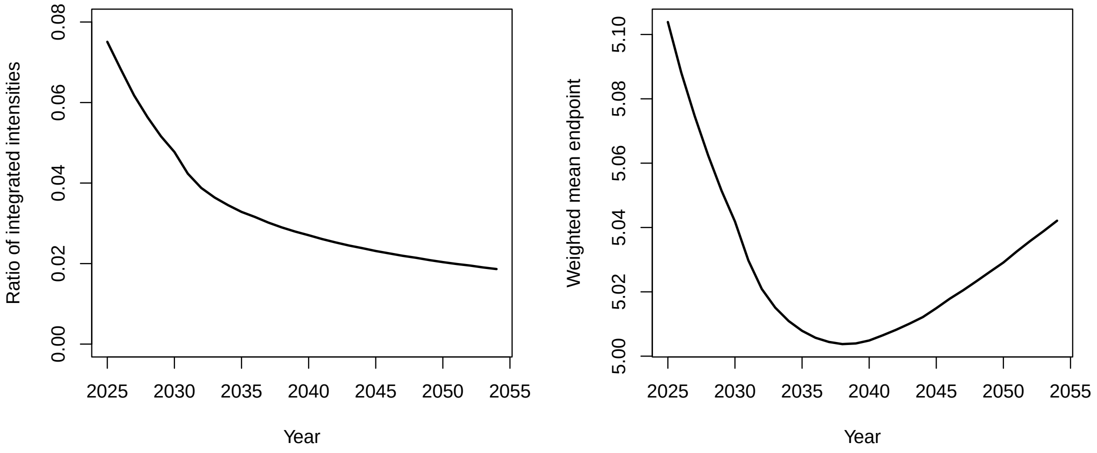

Figure C.1.1: Estimates of future earthquake properties for $T = 2025 - 2054$: [left] the probability mass function for an arbitrarily selected future earthquake's year T of occurrence, i.e., $\Gamma^{\mathcal{X}}(T)/\Gamma^{\mathcal{X}}(\mathcal{T}_F)$ and [right] weighted annual event per year with weights proportional to the spatial density of earthquake occurrences in that year, i.e., $e_{wm}(T|\mathcal{S}_F)$.

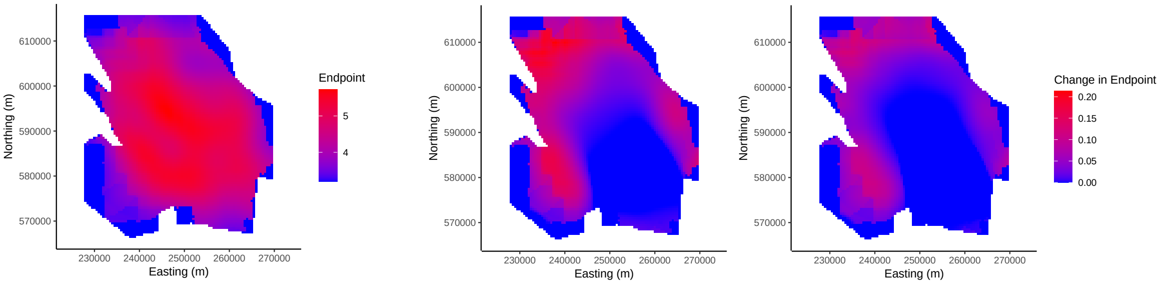

Figure C.1.2: Spatial plots of the temporal evolution of the estimated endpoints: [left] endpoint estimates of $e(\boldsymbol{x}, t|\mathcal{S}_F)$ for January 2025 and [centre and right] changes in endpoint estimates from January 2025- January 2040, and January 2040 - January 2055 respectively.

Appendix D

Supplementary materials to Chapter 6

D.1 Additional figures for Section 6.3

In this section, we present additional figures for Section 6.3, concerned with challenges C1 and C2. Figures D.1.1-D.1.3 support the exploratory analysis for challenges C1 and C2. We explore the within-year seasonality of the response variable Y in Figure D.1.1, looking at the distribution of Y per month and across the two seasons. This shows that there is a significant difference in the distribution of Y between seasons 1 and 2, but within each season there is little difference across months.

Figure D.1.2 shows a scatter plot of Y against each covariate $V_1, \ldots, V_8$, with quantile regression lines plotted in blue at probabilities $0.5, 0.6, \ldots, 0.9$ excluding V_6 which corresponds to season. Covariates V_1, V_2, V_4 and V_8 do not show any clear relationship with Y at any quantile level. However, Y shows dependence with the remaining covariates. In particular, the observed relationship with V_3 appears complex and non-linear across all quantiles. There is also evidence of relationships between Y and both V_6 (wind speed) and V_7 (wind direction). V_6 appears to show a somewhat linear relation-

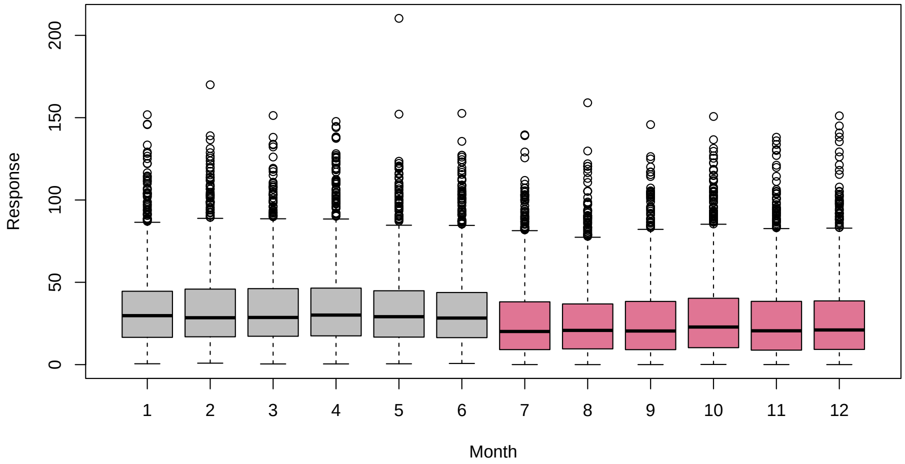

Figure D.1.1: Box plots of the response variable Y with each month and season (season 1 in grey and season 2 in red).

ship at all quantile levels, although we choose to allow more flexibility by using a spline on V_6 in our models. Y shows evidence of sinusoidal variation with V_7 which could be incorporated in our models, however, due to the dependence between wind speed and direction and with the goal of keeping our models parsimonious, we chose to omit this covariate from the analysis.

We also explore temporal dependence in Figure D.1.3 that details the auto-correlation function (acf) values for the response Y and explanatory variables $V_1, \ldots, V_4, V_6, \ldots, V_8$, up to a lag of 60. All variables have negligible acf values across all lags, except V_6 (wind speed), V_7 (wind direction) and V_8 (atmosphere). Covariates V_6 and V_7 show moderately strong temporal dependence across all lags while V_8 shows very strong correlation at early lags which gradually diminishes with increasing lag to negligible values at the largest lags.

Figure D.1.4 shows the QQ-plots corresponding to a standard GPD model fitted to the excesses of Y above a constant (left) and seasonally-varying threshold (right). 95% tolerance bounds (grey) show a lack of agreement between observations and the standard GPD model above a constant threshold. The second plot demonstrates a

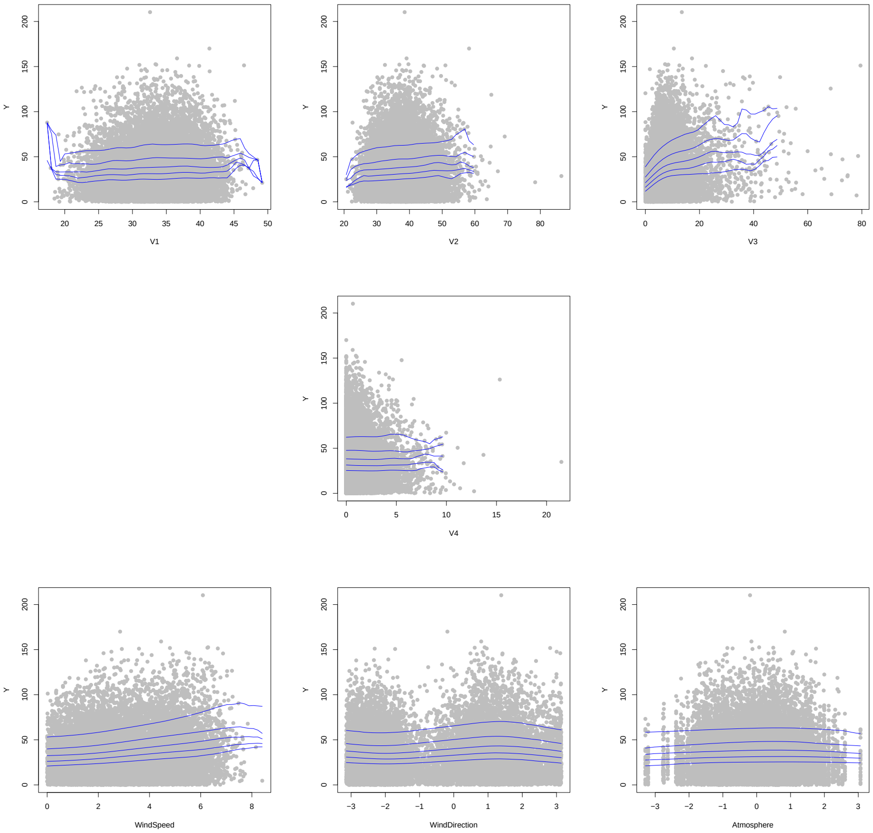

Figure D.1.2: Scatter plots of explanatory variables $V_1, \ldots, V_4$, wind speed (V_6), wind direction (V_7) and atmosphere (V_8), from top-left to bottom-right (by row), against the response variable Y. Quantile regression lines at probabilities $0.5, 0.6, \ldots, 0.9$ as blue lines.

significant improvement in model fit.

Figure D.1.5 shows a detailed summary of the pattern of missing data in the data and can be produced using the `missing_pattern` function in the `finalfit` package in `R` (Harrison et al., 2023). To interpret the figure note that blue and red squares represent observed and missing variables, respectively. The number on the right indicates the number of missing predictor variables (i.e., the number of red squares in the row), while the number on the left is the number of observations that fall into the row category. On the bottom, we have the number of observations that fall into the column category. For example, 18,545 observations are fully observed (denoted by the first row); there are 407 observations where only $V4$ is missing (denoted by the second row), 13 observations

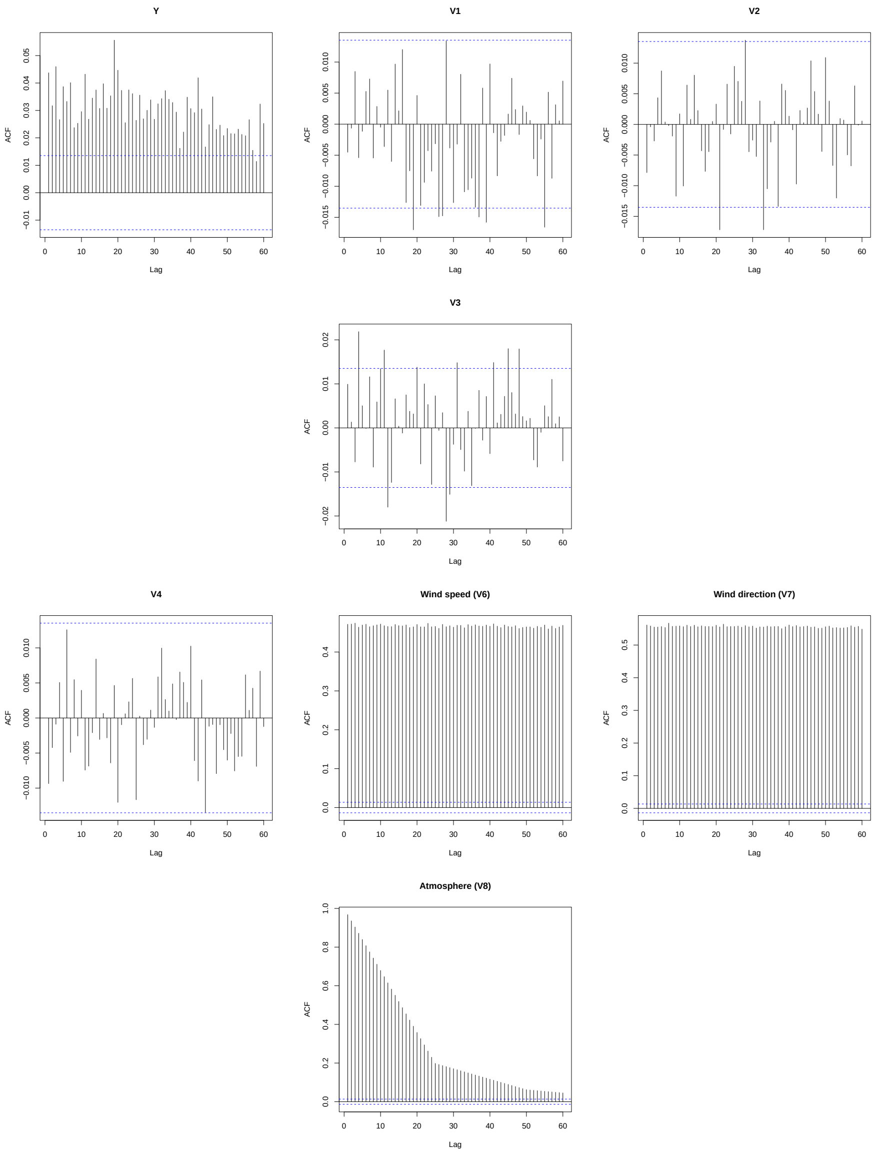

Figure D.1.3: Autocorrelation function plots (with lag 0 removed) for the response variable Y and explanatory variables $V1, \ldots, V4$, wind speed ($V6$), wind direction ($V7$) and atmosphere ($V8$), from top-left to bottom-right (by row).

where both $V4$ and $V6$ are missing (denoted by the fourth row), 456 observations where $V4$ and at least one other predictor is missing (denoted by the last column in the table), etc. There are very few observations where more than one predictor is missing.

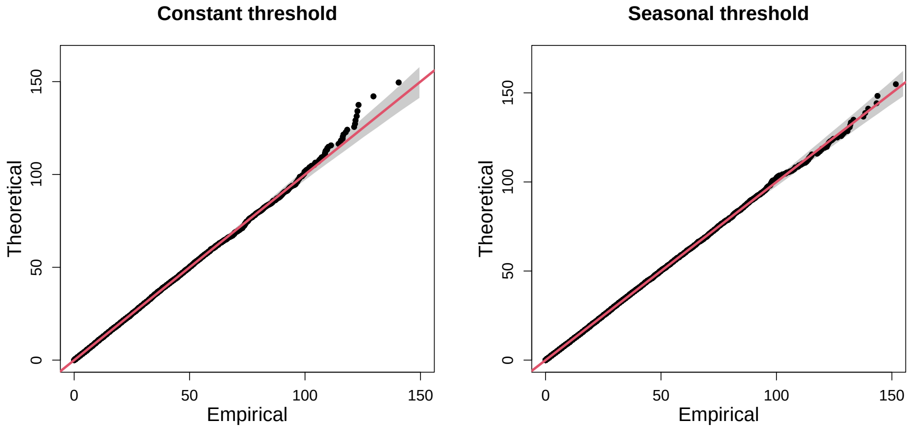

Figure D.1.4: QQ-plots showing standard GPD model fits with 95% tolerance bounds (grey) above a constant (left) and stepped-seasonal (right) threshold.

D.1.1 Additional figures for Section 6.4

In this section, we present additional plots related to Section 6.4. Figure D.1.6 illustrates the time series of both covariates for the first 3 years of the observation period. It can be seen how the seasons vary periodically over each year, as well as the discrete nature of the atmospheric covariate.

Bootstrapped χ estimates for the groups $G^A_{I,k}, k \in \{1, \dots, 10\}$, $I \in \mathcal{I} \setminus \{1, 2, 3\}$ and $G^S_{I,k}, k \in \{1, 2\}$, $I \in \mathcal{I}$ are given in Figures D.1.7 - D.1.10. These estimates illustrate the impact of atmosphere on the dependence structure.

Bootstrapped χ estimates for the groups $G^A_{I,k}, k \in \{1, \dots, 10\}$, $I \in \mathcal{I} \setminus \{1, 2, 3\}$ and $G^S_{I,k}, k \in \{1, 2\}$, $I \in \mathcal{I}$ are given in Figures D.1.7 - D.1.10. These estimates illustrate the impact of atmosphere on the dependence structure.

For a 3-dimensional random vector, the angular dependence function, denoted $\lambda(\cdot)$, is defined on the unit-simplex $\boldsymbol{S}^2$ and describes extremal dependence along different rays $\boldsymbol{\omega} \in \boldsymbol{S}^2$. As noted in Section 6.4.2, we can associate each of the probabilities from C3, p_1 and p_2, with points on $\boldsymbol{S}^2$, denoted $\boldsymbol{\omega}^1$ and $\boldsymbol{\omega}^2$ respectively. With $I = \{1, 2, 3\}$, we consider $\lambda(\boldsymbol{\omega}^1)$ and $\lambda(\boldsymbol{\omega}^2)$ over the subsets $G^S_{I,k},\ k \in \{1, 2\}$ and $G^A_{I,k},\ k \in \{1, \dots, 10\}$. We note that $\lambda(\boldsymbol{\omega}^1)$ is analogous with the coefficient of tail dependence $\eta \in (0, 1]$

(Ledford and Tawn, 1996), with $\eta = 1/3\lambda(\boldsymbol{\omega}^1)$; this corresponds with the region where all variables are simultaneously extreme. Furthermore, $\lambda(\boldsymbol{\omega}^2)$, which corresponds to a region where only two variables are extreme, is only evaluated after an additional marginal transformation of the third Coputopia time series; see Section 6.4.2.

Estimation of $\lambda(\cdot)$ for each simplex point and subset was achieved using the Hill estimator (Hill, 1975) at the 90% level, with uncertainty subsequently quantified via bootstrapping. The results shown in Figures D.1.11 - D.1.14 provide further evidence of a relationship between the extremal dependence structure and the covariates.

To illustrate the estimated trend in dependence, Figure D.1.15 shows the estimated scale functions, $\sigma\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right)$, over atmosphere for parts 1 and 2. Under the assumption of asymptotic normality in the spline coefficients, 95% confidence intervals are obtained via posterior sampling; see Wood (2017) for more details. We observe that σ tends to increase and decrease over atmosphere for parts 1 and 2, respectively, although the trend is less pronounced for the latter. Under our modelling framework, we note that higher values of σ are associated with less positive extremal dependence in the direction $\boldsymbol{\omega}$ of interest; to see this, observe that the survivor function of the GPD with fixed ξ is negatively associated with σ. Considering the trend in $\sigma\left(\boldsymbol{\omega}; \boldsymbol{x}_t\right)$, our results indicate a decrease in dependence in the region where all variables are extreme.

D.1.2 Additional figures for Section 6.5

In this section, we present additional plots related to Section 6.5 an we refer to p_1 and p_2 as parts 1 and 2 of C4, respectively. Figure D.1.16 shows a heat map of empirically estimated $\eta(\cdot)$ dependence coefficients and provides further evidence of the existence of the 5 dependence subgroups identified in our exploratory analysis for challenge C4. It also suggests that our modelling assumptions are reasonable; specifically that there is in-between group independence, and that the extremes within each group do not occur simultaneously.

Figure D.1.17 shows the bootstrapped estimated individual group and overall probabilities with respect to conditioning threshold quantile for part 1 of challenge C4. Similarly, Figure D.1.18 shows the bootstrapped estimated individual group and overall probabilities with respect to conditioning threshold quantile for part 2.

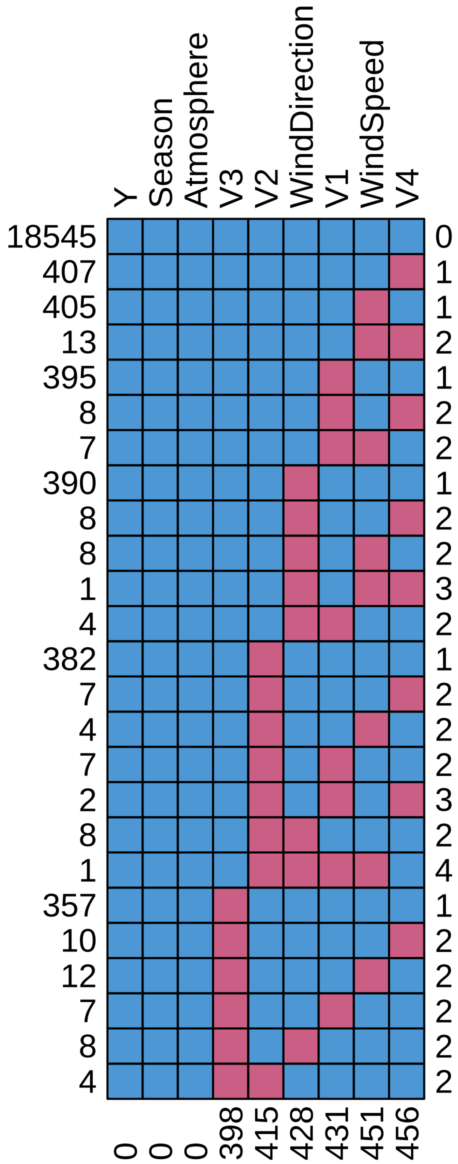

Figure D.1.5: Detailed pattern of missing predictor variables in the Amaurot data set.

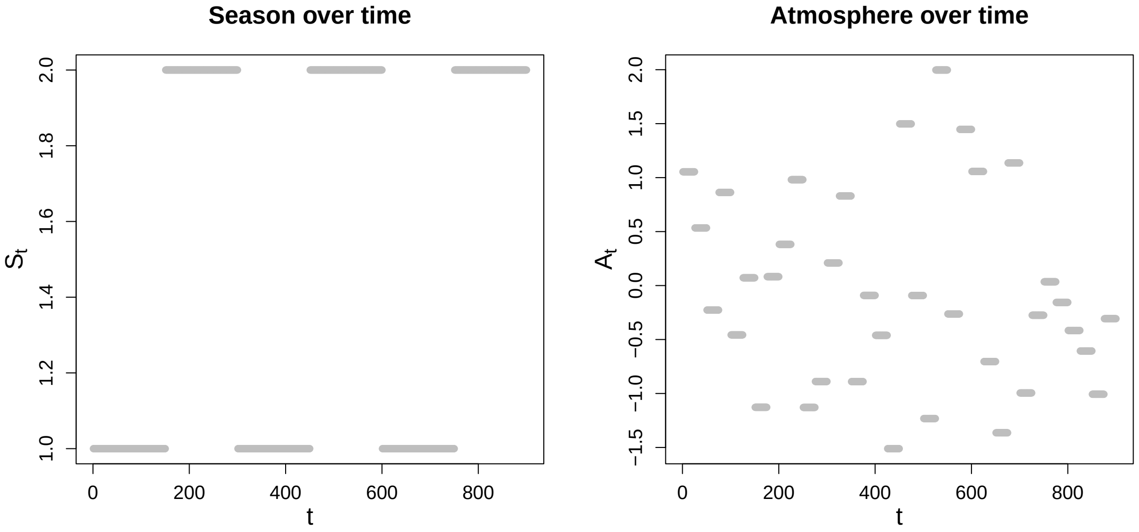

Figure D.1.6: Plots of S_t (left) and A_t (right) against t for the first 3 years of the observation period.

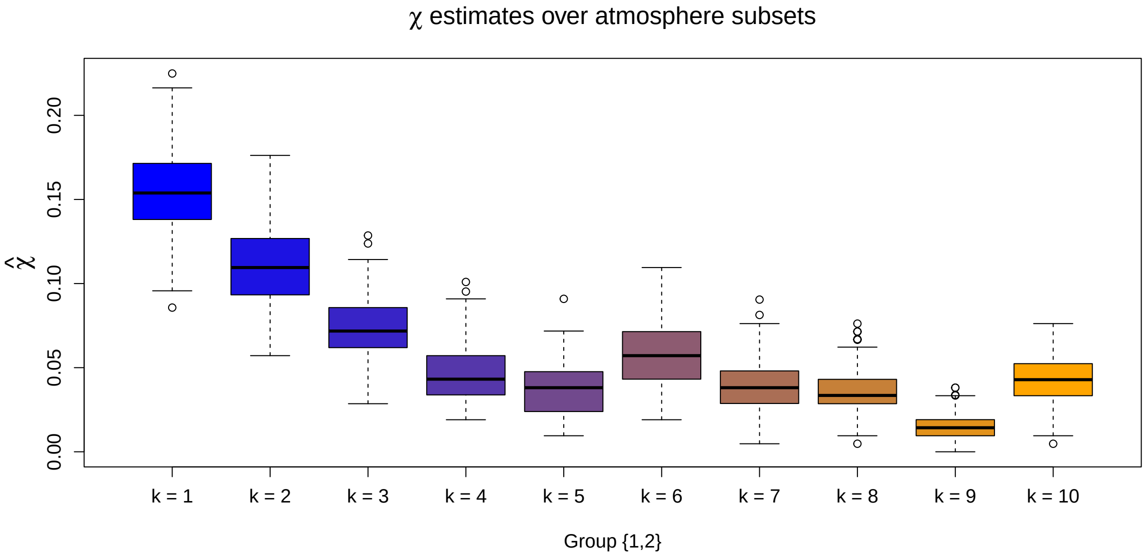

Figure D.1.7: Boxplots of empirical χ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2\}$. The colour transition (from blue to orange) over k illustrates the trend in χ estimates as the atmospheric values are increased.

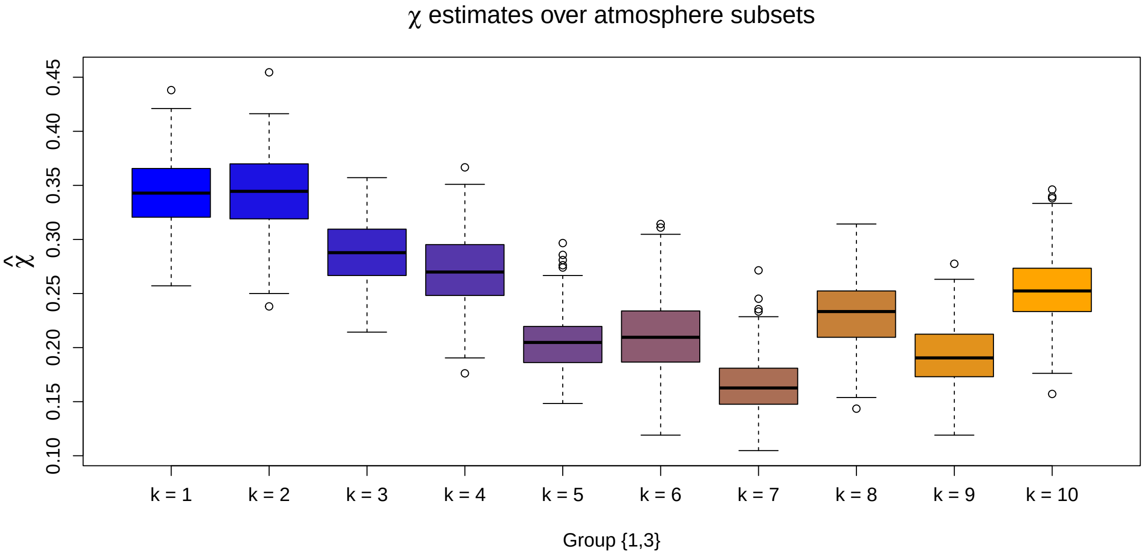

Figure D.1.8: Boxplots of empirical χ estimates obtained for the subsets $G^{A}_{I,k}$, with $k = 1, \dots, 10$ and $I = \{1, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in χ estimates as the atmospheric values are increased.

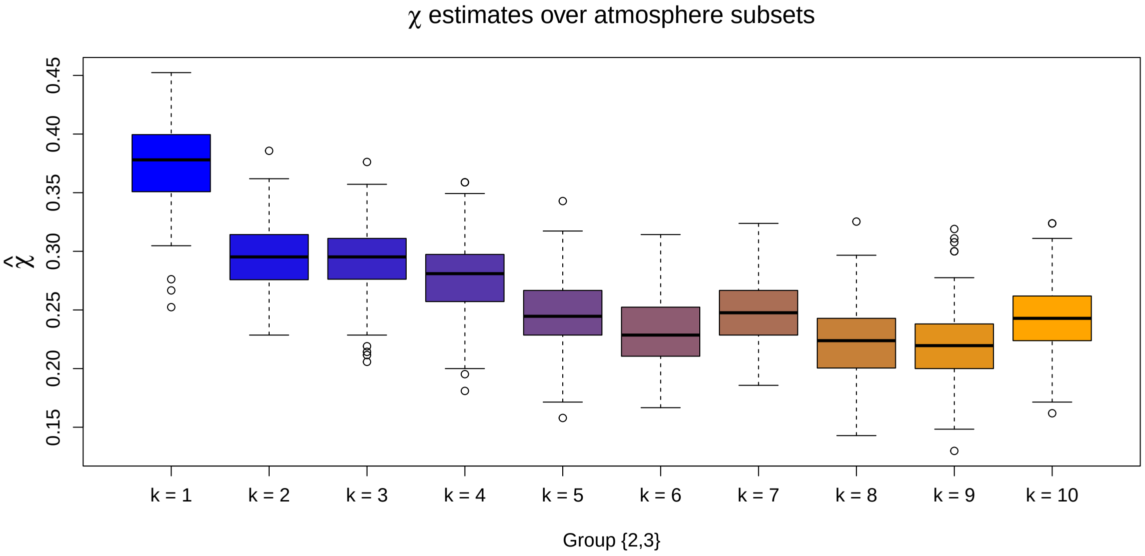

Figure D.1.9: Boxplots of empirical χ estimates obtained for the subsets $G^{A}_{I,k}$, with $k = 1, \dots, 10$ and $I = \{2, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in χ estimates as the atmospheric values are increased.

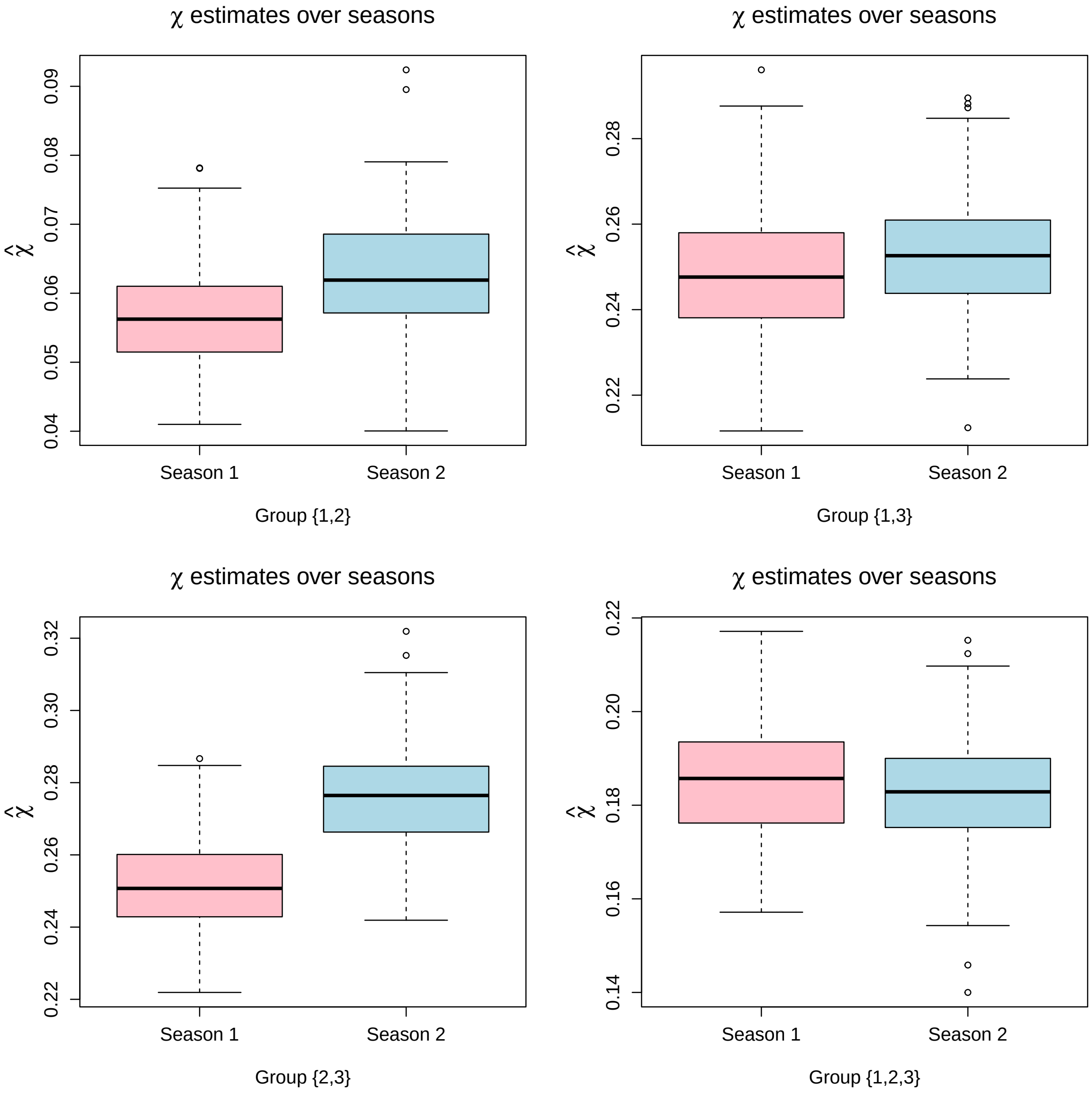

Figure D.1.10: Boxplots of empirical χ estimates obtained for the subsets $G^S_{I,k}$, with $k = 1, 2$. In each case, pink and blue colours illustrate estimates for seasons 1 and 2, respectively. From top left to bottom right: $I = \{1, 2, 3\}$, $I = \{1, 2\}$, $I = \{1, 3\}$, $I = \{2, 3\}$.

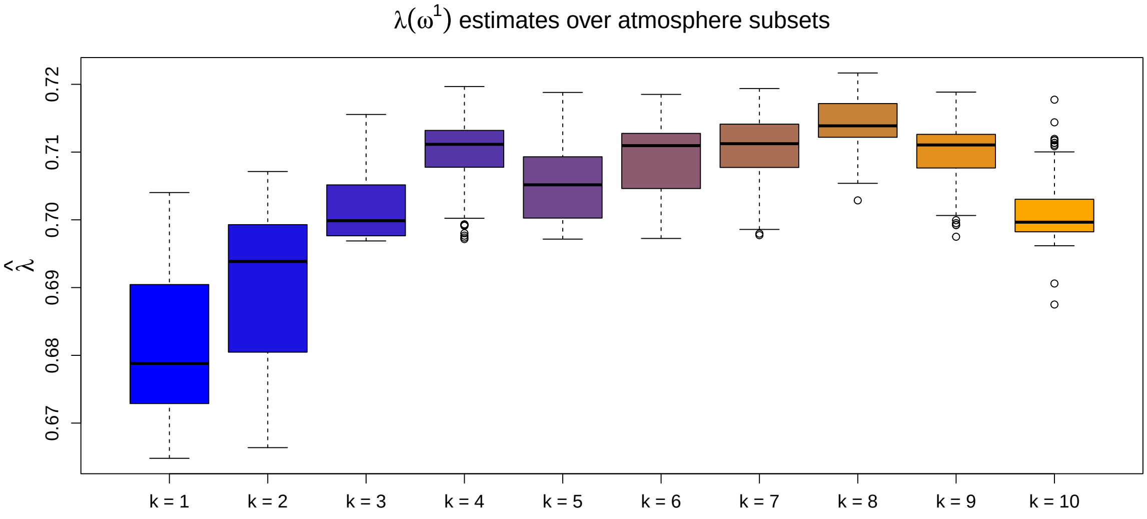

Figure D.1.11: Boxplots of empirical $\lambda(\boldsymbol{\omega}^1)$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in λ estimates as the atmospheric values are increased.

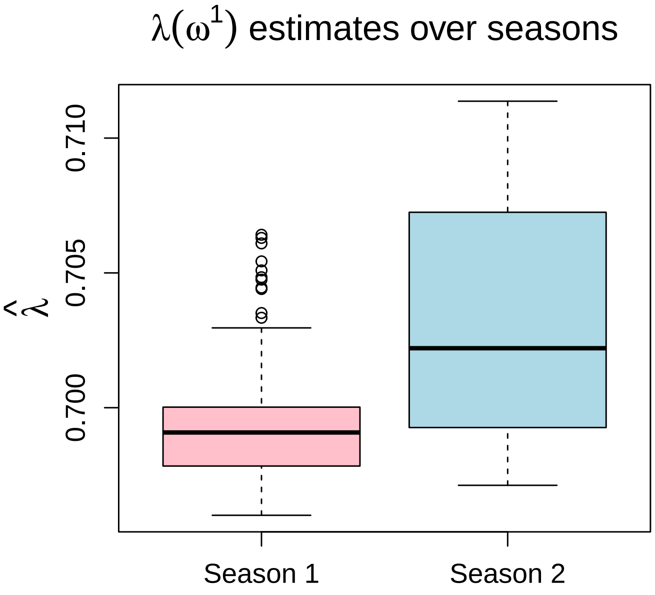

Figure D.1.12: Boxplots of empirical $\lambda(\boldsymbol{\omega}^1)$ estimates obtained for the subsets $G^S_{I,k}$, with $k = 1, 2$ and $I = \{1, 2, 3\}$. In each case, pink and blue colours illustrate estimates for seasons 1 and 2, respectively.

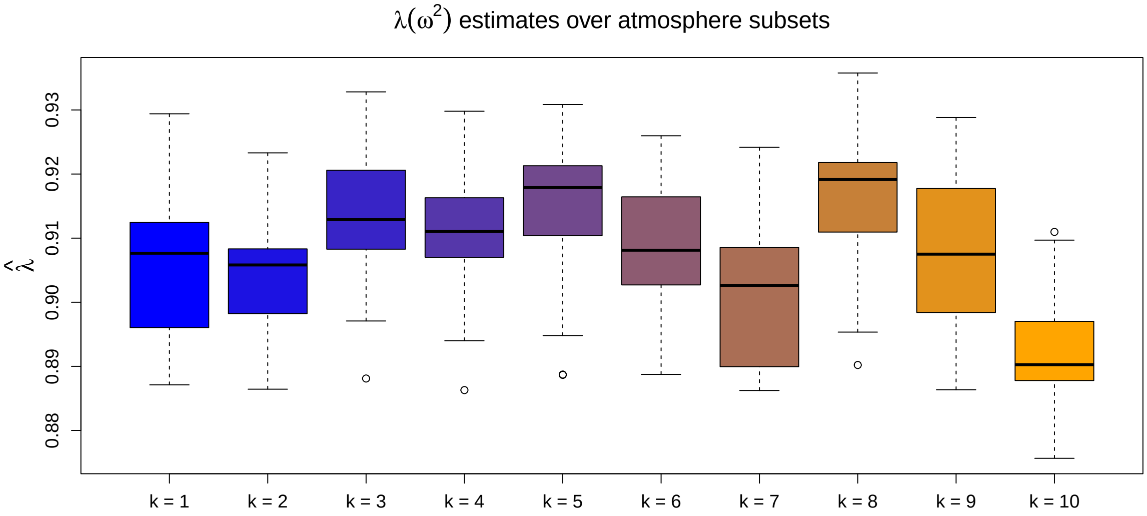

Figure D.1.13: Boxplots of empirical $\lambda(\boldsymbol{\omega}^2)$ estimates obtained for the subsets $G^A_{I,k}$, with $k = 1, \ldots, 10$ and $I = \{1, 2, 3\}$. The colour transition (from blue to orange) over k illustrates the trend in λ estimates as the atmospheric values are increased.

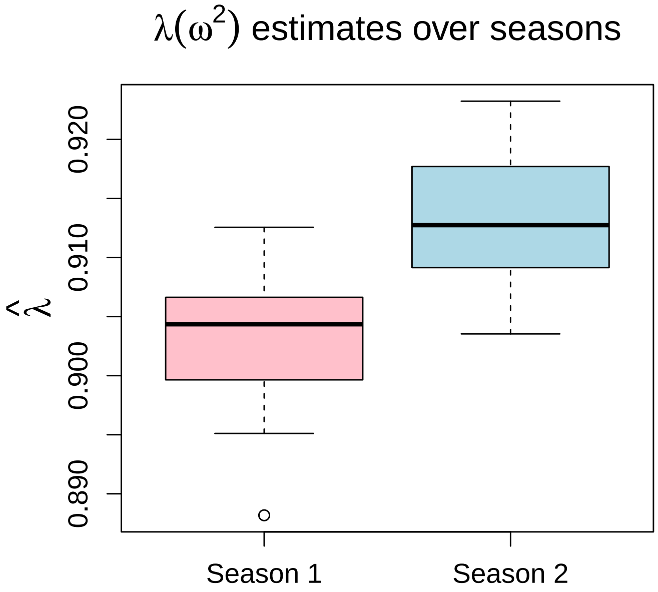

Figure D.1.14: Boxplots of empirical $\lambda(\boldsymbol{\omega}^2)$ estimates obtained for the subsets $G^S_{I,k}$, with $k = 1, 2$ and $I = \{1, 2, 3\}$. In each case, pink and blue colours illustrate estimates for seasons 1 and 2, respectively.

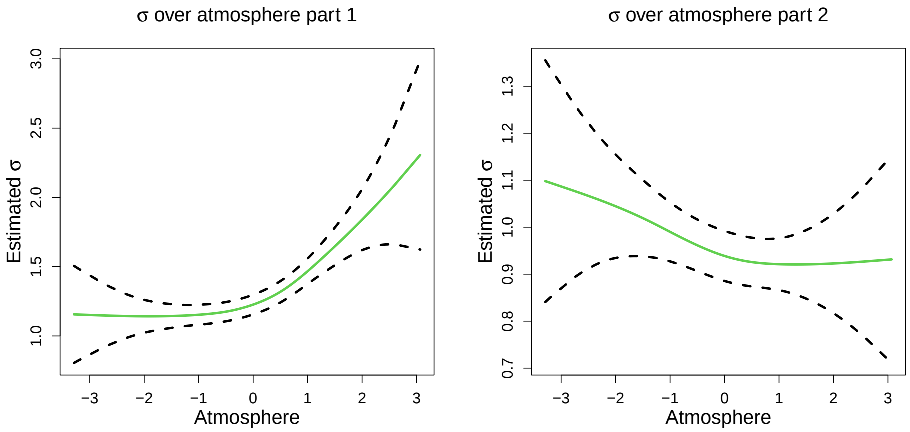

Figure D.1.15: Estimated σ functions (green) over atmosphere for part 1 (left) and 2 (right). In both cases, the regions defined by the black dotted lines represent 95% confidence intervals obtained using posterior sampling.

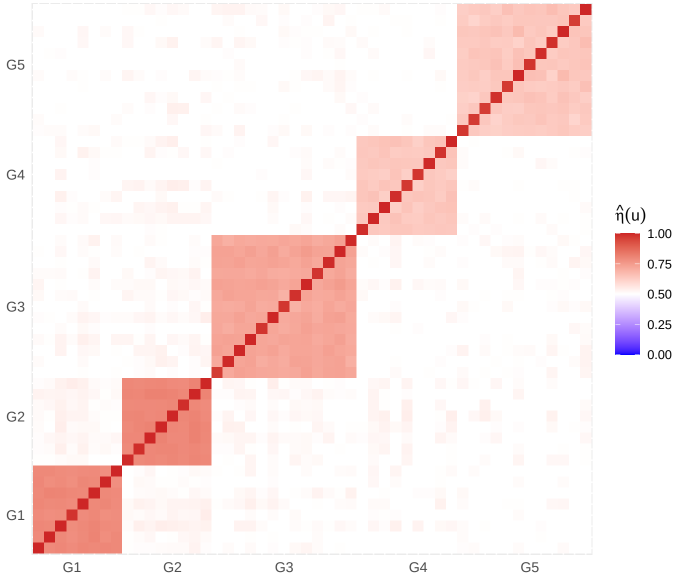

Figure D.1.16: Heat map of estimated empirical pairwise $\eta(u)$ extremal dependence coefficients with $u = 0.95$.

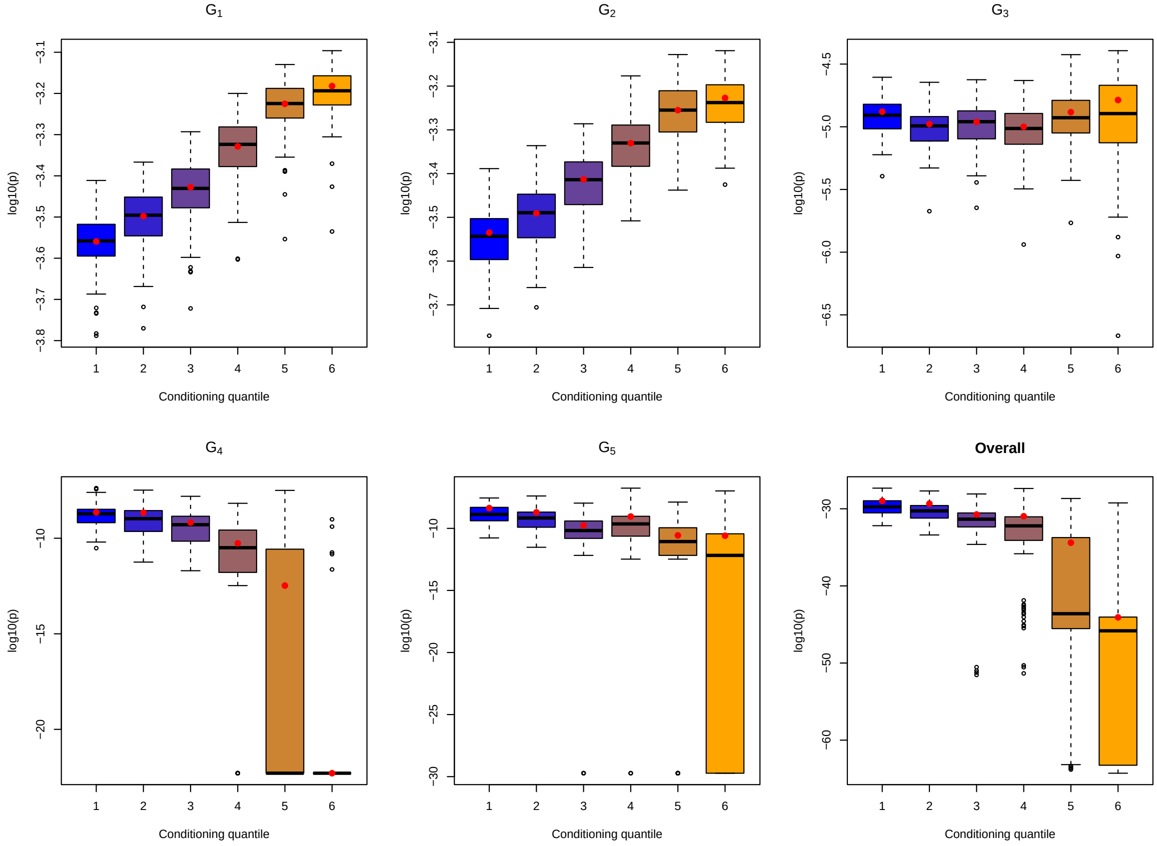

Figure D.1.17: Part 1 subgroup and overall bootstrapped probability estimates on the log scale. The red points indicate the original sample estimates and the colouring of the boxplots indicates the choice of conditioning threshold, with the conditioning quantile indices 1-6 referring to the quantile levels $\{0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$, respectively.

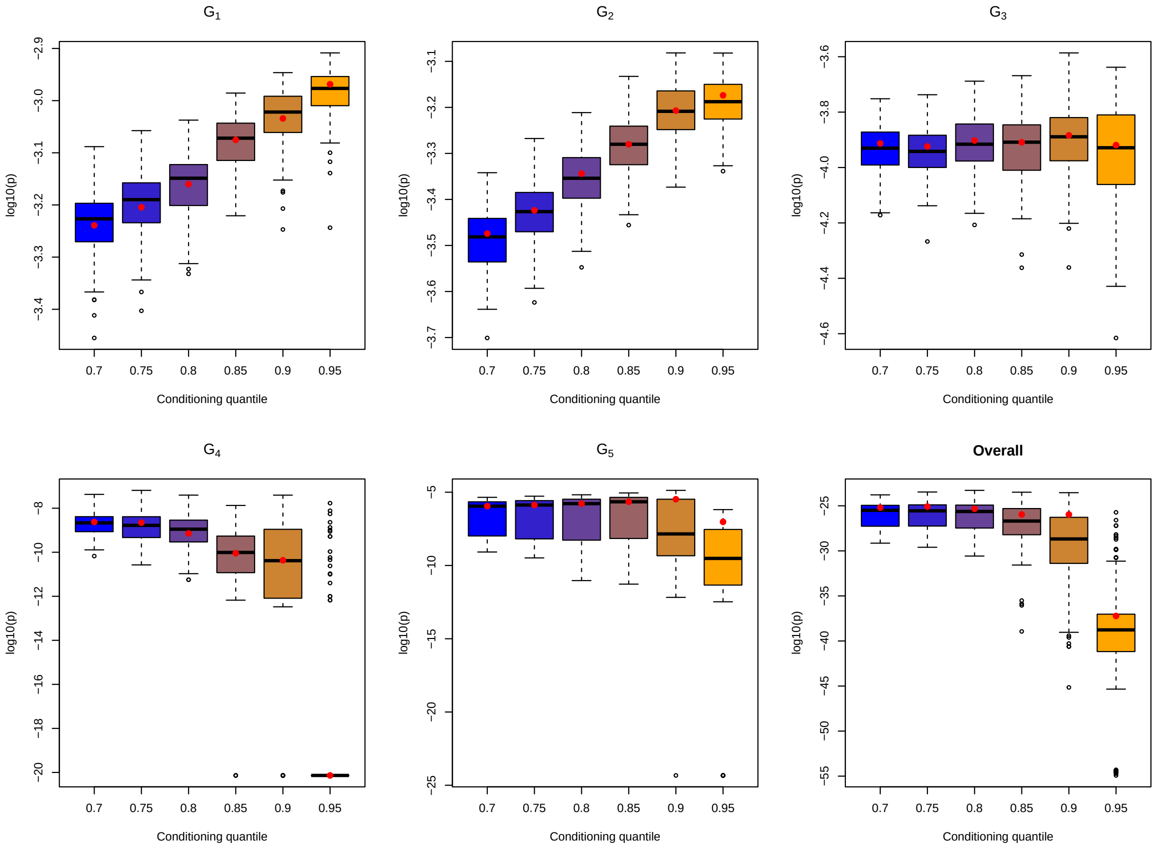

Figure D.1.18: Part 2 subgroup and overall bootstrapped probability estimates on the log scale for C4. The red points indicate the original sample estimates and the colouring of the boxplots indicates the choice of conditioning threshold, with the conditioning quantile indices 1-6 referring to the quantile levels $\{0.7, 0.75, 0.8, 0.85, 0.9, 0.95\}$, respectively.

Bibliography

Amin, M. (1982). On analysis and forecasting of surges on the West Coast of Great Britain. *Geophysical Journal International*, 68(1):79–94.

André, L. M., Campbell, R., D'Arcy, E., Farrell, A., Healy, D., Kakampakou, L., Murphy, C., Murphy-Barltrop, C. J. R., and Speers, M. (2025). Extreme value methods for estimating rare events in Utopia: EVA (2023) conference data challenge: team Lancopula Utopiversity. *Extremes*, 28:23–45.

Bader, B., Yan, J., and Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *Annals of Applied Statistics*, 12(1):310–329.

Baker, J., Bradley, B., and Stafford, P. (2021). *Probabilistic Seismic Hazard and Risk Analysis*. Cambridge University Press.

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2:792–804.

Bauer, R. A., Will, R., Greenberg, S. E., Whittaker, S. G., Davis, T., Landrø, M., Wilson, M., et al. (2019). *Geophysics and Geosequestration*, chapter Illinois basin–Decatur Project, pages 339–369. Cambridge University Press Cambridge, United Kingdom.

Behrens, C. N., Lopes, H. F., and Gamerman, D. (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227–244.

Beirlant, J., Kijko, A., Reynkens, T., and Einmahl, J. H. (2019). Estimating the maximum possible earthquake magnitude using extreme value methodology: the Groningen case. *Natural Hazards*, 98(3):1091–1113.

Belzile, L., Dutang, C., Northrop, P., and Opitz, T. (2023). A modeler's guide to extreme value software. *Extremes*, 26:1–44.

Bernard, E., Naveau, P., Vrac, M., and Mestre, O. (2013). Clustering of maxima: Spatial dependencies among heavy rainfall in France. *Journal of Climate*, 26:7929–7937.

Bommer, J. J., Dost, B., Edwards, B., Kruiver, P. P., Ntinalexis, M., Rodriguez-Marek, A., Stafford, P. J., and Van Elk, J. (2017). Developing a model for the prediction of ground motions due to earthquakes in the Groningen gas field. *Netherlands Journal of Geosciences*, 96(5):s203–s213.

Bommer, J. J. and Stafford, P. J. (2016). Seismic hazard and earthquake actions. In *Seismic Design of Buildings to Eurocode 8*, pages 21–54. CRC Press.

Bourne, S. and Oates, S. (2017). Extreme threshold failures within a heterogeneous elastic thin sheet and the spatial-temporal development of induced seismicity within the Groningen gas field. *Journal of Geophysical Research: Solid Earth*, 122(12):10–299.

Bourne, S., Oates, S., and Van Elk, J. (2018). The exponential rise of induced seismicity with increasing stress levels in the Groningen gas field and its implications for controlling seismic risk. *Geophysical Journal International*, 213(3):1693–1700.

Bourne, S. J. and Oates, S. (2020). Stress-dependent magnitudes of induced earthquakes in the Groningen gas field. *Journal of Geophysical Research: Solid Earth*, 125(11):e2020JB020013.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Caballero-Megido, C., Hillier, J., Wyncoll, D., Bosher, L., and Gouldby, B. (2018). Technical note: comparison of methods for threshold selection for extreme sea levels. *Journal of Flood Risk Management*, 11(2):127–140.

Chatfield, C. (2013). *The Analysis of Time Series: Theory and Practice*. Springer.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C*, 54(1):207–222.

Chen, Q., Wang, L., and Tawes, R. (2008). Hydrodynamic response of Northeastern Gulf of Mexico to hurricanes. *Estuaries and Coasts*, 31:1098–1116.

Choulakian, V. and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, 43(4):478–484.

Code, P. (2005). Eurocode 8: Design of structures for earthquake resistance-part 1: general rules, seismic actions and rules for buildings. *Brussels: European Committee for Standardization*, 10.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, New York.

Coles, S. G. and Pericchi, L. R. (2003). Anticipating catastrophes through extreme value modelling. *Journal of the Royal Statistical Society: Series C*, 52(4):405–416.

Coles, S. G., Pericchi, L. R., and Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, 273(1-4):35–50.

Coles, S. G. and Tawn, J. A. (1994). Statistical methods for multivariate extremes: An application to structural design. *Applied Statistics*, 43:1–48.

Coles, S. G. and Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45(4):463–478.

Collings, T. P., Murphy-Barltrop, C. J. R., Murphy, C., Haigh, I. D., Bates, P. D., and Quinn, N. D. (2025). Automated tail-informed threshold selection for extreme coastal sea levels. *Submitted*.

Collings, T. P., Quinn, N. D., Haigh, I. D., Green, J., Probyn, I., Wilkinson, H., Muis, S., Sweet, W. V., and Bates, P. D. (2024). Global application of a regional frequency analysis to extreme sea levels. *Natural Hazards and Earth System Sciences*, 24(7):2403–2423.

Curceac, S., Atkinson, P. M., Milne, A., Wu, L., and Harris, P. (2020). An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales. *Journal of Hydrology*, 585:124845.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 76(2):226–248.

Danielsson, J., Ergun, L., de Haan, L., and de Vries, C. G. (2019). Tail index estimation: quantile-driven threshold selection. Staff Working Papers 19-28, Bank of Canada.

Das, R., Wason, H., and Sharma, M. L. (2012). Temporal and spatial variations in the magnitude of completeness for homogenized moment magnitude catalogue for northeast india. *Journal of Earth System Science*, 121:19–28.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society: Series B*, 52(3):393–425.

Demuth, A., Ottemöller, L., and Keers, H. (2016). Ambient noise levels and detection threshold in norway. *Journal of Seismology*, 20:889–904.

Dey, D. K. and Yan, J. (2016). *Extreme Value Modeling and Risk Analysis: Methods and Applications*. CRC Press.

Dong, X., Zhang, S., Zhou, J., Cao, J., Jiao, L., Zhang, Z., and Liu, Y. (2019). Magnitude and frequency of temperature and precipitation extremes and the associated atmospheric circulation patterns in the Yellow River basin (1960–2017), China. *Water*, 11(11):2334.

Dost, B. and Kraaijpoel, D. (2013). The August 16, 2012 earthquake near Huizinge (Groningen). *KNMI, de Bilt, the Netherlands*.

Dupuis, D. (1999). Exceedances over high thresholds: A guide to threshold selection. *Extremes*, 1:251–261.

Durocher, M., Mostofi Zadeh, S., Burn, D. H., and Ashkar, F. (2018). Comparison of automatic procedures for selecting flood peaks over threshold based on goodness-of-fit tests. *Hydrological Processes*, 32(18):2874–2887.

D'Arcy, E., Tawn, J. A., Joly, A., and Sifnioti, D. E. (2023). Accounting for seasonality in extreme sea-level estimation. *The Annals of Applied Statistics*, 17:3500–3525.

Eastoe, E. F. and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):25–45.

Eastoe, E. F. and Tawn, J. A. (2012). Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika*, 99(1):43–55.

Ellsworth, W. L. (2013). Injection-induced earthquakes. *Science*, 341(6142):1225942.

Fawcett, L. and Walshaw, D. (2006). A hierarchical model for extreme wind speeds. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 55(5):631–646.

Fawcett, L. and Walshaw, D. (2007). Improved estimation for temporally clustered extremes. *Environmetrics*, 18(2):173–188.

Ferro, C. A. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):545–556.

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

Freudenreich, Y., Oates, S. J., and Berlang, W. (2012). Microseismic feasibility studies–assessing the probability of success of monitoring projects. *Geophysical Prospecting*, 60(6):1043–1053.

Fyodorov, Y. V. and Bouchaud, J.-P. (2008). Freezing and extreme-value statistics in a random energy model with logarithmically correlated potential. *Journal of Physics A: Mathematical and Theoretical*, 41(37):372001.

Galis, M., Ampuero, J. P., Mai, P. M., and Cappa, F. (2017). Induced seismicity provides insight into why earthquake ruptures stop. *Science Advances*, 3(12):eaap7528.

Gaucher, E. (2016). Earthquake detection probability within a seismically quiet area: application to the Bruchsal geothermal field. *Geophysical Prospecting*, 64(2):268–286.

Gelfand, A. E. (1996). Model determination using sampling-based methods. *Markov Chain Monte Carlo in Practice*, 4:145–161.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Goertz, A., Riahi, N., Kraft, T., and Lambert, M. (2012). Modeling detection thresholds of microseismic monitoring networks. In *SEG International Exposition and Annual Meeting*, pages SEG–2012. SEG.

Gomes, M. I. (1994). Penultimate behaviour of the extremes. *Extreme Value Theory and Applications: Proceedings of the Conference on Extreme Value Theory and Applications, Gaithersburg Maryland 1993*, 1:403–418.

Gomes, M. I. and Guillou, A. (2015). Extreme value theory and statistics of univariate extremes: A review. *International Statistical Review*, 83:263–292.

Guerrero, M. B., Huser, R., and Ombao, H. (2023). Conex–Connect: Learning patterns in extremal brain connectivity from multichannel EEG data. *The Annals of Applied Statistics*, 17:178–198.

Gutenberg, B. and Richter, C. F. (1956). Earthquake magnitude, intensity, energy, and acceleration: (second paper). *Bulletin of the Seismological Society of America*, 46(2):105–145.

Haigh, I. D., Marcos, M., Talke, S. A., Woodworth, P. L., Hunter, J. R., Hague, B. S., Arns, A., Bradshaw, E., and Thompson, P. (2023). GESLA version 3: A major update to the global higher-frequency sea-level dataset. *Geoscience Data Journal*, 10(3):293–314.

Haigh, I. D., Wadey, M. P., Wahl, T., Ozsoy, O., Nicholls, R. J., Brown, J. M., Horsburgh, K., and Gouldby, B. (2016). Spatial and temporal analysis of extreme sea level and storm surge events around the coastline of the UK. *Scientific Data*, 3(1):1–14.

Harrison, E., Drake, T., and Ots, R. (2023). *finalfit: Quickly Create Elegant Regression Results Tables and Plots when Modelling.* R package version 1.0.7.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer, New York.

Healy, D., Tawn, J. A., Thorne, P., and Parnell, A. (2025). Inference for extreme spatial temperature events in a changing climate with application to Ireland (with discussion). *Journal of the Royal Statistical Society: Series C: Applied Statistics*, 74(2):275–330.

Heffernan, J. E. and Tawn, J. A. (2003). An extreme value analysis for the investigation into the sinking of the MV Derbyshire. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 52(3):337–354.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B*, 66(3):497–546.

Hiles, C. E., Robertson, B., and Buckham, B. J. (2019). Extreme wave statistical methods and implications for coastal analyses. *Estuarine, Coastal and Shelf Science*, 223:50–60.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

Hutton, K., Woessner, J., and Hauksson, E. (2010). Earthquake monitoring in southern California for seventy-seven years (1932–2008). *Bulletin of the Seismological Society of America*, 100(2):423–446.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall/CRC, New York.

Jonathan, P. and Ewans, K. (2007a). The effect of directionality on extreme wave design criteria. *Ocean Engineering*, 34(14-15):1977–1994.

Jonathan, P. and Ewans, K. (2007b). Uncertainties in extreme wave height estimates for hurricane-dominated regions. *Journal of Offshore Mechanics and Arctic Engineering*, 129(4):300–305.

Jonathan, P., Randell, D., Wu, Y., and Ewans, K. (2014). Return level estimation from non-stationary spatial data exhibiting multidimensional covariate effects. *Ocean Engineering*, 88:520–532.

Jones, D. (1997). Plotting positions via maximum-likelihood for a non-standard situation. *Hydrology and Earth System Sciences*, 1(2):357–366.

Kagan, Y. Y. and Jackson, D. D. (2016). Earthquake rate and magnitude distributions of great earthquakes for use in global forecasts. *Geophysical Journal International*, 206(1):630–643.

Kaveh, H., Batlle, P., Acosta, M., Kulkarni, P., Bourne, S. J., and Avouac, J. P. (2024). Induced seismicity forecasting with uncertainty quantification: Application to the Groningen gas field. *Seismological Research Letters*, 95(2A):773–790.

Keef, C., Papastathopoulos, I., and Tawn, J. A. (2013a). Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *Journal of Multivariate Analysis*, 115:396–404.

Keef, C., Tawn, J. A., and Lamb, R. (2013b). Estimating the probability of widespread flood events. *Environmetrics*, 24:13–21.

KNMI (2020). Aardbevings catalogus. (https://www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus).

Koh, J., Pimont, F., Dupuy, J.-L., and Opitz, T. (2023). Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. *Annals of Applied Statistics*, 17(1):560–582.

Kyselý, J., Picek, J., and Beranová, R. (2010). Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold. *Global and Planetary Change*, 72:55–68.

Leadbetter, M. R. (1991). On a basis for 'peaks over threshold'modeling. *Statistics & Probability Letters*, 12(4):357–362.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (2012). *Extremes and Related Properties of Random Sequences and Processes*. Springer Science & Business Media, New York.

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Ledford, A. W. and Tawn, J. A. (2003). Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):521–543.

Li, Y., Cai, W., and Campbell, E. (2005). Statistical modeling of extreme rainfall in southwest Western Australia. *Journal of Climate*, 18(6):852–863.

Lin, Q. and Newberry, M. (2023). Seeing through noise in power laws. *Journal of the Royal Society Interface*, 20(205):20230310.

Mackay, E. and Jonathan, P. (2020). Assessment of return value estimates from stationary and non-stationary extreme value models. *Ocean Engineering*, 207:107406.

Majer, E. L., Baria, R., Stark, M., Oates, S., Bommer, J., Smith, B., and Asanuma, H. (2007). Induced seismicity associated with enhanced geothermal systems. *Geothermics*, 36(3):185–222.

McGarr, A. (2014). Maximum magnitude earthquakes induced by fluid injection. *Journal of Geophysical Research: Solid Earth*, 119(2):1008–1019.

Mhalla, L., de Carvalho, M., and Chavez-Demoulin, V. (2019). Regression-type models for extremal dependence. *Scandinavian Journal of Statistics*, 46(4):1141–1167.

Mignan, A., Werner, M., Wiemer, S., Chen, C., and Wu, Y. (2011). Bayesian estimation of the spatially varying completeness magnitude of earthquake catalogs. *Bulletin of the Seismological Society of America*, 101(3):1371–1385.

Mignan, A. and Woessner, J. (2012). Estimating the magnitude of completeness for earthquake catalogs. *Community Online Resource for Statistical Seismicity Analysis*.

Moftakhari, H. R., AghaKouchak, A., Sanders, B. F., Allaire, M., and Matthew, R. A. (2018). What is nuisance flooding? Defining and monitoring an emerging challenge. *Water Resources Research*, 54(7):4218–4227.

Murphy, C., Tawn, J. A., and Varty, Z. (2025). Automated threshold selection and associated inference uncertainty for univariate extremes. *Technometrics*, 67(2):215–224.

Murphy, C., Tawn, J. A., Varty, Z., and Towe, R. (2023). Software for threshold selection. `https://github.com/conor-murphy4/automated_threshold_selection`.

Murphy-Barltrop, C. (2024). ONR-RRR-079. Technical report, Office for Nuclear Regulation.

Murphy-Barltrop, C. and Wadsworth, J. (2024). Modelling non-stationarity in asymptotically independent extremes. *Computational Statistics & Data Analysis*, 199:108025.

Murphy-Barltrop, C. J. R., Mackay, E., and Jonathan, P. (2024). Inference for bivariate extremes via a semi-parametric angular-radial model. *arXiv*, 2401.07259.

NAM (2022). Report on the second workshop on mmax for seismic hazard and risk analysis in the Groningen gas field. https://nam-onderzoeksrapporten.data-app.nl/reports/download/groningen/en/77951661-552a-46bc-9f2e-f1580cd6abc3.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769.

Northrop, P. J. and Attalides, N. (2020). *threshr: Threshold Selection and Uncertainty for Extreme Value Analysis*. R package version 1.0.3.

Northrop, P. J., Attalides, N., and Jonathan, P. (2017). Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society: Series C*, 66(1):93–120.

Northrop, P. J. and Coleman, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes*, 17(2):289–303.

Northrop, P. J. and Jonathan, P. (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics*, 22(7):799–809.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27.

Olbert, A. I. and Hartnett, M. (2010). Storms and surges in Irish coastal waters. *Ocean Modelling*, 34(1):50–62.

Ossberger, J. (2020). *tea: Threshold Estimation Approaches*. R package version 1.1.

Pan, X. and Rahman, A. (2022). Comparison of annual maximum and peaks-over-threshold methods with automated threshold selection in flood frequency analysis: a case study for Australia. *Natural Hazards*, 111:1219—-1244.

Panakkat, A. and Adeli, H. (2007). Neural network models for earthquake magnitude prediction using multiple seismicity indicators. *International Journal of Neural Systems*, 17(01):13–33.

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *Journal of Applied Probability*, 8(4):745–756.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89:1303–1313.

Powell, E. (2024a). Hurricane Helene post-storm summary report. https://climatecenter.fsu.edu/images/docs/Hurricane-Helene-Summary-Report.pdf. Accessed: 07/01/2025.

Powell, E. (2024b). Post-storm summary report on Hurricane Milton. https://climatecenter.fsu.edu/images/docs/Hurricane-Helene-Summary-Report.pdf. Accessed: 07/01/2025.

Quinn, N., Bates, P. D., Neal, J., Smith, A., Wing, O., Sampson, C., Smith, J., and Heffernan, J. (2019). The spatial dependence of flood hazard and risk in the United States. *Water Resources Research*, 55:1890–1911.

Raschke, M. (2015). Modeling of magnitude distributions by the generalized truncated exponential distribution. *Journal of Seismology*, 19(1):265–271.

Reiss, R.-D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values*. Springer, Second edition.

Resnick, S. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5:303–336.

Richards, J., Tawn, J. A., and Brown, S. (2022). Modelling extremes of spatial aggregates of precipitation using conditional methods. *The Annals of Applied Statistics*, 16(4):2693–2713.

Richter, G., Hainzl, S., Dahm, T., and Zöller, G. (2020). Stress-based, statistical modeling of the induced seismicity at the Groningen gas field, the Netherlands. *Environmental Earth Sciences*, 79:1–15.

Rohrbeck, C., Simpson, E. S., and Tawn, J. A. (2023). Editorial: EVA (2023) Conference Data Challenge. *Extremes*, 28.

Ross, G. (2016). Bayesian estimation of the ETAS model for earthquake occurrences. *Technical Report*.

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60.

Schlather, M. and Tawn, J. A. (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90(1):139–156.

Shaby, B. A. and Reich, B. J. (2012). Bayesian spatial extreme value analysis to assess the changing risk of concurrent high temperatures across large portions of European cropland. *Environmetrics*, 23(8):638–648.

Sharkey, P. and Tawn, J. A. (2017). A Poisson process reparameterisation for Bayesian inference for extremes. *Extremes*, 20(2):239–263.

Sigauke, C. and Bere, A. (2017). Modelling non-stationary time series using a peaks over threshold distribution with time varying covariates and threshold: An application to peak electricity demand. *Energy*, 119:152–166.

Simpson, E. S., Wadsworth, J. L., and Tawn, J. A. (2020). Determining the dependence structure of multivariate extremes. *Biometrika*, 107:513–532.

Sinclair, C., Spurr, B., and Ahmad, M. (1990). Modified Anderson Darling test. *Communications in Statistics - Theory and Methods*, 19(10):3677–3686.

Smith, J. D., Heimisson, E. R., Bourne, S. J., and Avouac, J.-P. (2022). Stress-based forecasting of induced seismicity with instantaneous earthquake failure functions: Applications to the Groningen gas reservoir. *Earth and Planetary Science Letters*, 594:117697.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–90.

Smith, R. L. (1987). Approximations in extreme value theory. Technical report, No. 205, Department of Statistics, Univeristy of North Carolina.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4(4):367–377.

Smith, R. L. (2003). Statistics of extremes, with applications in environment, insurance, and finance. In *Extreme Values in Finance, Telecommunications, and the Environment*, edited by Finkenstadt, B. and Rootzén, H., pages 20–97. Chapman and Hall/CRC.

Smith, R. L., Tawn, J. A., and Coles, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268.

Smith, R. L. and Weissman, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):515–528.

Solari, S., Egüen, M., Polo, M. J., and Losada, M. A. (2017). Peaks over threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value. *Water Resources Research*, 53(4):2833–2849.

Spearing, J., Tawn, J., Irons, D., and Paulden, T. (2023). A framework for statistical modelling of the extremes of longitudinal data, applied to elite swimming. *arXiv preprint arXiv:2306.12419*.

Stange, S. (2006). M_L determination for local and regional events using a sparse network in southwestern Germany. *Journal of Seismology*, 10:247–257.

Suckale, J. (2009). Chapter 2 - Induced seismicity in hydrocarbon fields. In *Advances in Geophysics*, volume 51, pages 55–106. Elsevier.

Sweet, W. V., Dusek, G., Obeysekera, J., and Marra, J. J. (2018). Patterns and projections of high tide flooding along the U.S. coastline using a common impact threshold. Technical report, National Oceanic and Atmospheric Administration. Accessed: 10/03/2025.

Sweet, W. V., Genz, A. S., Obeysekera, J., and Marra, J. J. (2020). A regional frequency analysis of tide gauges to assess Pacific Coast flood risk. *Frontiers in Marine Science*, 7:1–15.

Tancredi, A., Anderson, C. W., and O'Hagan, A. (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes*, 9(2):87–106.

Tang, C. and Hudson, J. A. (2010). *Rock Failure Mechanisms: Illustrated and Explained*. CRC Press.

Turkman, K. F., Turkman, M. A. A., and Pereira, J. M. (2010). Asymptotic models and inference for extremes of spatio-temporal data. *Extremes*, 13:375–397.

Varty, Z., Tawn, J. A., Atkinson, P. M., and Bierman, S. (2021). Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. *arXiv:2102.00884*.

Vere-Jones, D., Robinson, R., and Yang, W. (2001). Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation. *Geophysical Journal International*, 144(3):517–531.

Wadsworth, J., Tawn, J., and Jonathan, P. (2010). Accounting for choice of measurement scale in extreme value modeling. *Annals of Applied Statistics*, 4:1558–1578.

Wadsworth, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics*, 58(1):116–126.

Wadsworth, J. L. and Tawn, J. A. (2012). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *Journal of the Royal Statistical Society: Series B*, 74(3):543–567.

Wadsworth, J. L. and Tawn, J. A. (2013). A new representation for multivariate tail probabilities. *Bernoulli*, 19:2689–2714.

Walshaw, D. (1994). Getting the most from your extreme wind data: a step by step guide. *Journal of Research of the National Institute of Standards and Technology*, 99(4):399.

Wan, P. and Davis, R. A. (2019). Threshold selection for multivariate heavy-tailed data. *Extremes*, 22(1):131–166.

Weng, H., Ampuero, J.-P., and Buijze, L. (2021). Physics-based estimates of the maximum magnitude of induced earthquakes in the groningen gas field. In *EGU General Assembly Conference Abstracts*, pages EGU21–6144.

Wing, O. E., Quinn, N., Bates, P. D., Neal, J. C., Smith, A. M., Sampson, C. C., Coxon, G., Yamazaki, D., Sutanudjaja, E. H., and Alfieri, L. (2020). Toward global stochastic river flood modeling. *Water Resources Research*, 56:e2020WR027692.

Winter, H. C. and Tawn, J. A. (2017). k th-order Markov extremal models for assessing heatwave risks. *Extremes*, 20:393–415.

Woessner, J. and Wiemer, S. (2005). Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty. *Bulletin of the Seismological Society of America*, 95(2):684–698.

Wood, S. N. (2017). *Generalized Additive Models*. Chapman and Hall/CRC, New York.

Yadav, R., Huser, R., Opitz, T., and Lombardo, L. (2023). Joint modelling of landslide counts and sizes using spatial marked point processes with sub-asymptotic mark distributions. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(5):1139–1161.

Youngman, B. D. (2019). Generalized additive models for exceedances of high thresholds with an application to return level estimation for U.S. wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879.

Youngman, B. D. (2022). evgam: An R package for generalized additive extreme value models. *Journal of Statistical Software*, 103:1–26.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54:437–447.

Yue, Q., Guo, Y., Sayed, T., Zheng, L., Lyu, H., and Liu, P. (2025a). Bayesian hierarchical non-stationary hybrid modeling for threshold estimation in peak over threshold approach. *arXiv preprint arXiv:2503.14839*.

Yue, W., Tawn, J. A., Towe, R., and Varty, Z. (2025b). Integrating experts' belief in upper tail inference for modelling of human-induced earthquake magnitudes. *Submitted*.

Zachry, B. C., Booth, W. J., Rhome, J. R., and Sharon, T. M. (2015). A national view of storm surge risk and inundation. *Weather, Climate, and Society*, 7(2):109 – 117.

Zaliapin, I., Gabrielov, A., Keilis-Borok, V., and Wong, H. (2008). Clustering analysis of seismicity and aftershock identification. *Physical Review Letters*, 101(1):018501.

Zang, A., Oye, V., Jousset, P., Deichmann, N., Gritto, R., McGarr, A., Majer, E., and Bruhn, D. (2014). Analysis of induced seismicity in geothermal reservoirs–an overview. *Geothermics*, 52:6–21.

Zhao, F., Lange, S., Goswami, B., and Frieler, K. (2024). Frequency bias causes overestimation of climate change impacts on global flood occurrence. *Geophysical Research Letters*, 51(16):e2024GL108855.

Zhuang, J., Harte, D. S., Werner, M. J., Hainzl, S., and Zhou, S. (2012). Basic models of seismicity: Temporal models. *Community Online Resource for Statistical Seismicity Analysis*.