Environmetrics

WILEY

INVITED ARTICLE OPEN ACCESS

Spatio-Temporal Modelling of Extreme Induced Seismicity in the Presence of An Evolving Measurement Network

Conor Murphy[1] | Jonathan A. Tawn[1] | Zak Varty[2] | Ross Towe[3] | Peter M. Atkinson[1]

[1]School of Mathematical Sciences, Lancaster University, Lancaster, UK | [2]Department of Mathematics, Imperial College London, London, UK | [3]Shell Information Technology International Ltd., London, UK

Correspondence: Conor Murphy (murphyconor148@gmail.com)

Received: 6 August 2025 | **Revised:** 4 March 2026 | **Accepted:** 16 March 2026

Keywords: earthquakes | extreme values | gas field | generalised Pareto distribution | geophones | magnitude of completion | threshold selection | uncertainty quantification

ABSTRACT

Earthquakes induced by injecting or extracting gas from underground reservoirs can pose a significant hazard to surrounding infrastructure and populations. Safeguarding against future seismic hazards requires accurate models for the upper tail of the earthquake magnitude distribution that are able to represent various intervention strategies. For these models, we need efficient inference methods and reliable estimates of uncertainty. We propose a novel extreme value modelling procedure, which uses known changes in the earthquake measurement network, to automatically select a parametric spatio-temporal extreme value threshold which accounts for undetected earthquake values. We introduce methods to propagate the uncertainties in the extreme value model parameters, the threshold parameters and the threshold functional formulation through to future hazard estimates. We apply our methodology to the earthquake catalogue from the Groningen gas field in the Netherlands, delivering clear improvements over existing analyses and providing the first quantification of the different sources of uncertainty in such estimates. The procedure has the potential to be useful for a broad range of extreme value contexts to account for threshold uncertainty when parametric threshold models are used, or where data are missing due to limitations in measurement equipment.

1 | Introduction

Extraction or injection of gases from or into underground reservoirs of porous rock causes 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 exist significant risks to public safety and of damage to infrastructure (Ellsworth 2013), so accurately estimating the distribution of induced earthquake magnitudes under future extraction or injection scenarios is of paramount importance. Such estimates can be used to inform the construction or reinforcement of infrastructure to 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, after a series of damaging earthquakes, extraction has now stopped (despite substantial remaining gas reserves), but earthquakes continue to occur, with one of the largest ever recorded occurring in November 2025, 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. We

---

focus our inferences on two particular quantities of practical interest. 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 a substantial body of geophysical techniques (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 they can cause significant damage in close proximity to their locations despite their size. 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 the reliable, available information. For Groningen, the spatial region has been defined by geophysicists taking into account the spatial extent of the hazard from induced earthquakes in the gas field.

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. An earthquake is detected only if its magnitude is large enough that its location may be identified by the geophone network. Investment in this network over time has resulted in much denser spatial coverage of geophones that 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 of geophones around location $\boldsymbol{x}$ at time t, it can only be estimated using the observed earthquake catalogue.

To avoid bias, only detected earthquakes with magnitudes that exceed the estimated magnitude of completion, denoted by $\hat{m}_c(\boldsymbol{x}, t)$, for that earthquake's 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 $m_c(\boldsymbol{x}, t)$ is vital. In estimating $m_c(\boldsymbol{x}, t)$, there is a bias-variance trade-off which affects design level and $M_{\max}$ inferences; too low an estimate incorporates a set of observed earthquake magnitudes whose distribution deviates from the distribution of true earthquakes due to some true earthquakes being undetected (which biases inferences); while over-estimation excludes valid data (leading to unnecessary variance).

As earthquake magnitudes are recorded on a logarithmic-scale, they have $\mathbb{R}$ as a domain. Existing methods of inference focus on modelling magnitude excesses above some constant value and, hence, have a domain of $\mathbb{R}_+$. Since the very smallest magnitude events are not of practical interest, we focus our presentation on the distribution of true positive-magnitude earthquakes (detected and undetected together) at any $(\boldsymbol{x}, t)$ through the choice of a parametric distribution function $F_0(y; \boldsymbol{x}, t, \boldsymbol{\theta}_{F_0})$ for $y > 0$ and with $\boldsymbol{\theta}_{F_0}$ the vector of unknown parameters for F_0. This choice of 0 as the constant is arbitrary provided that $m_c(\boldsymbol{x}, t) \geq 0$ over $\mathcal{X} \times \mathcal{T}$, and we will explain why it makes no difference to our end inferences later in the paper. The most common assumption is that true earthquake magnitudes are mutually independent and that both F_0 and the ability to detect earthquakes are both invariant to all $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}$ (Ogata 1988). Hence, true earthquakes have independent and identically distributed (IID) magnitudes and $m_c(\boldsymbol{x}, t) = m_c$ over $\mathcal{X} \times \mathcal{T}$. The undetected earthquakes below m_c make fitting F_0 to the detected data above 0 subject to substantial bias due to the informative nature of the missing data. So under these idealised assumptions, m_c and F_0 are estimated as follows. For a choice of parametric family for F_0 and a candidate selection $m > 0$ for m_c, then the sample of earthquake magnitudes which exceed m are realisations of IID variables with conditional distribution function F_m where

$$F_m(y; \boldsymbol{\theta}_{F_0}) = \frac{F_0(m + y; \boldsymbol{\theta}_{F_0}) - F_0(m; \boldsymbol{\theta}_{F_0})}{1 - F_0(m; \boldsymbol{\theta}_{F_0})} \text{ for } y > 0.$$

Then, $\hat{m}_c$ is selected as the value of m for which there is deemed to be the best fit between $\hat{F}_m$ and the observed excesses of m, whilst accounting for the trade-off between bias and variance. Once $\hat{m}_c$ has been obtained, $\hat{F}_{\hat{m}_c}$ together with the empirical rate per year of magnitude exceedances of $\hat{m}_c$ are used for subsequent inferences for risk assessments.

Even in the above idealised setting, inference for $\hat{m}_c$ and $\hat{F}_{\hat{m}_c}$ are complicated as (i) the parametric family for F_0 is unknown, (ii) no automated selection procedure exists for obtaining $\hat{m}_c$ which balances bias and variance considerations, and (iii) the inferences of $\hat{F}_{\hat{m}_c}$ do not take account of the uncertainty in $\hat{m}_c$ and so will typically lead to under-estimated uncertainty. Additionally, this idealised formulation does not exploit the substantial investment to the geophone network across $\mathcal{T}$, as there is a strong justification for spatio-temporal changes in $m_c(\boldsymbol{x}, t)$. Furthermore, it does not allow for F_0 to change with $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}$; an identically distributed F_0 is too simplistic for induced earthquake magnitudes where there is evidence of space-time variation in F_0 due to changes in the incremental stress field (Bourne and Oates 2020; Richter et al. 2020). Our paper addresses all of these limitations.

Of the issues raised above, first we discuss the fundamental choice of parametric family for F_0. The typical, and essentially the default, choice for F_0 is an exponential distribution as motivated by the Gutenberg-Richter law (Gutenberg and Richter 1956). The memoryless property of the exponential implies that F_m will also be exponentially distributed with the same rate parameter. Natural extensions of the exponential distribution are lower-tail bounded distributions, such as the Weibull and gamma distributions. Both of these have exponential upper tail decays so, as a model for F_m, they do not offer much additional flexibility. Other work has addressed concerns about the exponential distribution being unbounded in its upper tail, which does not align with the physical understanding that the seismic energy that 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), the tapered Gutenberg-Ritcher distribution (Vere-Jones et al. 2001), and the extended generalised Pareto distribution (Papastathopoulos and Tawn 2013). None of these distributions has 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. Critically, unlike the exponential distribution, they do not possess the threshold stability closure property that F_m is of the same parametric form (with potentially different parameters) as F_0 for any $m > 0$. Since we propose a model for F_0 above an arbitrary baseline of 0 and use the excesses of the magnitude of completion to fit this model, a lack of threshold stability substantially undermines each of the above listed alternative families of distributions as a feasible model in this instance.

The only entire family of continuous distributions which possesses the required threshold stability closure property is the generalised Pareto distribution (GPD), with this family containing the exponential distribution as a special case (Davison and Smith 1990; Coles 2001). This is the distribution that both Varty et al. (2021) and Yue, Tawn, et al. (2025) used for modelling the Groningen magnitude data (under the assumption that the true earthquakes are IID) based on the following additional arguments: (i) the GPD allows for lighter and heavier tails than the exponential distribution in a flexible and parsimonious manner including the finite upper-endpoint case which is consistent with physical knowledge of $M_{\max}$; (ii) there is good empirical evidence for this choice, both for Groningen and other regions (Shcherbakov et al. (2019)); (iii) it allows accurate estimation of the upper tail and extrapolation beyond what has previously been observed; and (iv) it has a strong asymptotic motivation from extreme value methods for upper tail modelling. We adopt the GPD for the distribution F_0, of magnitudes over the arbitrary level of 0, primarily due to its unique property of threshold stability with the above factors being strong secondary reasons. In an extension to Varty et al. (2021), we exploit the knowledge from Bourne and Oates (2020) that the parameters of F_0 vary over $(\boldsymbol{x}, t)$ with the changing stresses due to gas extraction through the use of a spatio-temporal stress covariate, which has been derived by geophysicists using a computationally-intensive numerical model. Using this covariate, we can estimate the occurrence rate of earthquakes exceeding different magnitudes into the future under various extraction scenarios.

The selection of an appropriate magnitude of completion function $m_c(\boldsymbol{x}, t)$ is the next challenge. The vast majority of previous approaches take this function to be a constant over $\mathcal{X} \times \mathcal{T}$ and assume that earthquakes are IID with F_0 being exponential (e.g., Mignan and Woessner (2012)). Other simplifying assumptions for $m_c(\boldsymbol{x}, t)$ have been used, for example, Hutton et al. (2010) and Das et al. (2012) used a piecewise-constant function with changepoints at pre-determined times and spatial regions, respectively. These methods are likely to be inefficient as (a) knowledge of the geophone network is not exploited, (b) the results will be sensitive to the choice of the pre-determined neighbourhoods, (c) the exponential distribution has practical limitations, and (d) the assumption of being identically distributed may not hold. One notable exception is Mignan et al. (2011), who focused on estimating $m_c(\boldsymbol{x}, t)$ as a spatial function $m_c(\boldsymbol{x})$, and overcame inefficiencies (a)–(c) above by using the surface distance from $\boldsymbol{x}$ to the ith nearest geophone as a covariate, for $i = 3 - 5$ and avoiding using an explicit model of F_0 through use of the non-parametric maximum curvature method (Wyss et al. 1999).

Varty et al. (2021) proposed a range of new developments for estimating $m_c(\boldsymbol{x}, t)$ by taking it to be the threshold for a GPD. At first sight, the selection of the magnitude of completion and the GPD threshold may appear very different issues. In particular, the magnitude of completion is the smallest level m for which there is not a statistically significant departure of the detected magnitude excesses of m from distribution F_m (as derived from F_0), while the GPD threshold is the smallest level u that provides threshold-stability in the GPD fit at all higher levels than u. However, these are identical problems when F_0 is assumed to be a GPD. In particular, Varty et al. (2021) proposed using an automated method for the selection of the threshold for a GPD with a solely temporally-varying function, that is, $m_c(\boldsymbol{x}, t) := m_c(t)$. They assumed that $m_c(t)$ follows a parametric sigmoid function over time, with four parameters. This choice of parametric form is somewhat arbitrary, but accounts for empirical evidence of a smooth transition between time periods with constant values. Although information about changes in the Groningen geophone network is not used in their analysis, the estimated magnitude of completion is consistent with periods of known investment into the network and almost doubled the number of excesses over $m_c(t)$ relative to using the established best constant estimate of this function. ReferenceMurphy et al. (2025) showed that a variant of the Varty et al. (2021) threshold selection method outperforms the leading automated techniques for fitting a GPD in the IID context.

Our research builds upon that of Varty et al. (2021) and Murphy et al. (2025) to provide an automated threshold selection procedure which overcomes each of the inefficiencies (a)–(d) of previous approaches. We focus on developing the first automated method for making inference on the spatio-temporal function $m_c(\boldsymbol{x}, t)$, which accounts explicitly for the known details of the evolving geophone network for the Groningen gas-field. We propose to estimate $m_c(\boldsymbol{x}, t)$ as the threshold of a covariate-dependent GPD. To simplify notation and align with standard terminology for the GPD threshold, we subsequently use $u(\boldsymbol{x}, t)$ in place of $m_c(\boldsymbol{x}, t)$ for our inference, but refer to existing methods as estimating $m_c(\boldsymbol{x}, t)$. Within our methodology, we allow for $u(\boldsymbol{x}, t)$ to be determined by a range of functional forms of the covariate $V_i(\boldsymbol{x}, t)$, which is the three-dimensional distance from $\boldsymbol{x}$ to the ith nearest geophone available at time t, and the choice of i for $1 \leq i \leq 4$. Hence, we use a more informative spatio-temporal set of covariates than Mignan et al. (2011). We find statistical and practically significant impact from using both $V_i(\boldsymbol{x}, t)$ as a covariate for the threshold and a gas extraction covariate in the GPD. Our fitted model provides an excellent fit to the data with a lower threshold than previously and the novel ability to predict earthquake risk under different future gas extraction scenarios.

Critically this paper is novel 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 paper is structured as follows. In Section 2, we present details of the Groningen seismic data and the geophone network. Section 3 reviews the extreme value methods that underpin the use of the GPD and outlines the associated threshold selection approaches that we build on. In Section 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 $\mathcal{X} \times \mathcal{T}$. In Section 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 6. We finish with a discussion in Section 7. A Supporting Information document provides additional information as well as a glossary of the key notation used in this paper.

2 | Groningen Data

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 geophysicists. 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 1. The catalogue includes the event hypocentre (a three-dimensional location of surface position and depth), time, and magnitude denoted, respectively, by $(\boldsymbol{x}_k, t_k, y_k)$ for $k = 1, \dots, n$. 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 being 3-4 km. Figure 1 shows the event magnitudes over time and their spatial locations separately. The magnitudes are recorded on the local magnitude ($\mathrm{M_L}$) scale, a logarithmic scale used to measure the energy released by an earthquake; hence the few negative recorded values shown in the figure, which arise when an earthquake has a hypocentre very close to a geophone. As discussed in Section 1, we model F_0, the distribution of true magnitude excesses of $0\mathrm{M_L}$, so such negative magnitudes are excluded from our analysis, but the majority of the Groningen catalogue is relevant. Like Yue, Tawn, et al. (2025), we use a catalogue with magnitudes reported to at least two decimal places and the temporal precision is to a within a second of time resolution. Typically, catalogues report magnitudes to one decimal place, resulting in a dataset of rounded observations, which pose an added modelling challenge addressed by Varty et al. (2021). There is uncertainty in the measurement of $\boldsymbol{x}_k$ and y_k, but not t_k, with this uncertainty potentially changing over time due to improvements in the precision of individual geophones in the network. We ignore these features in this analysis as geophysicists anticipate that such uncertainty will be small by comparison to the uncertainties addressed in this paper.

Figure 1 (left) shows that over time the number of recorded earthquakes has increased as has the proportion of small earthquakes. We model both of these aspects in Section 4. There are other noteworthy features, such as temporal clustering of events, for example, just before 2010 there is a small cluster of events which correspond to multiple aftershocks following a large earthquake—we do not attempt to model this feature here. Figure 1 (left) also shows the lines corresponding to 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, Tawn, et al. 2025). The former is accepted as a conservative estimate of m_c, after accounting for rounding, whereas the latter is a boundary case of the sigmoid function of Varty et al. (2021) as used by Yue, Tawn, et al. (2025). 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 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

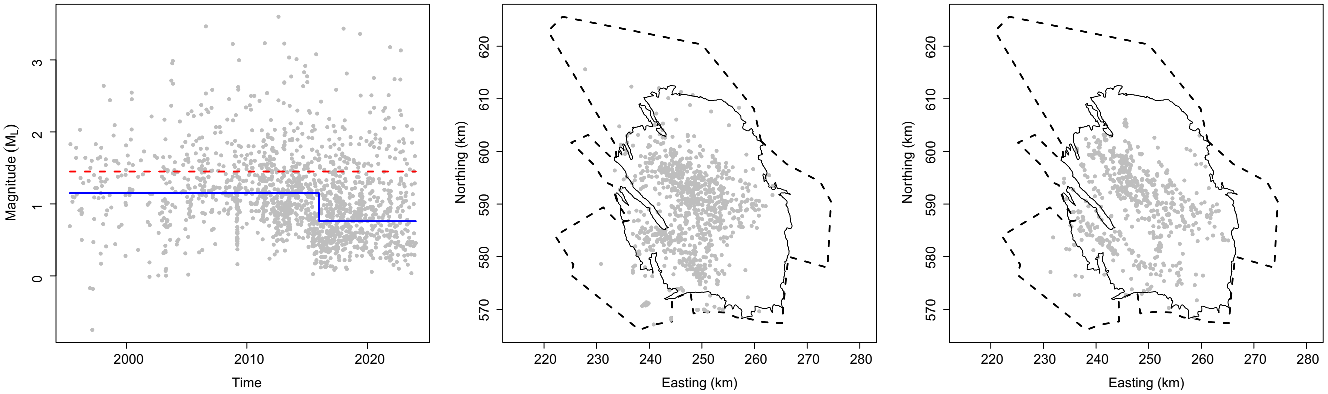

FIGURE 1 | Temporal and spatial features of the induced earthquakes in the Groningen catalogue: [left] magnitudes (in $\mathrm{M_L}$) against date of occurrence with two previously used $m_c(t)$: a conservative level of $m_c = 1.45$ (red-dashed line) and a changepoint (solid blue line). Occurrence locations of events in the region of interest $\mathcal{X}$ (dashed line) are shown for before [centre] and after [right] the time of the temporal changepoint that is shown in the left panel. The boundary of gas field $\mathcal{G}$ (solid line) is also shown for reference.

0.187 (0.013) and 0.448 (0.020) respectively. There are two potential reasons for this increase: improvements to the geophone network and the impact of gas extraction. We explore each of these aspects in Sections 2.2 and 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 panels of Figure 1 show that there is no discernible substantial change in the locations of earthquakes occurring before and after December 25, 2015. They also show that most of the detected earthquakes occurring within $\mathcal{X}$ are located within $\mathcal{G}$ and there are clear sub-regions of seismic activity in $\mathcal{G}$ where the vast majority of the earthquake activity is focused.

2.2 | The Geophone Network and the Magnitude of Completion

The Royal Netherlands Meteorological Institute (KNMI) measures seismic activity across the Netherlands through an extensive network of geophones (KNMI 2020). For this study, a geophone dataset was available containing their locations (in RD coordinates), depths and dates of operation over the Netherlands. We use a subset of the country, denoted by $\mathcal{R}$, as the plotting domain in Figure 2, with $\mathcal{X} \subset \mathcal{R}$. This dataset was not available to Varty et al. (2021); they had access only to knowledge of the time-window of the major developments across the network between 2014–2017. Figure 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 slow growth in the number of geophones prior to 2014, then massive growth in the period 2014–2017. There have also been some smaller changes post-2017 for $\mathcal{G}$ and $\mathcal{X}$, and ongoing evolution in the network in $\mathcal{R} \setminus \mathcal{X}$. To illustrate 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 that 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 1 and 3, respectively. Some geophones even exist outside of the region $\mathcal{X}$, with their locations selected to improve detection of events occurring in $\mathcal{X}$. Individual geophones can cease operating and so the total number of active geophones in a year can decrease from previous years, for example, as can be seen from Figure 2 with operating numbers decreasing from previous years in 2005, 2014, 2015, and after 2020.

We exploit understanding of the spatio-temporal variation in detection capability by measuring distances in three-dimensional space and incorporating the operation times of individual geophones, with our measure given by $V_i(\boldsymbol{x}, t)$ for the distance to the i^{th} nearest geophone from an earthquake at $(\boldsymbol{x}, t)$ for $1 \leq i \leq 4$. The inclusion of the depth data is particularly helpful in capturing differences in the distances to geophones when earthquakes are close to a geophone in surface distance. We use $V_i(\boldsymbol{x}, t)$, over $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}$ for a range of i, as a covariate for estimating $u(\boldsymbol{x}, t)$. Unlike Mignan et al. (2011), we use V_i in our inference for a parametric model of the distribution of exceedances of $u(\boldsymbol{x}, t)$ via its impact on $u(\boldsymbol{x}, t)$ itself. We also consider $i = 1, 2$ as these may better reflect the spatio-temporal changes in the magnitude of completion, particularly for the early period of $\mathcal{T}$ when there were very few geophones. Using $V_i(\boldsymbol{x}, t)$ in this way opens up the first possibility for spatio-temporal inference for $u(\boldsymbol{x}, t)$. This enables a lowering of $u(\boldsymbol{x}, t)$ for some t and $\boldsymbol{x}$ relative to previous studies at Groningen which improves the efficiency of all subsequent inferences based on exceedances of the $u(\boldsymbol{x}, t)$ function.

2.3 | Extraction Stress Covariate

Following Bourne and Oates (2020), Smith et al. (2022) and Kaveh et al. (2024), we use a key physical covariate in our analysis to describe smooth changes in the distribution and occurrence rates of large magnitude earthquakes as a result of gas extraction. The covariate, that we denote by $s(\boldsymbol{x}, t)$, is the Kaiser stress (KS) which, for each $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}$, is the maximum difference up to time t of the vertically averaged stress at $\boldsymbol{x}$ from the initial stress state at the start of gas extraction in $\mathcal{G}$. The KS covariate is available on a 500 m × 500 m resolution spatial grid over $\mathcal{X}$ (at a fixed depth of 3 km) on a monthly basis over $\mathcal{T} \cup \mathcal{T}_F$, where $\mathcal{T}_F$ is the period from February 2024–January 2055. These KS values are derived from complex computationally intensive

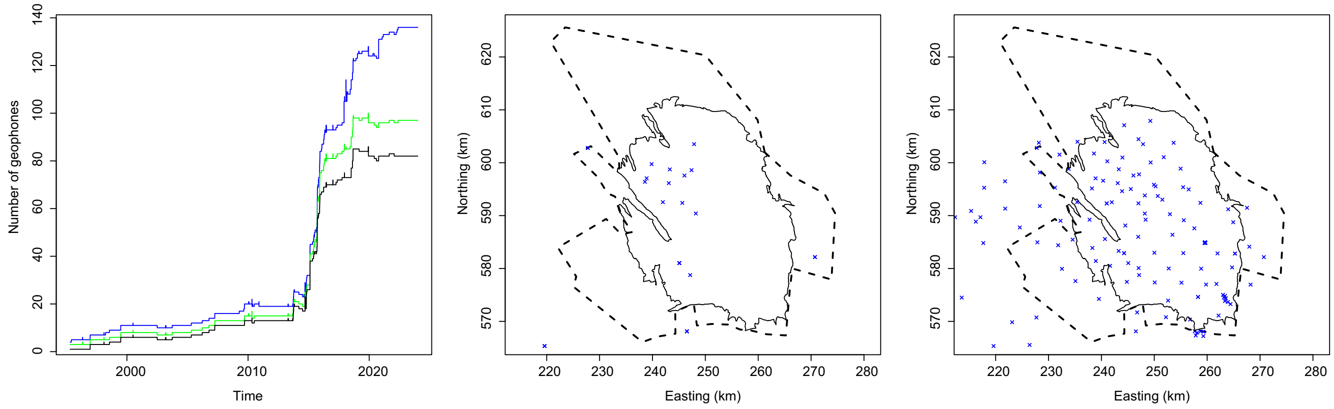

FIGURE 2 | Features of the Groningen region geophone network present over time and space: [left] number of geophones in operation daily throughout $\mathcal{T}$ within the regions $\mathcal{R}$ (blue), $\mathcal{X}$ (green) and $\mathcal{G}$ (black); [centre] and [right] the respective locations (blue crosses) of geophones in operation in the years 2010 and 2020.

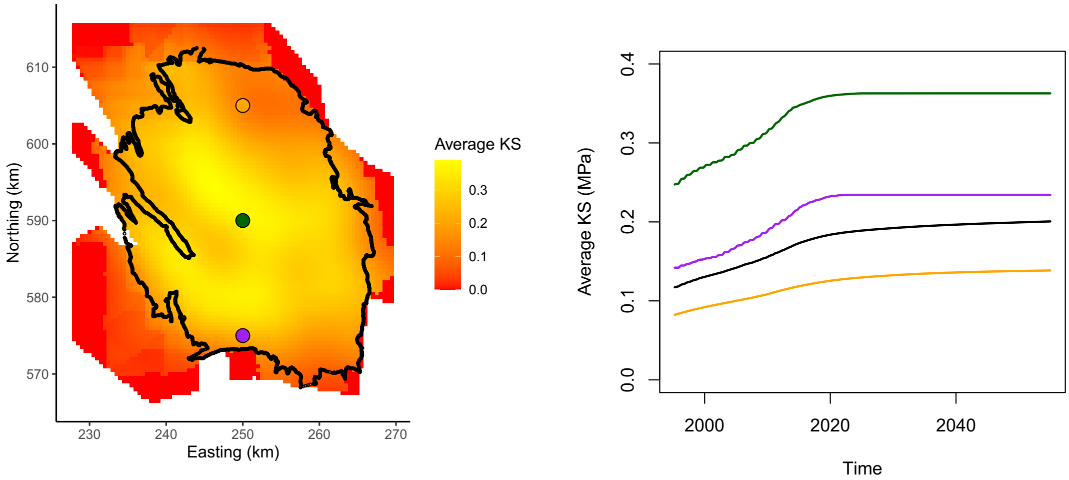

FIGURE 3 | Features of the Kaiser stress (KS) covariate field: [left] Average KS for 2020, for a fixed depth of 3 km, across all of $\mathcal{G}$ (solid black line) and earthquake active region $\mathcal{X} \setminus \mathcal{G}$. [right] Temporally varying spatial average of KS over region shown in the left panel (black) and location specific $s(\boldsymbol{x}, t)$ series for three sites shown as dots on left panel: top (red), middle (green) and bottom (blue), with the colours of lines matching the colours of location dots on left panel.

geophysics-based rock reservoir models which are driven by the known spatio-temporal data on gas extraction for Groningen. The KS values are calibrated for the period $\mathcal{T}$ by using measurements of the time varying stress fields taken using boreholes and seismic imaging. The projected values for $\mathcal{T}_F$ that we use are derived under the assumption that there is no further extraction from the gas-field, but KS for a range of other extraction scenarios can also be derived.

Tang and Hudson (2010) and Zang et al. (2014) provided arguments for KS being an informative covariate for earthquake occurrences and their magnitudes. They point out that no seismic activity can occur at a particular location until the maximum previous stress level at that location is exceeded, since that maximum stress will have already caused all feasible earthquakes. There is a limitation in that argument, as KS does not account for local induced stresses in time and space caused by earthquakes themselves, which can cause aftershocks. This is not of major concern for Groningen, as only 20% of all earthquakes were classified as aftershocks using the classification method of Zaliapin et al. (2008). We ignore such local changes in KS in our analysis. A simple exploratory analysis, provided in Figure S.1 in the Supporting Information, shows evidence that KS accounts for differences in the distribution of the observed magnitudes of earthquakes, justifying its consideration in our statistical modelling in Section 4.2.

We denote the KS field for the operation time period by $\mathcal{S} = \{s(\boldsymbol{x}, t) : s(\boldsymbol{x}, t) \geq 0, \boldsymbol{x} \in \mathcal{X}, t \in \mathcal{T}\}$ and the forecast period by $\mathcal{S}_F = \{s(\boldsymbol{x}, t) : s(\boldsymbol{x}, t) \geq 0, \boldsymbol{x} \in \mathcal{X}, t \in \mathcal{T}_F\}$. 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 $k = 1, \dots, n$), 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 3 [left] provides KS averaged over the year 2020 for a fixed depth of 3 km, presented for all $\mathcal{G}$, with some values in $\mathcal{X} \setminus \mathcal{G}$ not shown as these values are zero. A comparison of the regions of highest KS with the spatial locations for the catalogue events shown in Figure 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 3 [right] shows KS at a fixed depth of 3 km averaged over the spatial region shown in Figure 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.

3 | Underpinning Extreme Value Methods

Consider a univariate random variable Y with continuous distribution function F, 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 over a high threshold u, that is, $(Y - u) | (Y > u)$, where $F_u(y) = [F(u + y) - F(u)]/[1 - F(u)]$ for $y > 0$. Specifically, Pickands (1975) showed 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

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

with $y > 0$, $w_+ = \max(w, 0)$, the shape parameter $\xi \in \mathbb{R}$ determined by F, and the scale parameter $\sigma \in \mathbb{R}_+$. This distribution is denoted by $\text{GPD}(\sigma, \xi)$. A key property of the GPD for our modelling is its distributional stability when the threshold is changed. Suppose that for some threshold u, $(Y - u) | (Y > u) \sim \text{GPD}(\sigma_u, \xi)$. Then, for any threshold v, with $u \leq v < y^F$, it follows that

$$(Y - v) | (Y > v) \sim \text{GPD}(\sigma_v, \xi), \text{ with } \sigma_v = \sigma_u + \xi(v - u), \tag{2}$$

see Coles (2001). So, the GPD family and the shape parameter are stable to a change in threshold, although the scale parameter changes. The GPD is the only family of univariate continuous distributions that satisfies such threshold stability, with the $\xi = 0$ case (taken as the limit as $\xi \to 0$) of the exponential distribution resulting in the stronger lack of memory property.

A key 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 can result in very few threshold excesses to estimate parameters, leading to unnecessary variance (Coles 2001). Even in the context of IID data, this is not an easy task, and has been the focus of much research e.g., Scarrott and MacDonald (2012), with a recent review of the variety of methods, some automated, provided by Belzile et al. (2023).

Applying the limit distribution G as an approximation for the excesses over a selected 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:

$$F(y) = 1 - \lambda_u[1 - G_u(y - u; \sigma_u, \xi)], \tag{3}$$

for $y > u$ with unknown threshold exceedance (intensity) 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: $\xi = 0$ as stated above 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 $(1-p)$th-quantile of Y, denoted by $y_p(\boldsymbol{\theta}_u)$, where Y has a GPD tail, satisfies $F(y_p(\boldsymbol{\theta}_u)) = 1 - p$. So,

$$y_p(\boldsymbol{\theta}_u) = \begin{cases} u - \sigma_u\left[1 - (p/\lambda_u)^{-\xi}\right]/\xi & \text{for } \xi \neq 0, \\ u - \sigma_u \log(p/\lambda_u) & \text{for } \xi = 0, \end{cases} \tag{4}$$

when $p < \lambda_u$, that is, $y_p(\boldsymbol{\theta}_u) > u$. Inference for the GPD above threshold u is well established using likelihood (Davison and Smith 1990) and Bayesian (Coles and Tawn 1996) approaches, with comparisons made recently by Yue, Tawn, et al. (2025) for incorporating a penalty function to account for experts' knowledge on the distribution of M_{max}, that is, an upper bound for the GPD upper endpoint y^F. Specifically, Yue, Tawn, et al. (2025) 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.

There are also well-established extreme value methods for non-identically distributed variables and, in particular, when covariates $\boldsymbol{Z}$ exist, observed as $\boldsymbol{z}$, and affect the parameters of the GPD above a threshold function, $u(\boldsymbol{z})$, which can vary with $\boldsymbol{z}$ (Kyselý et al. 2010; Northrop and Jonathan 2011; Yue, Guo, et al. 2025). Specifically, it is assumed that, for $y > u(\boldsymbol{z})$,

$$\Pr(Y > y | \boldsymbol{Z} = \boldsymbol{z}) = 1 - \lambda_u(\boldsymbol{z})\left(1 + \xi(\boldsymbol{z})\frac{y - u(\boldsymbol{z})}{\sigma_u(\boldsymbol{z})}\right)_+^{-1/\xi(\boldsymbol{z})}, \tag{5}$$

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

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

For a given threshold function, the typical approaches for modelling the functional forms of the parameters use 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, although Eastoe and Tawn (2009) showed that for the threshold stability property (2) to hold across covariate values, the identity link is required. It is relatively standard to assume that $\xi(\boldsymbol{z})$ is constant, that is, $\xi(\boldsymbol{z}) = \xi$ 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).

Our spatio-temporal approach for threshold selection extends the purely temporal analysis of Varty et al. (2021). Their method was developed specifically for the Groningen catalogue, in which the underlying variables were assumed to be IID by a GPD, with the missing data accounted for temporally through a time-varying threshold. Varty et al. (2021) addressed 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 and approximate the integrated absolute error (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). Furthermore, in the context of IID GPD data, Murphy et al. (2025) proposed a novel automated selection procedure for a constant threshold, termed the *expected quantile discrepancy* (EQD), which selected the threshold value, by similarly approximating the IAE, and selecting 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. A brief description of the EQD method is provided in Supporting Information S:3. The adjustments we make to the EQD approach for our modelling context are discussed in Section 4.

4 | Statistical Modelling and Inference

4.1 | Threshold Model

To properly account for 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 structural model form for $u(\boldsymbol{x}, t)$. The magnitude of completion varies as a function of the Euclidean distance $r \geq 0$ from the hypocentre of an earthquake to a geophone. With a slight abuse of notation, we denote this relationship by $m_c(r)$. In idealised conditions, such as uniform rock types, and with a single geophone detector, geophysicists have identified that, for $r > 0$,

$$m_c(r) = \phi_0 + \phi_1 \log r + \phi_2 r \tag{6}$$

for constants (ϕ_0, ϕ_1, ϕ_2) determined by the rock properties and the shock wave amplitude (Freudenreich et al. 2012; Gaucher 2016). The log and linear terms are linked to shock-wave attenuation and geometric spreading, respectively. Depending on the conditions, simplifications of relationship (6) have been proposed, with Goertz et al. (2012) and Demuth et al. (2016) setting $\phi_0 = 0$ and $\phi_1 = 0$, respectively.

Here, we choose to consider the components of relationship (6), $\log r$ and r, individually. Additionally, we consider an intermediate form in $r^{1/2}$. Combined with these choices of transformations, we take $r = V_i(\boldsymbol{x}, t)$ where $V_i(\boldsymbol{x}, t)$ is the spatio-temporal covariate, introduced in Section 2.2, corresponding to the three-dimensional Euclidean distance from $\boldsymbol{x}$ to the ith nearest geophone in the network at time t. Further reasoning for these choices is given in Supporting Information S: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, $V_i(\boldsymbol{x}, t)$, $i = 1 - 4$, as input to the covariate model for the GPD threshold $u(\boldsymbol{x}, t)$. Specifically, we consider the threshold functions:

$$\begin{aligned} A_i &: \; u(\boldsymbol{x}, t) = \alpha_0 + \alpha_1 V_i(\boldsymbol{x}, t); \\ B_i &: \; u(\boldsymbol{x}, t) = \alpha_0 + \alpha_1 \log(V_i(\boldsymbol{x}, t)); \\ C_i &: \; u(\boldsymbol{x}, t) = \alpha_0 + \alpha_1 \sqrt{V_i(\boldsymbol{x}, t)}, \end{aligned} \tag{7}$$

for $i = 1 - 4$, where $(\alpha_0, \alpha_1) \in \mathbb{R} \times \mathbb{R}_+$, with the restriction $\alpha_1 \geq 0$ reflecting that threshold does not increase as the network becomes denser.

Each of the resulting 12 (types $A - C$ and $i = 1 - 4$) threshold models (7) captures the changing magnitude of completion over time, in a broadly similar way to the sigmoid model of Varty et al. (2021). Critically, our formulation provides a physical basis for the spatio-temporal variations in the threshold function through covariates that capture the physical processes of earthquake measurement. Consequently, it is more parsimonious, with two parameters to estimate rather than the four of the sigmoid of Varty et al. (2021), despite our threshold capturing richer spatio-temporal variation and allowing for non-identically distributed excesses. For subsequent inference, we face the dual challenges of fitting the spatio-temporal parametric threshold while also accounting for model choice uncertainty over the 12 different possible formulations.

4.2 | Distributional Model within the Observed Period $\mathcal{T}$

We require a single distributional model of the magnitudes of the true earthquakes. The term "true earthquakes" includes all detected and undetected earthquakes. For the reasons discussed in Section 1, we assume that the distribution of true earthquake magnitudes above 0, that is, F_0, is in the GPD parametric family (1). This choice of $0\mathrm{M_L}$ is made for mathematical convenience, but in practice, any constant threshold below the minimum of $u(\boldsymbol{x}, t)$ over $\mathcal{X} \times \mathcal{T}$ will give identical inferences to our choice due to the threshold stability property (2) of this distribution. We use $Y_0(\boldsymbol{x}, t)$ to denote the magnitude of true earthquake excesses of $0\mathrm{M_L}$ for hypocentre at $(\boldsymbol{x}, t)$ and only fit this model to data excesses of the spatio-temporal threshold $u(\boldsymbol{x}, t)$ as it is only above this threshold that the detected magnitudes are deemed to correspond to all of the true magnitudes.

Unlike Ogata (1988), and the majority of the literature referenced in Section 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 2.3 of the recent use of the spatio-temporal covariate KS and our findings from the exploratory analysis presented in Figure 3, we propose a simple formulation for the inclusion of $s(\boldsymbol{x}, t)$ in the GPD parameterisation. As discussed in Section 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 F_0 that the true earthquake magnitude distribution above 0 is given by

$$\begin{aligned} &Y_0(\boldsymbol{x}, t) \sim \mathrm{GPD}(\sigma_0(\boldsymbol{x}, t), \xi), \text{ with } \sigma_0(\boldsymbol{x}, t) = \beta_0 + \beta_1 s(\boldsymbol{x}, t), \\ &\quad \text{for } (\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}, \end{aligned} \tag{8}$$

where $(\beta_0, \beta_1, \xi) \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}$, where $\beta_1 \geq 0$ since extraction leads to a stochastic increase in earthquake magnitudes.

Model (8) 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 $Y_0(\boldsymbol{x}, t)$. We focus on the excesses of $u(\boldsymbol{x}, t)$, with this threshold function given by one of the set of formulations (7). Exploiting the threshold stability property (2) results in the following conditional distribution of the excesses of the threshold $u(\boldsymbol{x}, t)$:

$$\begin{aligned} &[Y(\boldsymbol{x}, t) - u(\boldsymbol{x}, t)] | [Y(\boldsymbol{x}, t) > u(\boldsymbol{x}, t)] \sim \mathrm{GPD}(\sigma_u(\boldsymbol{x}, t), \xi), \\ &\quad \text{for } (\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}, \end{aligned} \tag{9}$$

where $\sigma_u(\boldsymbol{x}, t) = \sigma_0(\boldsymbol{x}, t) + \xi u(\boldsymbol{x}, t)$, with the scale and threshold given by models (8) and (7), respectively, and $\sigma_u(\boldsymbol{x}, t)$ is a function of parameters $(\alpha_0, \alpha_1, \beta_0, \beta_1) \in \mathbb{R}_+^4$. All parameters of our model for the threshold function and the GPD excesses of this function are entirely identifiable, though at first sight this may not appear to be the case. Concerns may arise given the structure of $\sigma_u(\boldsymbol{x}, t)$. In particular, its intercept is $\alpha_0 + \xi \beta_0$. However, we have extra information about the individual parameters (α_0, ξ) from inferences of the threshold function $u(\boldsymbol{x}, t)$ and the GPD distribution shape parameter. Also, in terms of the regression parameters, as $V_i(\boldsymbol{x}, t)$ and $s(\boldsymbol{x}, t)$ appear in separate components of the model and have sample correlation of -0.5, there are no identifiability concerns there either.

4.3 | Intensity Model for Threshold Exceedances

In addition to expressing parametric models for $u(\boldsymbol{x}, t)$ and F_0, in terms of parameters (α_0, α_1) and (β_0, β_1, ξ), respectively, we need to model the spatio-temporal rate of exceedance of earthquake magnitudes over $u(\boldsymbol{x}, t)$, which we denote by $\lambda_u(\boldsymbol{x}, t)$. As we know only that $u(\boldsymbol{x}, t) \geq 0$, we need both a model for the baseline intensity of true earthquakes above $0\mathrm{M_L}$, which we denote by $\lambda_0(\boldsymbol{x}, t)$, and one for the rate of true earthquakes, occurring at $(\boldsymbol{x}, t)$, when they have magnitudes exceeding $u(\boldsymbol{x}, t)$. We consider these elements of our model for $\lambda_u(\boldsymbol{x}, t)$ below, but first we give the established approach for modelling the intensity of true earthquakes.

Bourne et al. (2018) present a parametric geophysics-based covariate model formulation for the intensity function $\lambda_m(\boldsymbol{x}, t)$ of true earthquakes over a constant magnitude, m. When the gas reservoir is of constant thickness over $\mathcal{X}$, which is a reasonable first order approximation for the Groningen gas field, their model simplifies to

$$\lambda_m(\boldsymbol{x}, t; \gamma_0, \gamma_1) = \frac{\partial s(\boldsymbol{x}, t)}{\partial t} \exp[\gamma_0 + \gamma_1 s(\boldsymbol{x}, t)], \quad \text{for } (\boldsymbol{x}, t) \in \mathcal{X} \times (\mathcal{T} \cup \mathcal{T}_F), \tag{10}$$

with parameters $(\gamma_0, \gamma_1) \in \mathbb{R}^2$, where the temporal partial derivative term for the KS covariate $s(\boldsymbol{x}, t)$ is evaluated using finite differencing. For a fixed hypocentre, the intensity model (10) increases exponentially with $s(\boldsymbol{x}, t)$ when its temporal derivative is positive. To fit this intensity model, they avoid issues with undetected earthquakes by taking m to be the conservative estimate of the magnitude of completion, that is, $m = 1.45\mathrm{M_L}$, and they estimate the parameters (γ_0, γ_1) using the subset $\{(t_k, \boldsymbol{x}_k) : y_k > 1.45, k = 1, \ldots, n\}$ of catalogue values, by making the assumption that these are realisations of a Poisson process with intensity $\lambda_{1.45}(\boldsymbol{x}, t; \gamma_0, \gamma_1)$ over $\mathcal{X} \times \mathcal{T}$. To illustrate the implications of intensity model (10), Figure 3 [right] shows how KS varies over time for three selected locations, with KS being constant since 2021 for the southerly and centre locations. Hence, whatever values of (m, γ_0, γ_1) are used, the resulting intensity $\lambda_m(\boldsymbol{x}, t; \gamma_0, \gamma_1) = 0$ at these locations since $\partial s(\boldsymbol{x}, t)/\partial t = 0$, that is, no earthquakes with a magnitude above $0\mathrm{M_L}$ could occur at these locations after 2021. We model the baseline intensity $\lambda_0(\boldsymbol{x}, t)$ using formulation (10) with $m = 0$ and if we detect an earthquake at $(\boldsymbol{x}, t)$ with $\lambda_0(\boldsymbol{x}, t; \gamma_0, \gamma_1) = 0$, we deem that earthquake to be an aftershock and omit it from the subsequent inference for the intensity function.

We cannot estimate the function $\lambda_0(\boldsymbol{x}, t; \gamma_0, \gamma_1)$ directly using the Poisson process inference methods of Bourne et al. (2018) because $u(\boldsymbol{x}, t) \geq 0$ so any such inference would be biased due to any undetected magnitudes in region $\{0 < y < u(\boldsymbol{x}, t) : (\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}\}$. To overcome this limitation of the data, we estimate $\lambda_0(\boldsymbol{x}, t; \gamma_0, \gamma_1)$ using only magnitudes above $u(\boldsymbol{x}, t) \geq 0$. Specifically, the intensity function of exceedances of $u(\boldsymbol{x}, t)$ reduces from $\lambda_0(\boldsymbol{x}, t; \gamma_0, \gamma_1)$ by the factor $1 - F_0(u(\boldsymbol{x}, t); \boldsymbol{x}, t)$, that is, the probability that an earthquake with a magnitude above 0, at hypocentre $(\boldsymbol{x}, t)$, exceeds $u(\boldsymbol{x}, t)$, Hence, our intensity model for hypocentres and occurrence times associated with earthquakes with magnitudes exceeding our spatio-temporal threshold $u(\boldsymbol{x}, t)$ is given by

$$\begin{aligned} \lambda_u(\boldsymbol{x}, t; \boldsymbol{\theta}) &= \lambda_0(\boldsymbol{x}, t; \gamma_0, \gamma_1)[1 - F_0(u(\boldsymbol{x}, t)] \\ &= \lambda_0(\boldsymbol{x}, t; \gamma_0, \gamma_1) \left[1 + \xi \frac{u(\boldsymbol{x}, t)}{\sigma_0(\boldsymbol{x}, t)}\right]_+^{-1/\xi}, \end{aligned} \tag{11}$$

where $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \beta_0, \beta_1, \xi, \gamma_0, \gamma_1)$ and the second equality is due to our modelling of F_0 by the GPD model formulation (8).

4.4 | Inference and Diagnostics

Here, we detail our inferences for the parameters $\boldsymbol{\theta}$ 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 levels; the GPD parameters (β_0, β_1, ξ); the threshold parameters (α_0, α_1); and the threshold formulation $(A - C, i = 1 - 4)$ and for the latter, we optimise over parameters (γ_0, γ_1) of the baseline intensity model formulation (10). Given the threshold function choice, inferences on (β_0, β_1, ξ) and (γ_0, γ_1) are orthogonal and so it is possible to make inference on these separately.

For the combined threshold-distributional model, the inference strategy is as follows. Firstly, we select one of the 12 specific functional threshold formulations (7) and minimise the EQD metric, of Section 3, 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 $K_u = \{k \in \{1, \ldots, n\} : y_k > u_k\}$, where $u_k := u(\boldsymbol{x}_k, t_k)$, and the vector of the associated exceedances $\boldsymbol{y}_u = \{y_k : k \in K_u\}$. Letting $\sigma_{u,k} := \sigma_u(\boldsymbol{x}_k, t_k)$, then for a given choice of (α_0, α_1), the GPD likelihood for (β_0, β_1, ξ) is

$$L((\beta_0, \beta_1, \xi); \boldsymbol{y}_u, (\alpha_0, \alpha_1)) = \prod_{k \in K_u} \left\{ \frac{1}{\sigma_{u,k}} \left(1 + \xi(y_k - u_k)/\sigma_{u,k}\right)_+^{-1-1/\xi} \right\}. \tag{12}$$

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 $\hat{\xi}_u$. Despite our strong belief about the existence of a finite upper endpoint, and hence $\xi < 0$, we do not impose this constraint on the parameter space when maximising the likelihood (12). Our reasons for this choice are that we wanted to find the evidence in the data alone that suggests that the upper tail is shorter than the established exponential distribution for magnitude modelling. In Section 7, we discuss possible ways of imposing a finite endpoint if that was desired.

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). Here, we outline our approach for evaluation of the EQD metric for the parameter models detailed in Sections 4.1 and 4.2. 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 $(X_u, \boldsymbol{t}_u, \boldsymbol{s}_u) := \{(\boldsymbol{x}_k, t_k, s(\boldsymbol{x}_k, t_k) : k \in K_u\}$. We resample with replacement the rows of the array $(\boldsymbol{y}_u, X_u, \boldsymbol{t}_u, \boldsymbol{s}_u)$ to obtain bootstrapped samples (Davison and Hinkley 1997), 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 the likelihood (12) to obtain parameter estimates $(\hat{\beta}_0^b, \hat{\beta}_1^b, \hat{\xi}_u^b)$ 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 (9), 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 kth component $y_{u,k}^{E,b} = F_{Exp}^{-1}\{G(y_{u,k}^b - u_k; \hat{\sigma}_{u,k}^b, \hat{\xi}_u^b)\}$, where F_{Exp}^{-1} is the inverse distribution function of a standard exponential and G is the GPD distribution function (1). If the threshold is 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

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

$$d_b(\alpha_0, \alpha_1) = \frac{1}{m}\sum_{j=1}^{m}\left|F_{Exp}^{-1}\left(\frac{j}{m+1}\right) - Q\left(\frac{j}{m+1};\ \boldsymbol{y}_u^{E,b}\right)\right|, \tag{13}$$

in the EQD method, where $Q(j/(m+1); \boldsymbol{y}_u^{E,b})$ denotes the $j/(m+1)$ empirical quantile of $\boldsymbol{y}_u^{E,b}$. 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 $\mathcal{X} \times \mathcal{T}$, that is, the elements of the set K_u need not be nested across different choices of (α_0, α_1). Although not immediately obvious, the EQD penalises candidate thresholds that are unnecessarily high as the metric refits the GPD parameters for each bootstrapped data sample above that threshold. When the candidate threshold is too high, these bootstrapped parameter estimates create high variability in the values of $y_{u,k}^{E,b}$ leading to the EQD value being larger than when a lower, more appropriate threshold is used.

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, we detail our procedure for accounting for the various uncertainties in this inference procedure, that is, the uncertainty in the GPD parameter estimates, the uncertainty in the threshold parameter estimation, and the uncertainty due to the selection of the functional formulation of the threshold.

Using the estimated values $\hat{\boldsymbol{\theta}}_{-\gamma} := (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}_0, \hat{\beta}_1, \hat{\xi}_{\hat{u}})$ obtained, we estimate the parametric intensity model for $\lambda_u(\boldsymbol{x}, t; \boldsymbol{\theta})$, given by expression (11) using only earthquake data with magnitudes above $\hat{u}(\boldsymbol{x}, t)$. Specifically, letting $\hat{\boldsymbol{\theta}}_\gamma = (\hat{\boldsymbol{\theta}}_{-\gamma}, \gamma_0, \gamma_1)$ we have

$$\lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}_\gamma) = \lambda_0(\boldsymbol{x}, t; \gamma_0, \gamma_1)\left[1 + \hat{\xi}_{\hat{u}}\frac{\hat{u}(\boldsymbol{x}, t)}{\hat{\sigma}_0(\boldsymbol{x}, t)}\right]_+^{-1/\hat{\xi}_{\hat{u}}}, \tag{14}$$

where the estimated functions $\hat{u}$ and $\hat{\sigma}_0$ are functions of $(\hat{\alpha}_0, \hat{\alpha}_1)$ and $(\hat{\beta}_0, \hat{\beta}_1)$, 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 $\hat{u}$, that is, using the likelihood

$$L(\gamma_0, \gamma_1) \propto \left(\prod_{k=1}^{n}\lambda_{\hat{u}}(\boldsymbol{x}_k, t_k; \hat{\boldsymbol{\theta}}_\gamma)^{I(y_k > \hat{u}_k)}\right)$$

$$\exp\left(-\int_{\boldsymbol{x}\in\mathcal{X}}\int_{t\in\mathcal{T}} \lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}_\gamma)\ \mathrm{d}t\mathrm{d}\boldsymbol{x}\right), \tag{15}$$

with $\lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}}_\gamma)$ given by expression (14) and $I(A)$ the indicator function of event A. Then, $(\hat{\gamma}_0, \hat{\gamma}_1)$ are obtained by maximising the likelihood (15). 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 2021), but, as discussed in Section 4.3, we do not incorporate aftershocks into this likelihood.

The estimated magnitude of completion function $\hat{u}(\boldsymbol{x}, t)$ defines a dataset of its exceedances with catalogue values indexed by $K_{\hat{u}}$, as defined in Section 4.4. 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 spatio-temporal density of earthquakes as these are of key importance. We use only the subset indexed by $K_{\hat{u}}$ 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 (9). We define the transformed values by $\{\hat{y}_k^E : k \in K_{\hat{u}}\}$, where

$$\hat{y}_k^E = -\log\{[1 + \hat{\xi}(y_k - \hat{u}_k)/\hat{\sigma}_{u,k}]_+^{-1/\hat{\xi}_{\hat{u}}}\} \text{ for } k \in K_{\hat{u}}, \tag{16}$$

which will be a realisation of an approximately exponential(1) distributed variable if the fit is adequate. Hence, we can assess the fit of the model through standard quantile-quantile (QQ) plot techniques. To assess the performance of $\lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}})$, our estimated spatio-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

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

$$\text{and } \Lambda_{\hat{u}}(\boldsymbol{x}, T; \hat{\boldsymbol{\theta}}) = \int_{t\in T} \lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}})\ \mathrm{d}t. \tag{17}$$

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.

4.5 | Inference for Future Extreme Magnitude Events

Section 4.4 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, although 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, that is, the period $\mathcal{T}_F$. For this scenario, we have access to the geophysical model-based predictions for the KS covariate over the period $\mathcal{T}_F$, that is, $\mathcal{S}_F$, although, as noted in Section 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 $\sigma_0(\boldsymbol{x}, t|\mathcal{S}_F)$ and $\lambda_0(\boldsymbol{x}, t|\mathcal{S}_F)$, respectively.

Under an assumption of temporal stationarity and ignoring any spatial variation, Beirlant et al. (2019) estimated the upper-endpoint of the magnitude distribution. Under an identically distributed GPD assumption, the upper endpoint corresponds to the finite value of $u - \sigma_u/\xi$ only when $\xi < 0$. Under our GPD covariate model of Section 4.2, the upper endpoint varies temporally and spatially into the future with form

$$\begin{aligned} e(\boldsymbol{x}, t|\mathcal{S}_F) &:= u(\boldsymbol{x}, t) - \sigma_u(\boldsymbol{x}, t|\mathcal{S}_F)/\xi \\ &= -\sigma_0(\boldsymbol{x}, t|\mathcal{S}_F)/\xi = -[\beta_0 + \beta_1 s(\boldsymbol{x}, t)]/\xi, \text{ for } \xi < 0, \end{aligned} \tag{18}$$

and $e(\boldsymbol{x}, t|\mathcal{S}_F) = \infty$ for $\xi \geq 0$ for all $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}_F$. To provide endpoint values which are practically useful for planning infrastructure maintenance and reinforcement, we consider two summaries of the endpoint function (18) through its maximum $e_{\max}(\mathcal{S}_F)$ and a weighted average $e_{\text{wm}}(\mathcal{S}_F)$. The definition of these quantities are given in Supporting Information S:8.

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. We consider earthquakes with magnitudes exceeding level v in the 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 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.

We focus on the extreme event of the form $R_v(\mathcal{X}, \mathcal{T}_F) = \{\boldsymbol{x} \in \mathcal{X}, t \in \mathcal{T}_F, y \in \mathbb{R}_+ : y > v\}$, as 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 $R_v(\mathcal{X}, \mathcal{T}_F)$ is given by

$$\Lambda_v(\mathcal{X}, \mathcal{T}_F|\mathcal{S}_F) := \int_{\boldsymbol{x}\in\mathcal{X}} \int_{t\in\mathcal{T}_F} \lambda_0(\boldsymbol{x}, t|\mathcal{S}_F) \left[1 + \xi \frac{v}{\sigma_0(\boldsymbol{x}, t|\mathcal{S}_F)}\right]_+^{-1/\xi} \mathrm{d}t \, \mathrm{d}\boldsymbol{x}. \tag{19}$$

Letting $N_v(\mathcal{X}, \mathcal{T}_F|\mathcal{S}_F)$ 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 $\mathcal{X} \times \mathcal{T}_F$ is given by

$$\Pr(N_v(\mathcal{X}, \mathcal{T}_F|\mathcal{S}_F) = 0) = \exp[-\Lambda_v(\mathcal{X}, \mathcal{T}_F|\mathcal{S}_F)]. \tag{20}$$

Hence, for a level of risk specified by Code (2005), we require that v is such that $\Lambda_v(\mathcal{X}, \mathcal{T}_F|\mathcal{S}_F) = -\log(0.9387)$. We estimate the level v by solving this equation with the parameters of the statistical models replaced by values estimated using the methods of Section 4.4.

5 | Uncertainty Quantification

Similarly to Murphy et al. (2025), we use bootstrapping methods (Davison and Hinkley 1997) 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 nested algorithms which capture uncertainty: (i) in the GPD parameter estimation and the parameters of the rate of exceedance of the threshold function, (ii) capturing the additional uncertainty from threshold parameters, and (iii) further incorporating the threshold formulation uncertainty. 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. Murphy et al. (2025) provided strong evidence that the proposed extensions of the basic bootstrap algorithm will provide useful quantifications of uncertainty. In particular, they demonstrated a substantial improvement in coverage probabilities of the resulting CIs for high quantiles in several IID case studies.

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, its value is unknown and subsequent tail inferences can be highly sensitive to its choice. Hence, we incorporate this additional source of uncertainty to ensure that CIs for seismic hazards are not too narrow. The algorithms detailed below provide methods for uncertainty quantification when estimating a summary of interest, which we denote by $w(\boldsymbol{\theta})$, for any of the quantities discussed in Section 4.5 or for other features of interest to geophysicists.

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 4.4. For each b of the B_{par} bootstraps, the number of exceedances $n_{\hat{u}}^b$ of the threshold function $\hat{u}$ is generated along with the corresponding hypocentre for each exceedance

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 independently the $n_{\hat{u}}^b$ hypocentres using density $\lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}})/\Lambda_{\hat{u}}(\hat{\boldsymbol{\theta}})$ 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 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 $w(\boldsymbol{\theta})$.
```

and the respective magnitude excess values. Here, $n_{\hat{u}}^b$ is a realisation of a Poisson $(\Lambda_{\hat{u}}(\hat{\boldsymbol{\theta}}))$ variable, where

$$\Lambda_{\hat{u}}(\hat{\boldsymbol{\theta}}) = \int_{\boldsymbol{x} \in \mathcal{X}} \int_{t \in \mathcal{T}} \lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}})\, \mathrm{d}t \mathrm{d}\boldsymbol{x} \tag{21}$$

and $\lambda_{\hat{u}}(\boldsymbol{x}, t; \boldsymbol{\theta})$ is given by expression (14). We sample $n_{\hat{u}}^b$ hypocentres independently according to the density $\lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}})/\Lambda_{\hat{u}}(\hat{\boldsymbol{\theta}})$, 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 (9) parameter estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\xi})$, to generate the parametric bootstrap sample of $n_{\hat{u}}^b$ 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 $\hat{\boldsymbol{\theta}}_\alpha^b = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}_0^b, \hat{\beta}_1^b, \hat{\xi}^b, \hat{\gamma}_0^b, \hat{\gamma}_1^b)$ with which we can estimate any summary of interest, $w(\hat{\boldsymbol{\theta}}_\alpha^b)$, and construct CIs as quantiles of the sample of bootstrap estimates.

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 (7). 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 (13) and employ Algorithm 1 to account for the uncertainty in the estimation of $(\beta_0, \beta_1, \xi, \gamma_0, \gamma_1)$. For the bth 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 $w(\boldsymbol{\theta})$.
```

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 $w(\boldsymbol{\theta})$.
```

The spatio-temporal formulation of the true threshold function is unknown, but, as motivated in Section 4.1, 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, allows the formulation of the threshold function (as given by expressions (7)) to vary for each bootstrapped resample of the observed data.

6 | Application to Groningen Earthquakes

6.1 | Threshold Model Selection and GPD Inference

We apply our developed model and inference methods to the Groningen earthquakes and covariate data described in Section 2. We use each of the 12 threshold function model formulations identified in Section 4.1, which cover all combinations of covariates ($V_i : i = 1, \dots, 4$) and transformations $A - C$. As only three geophones were active 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 $(m, B) = (500, 200)$, 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. Rerunning the threshold selection procedure with $B = 1000$ does not change the selected threshold formulation. We optimise the EQD metric, $d(\alpha_0, \alpha_1)$, over a fine grid of increments of 0.02 over the range $(\alpha_0, \alpha_1) \in [0, 1.5] \times [0, 0.8]$.

The EQD metric of fit for each of these 12 threshold function formulations is given in Table 1, presented as relative values in relation to the minimum obtained EQD metric. 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 forms B and C. The EQD values are very close suggesting that there is little 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*. A probable reason for the lack of discrimination between the V_i covariates is due to them being highly co-linear. For formulations $A - C$, Figure 4 [right] shows that the differences in the estimated threshold functions are small, apart from the period 1995–2000 when there were few earthquakes recorded. The estimated threshold functions are generally high in early periods of $\mathcal{T}$ due to the small number of geophones in operation then. However, there may be important differences between these estimated threshold models when inferences are made far into the tail of the distribution, so accounting for the uncertainty in the threshold functional form is vital for valid evaluation of the inference uncertainty. This is investigated in Section 6.3, but for now, we focus attention on models A_2, B_1, and C_2.

TABLE 1 | EQD values for each threshold model formulation, $A - C$ and V_i: $i = 1, \dots, 4$, relative to the best EQD value of 0.0321 for model A_2. The smallest EQD value for each formulation $A - C$ is given in italics.

Model	1	2	3	4
A	1.037	*1*	1.084	1.019
B	*1.019*	1.028	1.115	1.053
C	1.034	*1.025*	1.128	1.065

To help understand the form of the best fitted threshold function model A_2, Figures 4 and 5 provide temporal and spatial summaries, respectively. Figure 4 [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, that is, $\int_{\boldsymbol{x} \in \mathcal{G}} \hat{u}(\boldsymbol{x}, t)\, \mathrm{d}\boldsymbol{x}$ over $t \in \mathcal{T}$. As reference points, also included are the two previously studied conservative and changepoint thresholds, as given in Figure 1. 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 close 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 4 [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 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 sites' thresholds are closely aligned.

Figure 4 [right] identifies that the fitted threshold for A_2 is mostly above both B_1 and C_2. There is variability in both threshold differences throughout, with the differences diminishing after 2016, particularly for C_2. Despite these differences in the thresholds,

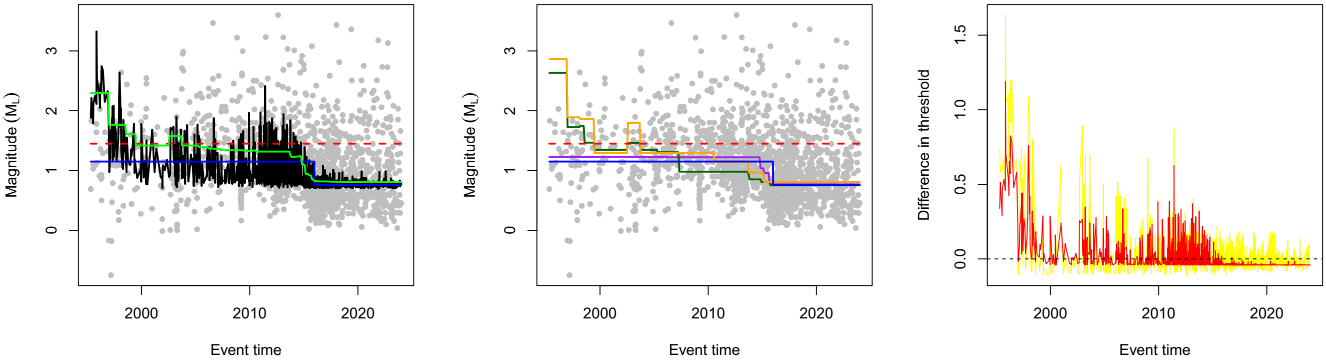

FIGURE 4 | Comparisons of fitted threshold functions. [Left] Best-performing threshold A_2 over time: (black) is $\hat{u}(\boldsymbol{x}_k, t_k)$ for the kth earthquake in the catalogue; (green) is the spatial average of the A_2 threshold across $\mathcal{G}$. Also on this plot is the conservative level 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 3: with the colours of the lines matching the Figure 3 [left] location dot colours. The conservative and changepoint thresholds are shown for reference. [right] Difference between thresholds: $A_2 - B_1$ (yellow) and $A_2 - C_2$ (red).

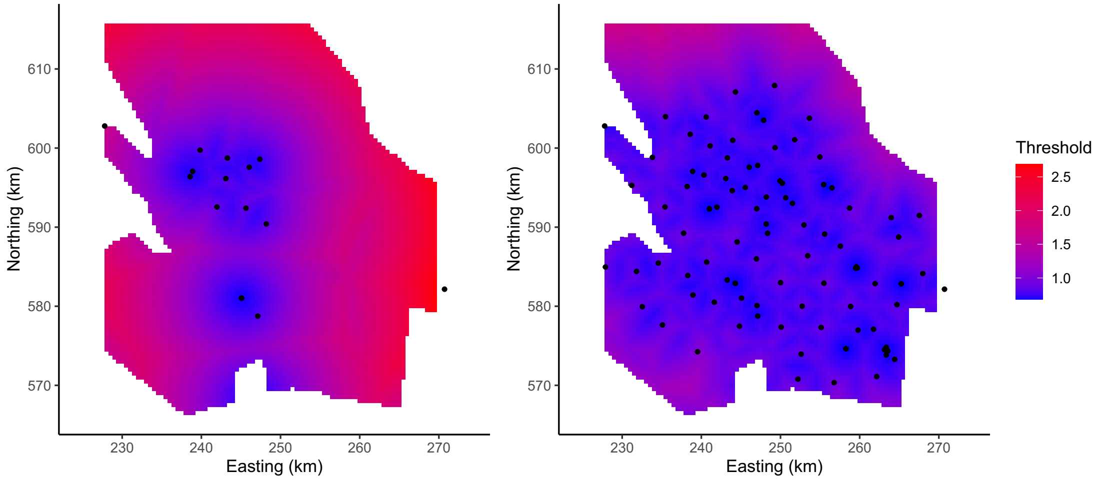

FIGURE 5 | Spatial plots of model A_2 threshold function $\hat{u}(\boldsymbol{x}, t)$ for $\boldsymbol{x} \in \mathcal{X}$ for the dates January 01, 2010 (left) and January 01, 2020 (right). Active geophones are shown as black dots.

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 of 1.45M$_\text{L}$ 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. See Section 6.3, and Figure 9 for a discussion of the CIs of the temporal features of these threshold estimates, with these CIs obtained using Algorithms 2 and 3 to account for different aspects of the model uncertainty.

Figure 5 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 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 1M$_\text{L}$. In the Supporting Information, Figure S.2 presents the estimated bootstrap standard errors for $\{\hat{u}(\boldsymbol{x}, t) : \boldsymbol{x} \in \mathcal{X}\}$ at the same two time points as in Figure 5, with the standard errors estimated using Algorithms 2 and 3 to account for different aspects of the model uncertainty. There is a slight increase in standard error found when using Algorithm 3 rather than Algorithm 2 as would be expected. The mean (max) ratios of the spatial standard errors for the two dates are 1.12 (3.20) and 1.36 (2.17), respectively. For both algorithms, the standard errors are at their lowest in rings surrounding the geophone locations at that time, due to a property of prediction in regression as explained in the Supporting Information.

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. The corresponding maximum likelihood estimates (and bootstrapped standard errors obtained using Algorithm 1) for ξ, a key parameter for extreme value inference, 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. The reduced standard errors show strong evidence that $\xi < 0$ and hence, provide evidence for the existence of a finite upper endpoint for the distribution of magnitudes.

As most previous analyses of earthquakes treat excesses of the threshold to be identically distributed, we also assess the significance of the KS covariate in the GPD scale parameter. Using the same methods and summaries as for ξ, we obtain that $\hat{\beta}_1^{A_2} = 0.984$ (0.375), $\hat{\beta}_1^{B_1} = 1.023$ (0.339) and $\hat{\beta}_1^{C_2} = 0.951$ (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. In the Supporting Information, Figure S.3 presents plots of $\hat{\sigma}_0(\boldsymbol{x}, t)$ over $\boldsymbol{x} \in \mathcal{X}$ for four values of $t \in \mathcal{T}$. These fitted values show a substantial practical impact from incorporating the KS covariate into the scale parameter model, which is found to vary by a factor of up to two.

To assess the global fit of the GPD of a given threshold function, we use standard QQ and probability-probability (PP) diagnostics, each with 95% tolerance intervals which show bounds on the stochastic variation on such plots if the data are exponential (1) or uniform samples, respectively, as is standard practice (Coles 2001). Here, we present the more informative QQ-plots, with the corresponding PP-plots for these three fits given in Figure S.4 of the Supporting Information. Figure 6 compares the QQ-plots for three different GPD threshold/covariate selection combinations. For the conservative threshold of $m_c = 1.45$M$_\text{L}$, we have two fits: under an IID assumption (left) and with the KS covariate included in the scale parameter formulation (centre). Lastly, the QQ-plot for our best fitting selected model is shown

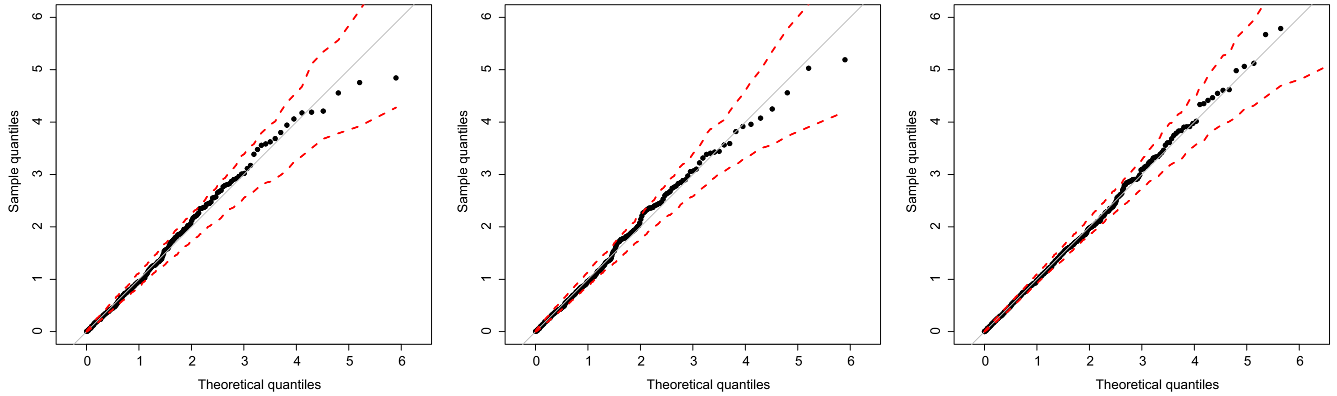

FIGURE 6 | QQ-plot assessments of various GPD model fits with excess values transformed onto standard Exponential margins: [left] IID GPD with conservative threshold 1.45$\mathrm{M_L}$, [centre] GPD with KS covariate also with threshold 1.45$\mathrm{M_L}$, and [right] GPD with KS covariate above the threshold function A_2. Pointwise 95% tolerance intervals are in red-dashed lines.

(right), that is, with the estimated threshold function A_2 and GPD model with the KS covariate. The IID fit with the conservative threshold acts as a baseline to compare against previous modelling of these data and provides a reasonable fit. Using the same constant threshold, but with the KS covariate included, leads to a visually similar quality of fit. For the threshold A_2, collectively the QQ- and PP-plots show an excellent fit across the whole distribution. These A_2 threshold tolerance intervals are much tighter than for the 1.45$\mathrm{M_L}$ threshold, so we have majorly increased the power for model selection. To recognise the full quality of the fit above the A_2 threshold, we need to recall that the fit uses only three parameters (β_0, β_1, ξ) to 849 data points, to capture the complex spatio-temporal structure in the data, so there is no risk of overfitting.

6.2 | Intensity Inference

Consider the fit of $\lambda_{\hat{u}}(\boldsymbol{x}, t; \hat{\boldsymbol{\theta}})$, the intensity model (14) for exceedances of the A_2 estimated threshold function. In Section 4.4, we outlined how we fitted model (11) using estimates from the GPD 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 identified in Section 2.3, when $s(\boldsymbol{x}, t)$ is constant between consecutive monthly values, model (10) gives a zero value, and we choose to interpret earthquakes at those times as after-shocks. Twenty-six such earthquakes were removed from our inference and diagnostics for the intensity model on this basis. With the remaining exceedances, we obtained estimates and associated bootstrap standard errors (in parentheses) of $\hat{\gamma}_0 = -0.4$ (0.2) and $\hat{\gamma}_1 = 15.6$ (0.7). We can compare our method to that of Bourne et al. (2018), which uses a 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 KS zero gradient events, 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 7 [left] and 8, respectively, show temporal and spatial intensity summaries $\Lambda^{\mathcal{X}}_{\hat{u}}(T; \hat{\boldsymbol{\theta}}_{\hat{u}})$ and $\Lambda_{\hat{u}}(\boldsymbol{x}, T; \hat{\boldsymbol{\theta}}_{\hat{u}})$ of Section 4.4 and the earthquakes that exceed the estimated threshold, excluding those deemed to be aftershocks. Figure 7 [left] compares the observed and expected annual earthquake counts above the threshold A_2 together with pointwise bootstrap 95% CIs for the expected value estimates found using Algorithms 2 and 3. Both sets of CIs are very similar, with those from Algorithm 3 slightly wider when the intensity peaks, and both intervals contain the observed counts throughout $t \in \mathcal{T}$. In particular, the fit captures the rapid growth in earthquakes above this threshold caused by a combination of gas extraction stresses increasing and expansion in the geophone network lowering the threshold. The estimates also follow closely the observed decline in excesses after 2015, despite the lowering of the threshold in this period. This reduction comes from the KS settling to a constant for each $\boldsymbol{x} \in \mathcal{X}$, due to the cessation of extraction. In Figure 7 [left] the close agreement between the observed and expected annual earthquake counts above estimated threshold A_2 suggests that A_2 is at least as large as the magnitude of completion over time and space. Figure 8 shows a similarly good fit on the intensity from the spatial perspective. Specifically, it presents $\Lambda_{\hat{u}}(\boldsymbol{x}, T; \hat{\boldsymbol{\theta}}_{\hat{u}})$, the integrated intensity over a year, for each grid box in $\mathcal{X}$ for the years 2010 and 2020. The estimated expected spatial intensity matches closely with the locations of earthquakes exceeding threshold A_2 in these years: observations are clustered around the two clear spatial peaks in the estimated intensity and, in 2020, a few events in the south-east of the region are centred on local intensity maxima.

In Figure 7 [centre], we compare the observed and expected numbers of earthquakes above 0$\mathrm{M_L}$, both excluding earthquakes with zero KS temporal gradient as mentioned above. The expected numbers shown are estimated under our model given by $\Lambda^{\mathcal{X}}_0(T; \hat{\boldsymbol{\theta}}_{\hat{u}})$, see expression (18), with the expected value being derived under the assumption that F_0 is GPD. We also provide pointwise bootstrap 95% CIs for these estimated expected values found using Algorithms 2 and 3, with little practical difference between these two types of intervals. The expected and observed counts per year T show a marked difference, even after accounting for the CIs for the expected value estimates, with the observed count less than the lower bound of the CI in all years because

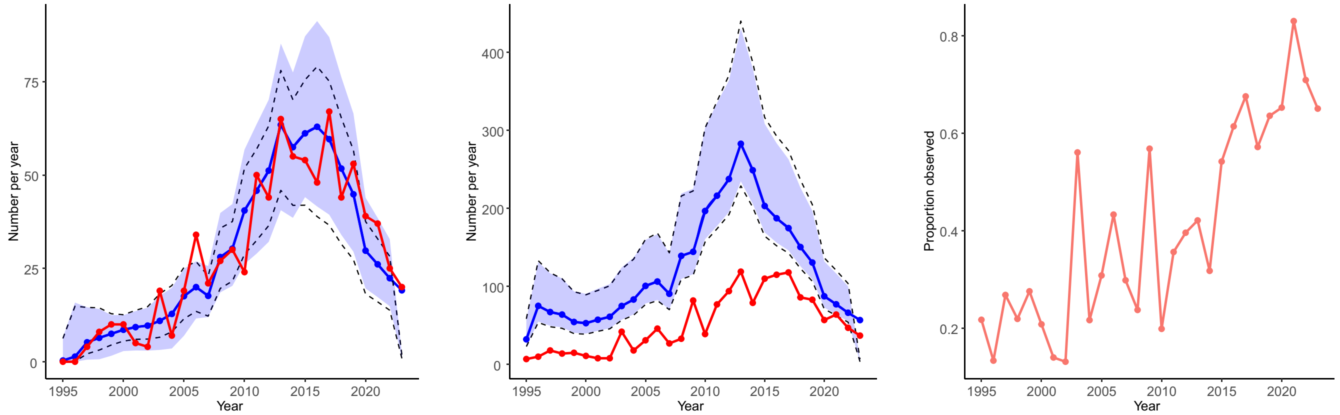

FIGURE 7 | Estimates of features of the occurrence properties of earthquakes for years $T = 1995, \ldots, 2023$: [left] observed and expected numbers of exceedances of threshold $\hat{u}(\boldsymbol{x}, t)$ per year based on estimated aggregated intensity $\Lambda_{\hat{u}}^{\mathcal{X}}(T)$ using model formulation A_2 (blue) and empirical estimate using 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 that an earthquake above $0\mathrm{M_L}$ is recorded, that is, the ratio of observed and expected annual earthquake counts above $0\mathrm{M_L}$ for each $T \in \mathcal{T}$. The left and centre panels show the pointwise bootstrap 95% CI for the model formulation estimate of $\Lambda_{\hat{u}}^{\mathcal{X}}(T)$ obtained using Algorithm 3 (as the shaded region) and using Algorithm 2 (as dashed lines).

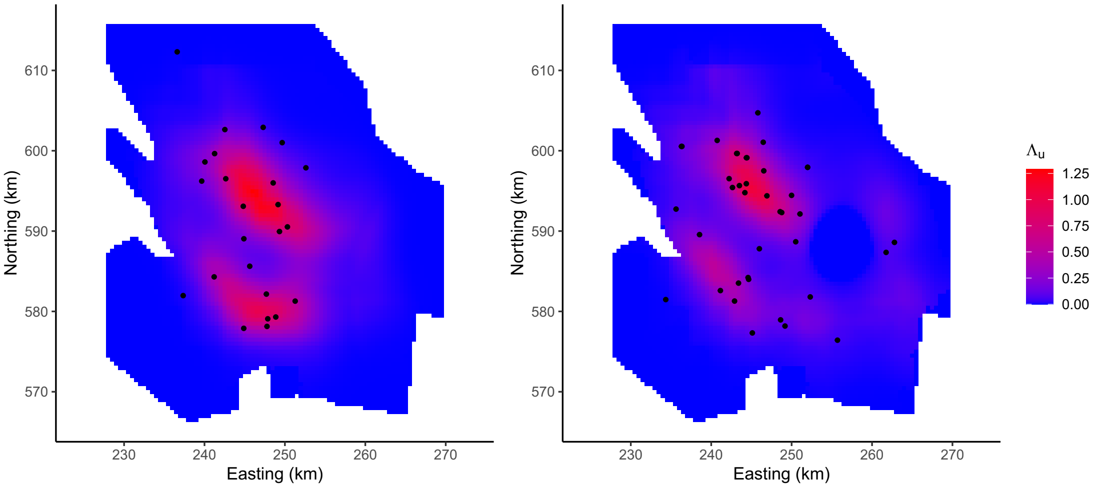

FIGURE 8 | Spatial plots of aggregated intensity $\Lambda_u(\boldsymbol{x}, T)$ per km^2 for years: (left) $T = 2010$ and (right) $T = 2020$. The locations of exceedances of the threshold A_2 occurring throughout each year shown as black dots.

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 7 [right] shows the ratio of the observed and expected annual earthquake counts above $0\mathrm{M_L}$ over T, again under the assumption that F_0 is GPD. 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 2, and it provides a novel estimator for the probability of detecting an earthquake, which is of interest to geophysicists.

Although our discussion and conclusions regarding Figure 7 [centre] and [right] are subject to the statistically untestable assumption that F_0 is GPD, we believe that if there is any error in such an assumption, the effect of this error will be secondary relative to the temporal pattern of failure to detect earthquakes between $0\mathrm{M_L}$ and our estimated magnitude of completion function. This belief is based on the following: (i) the geophysical experts' knowledge that earthquakes have gone undetected in the Groningen region; and (ii) the increasing rate of small magnitude events per number of detected earthquakes over years is consistent with the rate at which the geophone network was expanded. Furthermore, the GPD for F_0 is supported by strong empirical evidence in the earthquake modelling literature of true magnitudes being approximately exponential distributed based on the Gutenberg-Richter Law and our arguments in Section 1 regarding the GPD being the unique continuous distribution that possesses the required conditional threshold stability closure property. In Section 7, we discuss the inference for a related, but more specific, measure of detection which accounts for the values of the detected magnitudes, rather than solely their relation to the threshold.

6.3 | Threshold Uncertainty

We now explore the impact of threshold uncertainty as captured by Algorithm 3. For non-parametric bootstrap samples of the original earthquake data, the algorithm selects from the 12 different threshold function model formulations, that is, $A - C$ and $i = 1, \ldots, 4$. From the $B_{\text{nonpar}} = 200$ bootstrap replicated model selections 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 of the associated 200 selections for the distance covariates V_i 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. The percentages that each of the joint covariate-model formulations are selected are given in Table S.1 in the Supporting Information. 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, and that formulation A_4 is selected almost three times as often as A_2, even though the latter is the best fitting model to the observed data.

To illustrate the uncertainty in the A_2 threshold function, which was selected as the best threshold formulation, Figure 9 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, with the intervals from Algorithm 3 being wider, visually most notably prior to 2000. 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. In the periods when geophone numbers decreased, as identified from Figure 2 (left), the estimates and CI widths increase, with the most obvious example being for 2005.

To provide further insight into the variation in the bootstrapped threshold function estimates, we investigated how the number of exceedances $|K_{\hat{u}^b}|$ of the estimated threshold functions $\hat{u}^b$ varied across the $B_{\text{nonpar}} = 200$ 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 show that using Algorithm 3 leads to 59 more exceedances on average than Algorithm 2, and the maxima and standard deviations for both of 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.

Algorithms 2 and 3 are substantially more computationally intensive relative to the more standard Algorithm 1. Algorithm 2 requires $B_{\text{nonpar}} \times (B + B_{\text{par}})$ bootstrap samples through the combination of threshold parameter selection and model fitting for each of the B_{nonpar} bootstrap samples of the data. With the additional element of model selection, Algorithm 3 requires $B_{\text{nonpar}} \times (12B + B_{\text{par}})$ for our 12-model comparison. With our choices of $(B_{\text{par}}, B_{\text{nonpar}}, B, m)$, computation time for Algorithm 3 is in the order of 1-2 days when run in parallel (across B_{nonpar}) on a computing cluster with 100 cores. Computation time increases linearly at differing rates with each of these factors. While this computation time may be prohibitive for a large-scale simulation study, for a one-off analysis, the additional value derived from incorporating extra sources of uncertainty outweighs the cost in time. In particular, for this economically and politically important application, the inclusion of the additional uncertainty is of paramount importance and the computation time is trivial in comparison to the numerical models which have been used to derive the KS covariate.

6.4 | Inference and Uncertainty for Design Parameters

We first examine the inference uncertainty for shape parameter ξ since, with ξ assumed common over time, space and covariates, it underpins all aspects of our extrapolations, and recall that we do not impose $\xi < 0$. For the selected best threshold function model A_2, we obtained $\hat{\xi}^{A_2} = -0.154$. Now, we explore the uncertainty in ξ in terms of estimated 95% CIs using the bootstrapping Algorithms 2 and 3, with the intervals given in Table 2, with $B_{\text{par}} = B_{\text{nonpar}} = 200$, giving 40,000 bootstrapped estimates of ξ. When the threshold function is 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, reduces slightly 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, as we found for the threshold uncertainty in Figure 9. However, by allowing the threshold function form to vary across bootstrapped samples, we obtain less variable shape parameter estimates, a feature we discuss below in the context of inference for $e_{\max}(\mathcal{S}_F)$.

TABLE 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 M_L.

Parameter	Estimate	Alg 1	Alg 2	Alg 3
ξ	−0.154	(−0.221, −0.099)	(−0.240, −0.066)	(−0.236, −0.064)
$e_{\max}(\mathcal{S}_F)$	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)

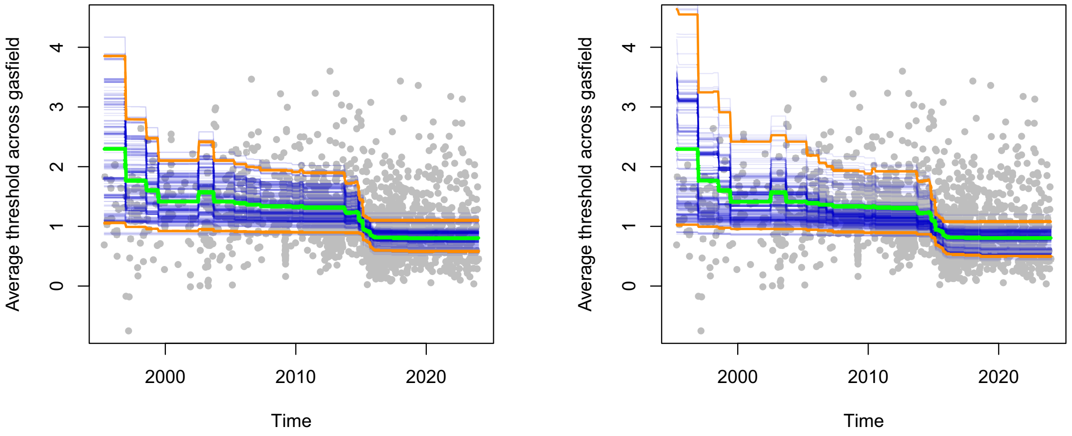

FIGURE 9 | Uncertainty in the spatial average threshold function over $\mathcal{G}$ plotted over time: showing bootstrap based pointwise 95% confidence intervals and individual bootstrap estimates. The $B_{\text{nonpar}} = 200$ bootstrapped best threshold function estimates for Groningen according to 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. The confidence intervals are shown in orange and individual bootstrap samples in blue.

The 95% CIs do not reveal the occurrence rate, over the bootstraps, of $\hat{\xi}^b > 0$, that is, corresponding to estimated GPDs with no finite upper endpoint. For Algorithms 1–3, we observed the percentages of samples with $\hat{\xi}^b > 0$ of $0\%, 0.26\%$ 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%. Critically, this shows that imposing $\xi < 0$ would not change the CI results for our selected thresholds, in contrast to the conservative threshold.

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 of Yue, Tawn, et al. (2025), 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 $e_{\max}(\mathcal{S}_F)$ 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 Supporting Information S:8. The maximum $e_{\max}(\mathcal{S}_F)$ is comparable to the endpoint estimate in previous studies which ignored the spatio-temporal variation in the endpoint.

Table 2 presents point estimates and CIs for the above two summaries using Algorithms 1–3. Consider our point estimate and 95% CIs for $\hat{e}_{\max}(\mathcal{S}_F)$ using Algorithm 1. For background, Beirlant et al. (2019) used a constant threshold of $1.5\mathrm{M_L}$ with rounded earthquake data and presented 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 our results which were obtained using a lower threshold than theirs, and their endpoint 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. Such a difference in findings relative to Beirlant et al. (2019) are substantial given that the same threshold is used and the good quality of fit exhibited by our model with this threshold as seen in Figure 6. 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 CI. This suggests that there must be very strong assumptions used in the inferences of Beirlant et al. (2019) to achieve this level of extra precision. Such assumptions cannot be checked as no diagnostic analyses are presented.

Neither Varty et al. (2021) nor Yue, Tawn, et al. (2025) 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, drawn from the report NAM (2022), 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 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. Possible reasons for this slightly counter-intuitive finding include that Algorithm 3 results in 59 more threshold exceedances and that the changing model selections offset poor fits that arise by imposing A_2 automatically. 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 more closely reflect the occurrence rates of earthquake hypocentres. Figures S.5 and S.6 in the Supporting Information provide estimates of the annual behaviour of this endpoint summary and insight into the temporal development of the estimated endpoints across space, respectively.

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

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 2. Of the previous analyses, only Yue, Tawn, et al. (2025) (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. By taking the rate of exceedance as the highest observed yearly rate of exceedance of a threshold at 1.15M_L, that is, 2017, they estimate the 475-year return level to be 4.55M_L. The corresponding (unreported) estimate, using the rate of exceedance in 2023, was 4.38M_L. Both values will be over-estimates as they do 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.

7 | Discussion

We developed spatio-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 spatio-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 increased 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 led to more excess data being used for the analysis and, hence, less uncertainty in the parameter estimation. This also resulted in useful inferences for the tail behaviour of earthquake magnitudes into the future, both for design levels and upper limits, 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 any other future scenario.

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 developed. However, the methodology is generic and so has potential for wide use in other cases of induced earthquakes, for example, 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. Thus, 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 expands substantially on previous work which focussed on realisations of IID variables (Murphy et al. 2025). Here, we 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).

Our focus was on accurately modelling 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 is illustrated in Figure 7 [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 $(\boldsymbol{x}, t) \in \mathcal{X} \times \mathcal{T}$. 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 $\lambda(\boldsymbol{x}, t, y)$ is the intensity of detected earthquakes of magnitude y at hypocentre and time $(\boldsymbol{x}, t)$, then

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

As $\lambda(\boldsymbol{x}, t, y)$ 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.

In Section 6.1, we selected the threshold formulation A_2 to be the best choice of the 12 threshold formulations considered. Although in Section 6.4 we used bootstrap methods to address the uncertainty of this threshold selection on the subsequent confidence interval inferences, we did not use a weighted point estimate to remove the model selection element in our point inferences. A further line of research could be to use Bayesian inference throughout and combine models using Bayesian model averaging (Hoeting et al. 1999). We envisage that using Bayesian inferences to underpin political decisions is likely to encounter concerns about the subjectivity of the prior selections.

As discussed in Section 6.4, for the upper endpoint of the distribution of earthquake magnitudes, our estimated upper 95% confidence interval limit is unfeasibly high, whereas the similar limit for the 50% interval is much more physically realistic. Rather than simply imposing that $\xi < -\epsilon$, for some arbitrary $\epsilon > 0$, a more natural approach is to impose some penalty or prior on the endpoint of the generalised Pareto distribution. An example of such an approach is given by Yue, Tawn, et al. (2025), who exploited the geophysicists' derived distribution for possible values of $M_{\max}$ for Groningen.

Other aspects of the modelling could be developed in future research. As mentioned in Section 4.5, we could extend the modelling to 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 of individual geophone technology has increased in time. Finally, when modelling earthquake baseline rates of occurrence, that is, in modelling $\lambda_0(\boldsymbol{x}, t)$, we did not account explicitly for the clustering of events due to the dependence between triggering main-shock earthquakes and after-shocks. Thus, the intensity modelling could be adapted to cover this feature through the use of the ETAS models of Ogata (1988).

Acknowledgments

We are grateful to Stephen Bourne (Shell Research Ltd.) and Steve Oates (Shell Research Ltd.) for providing their scientific and geophysical expertise to support this project. We thank the two referees for their detailed and insightful comments about the work and its presentation.

Funding

This work was supported by the Shell United Kingdom and Engineering and Physical Sciences Research Council (EP/S022252/1).

Disclosure

This paper is based on work completed while Conor Murphy was part of the EPSRC funded STOR-i centre for doctoral training (EP/S022252/1), with part-funding from Shell Research Ltd.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that supports the findings of this study are available in the linked Github repository. Furthermore, the earthquake catalogue and geophone datasets are both publicly available. These datasets have undergone some processing for our purposes to remove duplicates and constrain to the observation timeframe. The processed datasets for the earthquake catalogue, geophones and stress covariates are included in our public Github repository. The source code for the threshold selection method and code files to generate all figures in the main text and Supporting Information are available in the repository. There is a README.md file included to aid the reviewers.

References

Bauer, R. A., R. Will, S. E. Greenberg, et al. 2019. *Geophysics and Geosequestration, Chap. Illinois Basin–Decatur Project*, 339–369. Cambridge University Press.

Beirlant, J., A. Kijko, T. Reynkens, and J. H. Einmahl. 2019. "Estimating the Maximum Possible Earthquake Magnitude Using Extreme Value Methodology: The Groningen Case." *Natural Hazards* 98: 1091–1113.

Belzile, L., C. Dutang, P. Northrop, and T. Opitz. 2023. "A Modeler's Guide to Extreme Value Software." *Extremes* 26: 1–44.

Bommer, J. J., B. Dost, B. Edwards, et al. 2017. "Developing a Model for the Prediction of Ground Motions due to Earthquakes in the Groningen Gas Field." *Netherlands Journal of Geosciences* 96: s203–s213.

Bourne, S. J., and S. J. Oates. 2020. "Stress-Dependent Magnitudes of Induced Earthquakes in the Groningen Gas Field." *Journal of Geophysical Research - Solid Earth* 125: e2020JB020013.

Bourne, S. J., S. J. Oates, and J. Van Elk. 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: 1693–1700.

Chavez-Demoulin, V., and A. C. Davison. 2005. "Generalized Additive Modelling of Sample Extremes." *Journal of the Royal Statistical Society, Series C* 54: 207–222.

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, p. 10.

Coles, S. G. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer.

Coles, S. G., and J. A. Tawn. 1996. "A Bayesian Analysis of Extreme Rainfall Data." *Journal of the Royal Statistical Society: Series C: Applied Statistics* 45: 463–478.

Das, R., H. Wason, and M. L. Sharma. 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 D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Vol. No. 1. Cambridge University Press.

Davison, A. C., and R. L. Smith. 1990. "Models for Exceedances Over High Thresholds (With Discussion)." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 52: 393–425.

Demuth, A., L. Ottemöller, and H. Keers. 2016. "Ambient Noise Levels and Detection Threshold in Norway." *Journal of Seismology* 20: 889–904.

Dost, B., and D. Kraaijpoel. 2013. *The August 16, 2012 Earthquake Near Huizinge (Groningen)*. KNMI.

Eastoe, E. F., and J. A. Tawn. 2009. "Modelling Non-Stationary Extremes With Application to Surface Level Ozone." *Journal of the Royal Statistical Society. Series C, Applied Statistics* 58: 25–45.

Ellsworth, W. L. 2013. "Injection-Induced Earthquakes." *Science* 341: 1225942.

Freudenreich, Y., S. J. Oates, and W. Berlang. 2012. "Microseismic Feasibility Studies–Assessing the Probability of Success of Monitoring Projects." *Geophysical Prospecting* 60: 1043–1053.

Galis, M., J. P. Ampuero, P. M. Mai, and F. Cappa. 2017. "Induced Seismicity Provides Insight Into Why Earthquake Ruptures Stop." *Science Advances* 3: eaap7528.

Gaucher, E. 2016. "Earthquake Detection Probability Within a Seismically Quiet Area: Application to the Bruchsal Geothermal Field." *Geophysical Prospecting* 64: 268–286.

Goertz, A., N. Riahi, T. Kraft, and M. Lambert. 2012. "Modeling Detection Thresholds of Microseismic Monitoring Networks." In SEG International Exposition and Annual Meeting, SEG.

Gutenberg, B., and C. F. Richter. 1956. "Earthquake Magnitude, Intensity, Energy, and Acceleration: (Second Paper)." *Bulletin of the Seismological Society of America* 46: 105–145.

Healy, D., J. A. Tawn, P. Thorne, and A. Parnell. 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: 275–330.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial (With Comments by M. Clyde, David Draper and E. I. George, and a Rejoinder by the Authors)." *Statistical Science* 14: 382–417 Correction: vol. 15, pp. 193–195.

Hutton, K., J. Woessner, and E. Hauksson. 2010. "Earthquake Monitoring in Southern California for Seventy-Seven Years (1932–2008)." *Bulletin of the Seismological Society of America* 100: 423–446.

Kaveh, H., P. Batlle, M. Acosta, P. Kulkarni, S. J. Bourne, and J. P. Avouac. 2024. "Induced Seismicity Forecasting With Uncertainty Quantification: Application to the Groningen Gas Field." *Seismological Research Letters* 95: 773–790.

KNMI. 2020. "Aardbevings catalogus." https://www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus.

Koh, J., F. Pimont, J.-L. Dupuy, and T. Opitz. 2023. "Spatiotemporal Wildfire Modeling Through Point Processes With Moderate and Extreme Marks." *Annals of Applied Statistics* 17: 560–582.

Kyselý, J., J. Picek, and R. Beranová. 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.

Majer, E. L., R. Baria, M. Stark, et al. 2007. "Induced Seismicity Associated With Enhanced Geothermal Systems." *Geothermics* 36: 185–222.

Mignan, A., M. Werner, S. Wiemer, C. Chen, and Y. Wu. 2011. "Bayesian Estimation of the Spatially Varying Completeness Magnitude of Earthquake Catalogs." *Bulletin of the Seismological Society of America* 101: 1371–1385.

Mignan, A., and J. Woessner. 2012. "Estimating the Magnitude of Completeness for Earthquake Catalogs." Community Online Resource for Statistical Seismicity Analysis.

Murphy, C., J. A. Tawn, and Z. Varty. 2025. "Automated Threshold Selection and Associated Inference Uncertainty for Univariate Extremes." *Technometrics* 67: 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.

Northrop, P. J., and P. Jonathan. 2011. "Threshold Modelling of Spatially Dependent Non-Stationary Extremes With Application to Hurricane-Induced Wave Heights." *Environmetrics* 22: 799–809.

Ogata, Y. 1988. "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes." *Journal of the American Statistical Association* 83: 9–27.

Papastathopoulos, I., and J. A. Tawn. 2013. "Extended Generalised Pareto Models for Tail Estimation." *Journal of Statistical Planning and Inference* 143: 131–143.

Pickands, J. 1975. "Statistical Inference Using Extreme Order Statistics." *Annals of Statistics* 3: 119–131.

Raschke, M. 2015. "Modeling of Magnitude Distributions by the Generalized Truncated Exponential Distribution." *Journal of Seismology* 19: 265–271.

Richter, G., S. Hainzl, T. Dahm, and G. Zöller. 2020. "Stress-Based, Statistical Modeling of the Induced Seismicity at the Groningen Gas Field, the Netherlands." *Environmental Earth Sciences* 79: 1–15.

Ross, G. J. 2021. "Bayesian Estimation of the ETAS Model for Earthquake Occurrences." *Bulletin of the Seismological Society of America* 111: 1473–1480.

Scarrott, C., and A. MacDonald. 2012. "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification." *REVSTAT-Statistical Journal* 10: 33–60.

Shcherbakov, R., J. Zhuang, G. Zöller, and Y. Ogata. 2019. "Forecasting the Magnitude of the Largest Expected Earthquake." *Nature Communications* 10: 1–11.

Smith, J. D., E. R. Heimisson, S. J. Bourne, and J.-P. Avouac. 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.

Suckale, J. 2009. "Induced Seismicity in Hydrocarbon Fields." *Advances in Geophysics* 51: 55–106.

Tang, C., and J. A. Hudson. 2010. *Rock Failure Mechanisms: Illustrated and Explained*. CRC Press.

Turkman, K. F., M. A. A. Turkman, and J. M. Pereira. 2010. "Asymptotic Models and Inference for Extremes of Spatio-Temporal Data." *Extremes* 13: 375–397.

Varty, Z., J. A. Tawn, P. M. Atkinson, and S. Bierman. 2021. "arXiv:2102.00884."

Vere-Jones, D., R. Robinson, and W. Yang. 2001. "Remarks on the Accelerated Moment Release Model: Problems of Model Formulation, Simulation and Estimation." *Geophysical Journal International* 144: 517–531.

Weng, H., J.-P. Ampuero, and L. Buijze. 2021. "Physics-Based Estimates of the Maximum Magnitude of Induced Earthquakes in the Groningen Gas Field." In *EGU General Assembly Conference Abstracts*, EGU21-6144.

Wyss, M., A. Hasegawa, S. Wiemer, and N. Umino. 1999. "Quantitative Mapping of Precursory Seismic Quiescence Before the 1989, m 7.1 Off-Sanriku Earthquake, Japan." *Annals of Geophysics* 42: 851–869.

Yadav, R., R. Huser, T. Opitz, and L. Lombardo. 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: 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: 1865–1879.

Yue, Q., Y. Guo, T. Sayed, L. Zheng, H. Lyu, and P. Liu. 2025. "Bayesian Hierarchical Non-Stationary Hybrid Modeling for Threshold Estimation in Peak Over Threshold Approach." arXiv preprint arXiv:2503.14839.

Yue, W., J. A. Tawn, R. Towe, and Z. Varty. 2025. “Integrating Experts’ Belief in Upper Tail Inference for Modelling of Human-Induced Earthquake Magnitudes.” *Journal of Agricultural, Biological and Environmental Sciences*. https://doi.org/10.1007/s13253-025-00710-9.

Zaliapin, I., A. Gabrielov, V. Keilis-Borok, and H. Wong. 2008. “Clustering Analysis of Seismicity and Aftershock Identification.” *Physical Review Letters* 101: 018501.

Zang, A., V. Oye, P. Jousset, et al. 2014. “Analysis of Induced Seismicity in Geothermal Reservoirs – an Overview.” *Geothermics* 52: 6 – 21.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Supporting Information.