Integrating experts' belief in upper tail inference for modelling of human-induced earthquake magnitudes

Wanchen Yue[1], Jonathan Tawn[1], Ross Towe[2], and Zak Varty[3]

[1]Department of Mathematics and Statistics, Lancaster University LA1 4YF, United Kingdom.
[2]Shell Information Technology International Limited, London SE1 7NA, United Kingdom
[3]Department of Mathematics, Imperial College London SW7 2AZ, United Kingdom.

August 19, 2025

Abstract

Accurate estimation of the upper tail of a distribution is crucial in seismology, where estimating the probability of extreme earthquake magnitudes is vital for risk assessment and mitigation. Traditional statistical methods often overlook expert knowledge, particularly regarding physical upper bounds on earthquake magnitudes. This paper introduces a novel methodology for estimating the upper tail distribution, integrating experts' knowledge on the physical processes through a conservative bound on the worst possible earthquakes. The methodology combines rigorous statistical techniques with expert judgement, creating a hybrid model that complements existing data-driven methods and enhances the reliability of tail estimates. We demonstrate the benefits of incorporating experts' knowledge through the application to data on human-induced earthquakes in the Netherlands. Within this paper, we focus on seismological magnitude modelling, however, the proposed methodology has the potential to be implemented as a generic extreme value approach for multiple problem settings.

Keywords— Earthquake Modelling, Extreme Value Theory, Generalised Pareto Distribution (GPD), Magnitude of Completeness, Maximum Possible Earthquake Magnitude ($M_{\max}$).

1 Introduction

The rise in human-induced seismicity, particularly in regions undergoing extensive industrial activities such as gas extraction or the injection of captured carbon dioxide for storage, presents significant challenges for public safety and infrastructure integrity (Ellsworth 2013, Evans et al. 2012). One such region is the Groningen gas field in the Netherlands, where increased seismic activity within the area has been attributed to natural gas extraction. Concerns have since been raised about the potential for severe earthquakes and the subsequent impact on buildings, pipelines, and other critical infrastructure (Vlek 2019). Accurate inference and modelling of these earthquakes are essential to mitigate risks,

inform public policy, determine potential cash reserves needed to mitigate against future earthquake damage, and to ensure the safety and resilience of the affected communities.

Statistical modelling of earthquake magnitudes plays a pivotal role in understanding and predicting the behaviour of human-induced earthquakes. Due to constant monitoring of seismic sites, there are often large data sets that can be used to inform our models. However, as policyholders and risk analysts are often interested in determining the upper tail and the worst case scenario, inference often focuses on the extreme events with the largest magnitudes.

Seismic experts generally assume that actual earthquake magnitudes are independent and identically distributed (IID) over time (Ogata 1988). Many researchers (Aki 1965, Kagan 2002) have adopted the exponential distribution to model earthquake magnitude based on the Gutenberg-Richter law (Gutenberg & Richter 1944), which postulates a logarithmic relationship between earthquake magnitude and frequency. There are problems with implementing this exponential distribution in both tails. The smallest earthquakes cannot be guaranteed to be measured due to the quality of the recording network, which varies spatially, so there is a level for any given region, termed the magnitude of completeness M_c, below which earthquake data may be partially missing. Hence, observed earthquakes are inconsistent with an exponential distribution below M_c. To overcome this, the magnitudes' excesses over an estimated magnitude of completeness M_c are modelled as exponential, which is consistent with the Gutenberg-Richter law due to the lack of memory property of the exponential distribution. On the other hand, the lack of a finite upper endpoint of the exponential is at odds with the knowledge that the seismic energy stored in any region must have an upper bound and this imposes a finite upper bound, termed $M_{\max}$, on the largest possible earthquake in that region (Brune 1968, Chinnery 1969). To bridge this gap, a truncated exponential distribution has been proposed (Raschke 2015), with density function:

$$f(x;\theta,\tau) = \frac{\theta \exp(-\theta x)}{1-\exp(-\theta\tau)} \qquad 0 \le x \le \tau, \tag{1}$$

and 0 otherwise, where $\theta > 0$. This model introduces a step change in the density at the unknown upper endpoint τ, where $\tau \le M_{\max}$. However, there are drawbacks to this strategy as well. The maximum likelihood estimate (MLE) for τ is simply the largest value in the dataset, making the estimated model unreliable for magnitudes beyond the observed data. This inherent limitation highlights the need for more flexible models in capturing extreme seismic events. To address the limitations of a sharp cut-off in the large-event distribution tail of the truncated exponential, some researchers have introduced a gradual continuous 'taper' to the density function relative to an exponential density, which aligns more realistically with physical dynamic systems. An example of this approach is the tapered Gutenburg-Ritcher distribution (Vere-Jones et al. 2001). However, tapering introduces its own challenges, requiring the choice of the form of the taper function and the estimation of the associated additional unknown parameters. These choices are typically difficult to justify rigorously.

Statistical approaches have also been developed to estimate $M_{\max}$ directly as the endpoint of a distribution, along with associated measures of uncertainty (Beirlant et al. 2019). However, these methods often yield confidence intervals that are unrealistically narrow—much tighter than those provided by seismic experts. This suggests they may not adequately capture the uncertainty in the extreme tail. Moreover, such approaches typically lack empirical diagnostics to assess the validity of their underlying assumptions. For these reasons, we do not adopt this type of method in our analysis.

A flexible statistical alternative is provided by the Generalised Pareto distribution (GPD) from peaks-over-threshold extreme value theory (Coles 2001, Embrechts et al. 2013). Applied to magnitudes exceeding a sufficiently high threshold u, the GPD introduces a shape parameter ξ that controls the tail behaviour: it reduces to the exponential case underlying the Gutenberg-Richter relationship when $\xi = 0$, but yields a finite upper endpoint x_e when $\xi < 0$, thereby accommodating physically bounded maximum magnitudes; heavier-than-exponential tails ($\xi > 0$) are also permitted if supported by the data. This approach, which we will follow, has been taken by Varty et al. (2021) with promising results for Groningen magnitude data.

Geophysicists have explored the upper limits of earthquake magnitudes from a physical perspective (McGarr 2014, Galis et al. 2017, Weng et al. 2021). These approaches leverage geophysical principles to estimate the maximum possible magnitude, i.e., $M_{\max}$, based on underlying processes such as fault mechanics and energy release. Different physical models do not always agree with one another regarding the value of $M_{\max}$, resulting in a distribution of values for $M_{\max}$ across seismic experts and their methodologies. A complication is that there is no public statement of the relative confidence the experts have in their estimates and some of the $M_{\max}$ estimates may be negatively biased ($M_{\max} < x_e$).

We propose a novel approach to address these challenges by incorporating experts' physical knowledge of the maximum possible earthquake magnitude ($M_{\max}$) into GPD modelling. Rather than directly letting $M_{\max}$ dictate the upper endpoint, we treat the distribution of $M_{\max}$ as a constraint that bounds the GPD upper endpoint from above, ensuring that the model respects physical limits without over-restricting statistical inference. With this construction, the knowledge of $M_{\max}$ cannot simply be incorporated into the analysis as a prior for the GPD parameters, as the relationship between $M_{\max}$ and x_e does not provide a simple functional form for use in the prior. By combining the flexibility of the GPD with a bounded upper limit derived from the $M_{\max}$ distribution, our approach aims to reduce the risk of overestimation of events in the upper tail while retaining the ability to model rare, high-magnitude events in a statistically robust and physically informed manner. As noted above, some estimates of $M_{\max}$ may be negatively biased so we also explore the effect of using the experts' $M_{\max}$ distribution as a prior for x_e. This leads to estimates of x_e that are above the smallest of the experts' $M_{\max}$ estimates, in contrast to the constrained $M_{\max}$ approach where x_e can be smaller than this level. Ultimately the best of these two approaches will come down to what operational risk managers prefer to believe, an underestimation or an overestimation of the experts' claimed worst-case risks. Their decision has to be driven by their beliefs about the reliability of the geophysical experts. Our role, as statisticians, is simply to provide to the managers with the best possible statistical inferences conditional on the different perspectives taken on the views of the experts' performance.

The paper is structured as follows. Section 2 presents the data for the Groningen gas field and the value of M_c and how it changes over time, as well as giving the distribution of $M_{\max}$ derived by seismological practitioners for this region. In Section 3 we introduce the GPD properties as well as its specific features for modelling earthquake magnitudes with a varying threshold to account for time-variation in M_c. Section 4 outlines our proposed methodology for incorporating expert beliefs into an extreme value framework, with a focus on estimating the upper tail of the magnitude distribution. This is approached using both likelihood-based estimation (Section 5) and Bayesian inference (Section 6), the latter including both a penalised-likelihood formulation and the use of expert-informed $M_{\max}$ distribution as a prior on for the GPD endpoint. Results for Groningen catalogue from the likelihood-based

methods are presented in Section 5, while Section 6.4 summarises the Bayesian results. Concluding remarks are given in Section 7.

2 Groningen Data Catalogue

2.1 Earthquake Catalogue and Magnitude of Completeness

The Groningen region in the Netherlands has one of the largest natural gas reserves, globally. Extensive gas extraction since 1963 from these deep reserves has significantly altered underground pressures as the remaining gas re-distributes itself across the region, causing subsidence and increased seismic activity. Starting in the mid-1990s, the Royal Netherlands Meteorological Institute (KNMI) has been measuring and monitoring earthquake occurrences in the Netherlands through a network of geophones (KNMI 2020). This induced earthquake dataset comprises detailed records of seismic events, including the time, location, magnitude, and depth of each earthquake; it is a critical resource for understanding the seismic activity in the Groningen region. The magnitude recorded for each seismic event is expressed on a logarithmic scale with units of local magnitude M_L. Here, we have a precise set of magnitudes, which avoids issues linked to rounded data that previous analyses of these data encountered. Figure 1 shows the values of earthquake magnitudes over time from April 1995 to 2024, capturing a total of 1859 recorded earthquakes during this period. A large number of small earthquakes in this region have not been recorded due to their magnitudes being below the magnitude of completeness at the time of their occurrence.

It is widely accepted that since 1995, a conservative estimate for the M_c in the Groningen region has been 1.5 M_L (Dost & Kraaijpoel 2013), with only excesses over time of this constant being used for estimation of the upper endpoint of the distribution of magnitudes by Beirlant et al. (2019). In Figure 1, the rate of recorded earthquake occurrences shows a clear upward trend over time for the majority of the period. Even if this rate change was accounted for, the figure also shows more small magnitude earthquakes occur at the later dates. This change reflects the investment in improving and expanding the geophone network in 2014-2017. As noted by Paleja et al. (2016), we now possess the ability to be more certain to detect smaller magnitude events, i.e., the magnitude of completeness needs to be a function of time, t, i.e., $M_c(t)$, which decreases through the observation period.

The distribution of excesses above $M_c(t)$ is well described by the Gutenberg-Richter law, i.e., the exponential distribution, and it is likely to be contained in the GPD family. Hence, we can estimate $M_c(t)$ as the temporally varying threshold for the GPD. This is exactly the approach Varty et al. (2021) took in estimating $M_c(t)$ for the Groningen catalogue. They adopted a sigmoid threshold function and optimised its parameters. Inference for this function showed that there was considerable uncertainty in the rate of transition from one asymptote to the other, but reliable estimates for the other parameters. For simplicity within our study, we adopt the special case of their estimated function, i.e., the step function for t, between 1995-04-01 and 2024-01-04,

$$M_c(t) = u_1 \mathbb{1}\{t \leq t_0\} + u_2 \mathbb{1}\{t > t_0\}, \tag{2}$$

where $\mathbb{1}(A)$ is the indicator function of event A, $(\hat{u}_1, \hat{u}_2) = (1.15, 0.76)$ and $\hat{t}_0 =$ 2015-12-25. Before and after the change in $\hat{M}_c(t)$, there were 1159 and 700 earthquake events observed, within which,

there were 593 and 396 exceedances above the thresholds respectively. This threshold function is shown in Figure 1. Subsequently, we treat the $\hat{M}_c(t)$ as known and we drop the estimate notation. Critically, for our purposes, the excesses of this estimated threshold function are fitted excellently by a GPD, as found through a QQ plot in later data analysis, see Figure 3 (top left panel).

We are confident in the use of threshold function (2), which was estimated using an automated optimising technique for extreme value threshold selection developed by Varty et al. (2021). Specifically, the parameters of a flexible class of threshold functions are estimated simultaneously with likelihood inference for the GPD model for excesses of this threshold function.The method uses a data-driven estimation approach that optimises the GPD fit by taking into account both bias and variance of the estimators. Murphy et al. (2025) have illustrated that, for IID variables, this approach systematically outperforms the leading established automated threshold methods as well as the well-established visual methods popularised by Coles (2001).

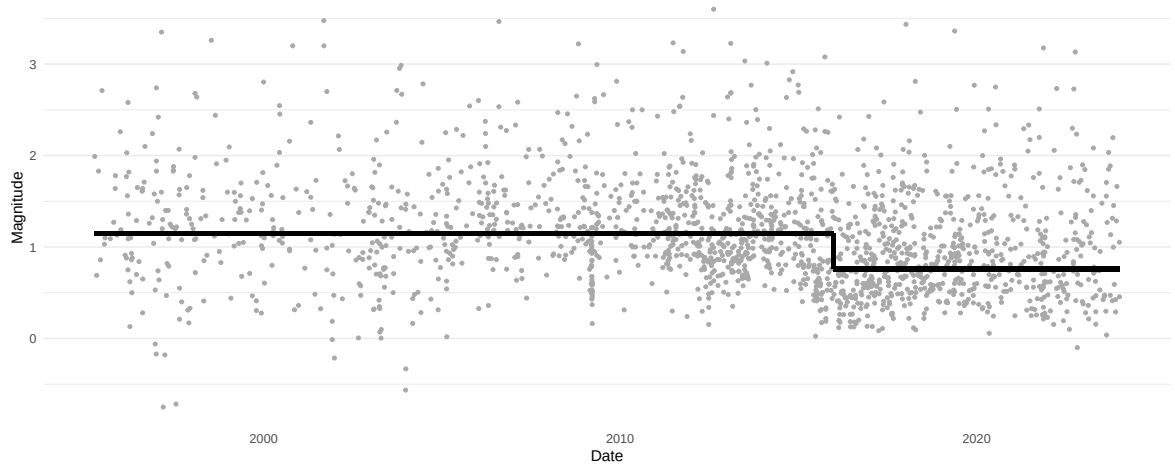

Figure 1: Groningen induced seismicity catalogues from April 1995 to January 2024 in units M_L. The step change in $M_c(t)$, given by expression (2), is shown by the horizontal black lines.

2.2 Expert View on $M_{\max}$ Distribution for Groningen

The maximum possible magnitude $M_{\max}$, represents the largest possible earthquake that can occur in a specific seismic region or along a particular fault system. This concept is crucial in seismology and earthquake engineering for various reasons. For several years, NAM (Nederlandse Aardolie Maatschappij) has been working on improving a seismic hazard and risk model in response to the induced earthquakes in the Groningen gas field. In line with these efforts, a workshop considering the statistical/hybrid and physical dimensions approaches was held in 2022 to determine $M_{\max}$ for the Groningen region with recommendations reported by NAM (2022). The resulting distribution for $M_{\max}$, ranging from 4.0 M_L to 6.5 M_L, was then obtained by combining across all elements from the two approaches (statistical and physical). The proposed $M_{\max}$ distribution is represented discretely by a probability mass function with values centred in 0.5 magnitude unit bins. The corresponding continuous cumulative distribution function (CDF) is illustrated in Table 1, where the CDF is constructed by assigning the probability mass in each discrete magnitude bin uniformly over the 0.5 magnitude unit bin width centred on the magnitudes.

$M_{\max}$(in M_L)	3.75	4.25	4.75	5.25	5.75	6.25	6.75
Cumulative Probability	0	0.27	0.675	0.8625	0.97	0.995	1

Table 1: The cumulative distribution function (CDF) of the assessed $M_{\max}$ distribution provided by the $M_{\max}$ report (NAM 2022).

To facilitate later analysis, where a differentiable model for the distribution of $M_{\max}$ is needed, we use a shifted truncated Gamma distribution over the interval $[a, b] = [3.75, 6.75]$ to approximate the true underlying distribution of the experts' views. Its distribution function, denoted by $F_{M_{\max}}(x)$, can be expressed in terms of the Gamma distribution function, F_G, as follows:

$$F_{M_{\max}}(x; \alpha, \beta) = \begin{cases} 0, & x < a, \\ \frac{F_G(x-a;\alpha,\beta)}{F_G(b-a,\alpha,\beta)}, & a \leq x \leq b, \\ 1, & x > b, \end{cases} \tag{3}$$

where α and β are the shape and scale parameters of a Gamma distribution. We use the estimates $(\hat{\alpha}, \hat{\beta}) = (1.64, 1.79)$, which were obtained by a maximum likelihood fit of distribution (3) using a large sample simulated from the piecewise linear distribution function of the discrete $M_{\max}$ distribution of Table 1. In the supplementary materials, we show that there are very little differences in the results of Section 5 when using either a different fitting method or with the above fitting method with $b = \infty$.

2.3 Strategy for Incorporating the $M_{\max}$ Distribution into Statistical Inferences

We can use estimates of $M_{\max}$ in our inference for the magnitude distribution based on the arguments in Section 1. Firstly we assume that underestimation of $M_{\max}$, i.e., that $M_{\max} < x_e$, is not an issue. Under this assumption, it is reasonable to treat $M_{\max}$ as an upper bound for the endpoint of the GPD used to model earthquake magnitudes, such that $x_e \leq M_{\max}$. If the distribution of the data generating mechanism is taken to follow a GPD above a fixed threshold and $M_{\max}$ has a known finite value, then the GPD shape parameter necessarily must satisfy $\xi < 0$.

Complexity comes from there not being a consensus on a single value for $M_{\max}$, as seen in Section 2.2, so a distribution of possible values for $M_{\max}$ exists for any region. While the $M_{\max}$ distribution is derived from experts' assessments, it is important to recognise that experts' judgements may vary due to different assumptions and methods. If we assumed that all experts were correct in their assessments, the minimum value from the distribution of $M_{\max}$ would provide a conservative upper bound for x_e, ensuring that it remains below the limit suggested by any expert. However, this is not a practical assumption. By having a distribution for $M_{\max}$, we account for the variability in expert opinions and allow for a probabilistic representation of the upper magnitude limit, which better reflects the uncertainty and enables more robust inference.

How we incorporate the experts' views into a statistical analysis must hinge of how much we trust that they do not under-estimate $M_{\max}$, because if they do then our bounding of x_e by these $M_{\max}$ values will lead to bias. So we also exploit the $M_{\max}$ distribution as a prior for x_e, but this risks over-estimation of the tail of the magnitude distribution as x_e must necessarily exceed the experts' smallest $M_{\max}$ value. These two approaches reflect different levels of trust in the experts' collective views. As

noted in Section 1, our role is to provide the best inferences possible based on these varying levels of confidence in experts' judgement and to leave the decision of how to the trade-off between under- and over-estimation to the policymakers based on their own confidence in the experts' reliability.

3 Magnitude modelling for Varying Threshold

3.1 Generalised Pareto Distribution

The focus on extreme events intuitively leads to the application of extreme value theory (Coles 2001, Embrechts et al. 2013), which provides a robust framework for modelling the tails of the earthquake magnitude distribution, where these rare, but high-impact, events occur. In particular, extreme value methods consider a continuous random variable with distribution function F. Under weak assumptions, an asymptotic argument justifies the use of the generalised Pareto distribution (GPD) as a model for the conditional distribution function, $F_u(x) = [F(u+x) - F(u)] \,/\, [1 - F(u)]$ for $x > 0$, of excesses by the random variable over a high threshold u. Specifically, as u tends to the upper endpoint of F, if there exists a function $a(u) > 0$ such that $F_u(a(u)x)$ is non-degenerate for $x > 0$, then $F_u(a(u)x)$ converges to a GPD (Pickands 1975, Davison & Smith 1990). Applying this limit distribution for the excesses over a threshold u, which is a high quantile of F, leads to the following model. For $x > 0$, the distribution function for $F_u(x)$ with threshold u is modelled by:

$$F_u(x; \sigma_u, \xi) = 1 - \left(1 + \frac{\xi x}{\sigma_u}\right)_+^{-1/\xi}, \tag{4}$$

where $\xi \in \mathbb{R}$ is the shape parameter, $\sigma_u > 0$ is the scale parameter, and $x_+ = \max(x, 0)$. Unlike the truncated exponential, the GPD allows for the modelling of extreme events beyond the largest observed magnitudes, offering a more natural framework for tail inference. In contrast to the tapered Gutenburg-Ritcher distribution, the GPD has a theoretical basis for its taper function. The upper endpoint, x_e, for the GPD is defined by

$$x_e = u - \frac{\sigma_u}{\xi} \qquad \text{when} \qquad \xi < 0, \tag{5}$$

while $x_e = \infty$ when $\xi \geq 0$. In the case where $\xi = 0$, the GPD is an exponential distribution (with $\theta = 1/\sigma_u$, $\tau = \infty$ in distribution (1)), which satisfies the Gutenberg-Richter law. This makes the GPD an apt choice of model as it can be arbitrary close to the exponential whilst having a finite upper endpoint when $\xi < 0$, with the value of ξ also determining the rate of decay of the taper. Establishing an appropriate threshold, u, is crucial, as it serves as the cutoff point for considering only the extreme events (Murphy et al. 2025).

3.2 GPD with Varying Threshold

Let $X_t > 0$ be an earthquake magnitude at time t. Following Ogata (1988), we assume the set $\{X_t\}$ consists of IID realisations of a random variable. Further we assume that $X_t \sim \text{GPD}(\sigma_0, \xi)$, where $\sigma_0 > 0$. However, due to the partially missing data of X_t below the time changing magnitude of completeness, $M_c(t)$, we could only fit this distributional model conditionally on $X_t > M_c(t)$ for any

t. Given the GPD threshold $u(t) = M_c(t)$, we have the property that

$$\{X_t - u(t)\}|\{X_t > u(t)\} \sim \text{GPD}(\sigma_{u(t)}, \xi), \tag{6}$$

where $\sigma_{u(t)} = \sigma_0 + \xi u(t)$, and this distribution is found by using the property of closure of the GPD to conditioning (Coles 2001). Property (6) ensures that the GPD remains consistent across different thresholds, with identical shape parameters and linked scale parameters determined entirely by σ_0 and ξ. Additionally, the upper endpoint x_e remains unaffected by the change in the threshold, so $x_e = x_e(\sigma_0, \xi) = -\sigma_0/\xi$, for $\xi < 0$ for all t. Although X_t are IID, property (6) shows that the excesses of X_t over $u(t)$ are no longer identically distributed but, of course, remain independent.

For Groningen earthquake data, we have $u(t)$ given by expression (2), which is piecewise constant with threshold $u_j > 0$ for periods $j = \{1, 2\}$, corresponding to the time period $\{t \leq t_0\}$ and $\{t > t_0\}$ respectively. Hence, following property (6), the excesses of the threshold are identically distributed within each period but not between periods, with scale parameter $\sigma_{u_j} = \sigma_0 + \xi u_j$ in period j ($j = \{1, 2\}$). For period j, denote the n_j excesses of u_j denoted by $y_{j,1}, y_{j,2}, \ldots, y_{j,n_j}$. The likelihood function for σ_0 and ξ is therefore given by:

$$L(\sigma_0, \xi) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} f_{u_j}(y_{j,i}; \sigma_{u_j}, \xi), \tag{7}$$

where f_{u_j} is the density function for GPD with threshold u_j derived from the CDF in expression (4).

Understanding the behaviour of this likelihood function is crucial for inference on the model parameters σ_0 and ξ, which in turn provides inference for the distribution of the upper tail. However, relying solely on the likelihood function can have limitations, as it does not account for expert knowledge or prior beliefs about the GPD parameters or about features of the distribution of magnitude, which can provide valuable insights and improve the accuracy of estimates. As discussed in Section 2.3, there are two core methodologies considered in this paper to incorporate expert knowledge depending on the different levels of trust in the experts' view.

One approach is to treat the expert-informed distribution for $M_{\max}$ as an upper bound for x_e, the upper endpoint of the GPD. In the field of biochemistry, Dryden & Zempléni (2006) explored how the presence of a known upper bound, influences the shape parameter of the distribution, when applied to muscle and DNA data. In their framework, x_e is unknown but constrained by a known constant upper bound. To address such scenarios, Dryden & Zempléni (2006) proposed two methods: fixed endpoint and constrained MLEs. In our seismology setting, the unknown upper endpoint x_e of the GPD satisfies $x_e \leq M_{\max}$ and we have knowledge about $M_{\max}$ through its distribution, so $M_{\max}$ is not a known constant as required by Dryden & Zempléni (2006). In Sections 5- 6, using likelihood and Bayesian inference paradigms respectively, we explore how to incorporate this expert-informed distribution of $M_{\max}$ into the GPD inference process using a penalised likelihood approach, which will enable us to account for the uncertainty in $M_{\max}$.

The other approach is by considering the $M_{\max}$ distribution as a prior for the upper endpoint of GPD (i.e., assume $x_e = M_{\max}$). As noted by Coles & Tawn (1996), including prior knowledge about a set of extreme quantiles (that may not correspond to the upper endpoint) can enhance the fitted model's performance by incorporating information that is not contained within the raw data.

4 Methodology

4.1 Penalised Likelihood Function

To incorporate the experts' belief about $M_{\max}$ being an upper bound for x_e into the inference approach for the GPD, we propose a novel method to use a penalty function within the likelihood estimation framework. Specifically, the experts' belief about the upper endpoint $x_e := x_e(\sigma_0, \xi)$, given by expression (5), can be represented through a penalty function $S(\cdot)$. This penalty function must possess certain properties to adequately capture the characteristics of $M_{\max}$: (i) no penalty should be applied when x_e is below the minimum plausible value of $M_{\max}$, as such x_e is consistent with the experts' beliefs on $M_{\max}$; (ii) the degree of penalty should increase monotonically with x_e across the range of the expert's distribution for $M_{\max}$, reflecting a growing deviation from the experts' beliefs; and (iii) the penalty should make it impossible for estimates of x_e to be larger than the maximum plausible value of $M_{\max}$.

Given the requirements on $S(\cdot)$, it is natural to think of the survival function of the experts' views over the distribution of $M_{\max}$ as the penalty function. Using the approximated distribution $F_{M_{\max}}$ given by expression (3) and the original likelihood function based on σ_0 (scale) and ξ (shape) by $L(\sigma_0, \xi)$ in expression (7), the penalised likelihood function L_p is expressed as

$$L_p(\sigma_0, \xi; \lambda) = L(\sigma_0, \xi) \times \left[1 - F_{M_{\max}}(x_e(\sigma_0, \xi); \hat{\alpha}, \hat{\beta})\right]^{\lambda}. \tag{8}$$

where $\lambda \geq 0$ is a tuning parameter that controls the weight of the penalty function in the overall penalised likelihood. A higher value of λ places more emphasis on adhering to the experts' belief about $M_{\max}$, while a lower value of λ reduces this emphasis, and no penalty is applied to the original likelihood function when $\lambda = 0$.

Figure 2 illustrates examples of the penalty function $\left[1 - F_{M_{\max}}(x; \hat{\alpha}, \hat{\beta})\right]^{\lambda}$, with different values of the weight parameter λ to the approximated Groningen $M_{\max}$ distribution. The figure shows that, for small values of λ, such as $\lambda = 0.1$, the penalty is relatively mild, allowing the likelihood to be driven primarily by the observed data apart from at the very upper endpoint of the $M_{\max}$ distribution. As λ increases to values like $\lambda = 1$ or $\lambda = 2$, the penalty function exerts a stronger influence, increasingly discouraging estimates of x_e that deviate toward the upper range of the $M_{\max}$ distribution, and even increasingly penalising values of x_e closer to the lower endpoint of the $M_{\max}$ distribution. By using this penalised likelihood approach, we are able to blend empirical data with expert opinion, leading to a more robust and informed estimation of the model and its corresponding upper tail. The selection of λ, which is critical for balancing the influence of empirical data and expert opinion, will be discussed at the end of Section 4.2 and in Section 5.2.

4.2 Fit Analysis and Penalty Weight Selection

To assess how well the resulting GPD model fits the data, we quantify goodness of fit via the expected quantile discrepancy (EQD) metric of Murphy et al. (2025). The EQD is defined as the expected value of the absolute differences between the observed quantiles and the corresponding theoretical quantiles from a specified distribution. For the GPD for threshold excesses of u_j, corresponding to the period j in our data, the theoretical quantiles $q_{p,j}$ with respect to the probability $p = \mathbb{P}(X_t < q_{p,j} \mid X_t > u_j)$,

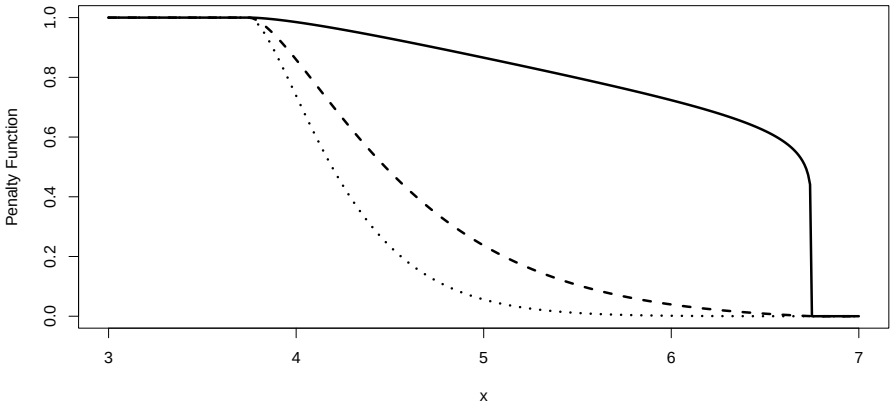

Figure 2: The multiplicative penalty function $\left[1 - F_{M_{\max}}(x; \hat{\alpha}, \hat{\beta})\right]^{\lambda}$ with different λ values. Here S is the survival function of the approximated $M_{\max}$ distribution for Groningen (given by the Section 2.2) and $\lambda = 0.1$ (solid); $\lambda = 1$ (dashed); $\lambda = 2$ (dotted).

for all t in period j, can be written as

$$q_{p,j} = u_j + \frac{\sigma_{u_j}}{\xi}\left[(1-p)^{-\xi} - 1\right].$$

Since both u_j and σ_{u_j} vary over $j = 1, 2$ in our application, to assess fit over all threshold excesses, we need to standardise the two distributions to a common form so that we can pool information on the fit of the model across all the data. The excesses above the time-varying threshold u_j of the observation magnitude data denoted by $y_{j,i}$ (for $j = 1, 2$ and $i = 1, 2, \ldots, n_j$) need to first be transformed onto same marginal distribution. Like Varty et al. (2021), in our analysis, this shared marginal distribution is selected to be the standard exponential distribution.

The transformed excesses can be derived using the inverse probability integral transform $z_{j,i} = -\log[1 - F_{u_j}(y_{j,i}; \hat{\sigma}_{u_j}, \hat{\xi})]$, where $F_u(\cdot)$ is the GPD distribution function (4). When pooling the $\boldsymbol{z}$ values together, we denote $\{z_{j,i} : j = 1, 2; i = 1, \ldots, n_j\}$ as $\boldsymbol{z} = \{z_1, z_2, \ldots, z_n\}$ with $z_i > z_{i-1}$, for all $i = 2, \ldots, n$ and where $n = n_1 + n_2$. The sample quantile function $Q(\cdot, \boldsymbol{z})$ is obtained by linear interpolations of the ordered empirical sample $\boldsymbol{z}$. The formulation described above applies to the EQD metric for a single realization, but to obtain a robust fit measure, following Murphy et al. (2025), we average this metric across the set of non-parametric bootstrapped datasets $\boldsymbol{y} = \{\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(k)}\}$, where $\boldsymbol{y}^{(b)} = \{y_{j,i}^{(b)} : j = 1, 2; i = 1, \ldots, n_j\}$ represents the b-th bootstrap sample ($b = 1, 2, \ldots, k$). Given the estimated GPD parameters, $\hat{\sigma}_{u_j}^{(b,\lambda)}$ and $\hat{\xi}^{(b,\lambda)}$, obtained from the penalised likelihood function with penalty weight λ, each transformed excess associated with threshold u_j is computed as:

$$z_{j,i}^{(b,\lambda)} = -\log[1 - F_{u_j}(y_{j,i}^{(b)}; \hat{\sigma}_{u_j}^{(b,\lambda)}, \hat{\xi}^{(b,\lambda)})].$$

For a set of equally-spaced evaluation probabilities $p_\ell = \ell/(m+1) : \ell = 1, \ldots, m$, the EQD metric $\hat{d}(\lambda)$ for the bootstrap excesses dataset $\boldsymbol{y}^{(b)}$ is calculated as:

$$\hat{d}(\lambda) = \frac{1}{k}\sum_{b=1}^{k} \hat{d}_b(\lambda), \quad \text{where } \hat{d}_b(\lambda) = \frac{1}{m}\sum_{\ell=1}^{m} \left| -\log(1-p_\ell) - Q\left(p_\ell, \boldsymbol{z}^{(b,\lambda)}\right)\right| \tag{9}$$

and $\boldsymbol{z}^{(b,\lambda)} = \{z_{j,i}^{(b,\lambda)} : j = 1, 2; i = 1, \ldots, n_j\}$ with $(m, k) = (500, 1000)$, i.e., values at least as large as Murphy et al. (2025) recommended. This metric allows us to capture both the fit of individual bootstrap realisations and the stability of fit across samples, ensuring that the EQD reflects both local and overall fit performance for the GPD model.

We guide the selection of an appropriate penalty weight λ by identifying the range of λ values ($\lambda > 0$) which have the benefit of incorporating the expert information about $M_{\max}$ but do not deviate substantially the quality of fit much relative to the fit when $\lambda = \lambda_0 = 0$. To determine an upper bound to this range of λ values, we use the metric,

$$T(\lambda) = 2[\log L(\hat{\sigma}_0(\lambda_0), \hat{\xi}(\lambda_0)) - \log L(\hat{\sigma}_0(\lambda), \hat{\xi}(\lambda))], \tag{10}$$

where $\hat{\xi}(\lambda)$ and $\hat{\sigma}_0(\lambda)$ are the penalised likelihood estimators of ξ and σ_0 respectively when the penalty weight value is λ and L is given by expression (7), so $T(\lambda)$ corresponds to the likelihood ratio for two choices of penalty weight λ and 0. This metric is analogous to the Akaike Information Criterion (AIC) as a function of λ. To guide this selection, we consider two specific critical values for the $T(\lambda)$ of: 3.84 and 6.64, which correspond to significance levels of 0.05 and 0.01, respectively, for 1 degree of freedom in a chi-squared test. The λ values that achieve these critical values for $T(\lambda)$ we denote by $\lambda_{0.95}$ and $\lambda_{0.99}$. These λ values provide a practical upper bound for controlling the degree of distortion of the GPD that it is reasonable to accept to incorporate the expert belief on $M_{\max}$.

5 Application of the Groningen Catalogue

5.1 Upper Tail Estimation

We use the piecewise constant threshold function (2) with $u_1 > u_2$, with its values being treated as known. Excesses of this time varying threshold will have a GPD with a time-varying scale parameter. To aid inference, we exploit the link between the scale parameters mentioned in Section 3.2 to give $\sigma_{u_1} := \sigma_{u_2} + \xi(u_1 - u_2)$ and reparameterise the penalised likelihood (8) from (σ_0, ξ) to (σ_{u_2}, ξ).

Before exploring the effect of penalty weighting λ on the GPD parameters estimation, we examine the MLE results obtained from the observed sample without any penalisation on the upper endpoint (i.e., $\lambda = 0$). We get $\hat{\xi} = -0.12$ with 95% confidence interval for $\hat{\xi}$ of $[-0.17, -0.08]$, obtained using asymptotic normality. This indicates that the shape parameter is almost certain to be negative, and hence the distribution has a finite upper endpoint but, as anticipated, the tail is not far from an exponential. The fact that we can draw such strong conclusions about the estimated value of ξ arises from the work of Varty et al. (2021) in optimising the threshold function to use as many excesses as possible, without diminishing the quality of the GPD fit.

5.2 Goodness of Fit and Selection of λ

To evaluate the goodness of fit for the Groningen dataset, we computed the EQD and $T(\lambda)$ metrics for the penalised GPD model fits for five different values of λ, with Table 2 giving the results. As λ increases, $\hat{d}(\lambda)$ and $T(\lambda)$ both gradually grow, indicating a deviation in the model's overall fit to the data, reflecting the increased influence of the penalty. As $T(0.5) = 3.86$ and $\lambda = 0.5 \approx \lambda_{0.95}$, similarly as $T(1) = 6.61$, then $\lambda_{0.99} \approx 1$. These values for λ provide the largest penalty values weighting in

λ	0	0.1	$\lambda_{0.95} \approx 0.5$	$\lambda_{0.99} \approx 1$	2
$\hat{d}(\lambda)$	0.03	0.03	0.04	0.04	0.05
$T(\lambda)$	0	1.03	3.85	6.61	9.98

Table 2: Goodness of fit metric values for a range of λ values.

favour of the expert information while maintaining a reasonable fit the data. For $\lambda = 2$, we observe a significant distortion in the model fit, as evidenced by the larger $\hat{d}(2) = 0.05$ and $T(2) = 9.98$ values. This substantial degradation highlights that overly strong penalties can force the fitted model to be too biased in favour of the expert information at the cost of distorting the GPD fit to data above $M_c(t)$. Therefore, the choice of λ involves a trade-off: ensuring consistency with expert-informed tail constraints while avoiding excessive distortion of the data-driven GPD fit. Our goal is not simply to minimise EQD, but to strike a balance that yields tail estimates consistent with physical knowledge without sacrificing the model fit.

Both EQD and $T(\lambda)$ provide information only on the overall fit of the distribution and so not about the fit quality at specific quantiles (e.g., far into the upper tail). A more comprehensive evaluation includes examining Q-Q and P-P plots. Figure 3 provides the Q-Q plots for the Groningen dataset with both 95% pointwise confidence intervals and tolerance intervals. They are of more use for assessing upper tail fit than P-P plots, which we present in the supplementary materials. Both sets of plots show that for both $\lambda = 0$ and 0.1, the model slightly overestimates extreme values at higher quantiles, as indicated by deviations of sample quantiles below the diagonal. When λ is increased to 0.5 and 1, corresponding to $\lambda_{0.95}$ and $\lambda_{0.99}$ respectively, the higher quantile fit improves, with confidence intervals aligning more closely to the diagonal. When $\lambda = 2$, the model fit has sample quantiles exceeding model quantiles at the higher end, reflecting the risk of over-regularisation when placing too much emphasis on experts' beliefs. Additionally, penalising the upper endpoint with increasing λ distorts the fit in the body of the distribution with reduced overlap between the 95% confidence interval and tolerance bounds. For the rest of the results shown in this paper, we will only present cases with $\lambda = 0$, 0.1, 0.5, and 1, i.e., in the range $\lambda \in [0, \lambda_{0.99}]$, as they represent a practically feasible range of penalty values.

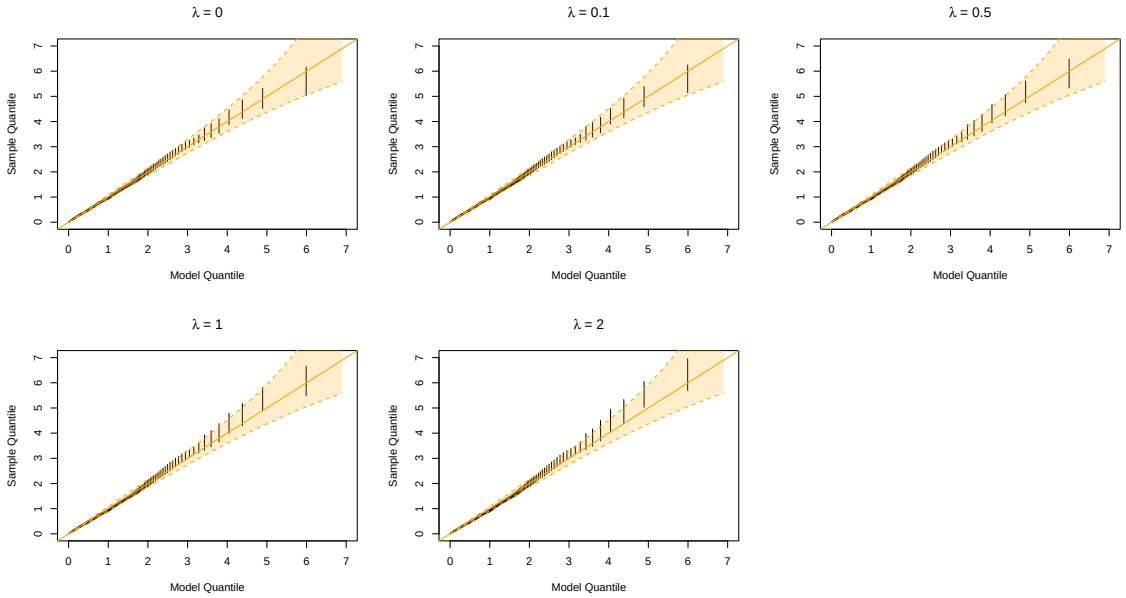

Figure 3: Quantile-Quantile plots of the model fit for different λ, the quantile values are in an exponential scale: pointwise 95% confidence intervals (vertical black bars); 95% tolerance intervals (orange shaded region).

5.3 Results of MLE

We applied the nonparametric bootstrap method with 1000 replicate samples $\{\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(1000)}\}$. These replicates were generated by randomly sampling data with replacement from the original excess data above the corresponding thresholds u_j, $j \in \{1, 2\}$. For each bootstrap sample, the penalised MLEs for the GPD scale, shape parameters and the upper endpoint are evaluated, with $\hat{\xi}^{(b)} < 0$ and so $\hat{x}_e^{(b)}$ is finite and is given by expression (5), i.e., $\hat{x}_e^{(b)} = u_2 - \hat{\sigma}_{u_2}^{(b)}/\hat{\xi}^{(b)}$, for all bootstrap datasets $b = 1, \ldots, 1000$.

Figure 4 shows the resulting bootstrap estimate of the sampling distribution for $\hat{x}_e$ with different λ choices ($\lambda = \{0.1, 0.5, 1\}$). For comparison, we include the case where no expert information on the distribution for $M_{\max}$ is applied ($\lambda = 0$), as well as when $\lambda > 0$ so that the approximated expert distribution for the $M_{\max}$ is used, with this imposing an upper limit of $b = 6.75$ on the sampling distribution. We observe that the sampling distribution when $\lambda = 0$ gives values of x_e noticeably larger than the expert-informed $M_{\max}$ distribution. From the bootstrap results, we find that 14% of the estimated upper endpoints ($\hat{x}_e^{(b)}$) exceed the upper bound of the $M_{\max}$ distribution. As λ increases, the penalty term exerts a stronger influence on the estimation process resulting in the densities of the sampling distribution becoming more concentrated, as well as the sampling distribution mean decreasing, approaching the mean of $M_{\max}$ distribution which is around 4.6 M_L when $\lambda = 1$.

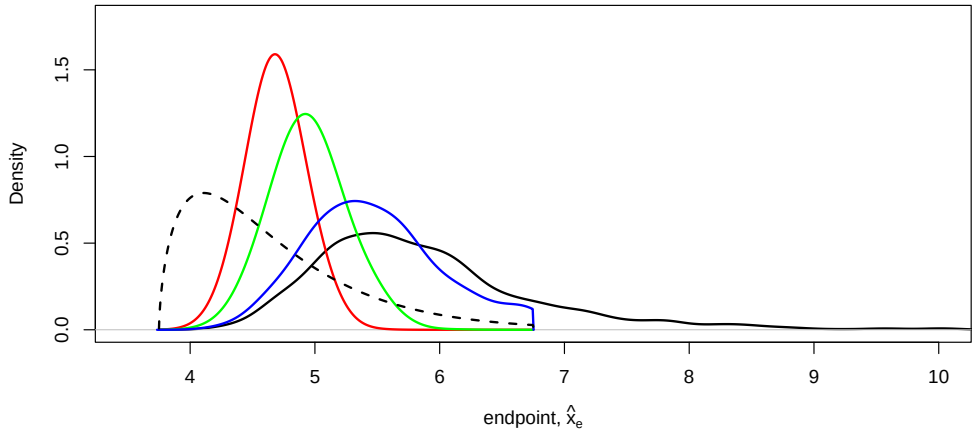

Figure 4: The bootstrap estimate of the sampling distribution of $\hat{x}_e$ (M_L), from the penalised GPD model, based on 1000 bootstraps. Penalty weighting values are shown for $\lambda = 0$ (black); $\lambda = 0.1$ (blue); $\lambda = 0.5$ (green); $\lambda = 1$ (red). The approximated expert $M_{\max}$ distribution is denoted in dashed black line.

As discussed earlier, the focus of our analysis is not solely on estimating x_e, but on understanding the entire upper tail of the magnitude distribution. This is particularly critical for long-term seismic hazard assessments, where rare, extreme events play a central role. Specifically, the 475-year return level is widely used in engineering and policy contexts to design structures as then, in the case of IID earthquake processes, such structures can withstand earthquakes with a 90% probability of occurrence over a 50−year span (Code 2005). In our dataset, the annual number of exceedances of a level u in year y, denoted by $m_u(y)$, varies with y due to changes in extraction rates, however, the magnitudes of excesses over any threshold u, with $u \geq \max_t M_c(t) = u_1$, are IID. Following (Coles 2001), the N-year return level, denoted by $q_N(y)$, satisfies

$$\hat{q}_N(y) = \begin{cases} u_2 + \frac{\hat{\sigma}_{u_2}}{\hat{\xi}} \left[(N m_{u_2}(y))^{\hat{\xi}} - 1 \right], & \hat{\xi} \neq 0; \\ u_2 + \hat{\sigma}_{u_2} \log \left[N m_{u_2}(y) \right], & \hat{\xi} = 0. \end{cases} \tag{11}$$

We focus on presenting a conservative estimate of the return level, by studying the estimate for $y = 2017$, the year with largest number of exceedances of u_1 and u_2, with $m_u(2017) = 78$.

For different λ, Figure 5 plots the return level estimates against return periods (years) on a logarithmic scale, together with the interquartile ranges of the sampling distribution of the estimates. As λ increases from 0.01, the return level at the 475-year period sharply drops from 4.47 M_L to the values estimated at λ values of interest, i.e., $\lambda_{0.95}$ $(\hat{q}_N = 4.28 M_L)$ and $\lambda_{0.99}$ $(\hat{q}_N = 4.16 M_L)$. For any given return period, we observe the mean of the estimated return level and the uncertainty both decrease as λ increases, indicating that the more we trust the experts on $M_{\max}$ the more confident we are about the upper tail in general.

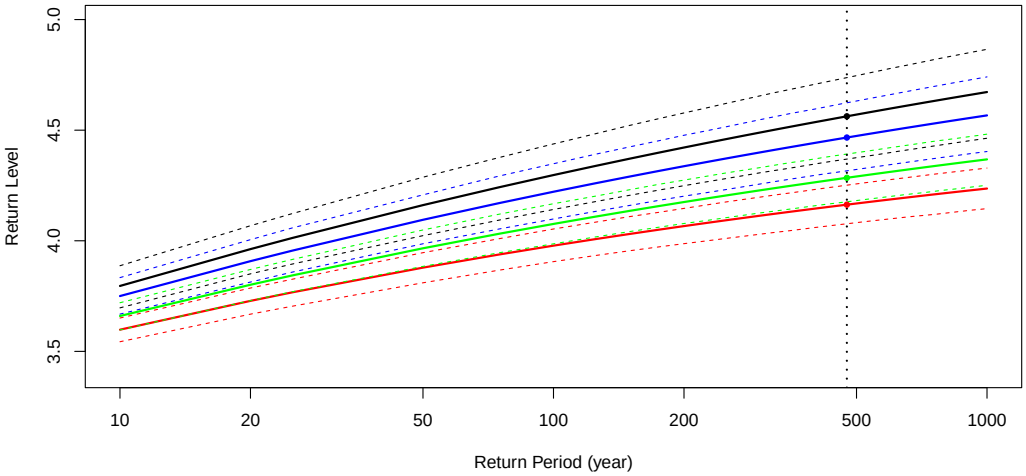

Figure 5: Estimated return level plot for different λ: $\lambda = 0$ (black); $\lambda = 0.1$ (blue); $\lambda_{0.95} = 0.5$ (green); $\lambda_{0.99} = 1$ (red). Solid lines represent the mean and the dashed lines denote the 25th and 75th percentiles of the sampling distribution of the return level estimates. The 475 year return level estimation is indicated with the dotted vertical line.

5.4 Simulation Study

To validate our inference procedures, we conducted a simulation study based on parameter estimates derived from our fit to the Groningen earthquake catalogue and using the experts' distribution for $M_{\max}$. Synthetic datasets were generated under conditions replicating the observed earthquake environment, using piecewise constant thresholds ($u_1 = 1.15$ and $u_2 = 0.76$), with 593 exceedances above u_1 and 396 exceedances above u_2. For the GPD model, the parameters used were $\xi = -0.12$ and $\sigma_{u_2} = 0.64$. These parameter choices reflect the MLEs obtained when no penalty was applied. The GPD endpoint used in the simulation is 5.74, which lies within the domain of the $M_{\max}$ distribution, $[a, b] = [3.75, 6.75]$. We conducted 1000 replicate simulations. Using these simulated data, we found that 15% of the estimated upper endpoints ($\hat{x}_e$) exceed the upper bound of the $M_{\max}$ distribution when using MLE without any penalty. This coincides with the 14% observed in the analysis of the Groningen data in the previous sections, indicating that the simulated data reflects the real data bootstraps quite well. These datasets were analysed and compared using three different approaches: MLE without incorporating expert knowledge on $M_{\max}$, the penalised likelihood method incorporating experts' knowledge on $M_{\max}$ with varying penalty weights ($\lambda = 0.1, 0.5, 1$), and using the truncated exponential distribution (1).

We assess the accuracy of the upper endpoint estimates ($\hat{x}_e$) and high quantile estimates, focusing on the 100-year and 475-year return levels. A boxplot comparison of the upper endpoint estimates is provided in Figure 6. We observe that as the penalty weight increases, the penalised likelihood method tends to underestimate the upper endpoint. However, the variance of the estimates decreases significantly once we put any weight on the penalty compared to when $\lambda = 0$. The estimates obtained by using truncated exponential show the worst deviation from the true upper endpoint.

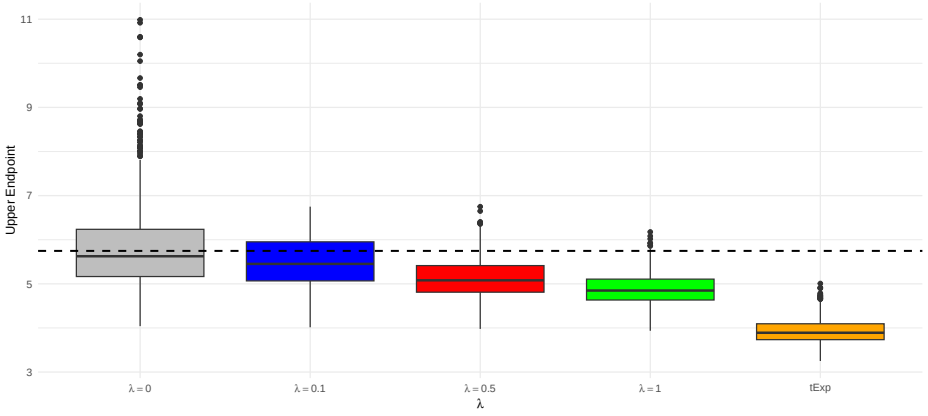

Figure 6: Comparison of the estimated upper endpoint for different approaches and penalty weights λ, the truncated exponential approach is denoted as tExp. The true upper endpoint is indicated by the horizontal dashed line.

The accuracy of these estimates is further quantified in terms of root mean squared error (RMSE), summarised in Table 3. For $\hat{x}_e$, the penalised likelihood methods yield lower RMSE compared to the MLE approach ($\lambda = 0$) for all choices of the penalty weight ($\lambda = 0.1, 0.5, 1$), demonstrating the value of incorporating experts' knowledge. The truncated exponential approach shows the highest RMSE for x_e, indicating its limited performance relative to the other methods.

For the return level estimates we considered two scenarios: one based on recent threshold excess rates in the Groningen region during 2023 and another on the highest annual threshold excess rates, i.e., in 2017. These scenarios represent a more conservative case, using the worst excess rates, and a more recent trend reflecting current conditions. The boxplots for the 100-year and 475-year return level estimates are provided in the supplementary materials. From Table 3, we observe the penalised likelihood methods with a small penalty weight ($\lambda = 0.1$) performed best, achieving the smallest RMSE across both scenarios. Conversely, methods with larger penalty weights ($\lambda = 0.5, 1$) underestimated the return levels, particularly under the worst-case scenario, leading to higher RMSE values. This pattern is consistent across both return periods, underscoring the sensitivity of the penalised likelihood method to the choice of λ. The truncated exponential method also performed poorly, with higher RMSE values than the penalised likelihood approach for most estimates.

	No Penalty	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	tExp
RMSE $\hat{x}_e$	1.05	0.67	0.94	1.15	1.84
RMSE $\hat{q}_{100}$ (Recent)	0.18	0.16	0.27	0.29	0.26
RMSE $\hat{q}_{100}$ (Worst)	0.22	0.20	0.29	0.36	0.38
RMSE $\hat{q}_{475}$ (Recent)	0.24	0.21	0.36	0.40	0.45
RMSE $\hat{q}_{475}$ (Worst)	0.29	0.24	0.37	0.46	0.61

Table 3: RMSE for the estimated upper endpoint $\hat{x}_e$ and estimated return levels. Recent: using the threshold excess rates for 2023. Worst: using the highest threshold excess rates since 1995.

6 Bayesian Inference for the Groningen Catalogue

6.1 Overview of Strategy

In addition to the frequentist approach introduced in Section 5, we adopt a Bayesian inference framework. This approach offers several advantages, including the flexibility to incorporate prior information on the GPD parameters σ_0 and ξ, as well as on the upper endpoint x_e. Such flexibility is particularly valuable in extreme value contexts, where data are typically scarce in the tail and uncertainty is high. As emphasised by Coles & Tawn (1996), prior assumptions can substantially influence posterior inference in these settings, making a well-justified prior specification especially important.

In this context, we again make use of expert knowledge regarding the $M_{\max}$. As discussed in Section 2.2, NAM (2022) provide a subjective distribution for $M_{\max}$ based on expert judgements, which may be viewed with varying degrees of confidence depending on the end user's needs and their assessment of the reliability of the experts' assessments. To accommodate this, we consider two different Bayesian inference strategies that reflect different ways of incorporating the expert information. In Section 6.2, we retain the penalised-likelihood formulation from the frequentist analysis, but carry out inference under a Bayesian framework by placing a prior on the GPD parameters (σ_0, ξ). The expert-informed distribution over $M_{\max}$ is incorporated indirectly via a penalty term in the likelihood, while the prior remains otherwise uninformative. In contrast, in Section 6.3 we take a fully Bayesian approach in which the endpoint x_e is treated explicitly as a model parameter and assigned the expert-informed distribution for $M_{\max}$ as its prior. Thus the penalty term of the penalised-likelihood is replaced by the prior on x_e, providing an alternative way to incorporate expert knowledge within the Bayesian framework. In both approaches, we implement inference using a Metropolis-Hastings (MH) algorithm within a Gibbs sampling framework. The full details of the sampling algorithms are provided in the supplementary materials. Section 6.4 presents the posterior inference results under both formulations for the Groningen earthquake catalogue.

6.2 Penalised-Likelihood With Principled Prior

The posterior distribution we aim to sample from is proportional to the product of the penalised likelihood function $L_p(\sigma_0, \xi; \lambda)$, as defined in expression (8), and a principled prior distribution $\pi(\sigma_0, \xi)$ for the parameters $\sigma_0 > 0$ and $\xi \in \mathbb{R}$. To account for the uncertainty in both the shape and scale parameters of the GPD, we adopt the prior structure $\pi(\sigma_0, \xi) = \pi(\xi)\pi(\sigma_0|\xi)$, where $\pi(\xi)$ represents the prior on the shape parameter ξ, and $\pi(\sigma_0|\xi)$ denotes the conditional prior on σ_0, given ξ. The remainder of this section is devoted to developing principled prior forms for these two components of the joint prior, which reflect our lack of additional understanding about these parameters given our knowledge about the GPD endpoint x_e being less than $M_{\max}$ is already accounted for in $L_p(\sigma_0, \xi; \lambda)$.

First consider the specification of $\pi(\xi)$. Although we have no specific prior knowledge about ξ for the Groningen data, as discussed in Section 1, in earthquake modelling the magnitudes are often modelled using the exponential distribution. So, although the GPD offers greater flexibility for capturing tail behaviour than the exponential, from the context we believe that ξ is more likely to be near zero than far away from it. To quantify the increased complexity from using the GPD instead of the exponential distribution, we employ the Kullback-Leibler divergence ($\mathrm{D_{KL}}$) (Kullback

& Leibler 1951), which motivates the use of a penalised complexity (PC) prior (Simpson et al. 2017). A PC prior incorporates a $\mathrm{D_{KL}}$-based penalty term that favors simpler, more interpretable models. By balancing model complexity and data-driven inference, the PC prior helps to achieve a more robust and interpretable analysis. We denote f_ξ as the density function for the GPD with shape parameter ξ and f_0 as the density function for the exponential distribution (i.e., a GPD with $\xi = 0$). Then $\mathrm{D_{KL}}$ can be calculated as follows, with its derivation provided in the supplementary materials:

$$\mathrm{D_{KL}}\left(f_\xi \| f_{\xi_0}\right) = \int f_\xi(x;\sigma,\xi)\log\left(\frac{f_\xi(x;\sigma,\xi)}{f_0(x;\sigma)}\right)dx = \frac{\xi^2}{1-\xi} \quad \text{for } \xi < 1, \tag{12}$$

and $\mathrm{D_{KL}} = 0$ otherwise. The distance between two models is then defined as

$$d(f_\xi, f_0) = \{2\mathrm{D_{KL}}(f_\xi \| f_0)\}^{1/2} = \sqrt{2}|\xi|(1-\xi)^{-1/2} \quad \text{for } \xi < 1.$$

By penalising the deviation from the exponential model by the distance $d(f_\xi, f_0)$, the PC prior employs a constant decay rate (Simpson et al. 2017). Therefore, $d(f_\xi, f_0)$ follows an exponential distribution with rate ϕ_D, where ϕ_D controls the shape of the prior: $\pi_{d(f_\xi,f_0)}(d) = \phi_D \exp(-\phi_D d)$ for $d > 0$ and $\phi_D > 0$. Transforming the prior to the $\xi-$scale, we have the exact PC prior:

$$\pi_\xi(\xi) \propto \exp\left\{-\phi\frac{|\xi|}{(1-\xi)^{1/2}}\right\}\left\{\frac{1-\xi/2}{(1-\xi)^{3/2}}\right\}, \quad \text{for } \xi < 1, \tag{13}$$

where the penalisation rate parameter $\phi = \sqrt{2}\phi_D$, is the hyper-parameter of this prior. Increasing ϕ leads to a stronger penalisation of ξ values away from 0, ultimately favouring values of ξ closer to 0. We also explore an approximation of $\mathrm{D_{KL}}$ function by ξ^2, which is the first order behaviour of $\mathrm{D_{KL}}$, given by expression (12), as $|\xi| \to 0$ (Opitz et al. 2018). Therefore, we have $d(f_\xi, f_0) = \sqrt{2}|\xi|$ and the approximated PC prior corresponds to a Laplace distribution prior as follows:

$$\pi_\xi(\xi) = \frac{\phi}{2}\exp(-\phi|\xi|), \quad \text{for } \xi \in \mathbb{R}. \tag{14}$$

The PC prior using the exact $\mathrm{D_{KL}}$ gives zero probability to $\xi \geq 1$, while the approximated PC prior gives positive density to all $\xi \in \mathbb{R}$ but locally around $\xi = 0$ the two PC priors (with densities (13) and (14)) are approximately the same. These features are illustrated in Figure 7 for different choices of the hyper-parameter ϕ. When ξ is negative, these two priors have very similar shapes. Additionally, when ϕ is large and ξ is in the range $(-1/2, 1/2)$, the priors also exhibit a similar behaviour. For both of the priors, they pull the posterior distribution mass for tail index ξ towards $\xi = 0$, relative to a posterior with a flat prior to ξ. As a consequence of the strong prior weight near 0, posterior mass for large values of $|\xi|$ will only be observed if the data strongly indicate heavy tails ($\xi \geq 0$), or the bulk of the distribution appears to be near the upper bound ($\xi < 0$), with the penalty function $\left[1 - F_{M_{\max}}(x;\hat{\alpha},\hat{\beta})\right]^\lambda$ ruling out the former in our case due to our $M_{\max}$ knowledge when $\lambda > 0$. When using the exact PC prior model, we notice that the prior distribution becomes increasingly concentrated around $\xi = 1$ as ϕ tends to 0, which is counterintuitive. To avoid this issue, we opt for the approximate PC prior for ξ in the rest of the study, as it maintains a similar shape in most areas and prevents these undesirable behaviours near $\xi = 1$, though in practice it would make little difference as the data pulls the posterior mass to $\xi < 0$.

When selecting a value of the hyper-parameter $\phi > 0$ we need to be aware of the potential for a trade-off between ϕ and the penalty weight λ. An increase in the former concentrates more posterior

mass near $\xi = 0$ while increasing the latter enforces a stronger upper bound and $\xi < 0$ further from zero. We explored how these two components jointly influence the posterior inference. While both parameters affect the posterior inferences, we found that their interaction is weak when studying the Groningen data, suggesting that in our analysis each parameter can be adjusted without needing to compensate for changes in the other. Further illustrations of this behaviour are provided in the supplementary materials, including contour plots of posterior summaries over a (λ, ϕ) grid.

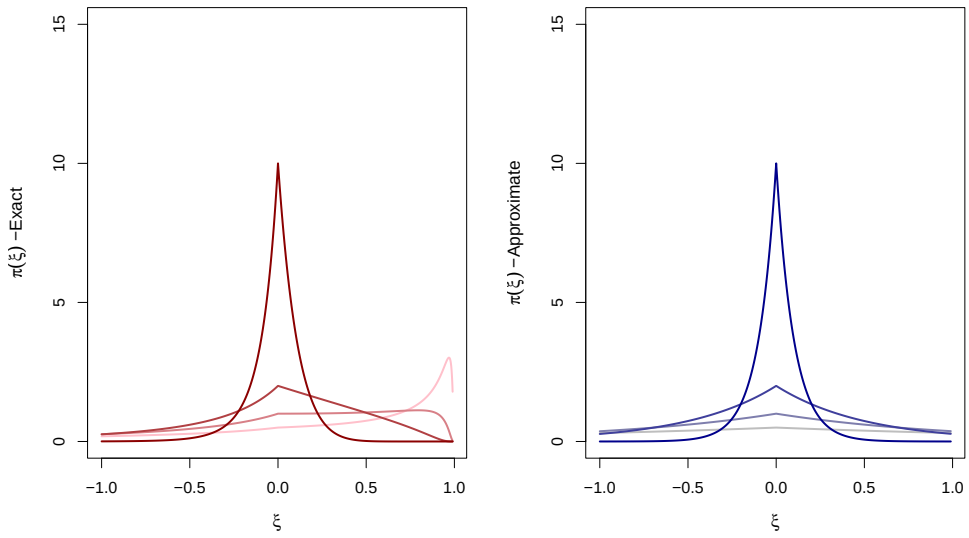

Figure 7: PC prior for ξ using exact KLD function $\xi^2/(1-\xi)$ (left) and approximate KLD function ξ^2 (right) with different values of the hyper-parameter ϕ: $\phi = (0.5, 1, 2, 10)$ (light $\rightarrow$ dark).

As there is no prior knowledge about the GPD scale parameter σ_0, a Jeffreys non-informative prior is used as it reflects prior ignorance which is invariant under reparameterisation. Leonelli & Gamerman (2020) derived the Jeffreys prior for σ_0 given ξ to be:

$$\pi(\sigma_0|\xi) \propto \frac{1}{\sigma_0\sqrt{1+2\xi}} \qquad \text{for} \quad \sigma_0 > 0 \quad \text{when} -\frac{1}{2} < \xi < \frac{1}{2}, \tag{15}$$

and $\pi(\sigma_0|\xi) = 0$ outside this domain range. The range limits for ξ reflects the requirement for a finite variance for GPD, which exists only if $\xi < 1/2$, and it corresponds to insights in Smith (1985) on non-regular likelihood inference asymptotics when $\xi \leq -1/2$. Combining the priors $\pi(\xi)$ and $\pi(\sigma_0|\xi)$, our joint prior density for σ_0 and ξ gives:

$$\pi(\sigma_0, \xi) \propto \frac{1}{\sigma_0\sqrt{1+2\xi}} \exp(-\phi|\xi|) \qquad \text{for} \quad \sigma_0 > 0 \quad \text{and} -\frac{1}{2} < \xi < \frac{1}{2}, \tag{16}$$

which corresponds to independent priors for σ_0 and ξ. A slice of the joint prior distribution with respect to ξ is illustrated in the supplementary materials, with this highlighting the behaviour of the joint prior under varying levels of ϕ.

6.3 Experts' Endpoint Prior

Now suppose that we apply the experts' beliefs about $M_{\max}$ directly to the GPD upper endpoint x_e, so that the prior density for x_e is $f_{M_{\max}}(x_e; \hat{\alpha}, \hat{\beta})$, corresponding to the density function for the expert-informed $M_{\max}$ distribution function (3). This formulation enables a direct regularisation of the upper tail of the magnitude distribution by embedding expert uncertainty within the prior for the endpoint, thus allowing use of the GPD likelihood, without needing the penalised-likelihood. Under this prior, we know that $a < x_e < b$, so when $b < \infty$, it follows that $\xi < 0$, with $x_e = -\sigma_0/\xi$, where $\sigma_0 > 0$ is the GPD scale parameter for a threshold of zero.

As Dryden & Zempléni (2006) did for MLE, we consider a reparameterised GPD model with the endpoint as one of the specified parameters but implemented within a Bayesian framework. Specifically, we reparameterise in terms of (ξ, x_e), with $\sigma_0 = -\xi x_e$. Under this formulation, the threshold-specific GPD scale parameters become

$$\sigma_{u_j}(\xi, x_e) = \sigma_0 - \xi(x_e - u_j), \qquad \text{for } j = 1, 2,$$

corresponding to the two periods of the Groningen catalogue with thresholds u_1 and u_2. In this context, the reparameterisation of likelihood (7) becomes:

$$L_R(\xi, x_e) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} f_{u_j}(y_{j,i}; \sigma_{u_j}(\xi, x_e), \xi) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \frac{1}{|\xi|(x_e - u_j)} \left(1 - \frac{y_{j,i}}{x_e - u_j}\right)_+^{-1/\xi - 1}. \tag{17}$$

We additionally take the prior for ξ as the approximated PC prior (14) restricted to $\xi < 0$. It is reasonable to assume that the priors for (ξ, x_e) are independent as knowing $x_e < \infty$ provides no information about ξ other than $\xi < 0$. The posterior joint density of (ξ, x_e) is then proportional to $L_R(\xi, x_e) \cdot \pi_\xi(\xi) \cdot f_{M_{\max}}(x_e; \hat{\alpha}, \hat{\beta})$ for $\sigma_0 > 0$ and $\xi < 0$.

6.4 Bayesian Results

We apply the Bayesian approaches of Sections 6.2 and 6.3 to the Groningen earthquake catalogue. To ensure robust posterior estimation, we use a Metropolis-within-Gibbs sampler with a burn-in period of 2000 iterations, followed by 150,000 further iterations, from which every 30th sample is retained. This yields effectively independent draws from the posterior. For both Bayesian approaches, in Figure 8 we present results for posterior distributions of x_e and of the 475-year return level, corresponding to a prior for the GPD shape parameter with hyper-parameter $\phi = 10$. Further details of the MCMC mixing, insensitivity to starting values, the posterior distributions of other parameters, and the lack of sensitivity to our choice of ϕ are given in the supplementary materials.

For the penalised likelihood formulation (Section 6.2), we use the principled joint prior $\pi(\sigma_{u_2}, \xi)$ (16) (with the prior for σ_{u_2} identical to that derived for σ_0) for a range of choices of $M_{\max}$ penalty weight, $\lambda = \{0, 0.1, 0.5, 1\}$. Figure 8 shows that the posterior distributions for x_e match the sampling distribution using MLEs for the associated λ values. For $\lambda = 0$, the posterior distribution, based only on the data and the principled prior, excluding any information on the distribution of $M_{\max}$, extends far beyond the range of the experts' $M_{\max}$ distribution. This indicates a lack of constraint on the upper tail, which is also reflected in higher and more uncertain posterior for 475-year return level. For $\lambda = 0.1$, the posterior distributions illustrate how putting a relatively small weight on the penalty

term changes the posterior distribution for x_e, particularly in its upper tail. Increasing λ also results in a more constrained return level estimates. For $\lambda_{0.95} \approx 0.5$ and $\lambda_{0.99} \approx 1$, a stronger regularisation effect is observed in the posterior distributions, resulting in lower estimates of the upper endpoint and reduced variability, with similarly findings for the 475-year return level. Posterior inferences for the standard GPD likelihood combined with experts' endpoint prior (Section 6.3) and the PC prior for ξ are also shown for the Groningen data in Figure 8. The posterior distributions for both the estimated endpoint and the 475-year return level are very closely aligned with those from the penalised model when $\lambda = 0.5$ (i.e., $\lambda_{0.95}$), with the most surprising feature being that both posterior densities are so similar in the lower tail, given that the penalised likelihood and the experts' endpoint prior push $x_e < a$ and $x_e > a$ respectively. This similarity in the two types of posteriors indicates that moderate penalisation yields posterior inferences that are broadly consistent with those obtained through direct incorporation of expert beliefs as a prior for x_e, suggesting a convergence between the two approaches under certain regularisation levels.

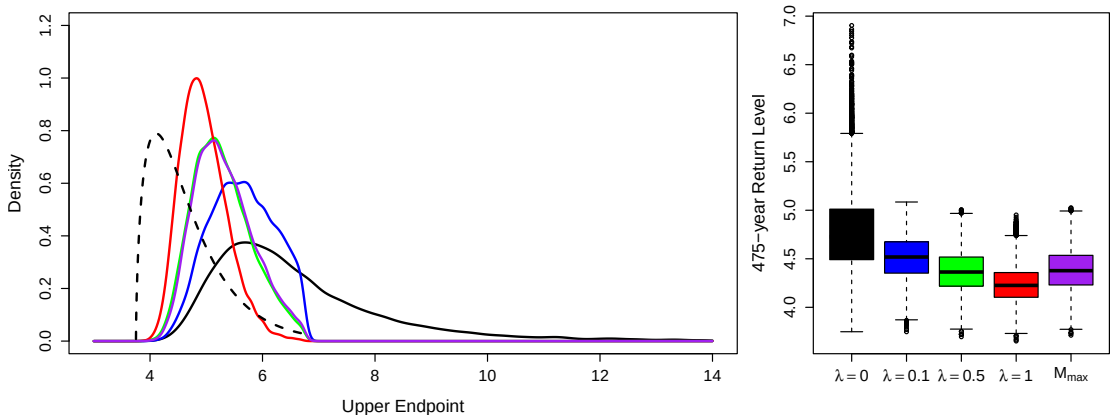

Figure 8: The estimated posterior distributions: density of x_e (left) and boxplots of the 475-year return level (right), for penalty weights $\lambda = 0, 0.1, 0.5, 1$ (black, blue, green, red). The purple curve and box correspond to the expert-informed beliefs about $M_{\max}$, being used as the prior for the endpoint x_e.

7 Conclusion and Discussion

This paper presents a methodology for enhancing the estimation of the upper tail of a distribution by incorporating experts' beliefs about an upper bound on the endpoint of the distribution of interest. The use of the GPD has been explored extensively due to its effectiveness in modelling the tail behaviour of extreme events. In the context of human-induced earthquakes, we find that by incorporating experts' opinions into the statistical framework, our approach refines the upper tail estimates of the magnitude distribution, which is crucial for seismic hazard assessments.

The proposed methodology integrates expert beliefs over $M_{\max}$ (a physical upper bound on earthquake magnitudes) into the likelihood function by penalising overestimated values of the GPD upper endpoint x_e. This approach leverages trust in expert knowledge about potential physical upper limits of earthquake magnitudes, contributing to more robust estimates of the upper tail of the magnitude distribution. Additionally, the Bayesian implementation using the penalised likelihood framework

allows the incorporation of prior distributions for the GPD parameters, in addition to incorporating information about $M_{\max}$. To further account for uncertainty in expert-specified bounds, we also explored the use of the expert-informed $M_{\max}$ distribution as a prior for x_e. This provides a complementary strategy for incorporating expert beliefs within the Bayesian framework, where we place greater trust in the expert's ability to produce tight bounds for x_e but at the risk of each of these experts' bounds is higher than the true endpoint.

A crucial aspect of this study is bridging the gap between expert-informed $M_{\max}$ constraints and empirical approaches based on the asymptotically justified GPD statistical model. While expert knowledge provides valuable insights into physical limits, relying solely on $M_{\max}$ without data-driven validation can lead to overconfidence in extreme magnitude estimates. Conversely, the empirical approach without upper constraints may lead to bias estimates of the risk of extreme earthquakes. By integrating experts' constraints into the statistical framework, our methodology provides a balanced solution that maintains empirical flexibility while preventing unrealistic extrapolations. Furthermore, our use of the two different ways to incorporate the experts' knowledge enables an assessment of the sensitivity to the results to the level of trust we place on expert judgement.

Application to the Groningen earthquake catalogue illustrates the practical utility of the methodology. Using both likelihood and Bayesian inferences, we obtained estimates for x_e and return levels. As the penalty weight λ given to the experts' beliefs about $M_{\max}$ increases, both the posterior distribution for x_e and the 475-year return level becomes more concentrated with decreasing mean values. We observed that the results obtained using MLE and Bayesian inference were very similar, even for high quantiles and endpoints. This similarity can be attributed to the use of bootstrap methods for MLE, rather than using the more typically used asymptotic normality results. When applying the $M_{\max}$ distribution as a prior for x_e, the resulting posterior estimates closely matched those from the penalised likelihood method with moderate regularisation ($\lambda = 0.5$). This suggests that the conclusions are not overly sensitive to the specific formulation used to incorporate expert knowledge.

To guide the selection of the penalty weight λ, we propose methods for determining a maximum degree of distortion for the model fit to include the experts' beliefs. The results from the simulation study highlight the importance of selecting an appropriate $\lambda > 0$ value when using penalised likelihood methods. Smaller values of λ provide a balance between incorporating expert knowledge and maintaining flexibility in the model, resulting in more accurate estimates for high quantiles and the upper endpoint. Conversely, larger λ values overly constrain the model, leading to underestimation of key parameters. This suggests that careful calibration of the penalty weight is critical for achieving reliable results in practical applications. Related to the choice of λ is the choice of threshold. Here we used a threshold function found by Varty et al. (2021), which was found to optimal in terms of EQD fit, see Section 4.2, for the GPD, i.e., when $\lambda = 0$ and provided an excellent fit for the GPD. Finding the optimal threshold function whilst allowing for $\lambda > 0$ is an interesting line of future research.

While our analysis assumes that the magnitude records are free from measurement error, this assumption does not hold in practice. Measurement error, which may vary over time/space due to changes in detection systems, could significantly influence the results (Lin & Newberry 2023). In future work, we aim to address this issue by incorporating a model for measurement error into the magnitude estimation process. By accounting for the temporal/spatial variation in the detection system's sensitivity, we could refine the analysis and improve the robustness of the magnitude modelling.

Overall, the proposed methodology contributes to the field of seismic risk assessment by demonstrating the value of incorporating expert knowledge into statistical modelling. Its applicability, however, extends beyond seismology. For instance, probable maximum precipitation in hydrology and meteorology plays a role analogous to $M_{\max}$ in seismology, representing the physical upper limit for extreme rainfall events. By applying this methodology to rainfall studies, experts' belief distributions could be incorporated to improve the assessment of extreme precipitation (Papalexiou & Koutsoyiannis 2006, Kunkel et al. 2013), demonstrating the potential for multi-disciplinary applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the two anonymous referees for their constructive comments and suggestions, which have significantly improved the clarity and presentation of the paper. The authors also would like to thank Stephen Bourne (Shell) for providing the support and supplying the earthquake catalogue. This paper is based on work completed while Wanchen Yue was part of the EPSRC funded STOR-i centre for doctoral training (EP/S022252/1).

References

Aki, K. (1965), 'Maximum likelihood estimate of b in the formula $\log 10N = a - bM$ and its confidence limits', *Bulletin of Earthquake Research* **43**(3), 237–239.

Beirlant, J., Kijko, A., Reynkens, T. & Einmahl, J. H. (2019), 'Estimating the maximum possible earthquake magnitude using extreme value methodology: the Groningen case', *Natural Hazards* **98**(3), 1091–1113.

Brune, J. N. (1968), 'Seismic moment, seismicity, and rate of slip along major fault zones', *Journal of Geophysical Research* **73**(2), 777–784.

Chinnery, M. A. (1969), 'Earthquake magnitude and source parameters', *Bulletin of the Seismological Society of America* **59**(5), 1969–1982.

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.

Coles, S. G. & 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.

Davison, A. C. & Smith, R. L. (1990), 'Models for exceedances over high thresholds', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **52**, 393–425.

Dost, B. & Kraaijpoel, D. (2013), 'The August 16, 2012 earthquake near Huizinge (Groningen)', KNMI, de Bilt, the Netherlands.

Dryden, I. L. & Zempléni, A. (2006), 'Extreme shape analysis', *Journal of the Royal Statistical Society Series C: Applied Statistics* **55**(1), 103–121.

Ellsworth, W. L. (2013), 'Injection-induced earthquakes', *Science* **341**(6142), 1225942.

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

Evans, K. F., Zappone, A., Kraft, T., Deichmann, N. & Moia, F. (2012), 'A survey of the induced seismic responses to fluid injection in geothermal and CO2 reservoirs in Europe', *Geothermics* **41**, 30–54.

Galis, M., Ampuero, J. P., Mai, P. M. & Cappa, F. (2017), 'Induced seismicity provides insight into why earthquake ruptures stop', *Science Advances* **3**(12), eaap7528.

Gutenberg, B. & Richter, C. F. (1944), 'Frequency of earthquakes in California', *Bulletin of the Seismological Society of America* **34**(4), 185–188.

Kagan, Y. Y. (2002), 'Seismic moment distribution revisited: I. statistical results', *Geophysical Journal International* **148**(3), 520–541.

KNMI (2020), 'Aardbevings catalogus', `https://www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus`.

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79–86.

Kunkel, K. E., Karl, T. R., Easterling, D. R., Redmond, K., Young, J., Yin, X. & Hennon, P. (2013), 'Probable maximum precipitation and climate change', *Geophysical Research Letters* **40**(7), 1402–1408.

Leonelli, M. & Gamerman, D. (2020), 'Semiparametric bivariate modelling with flexible extremal dependence', *Statistics and Computing* **30**(2), 221–236.

Lin, Q. & Newberry, M. (2023), 'Seeing through noise in power laws', *Journal of the Royal Society Interface* **20**(205), 20230310.

McGarr, A. (2014), 'Maximum magnitude earthquakes induced by fluid injection', *Journal of Geophysical Research: Solid Earth* **119**(2), 1008–1019.

Murphy, C., Tawn, J. A. & Varty, Z. (2025), 'Automated threshold selection and associated inference uncertainty for univariate extremes', *Technometrics* **67**(2), 215–224.

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.

Ogata, Y. (1988), 'Statistical models for earthquake occurrences and residual analysis for point processes', *Journal of the American Statistical Association* **83**(401), 9–27.

Opitz, T., Huser, R., Bakka, H. & Rue, H. (2018), 'INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles', *Extremes* **21**(3), 441–462.

Paleja, R., Bierman, S. & Jones, M. (2016), 'Impact of production shut-in on interevent time in Groningen. a statistical perspective. Shell report. Royal Dutch Shell (the Hague): 35pp'.

Papalexiou, S. M. & Koutsoyiannis, D. (2006), 'A probabilistic approach to the concept of probable maximum precipitation', *Advances in Geosciences* **7**, 51–54.

Pickands, J. (1975), 'Statistical inference using extreme order statistics', *The Annals of Statistics* **3**(1), 119–131.

Raschke, M. (2015), 'Modeling of magnitude distributions by the generalized truncated exponential distribution', *Journal of Seismology* **19**(1), 265–271.

Simpson, D., Rue, H., Riebler, A., Martins, T. G. & Sørbye, S. H. (2017), 'Penalising model component complexity: A principled, practical approach to constructing priors', *Institute of Mathematical Statistics* **32**, 1–18.

Smith, R. L. (1985), 'Maximum likelihood estimation in a class of nonregular cases', *Biometrika* **72**, 67–90.

Varty, Z., Tawn, J. A., Atkinson, P. M. & Bierman, S. (2021), 'Inference for extreme earthquake magnitudes accounting for a time-varying measurement process', arXiv:2102.00884.

Vere-Jones, D., Robinson, R. & Yang, W. (2001), 'Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation', *Geophysical Journal International* **144**(3), 517–531.

Vlek, C. (2019), 'Rise and reduction of induced earthquakes in the Groningen gas field, 1991–2018: statistical trends, social impacts, and policy change', *Environmental Earth Sciences* **78**(59).

Weng, H., Ampuero, J.-P. & Buijze, L. (2021), Physics-based estimates of the maximum magnitude of induced earthquakes in the Groningen gas field, *in* 'EGU General Assembly 2021', pp. EGU21–6144.