Statistical analysis of static and dynamic predictors for seismic b-value variations in the Groningen gas field

Dirk Kraaijpoel[1], Joana E. Martins[2], Sander Osinga[1], Bouko Vogelaar[2], and Jaap Breunese[2]

[1]TNO, Applied Geosciences, Utrecht, The Netherlands
[2]TNO, Advisory Group for Economic Affairs, Utrecht, The Netherlands

June 17, 2022

Abstract

We perform statistical analyses on spatiotemporal patterns in the magnitude distribution of induced earthquakes in the Groningen natural gas field. The seismic catalogue contains 336 earthquakes with (local) magnitudes above 1.45, observed in the period between 1995-01-01 and 2022-01-01. An exploratory moving-window analysis of maximum-likelihood b-values in both time and space does not reveal any significant variation in time, but does reveal a spatial variation that exceeds the 0.05 significance level.

In search for improved understanding of the observed spatial variations in physical terms we test five physical reservoir properties as possible b-value predictors. The predictors include two static (spatial, time-independent) properties: the reservoir layer thickness, and the topographic gradient (a measure of the degree of faulting intensity in the reservoir); and three dynamic (spatiotemporal, time-dependent) properties: the pressure drop due to gas extraction, the resulting reservoir compaction, and a measure for the resulting induced stress. The latter property is the one that is currently used in the seismic source models that feed into the state-of-the-art hazard and risk assessment.

We assess the predictive capabilities of the five properties by statistical evaluation of both moving window analysis, and maximum-likelihood parameter estimation for a number of simple functional forms that express the b-value as a function of the predictor. We find significant linear trends of the b-value for both topographic gradient and induced stress, but even more pronouncedly for reservoir thickness. Also for the moving window analysis and the step function fit, the reservoir thickness provides the most significant results.

We conclude that reservoir thickness is a strong predictor for spatial b-value variations in the Groningen field. We propose to develop a forecasting model for Groningen magnitude distributions conditioned on reservoir thickness, to be used alongside, or as a replacement, for the current models conditioned on induced stress.

Keywords— Groningen, Gutenberg-Richter law, spatiotemporal b-value variations, earthquake catalogue, induced seismicity, predictors

1 Introduction

The Groningen reservoir in the Netherlands is the largest gas field in western Europe. Since 2014, annual production volumes from the Groningen gas field are rapidly declining. Although approximately

one fifth of the initial ∼2900 billion m^3 is still technically recoverable, the Dutch government has decided to cease production by 2023 or 2024 to ensure public safety and well-being. Induced earthquakes at the Groningen natural gas reservoir have put local communities at risk to a level that has led to societal and political upheaval (Vlek, 2019). The residential population living and working near the contours of the Groningen gas field is about half a million people. About 25 thousand damage claims for masonry buildings have been filed in the past years of which a part can be associated to individual seismic events. Older homes and farms built with single brick walls are especially vulnerable; these are either taken down and rebuilt or structurally strengthened.

The 100-300 m thick gas-bearing Rotliegend sandstone reservoir at 2.6–3.2 km deep is overlain by a 1 km thick Zechstein salt formation that acts as a seal (De Jager & Visser, 2017). Around 700 faults have been mapped in and below the reservoir (Kortekaas & Jaarsma, 2017). The weight of the overburden causes compaction of the reservoir formation due to gas withdrawal. Differential compaction results in localized stress concentrations along existing faults resulting in ruptures and earthquakes (Van Thienen-Visser & Breunese, 2015). Groningen-specific models describing this process are given by (S. J. Bourne, Oates, van Elk, & Doornhof, 2014; S. Bourne & Oates, 2017; Candela et al., 2019; S. J. Bourne & Oates, 2020) and references therein.

The Gutenberg-Richter relation is a model for the magnitude distribution of earthquakes mostly characterized by its exponential coefficient, the b-value. Variations of the b-value express changes in the rate of occurrence of small earthquakes relative to large ones. Spatiotemporal b-value variations have been associated to various geophysical processes and circumstances for earthquake catalogues over natural tectonically active environments. Dependence of b-values on differential stress has been reported by C. Scholz (1968); Wyss (1973). More specifically, spatial b-value variations have been related to variations in tectonic regimes: thrust faulting systems being associated with relatively low values, normal faulting systems with relatively high values and strike-slip systems in between (Gulia & Wiemer, 2010; Nakaya, 2006; Schorlemmer, Wiemer, & Wyss, 2005). Also, the depth of occurrence of the events has been identified as a possible cause of b-value variations (Eaton, O'neill, & Murdock, 1970; Wyss, 1973; Wyss, McNutt, & Wyss, 1998; Mori & Abercrombie, 1997; Spada, Tormann, Wiemer, & Enescu, 2013; Gerstenberger, Wiemer, & Giardini, 2001; Wiemer & Benoit, 1996). Other possible factors that may be relevant to the Groningen situation include stress fluctuations or regimes (C. H. Scholz, 2015; Langenbruch & Shapiro, 2014; Wiemer & Wyss, 1997) and the (fractal) geometry of fault systems (Rundle, 1989; Mandal & Rastogi, 2005; Hirata, 1989).

In the context of induced seismicity, b-value variations have been attributed to several processes. These include: fluid-driven triggering mechanism possibly due to faults cutting into the reservoir and hydraulic connection between reservoir and basement at a CO2 injection site (Goertz-Allmann, Gibbons, Oye, Bauer, & Will, 2017); normalised shear stress instead of differential stress at the Basel Enhanced Geothermal System (Mukuhira, Fehler, Ito, Asanuma, & Häring, 2021); loading rates in the Lacq gas field (Lahaie & Grasso, 1999) and at the The Geysers geothermal field (Henderson, Barton, & Foulger, 1999); pore pressure, fluid content and injection activity at the Oklahoma oil field (Vorobieva, Shebalin, & Narteau, 2020; Rajesh & Gupta, 2021).

Despite the large and increasing number of proposed b-value dependencies, the physical context is not always well understood. This especially holds for induced seismicity cases, where the number of studies is significantly lower (possibly due to the limited size of event catalogues). Considering the highly non-stationary characteristics of the anthropogenic activities, hazard and risk assessment models could benefit greatly from a better understanding of b-value variations. The current state-of-the-art seismological source models for the Groningen gas field (S. Bourne & Oates, 2017; S. J. Bourne & Oates, 2020) forecast spatiotemporal variations in both activity rate and magnitude distribution in terms of an induced stress metric. S. J. Bourne and Oates (2020) provided an insightful separation of literature studies on stress dependent and not dependent b-value variations. The choice of induced stress as a predictor is based on a comparative analysis of a suite of possible predictors in S. Bourne and Oates (2017) for seismic activity rate. However, it appears that the authors did not make a separate attempt to determine the most appropriate predictor for the magnitude frequency model. We consider it possible that the best predictor for activity rate may not automatically be the best predictor for the magnitude distribution.

In this paper, we perform statistical analyses on the spatiotemporal variations of the b-value in the Groningen gas field. We use a relatively small catalogue of 336 events with (local) magnitudes above the estimated completeness level of 1.45 over the entire observation time span from 1995-01-01 to 2022-01-01. We investigate the predictive capabilities of five physical quantities (covariates) that describe static or dynamic properties of the Groningen field. The general goal of the investigation is to steer the development of forecasting models that allow a more precise assessment of the seismic hazard and risk analysis through more precise forecasts on the magnitude distribution.

2 Data and Methods

2.1 Earthquake catalogue

The first instrumentally recorded seismic event in the Groningen gas field was an earthquake with local magnitude 2.4 in 1991. When the event took place, the mean reservoir pressure had already decreased from the initial 350 bar to below 200 bar. With ongoing reservoir gas production the induced seismicity continued leading to the installation of a regional borehole seismic network by 1995 (Dost, Ruigrok, & Spetzler, 2017). In the following years, the annual number of recorded earthquakes fluctuated around ten to fifteen events per year until about 2003 when it started to increase, to nearly 120 in 2017. Figure 1 (top) shows the annual number of event in different magnitude categories. The largest magnitude observed to date is the 3.6 event near Huizinge in 2012.

[Figure 1 about here.]

The sensitivity of the monitoring network has not been uniform in space and time. A relatively safe (conservative upper bound) estimate of the completeness level over the entire period and region is a magnitude of 1.5 (Dost et al., 2017). For the purpose of the current study we adopt this level as the minimum threshold $m_{\min}$ for earthquakes to be considered. In fact, since we use unrounded magnitude values we slightly relax the threshold to $m_{\min} = 1.45$, as the value 1.5 was specified for magnitude values rounded to one decimal place. Figure 1 (bottom) shows the annual number of events in the curated catalogue.

We note that the current study might, in principle, benefit from an enlarged data set by taking into account a time-dependent completeness level (Dost et al., 2017; Varty, Tawn, Atkinson, & Bierman, 2021). However, by taking a conservative assumption here, we largely steer clear of discussions on how the spatiotemporal completeness level should be estimated and to what extent imperfections would affect the results (Herrmann & Marzocchi, 2020, e.g.). Also, since one of our objectives is to provide forecasting models for risks that occur only at magnitudes that are several units larger than the completeness magnitude (say, magnitude 4 and beyond), it is questionable whether we should go as low as we possibly can. After all, the lower we choose the range of input data for our model inference, the larger the distance we effectively have to extrapolate at a later stage.

We obtained the Groningen earthquake catalogue from the Seismological Service of the KNMI (KNMI, 2022b) through their FDSN Event Web Service (URL found in reference KNMI (2022a)), which provides origin times, locations in WSG-84 coordinates and unrounded local magnitudes. Epicentral coordinates have subsequently been transformed into the local Amersfoort/RD New coordinates system (epsg.io, 2022). The hypocenter depth is not used. We selected all earthquakes above $m_{\min} = 1.45$, in the time window between 1995-01-01T00:00 and 2022-01-01T00:00. To avoid interference of earthquakes due to other exploration activities in the vicinity as much as possible, the spatial extent of the catalogue is limited to the Groningen gas field outline (NAM, 2021). The total number of events in the curated catalogue is 336. Figure 2 shows the catalogue of all earthquakes of the Groningen gas field.

[Figure 2 about here.]

2.2 Gutenberg-Richter magnitude distribution

We employ the classical Gutenberg-Richter relation (Gutenberg & Richter, 1941, 1944), as a model for the magnitude distribution of induced earthquakes in the Groningen gas field. The survival function, or probability $P[M \geq m | M \geq m_{\min}]$ of a random magnitude sample M exceeding m, under the condition that it exceeds $m_{\min}$ is given by:

$$P[M \geq m | M \geq m_{\min}] = 10^{-b(m-m_{\min})}, \tag{1}$$

with b the exponential parameter, or b-value, and its probability density function as:

$$f_M(m) = -\frac{\mathrm{d}P[M \geq m | M \geq m_{\min}]}{\mathrm{d}m} \tag{2}$$

$$= b^* e^{-b^*(m-m_{\min})}, \tag{3}$$

where $b^* = b \log(10)$.

Unlike previous authors (S. J. Bourne & Oates, 2020), we do not consider a dedicated prescription of the high-magnitude tail of the distribution such as a truncation or a taper. The focus of our efforts is to find evidence for any significant spatiotemporal variation of the magnitude distribution. However, as argued in Marzocchi, Spassiani, Stallone, and Taroni (2019) and S. J. Bourne and Oates (2020), among others, failure to recognize and accommodate a truncated or tapered tail may lead to artifacts appearing in b-value estimates, especially if the tail starts close to the $m_{\min}$ considered. However, we argue that if a tail effect is relevant, in the sense that it is somehow exposed in the data, then such artifacts may actually help to reveal any spatiotemporal variation of this effect through the analysis of the b-value. We only need to keep in mind that any significant spatiotemporal variation in the assessed b-value does not necessarily have to be caused by a variation in the exponential character of the distribution, but may also be caused by variations in the tail behaviour that we do not model explicitly.

In this study we are interested in spatiotemporal variations of the b-value b, so that we can express it as a function of time t and space coordinates x. More specifically, we test prospective predictors that may act as a spatiotemporal covariate $c(t, x)$ for the b-value:

$$b = g_C(c(t, x), \theta), \tag{4}$$

where we use a generic functional form g_C that depends on covariate $c(t, x)$ and a, generally multivariate, parameter set θ, which represents, for example, the coefficients in functional form. We investigate a number of functional forms (i.e., models) with corresponding parameter sets, for which we infer information from the data.

The inference of model parameters starts with expressing the log-likelihood of the model, conditional on the data. The likelihood is defined as:

$$\mathcal{L}(\theta | \{(c_i, m_i), i = 1...N\}) = \prod_{i=1}^{N} f_M(m_i | c_i, \theta), \tag{5}$$

with m_i the magnitude, and $c_i = c(t_i, x_i)$ the covariate based on the time and location of each event i out of N events in the catalogue. Each event in the catalogue is associated to a specific b-value b_i:

$$b_i = g_C(c_i, \theta), \tag{6}$$

such that the log-likelihood equals:

$$\log(\mathcal{L})(\theta | \{(c_i, m_i), i = 1...N\}) = \sum_{i=1}^{N} [\log(b_i^*) - b_i^*(m_i - m_{\min})]. \tag{7}$$

Note that for a constant b-value a maximization of $\log(\mathcal{L})$ is achieved algebraically in closed form by the maximum likelihood estimate (MLE) b-value b_{MLE}:

$$b_{\mathrm{MLE}} = \frac{\log_{10} e}{\bar{m} - m_{\mathrm{min}}} \tag{8}$$

where $\bar{m}$ is the arithmetic mean of the magnitudes in the data set (Aki, 1965). For non-constant models the maximum-likelihood parameters can be found by maximizing the log-likelihood for model parameter vector θ to find θ_{MLE}. This search can be done using a gradient ascent algorithm or an exhaustive grid search.

Assuming a constant b-value we have determined the MLE b-value for the entire catalogue: $b_{\mathrm{MLE}} = 0.96$. The result is shown together with the empirical distribution in Figure 3.

[Figure 3 about here.]

2.3 Static and dynamic predictors

In S. J. Bourne and Oates (2020), the authors propose two models for the spatiotemporal evolution of the magnitude distribution, that are both conditioned on the induced stress ("Coulomb stress") predictor: one with a constant b-value and a stress-dependent high-magnitude taper, and one without a taper but with a stress-dependent b-value. On the basis of physical considerations and the result of some statistical tests the authors express a preference for the former model. As mentioned in section 1, they did not investigate or report alternative predictors for the magnitude distribution.

Similar to the approach of S. Bourne and Oates (2017) for activity rate prediction, we investigate a range of physical reservoir properties as predictors for the b-value throughout the Groningen gas field. These prospective predictors are related to the geological layout of the gas reservoir, the gas depletion process itself, or a combination of both.

- *Reservoir thickness* The reservoir thickness $h(x)$ is a 2-D spatial representation of the thickness of the Rotliegend reservoir formation, provided by the field operator (NAM, 2021) and is given in units of meters. The thickness is assumed to be static (time-invariant) and independent of the gas production. The (relatively small) compaction of the reservoir due to the gas extraction is considered as a separate prospective predictor below.

- *Topographic gradient* The topographic gradient $\Gamma(x)$ is a spatially smoothed, 2-D, static measure of the roughness of the topography of the top of reservoir. This roughness is largely due to faults with varying offsets (S. Bourne & Oates, 2017). It is calculated based on the locations and offsets of the pre-existing faults in the reservoir, and is controlled by two parameters: r_{max} and σ. The parameter r_{max} describes an upper cut-off value for the local fault offset-to-thickness ratio. Faults with an offset-to-thickness ratio larger than r_{max} are disregarded in the calculation. The final property is calculated by mapping the fault offsets that pass the r_{max} filter onto a regular grid, which is subsequently smoothed by a Gaussian kernel with kernel size σ. As a result, the property is roughly proportional to both fault offset and fault density. A ready-made topographic gradient grid with parameter values $r_{\mathrm{max}} = 1.1$ and $\sigma = 3500$ m is supplied by the operator (NAM, 2021; S. J. Bourne & Oates, 2020).

- *Pressure drop* The pressure drop $\Delta P(t, x)$ is at any time t a 2-D spatial representation of the vertically averaged pore pressure depletion in the reservoir with respect to the original (pre-production) gas pressure, provided by the operator (NAM, 2021). This property is dynamic (time-dependent) and naturally depends on the gas production.

- *Reservoir compaction* The reservoir compaction $\Delta h(t, x)$ is at any time t a 2-D spatial representation of the change in reservoir thickness as a result of the gas pressure decline. This covariate is defined as:

$$\Delta h(t, x) = \Delta P(t, x) \times c_m(x) \times h(x) \tag{9}$$

where $\Delta P(t,x)$ is the dynamic pressure drop, c_m is the poro-elastic coupling coefficient, and $h(x)$ is the reservoir thickness. These are all provided by the operator (NAM, 2021). Reservoir compaction is a dynamic (time-dependent) property due to its dependence on the gas production.

- *Induced stress* The induced stress at any time t is a 2-D spatial property representing the (spatially smoothed) change in Coulomb stress on pre-existing faults according to the thin-sheet model. It is calculated in accordance with S. Bourne and Oates (2017) as [1]:

$$\Delta C(t,x) = \Gamma(x) \times \Delta P(t,x) \times \frac{c_m(x)}{H_s^{-1} + c_m(x)} \tag{10}$$

where $\Gamma(x)$ is the topographic gradient, and H_s is a stiffness parameter. Both $\Gamma(x)$ and H_s are supplied by the operator (NAM, 2021). Induced stress is dynamic due to its dependence on the gas production. We note that the use of a spatial Gaussian smoothing kernel in the calculation of the topographic gradient makes it difficult to interpret the numerical values of this covariate field in terms of absolute stress changes. Rather, it is a metric that combines fault density of faults below a certain offset-to-thickness ratio and vertical compaction strain. In a stricter sense, it is perhaps best interpreted as a propensity-to-failure proxy, rather than Coulomb stress change.

Along with the above five prospective predictors representing physical properties of the reservoir, we also take time as a possible predictor to complete a set of six predictors that we submit to the same sequence of statistical tests. Figure 4 illustrates the spatial variation patterns of the six predictors. The figures come without explicit legend, serving as a visual representation of the spatial patterns. For the dynamic predictors (pressure drop, reservoir compaction and induced stress), the situation at the end of the observation period (i.e., 2022-01-01) has been chosen. The numerical value ranges of the predictors are shown in Table 1.

[Figure 4 about here.]

In the statistical analysis that follows, earthquakes are associated, or 'labelled', with the predictor values at the origin time and location of the earthquake according to the catalogue. For example, for the predictor reservoir thickness, we label each earthquake with the reservoir thickness at the location of the earthquake, while for the predictor induced stress, we label each earthquake with the induced stress at the location and the origin time of the earthquake. To facilitate the statistical analysis, after the earthquakes have been labelled, we perform a linear rescaling to the covariate values, such that the earthquake with the lowest covariate value receives a covariate value of 0 and the earthquake with the largest covariate value receives rescaled value of 1. In most figures, the minimum and maximum covariate values are simply labeled as 'min' and 'max' respectively.

The labelling of each event by the predictor values determines an ordering of observed magnitudes specific to that predictor. These predictor-specific orderings are displayed in Figure 5. The corresponding value ranges are given in Table 1. If a reservoir property has a predictive capacity with regard to the magnitude distribution, then the ordering (and spacing) of the magnitudes may be distinguishable from random orderings. In that case, the property has the potential to be used as a predictor in a seismic hazard and risk forecasting. In the following section we use this concept of ordering to define a null hypothesis.

[Figure 5 about here.]

[Table 1 about here.]

[1] Equation numbers in this footnote refer to S. Bourne and Oates (2017). From Equation 55 we obtain $\Delta C = \mu H_r \epsilon_{zz} - \gamma H \Gamma \varepsilon_{zz}$. For steeply dipping faults that offset the reservoir, $\Delta C = -\gamma H \Gamma \epsilon_{zz}$, which leads to Equation 57. The γ parameter is a field constant and is absorbed into the model parameter θ_2, which gives $\Delta C = H \Gamma \epsilon_{zz}$. Using Equation 15 $H\epsilon_{zz} = (H_r^{-1} + H_s^{-1})^{-1} \epsilon_{zz}$. Since $\epsilon_{zz} = \Delta P c_m$, we can write $\Delta C = (H_r^{-1} + H_s^{-1})^{-1} \Gamma \Delta P c_m$. Finally, using Equation 15 to obtain $H_r^{-1} = c_m$, we can write $\Delta C = \Gamma \Delta P c_m / (H_s^{-1} + c_m)$.

2.4 Statistical toolkit

2.4.1 Null hypothesis realization by random shuffling

If a predictor has predictive power for the magnitude distribution, then that implies that the ordering of magnitudes relative to that predictor may carry information on variations and trends. This also means that when the magnitudes of the events are randomly shuffled, i.e., all magnitudes are reassigned in the catalogue randomly to the predictor labels, that information will be erased. As a result, such a random shuffling represents a realization of a magnitude distribution that carries no relation to the predictor. In other words, the magnitude distribution is constant relative to the predictor. Random shufflings may therefore be regarded as realizations of a constant distribution.

In the following we consistently make use of random shufflings as realizations of a constant null hypothesis. If we observe a certain quantitative features in the data we may judge the feature significance from the occurrence frequency in random shufflings.

An alternative approach to realize samples of the null hypothesis would be to generate (new) magnitude samples from a (constant) magnitude distribution based on the Gutenberg-Richter distribution. However, by applying the random shuffling technique we make sure that the generated statistics are not contaminated by any possible deviation of our observations from an idealized Gutenberg-Richter distribution. By exactly honouring the empirical magnitude distribution we can focus on the spatiotemporal variations.

2.4.2 Moving window analysis

To gain some first-order insight into the spatiotemporal variations in the magnitude distribution as characterized by the MLE b-value, we apply moving window analysis. We perform a 1-D analysis on the set of six predictors that includes time, and a 2-D analysis on the spatial coordinates.

We define an algorithm that works consistently for both 1-D and 2-D cases and automatically adapts the algorithm resolution to the available data. We associate to each event in the catalogue a "window" of the N "nearest" neighbouring events. Then, for each window we determine the MLE b-value, under the assumption of a constant magnitude distribution, according to Equation 8. Note that we use the MLE b-value simply as a statistic on the data, without any reference to its uncertainty or evaluating the appropriateness of the Gutenberg-Richter model that it defines.

What remains is to define the concept of "nearest" that we apply. Although it is possible to use the Euclidean distance in terms of the covariate at hand (i.e., the difference in predictor values in 1-D, or the geometric distance in 2-D space), we have chosen instead to use the Euclidean distance in terms of the sequence number on each of the covariates. For space this means that we first define for each event a sequence number for both coordinates. This choice of nearest neighbourhood definition was made for a practical purpose. Using sequence numbers, the statistics of the null-hypothesis (i.e., constant magnitude distribution) are exactly the same for all 1-D analyses. Using the distance in terms of the covariate values, however, can lead to slightly different window configurations for each predictor and therefore slightly different statistics.

We use windows of $N = 51$ and $N = 101$ events. These choices are to a large extent arbitrary, but naturally represent a sort of compromise between resolution and stability. Within the resolution limits imposed by the size of the windows, the moving window analysis is sensitive to any type of b-value variation in the 1-D or 2-D domain considered and therefore is suitable for exploratory research. However, for the benefit of model building for forecasting purposes and hazard and risk assessment, it is more practical to consider simpler models with a limited number of parameters. These are introduced in the next section.

If there are spatiotemporal variations in the b-value, or if a prospective predictor does carry information on b-values variations, then we may expect the b-value variability among the collection of windows higher than if the b-value would actually be a constant. Like Schorlemmer, Neri, Wiemer, and Mostaccio (2003), we choose the difference between the maximum and the minimum b-value as a statistic representing these variations, and call it the min-max statistic. The min-max statistic

obtained for the observed catalogue can be compared with the distribution of this statistic under the null hypothesis to determine its significance.

2.4.3 Functional forms and maximum-likelihood estimation

If we can establish that a certain reservoir property carries information on the magnitude distribution we may try to exploit the predictive power of such property for more precise forecasting. We apply a total of 5 functional forms for the scalar predictor/covariate $c \in [0, 1]$, with one, two or three free parameters $(\theta_0, \theta_1, \theta_2)$:

- constant: $b = \theta_0$,
- linear: $b = \theta_0 + (\theta_1 - \theta_0)c$,
- quadratic: $b = \theta_0 + (\theta_1 - \theta_0)c + \theta_2 c(c - 1)$,
- step: $b = \theta_0 + (\theta_1 - \theta_0)H(c - \theta_2)$, and
- hyperbolic tangent: $b = \theta_0 + \theta_1(1 - \tanh(\theta_2 \times c))$,

with $H(c)$ the Heaviside step function. For our statistical investigations the constant model represents the null hypothesis. In all cases we determine the maximum-likelihood parameter estimates on both the observed data and on the synthetic realizations (by random shuffling) of the null hypothesis. The linear and the step model are subjected to significance tests. In particular, we use both the parameter difference $(\theta_1 - \theta_0)$ and the (maximum) likelihood as test statistics. The quadratic model is used in addition to compare the performance of the models in terms of their maximum-likelihood. We look at the added value of the extra quadratic term relative to the linear model and we compare the performance of the quadratic and step models in relation to their equal number of parameters.

Finally, the hyperbolic tangent (tanh) model is used, albeit only for the induced stress predictor, as it is used in the Groningen source model of S. J. Bourne and Oates (2020). We note that for the tanh model we apply a slightly different scaling for the covariate. To reproduce the model of S. J. Bourne and Oates (2020) we need to associate $c = 0$ with the zero induced stress rather than with the lowest value of stress in the catalogue as shown in Figure 5. The value $c = 1$ is still associated with the maximum value of stress in the catalogue.

2.4.4 Cramér-von Mises test

The maximum-likelihood regression of the step model effectively leads to a partitioning (in two parts) of both the predictor range and the catalogue ordered according to this predictor, each with a constant b-value. The significance of this partition can be further studied by a two-sample goodness-of-fit test. Such a test attempts to reject the null hypothesis that the two samples are actually generated by the same distribution. Commonly applied tests include Kolmogorov-Smirnov, Anderson-Darling, and Cramér-von Mises (Stephens, 1970; Darling, 1957). For our experiments we picked the Cramér-von Mises test as it turned out to be the most efficient in terms of computation time, while there was no particular reason to prefer one over the other.

Important to note, however, is that the p-value result from the goodness-of-fit test cannot be used without the following consideration. Due to the maximum-likelihood optimization of the step model, the two subsets are not completely independent anymore. In fact, the total likelihood of the step model benefits if both subsets are as dissimilar as possible. As a result, an overabundance of low p-values is to be expected even for the null hypothesis. Therefore, the Cramér-von Mises test needs to be recalibrated for this particular purpose. This is achieved by the experiment illustrated in Figure 6. The yellow curve shows that the Cramér-von Mises test works as expected for random partitions of a Groningen-sized constant b-value catalogue. The blue curve shows that introducing the optimization step compromises the test results. However, a correction is obtained relatively easily by applying the inverse CDF of the p-value distribution (blue curve). The formal test result appears on the horizontal axis, while the corrected result appears on the vertical axis. A formal Cramèr-von Mises p-value of 0.05 should be corrected to a p-value of 0.36 for this experiment (as indicated by the grey crosshair).

[Figure 6 about here.]

2.4.5 Likelihood ratio and the Akaike Information Criterion

As we investigate a total of six prospective predictors and five functional forms, we have quite a collection of statistical models for which we can assess the performance in terms of their maximum-likelihood with respect to the Groningen observations. At this point we immediately want to make the disclaimer that we do not intend to apply these maximum-likelihood models directly to hazard and risk assessments. For that purpose we prefer to apply the models in a Bayesian context, where we can take all uncertainties into account and make use of probability distributions of the model parameters rather than just the maximum-likelihood point estimates used in this study. Moreover, we would like to submit these models to pseudo-prospective testing and performance assessment before deciding on their use (Zechar, Gerstenberger, & Rhoades, 2010; S. J. Bourne et al., 2014, e.g.,).

That being said, model selection is commonly based on the likelihood-ratio test and information-theoretical extension such as the Akaike Information Criterion (Burnham, Anderson, & Huyvaert, 2011; Lewis, Butler, & Gilbert, 2011, e.g.,). These methods simply take maximum-likelihood results as an input.

The Akaike Information Criterion AIC_j for model j is defined as:

$$AIC_j = -2 \log \mathcal{L}_j + 2p_i, \tag{11}$$

where p_i is the number of model parameters (or degrees of freedom therein). According to this definition, a lower AIC corresponds to a better performance. Two models that differ by Δp degrees of freedom are considered to have an equal performance if their likelihoods differ by a factor $e^{-\Delta p}$. The inclusion of p_i in equation (11) is basically a bias correction that compensates for the higher likelihood values expected for models with a higher number of parameters under the null hypothesis that these parameters are not required.

The Akaike likelihood ratio $\mathcal{R}_{ij}$ between models i and j is basically a bias corrected likelihood ratio:

$$\mathcal{R}_{ij} = \exp\left(AIC_i - AIC_j\right)/2 = e^{(p_i - p_j)} \frac{\mathcal{L}_j}{\mathcal{L}_i}. \tag{12}$$

We compute AIC-corrected likelihood ratio's with respect to the constant b-value null hypothesis. This gives us relative measures of the model performance in terms of the relative likelihood. From an information-theory perspective these numbers indicate the relative probability that model j (relative to model i) is able to minimize the information loss inherent to the abstraction of reality in terms of a (mathematical) model.

We like to note that during the evaluation of the results of our analysis we found out that the Akaike formula (11) does adequately compensate the expected likelihood gain for the step model, as illustrated in Figure 7. Although the constant model is a nested model, i.e., special case of the step model, the likelihood ratio statistics are not chi-square distributed. It turns out that relative to the null hypothesis, the step model has an advantage that is higher than the number of its free parameters (3) would suggest. As a result, the AIC likelihood ratio's for the step model relative to the other models are expected to be inflated, i.e., biased by over-fitting. The performance results of the step models should therefore be interpreted with restraint. We speculate that the cause is related to the discontinuity of the model, and to the fact that the null hypothesis does not constrain the third parameter, i.e., the location of the discontinuity. Exploration of this specific hypothesis is beyond the scope of this paper. We expect that in a pseudo-prospective model performance testing procedure (Zechar et al., 2010), that we anticipate any new model to be subjected to before application in hazard and risk assessment, this issue will be treated adequately.

[Figure 7 about here.]

3 Results

The results of the analyses as described in Section 2 are shown in Figures 8 to 13, and in Tables 2 and 3.

Figure 8 shows the result of the spatial moving window analysis. A NW-SE trend can be seen in both the 51-event and the 101-event windows. The 51-event window results in a wider spread of MLE b-values. We determine the difference min-max statistic for both window sizes and compare it with the null hypothesis distribution. This comparison is visualized for both the spatial and the temporal moving windows in Figure 9. The Figure shows that the test statistic for temporal moving windows is not particularly special, as it is exceeded quite frequently in the null hypothesis distribution. As a result it cannot be used to reject the null hypothesis. The statistic for the spatial moving windows however is quite high and rarely exceeded for the null hypothesis. The corresponding p-values are 0.03 and 0.07 for the 51- and 101-event window sizes respectively. The moving window analysis does not resolve any trend or variation in time, but it is unlikely that the magnitude distribution is constant in space.

The moving window analyses for the other predictors follow the same procedure. The p-values are provided in Table 2, to be discussed below.

[Figure 8 about here.]

[Figure 9 about here.]

Figure 10 shows a visual summary of the moving window and maximum-likelihood analyses for all six considered covariates. The maximum-likelihood results are visualized in the spatial context in in Figures 11 and 12. Note that due to the static, spatial nature of these figures, the trends for the covariates reservoir thickness and topographic gradient, which are static spatial covariates, are easier to discern than for the dynamic covariates. It is interesting to see the difference in the empirical and MLE magnitude distributions for the catalogue partitioning that the step model effectively creates. These results are on display in Figure 13.

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

[Figure 13 about here.]

Table 2 summarizes the p-values of all statistical tests on the moving window and maximum-likelihood analyses. The table reveals that reservoir thickness as a predictor consistently scores lowest p-values, indicating a low probability that it does not carry any information on the magnitude distribution. To a lesser extent the same holds for topographic gradient on wide set of tests, and induced stress in particular for the linear trend model. In summary it appears that static factors are more informative than dynamic factors.

[Table 2 about here.]

Finally, Table 3 shows the relative likelihood of the various combinations of predictor and functional form. In this table we also include the results for the quadratic function for which we do not provide visualizations. The Table shows that the extra degree of freedom for quadratic models relative to the linear models does not lead to better results.The Table also shows an (apparently) exceptional performance of the step function, which we discuss in Section 2.4.5, and accompany with a warning to interpret with restraint. Overall, the Table shows that the maximum-likelihood models for reservoir thickness provide the best performance relative to the other predictors.

[Table 3 about here.]

4 Discussion

Previous analysis of variations of the magnitude distribution in the Groningen field focused on induced stress as the predicting covariate (S. J. Bourne & Oates, 2020). In this study, we have investigated the possibility of other covariates performing better as predictors of the observed earthquake magnitude distribution. Our exploratory moving window analysis indicated that while temporal variation in b-value appear to be statistically insignificant, there are significant spatial variations. In our subsequent search for a better b-value predictor, we tested a number of covariates, each of them with a distinct physical relation to the Groningen gas field. Reservoir thickness, a static property, unaffected by the gas production process, proves to be the statistically superior predictor of the b-value spatial variation.

In order to facilitate the comparison between different models, we have focused our attention on the relative performance of the MLE models. Although this provides a first-order indication of relative predictive performance, the performance assessment is not yet in line with the approach followed in seismic hazard and risk analysis. For future integration the models should be cast in a Bayesian framework, including the definition of a prior distribution for the model parameters and taking into account the full likelihood distribution of the model parameters conditional on the observation. In this framework a forecast is based on an integration of the posterior model parameter distribution. Therefore, ultimately, it is the performance of this integrated prediction that is of prime interest.

Nevertheless, our MLE based assessment provides valuable insights. Variations of b-value over the Groningen field seem to be predominantly controlled by static rather than dynamic factors, resulting in significant spatial variations, but (for the field as a whole) no significant changes in time. We find that the dynamic induced stress predictor does resolve a significant linear trend, albeit at lower significance level than reservoir thickness. Future work may investigate combinations of predictors to establish whether the static factors are sufficient predictors for b-value variations by itself, or that dynamic factors are able to contribute significantly.

Although our predictors were chosen because of their direct physical relation to the Groningen gas field, it should be noted that we are strictly looking at correlations. As such, the famous wisdom *correlation does not imply causation* also applies here. We do not propose a physical model that explains the observed effectiveness of the different predictors, nor do we claim that there necessarily is a direct causal relation between the more effective predictors (e.g. reservoir thickness) and the b-value. For reservoir thickness, we have shown that it is statistically very unlikely that the observed correlation is purely due to chance. However, it is possible (and perhaps even likely) that a confounding factor exists (i.e. that reservoir thickness *itself* is not the driving mechanism behind the b-value difference).

5 Conclusions

We have investigated spatiotemporal variations of magnitude distributions as characterized by the Gutenberg-Richter b-value in the Groningen gas field. We have found that spatial variations are more pronounced than variations in time. In addition we have investigated the predictability of observed variations in terms of a number of physical properties of the reservoir, including both static and dynamic properties, the latter being directly coupled to gas production.

We find evidence that observed variations are more likely to be controlled by static rather than by dynamic properties. Predictions in terms of the static properties topographic gradient and reservoir thickness lead to very low likelihood, around 2% and lower, of the null-hypothesis (i.e., no relation, or constant b-value) on a variety of statistical tests. Of the dynamic properties, induced stress is the most convincing predictor, still resolving a significant linear trend for the b-value.

In terms of relative likelihood, statistical models for b-values based on reservoir thickness outperform models based on the other predictor properties. An MLE linear model based on reservoir thickness outperforms the MLE linear model based on induced stress by a factor of 3. The hyperbolic tangent model based on induced stress, which is currently being applied in hazard and risk assessment models (S. J. Bourne & Oates, 2020), does not manage to improve on the linear trend with its extra parameter.

We find that step models, in general, reach higher likelihoods than linear and quadratic models, but we also note that these results may be inflated due to a level of over-fitting that is not adequately compensated for in the Akaike Information Criterion.

The main conclusion of this study is that reservoir thickness is a strong predictor for spatial b-value variations in the Groningen field. We propose to develop a forecasting model for Groningen magnitude distributions conditioned on reservoir thickness, to be used alongside, or as a replacement, for the current models conditioned on induced stress.

Acknowledgments

We thank the Groningen field operator NAM for providing the reservoir properties used as (input for) the prospective predictors investigated in this study. We thank KNMI for generating, maintaining and exposing the seismic monitoring database (KNMI, 2022b, 2022a), and the KEM subpanel (kemprogramma.nl) for several fruitful discussions about model performance and statistical methods. Most calculations and visualizations have been performed using Wolfram Mathematica (Wolfram Research, 2021).

References

Aki, K. (1965). Maximum likelihood estimate of b in the formula log n= a-bm and its confidence limits. *Bull. Earthq. Res. Inst., Tokyo Univ.*, *43*, 237–239.

Bourne, S., & Oates, S. (2017). Extreme threshold failures within a heterogeneous elastic thin sheet and the spatial-temporal development of induced seismicity within the Groningen gas field. *Journal of Geophysical Research: Solid Earth*, *122*(12), 10–299.

Bourne, S. J., & Oates, S. J. (2020). Stress-dependent magnitudes of induced earthquakes in the groningen gas field. *Journal of Geophysical Research: Solid Earth*, *125*(11), e2020JB020013. doi: https://doi.org/10.1029/2020JB020013

Bourne, S. J., Oates, S. J., van Elk, J., & Doornhof, D. (2014). A seismological model for earthquakes induced by fluid extraction from a subsurface reservoir. *Journal of Geophysical Research: Solid Earth*, *119*(12), 8991-9015.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011, January). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35. Retrieved from `http://link.springer.com/article/10.1007/s00265-010-1029-6` doi: 10.1007/s00265-010-1029-6

Candela, T., Osinga, S., Ampuero, J.-P., Wassing, B., Pluymaekers, M., Fokker, P. A., ... Muntendam-Bos, A. G. (2019). Depletion-induced seismicity at the groningen gas field: Coulomb rate-and-state models including differential compaction effect. *Journal of Geophysical Research: Solid Earth*, *124*(7), 7081-7104. doi: https://doi.org/10.1029/2018JB016670

Darling, D. A. (1957). The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, *28*(4), 823–838.

De Jager, J., & Visser, C. (2017). Geology of the groningen field – an overview. *Netherlands Journal of Geosciences*, *96*(5), s3–s15. doi: 10.1017/njg.2017.22

Dost, B., Ruigrok, E., & Spetzler, J. (2017). Development of seismicity and probabilistic hazard assessment for the groningen gas field. *Netherlands Journal of Geosciences*, *96*(5), s235–s245. doi: 10.1017/njg.2017.20

Eaton, J., O'neill, M., & Murdock, J. (1970). Aftershocks of the 1966 parkfield-cholame, california, earthquake: A detailed study. *Bulletin of the Seismological Society of America*, *60*(4), 1151–1197.

epsg.io. (2022). *Amersfoort / RD New - Netherlands.* `https://epsg.io/28992`, last accessed 2022-01-31.

Gerstenberger, M., Wiemer, S., & Giardini, D. (2001). A systematic test of the hypothesis that the b value varies with depth in california [Journal Article]. *Geophysical Research Letters*, *28*(1), 57-60.

Goertz-Allmann, B., Gibbons, S., Oye, V., Bauer, R., & Will, R. (2017). Characterization of induced seismicity patterns derived from internal structure in event clusters. *Journal of Geophysical Research: Solid Earth*, *122*(5), 3875–3894.

Gulia, L., & Wiemer, S. (2010). The influence of tectonic regimes on the earthquake size distribution: A case study for Italy [Journal Article]. *Geophysical Research Letters*, *37*(10).

Gutenberg, B., & Richter, C. (1941). *Seismicity of the earth* (Vol. 34). Geological Society of America.

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

Henderson, J., Barton, D., & Foulger, G. (1999). Fractal clustering of induced seismicity in the geysers geothermal area, california. *Geophysical Journal International*, *139*(2), 317–324.

Herrmann, M., & Marzocchi, W. (2020, 12). Inconsistencies and Lurking Pitfalls in the Magnitude–Frequency Distribution of High-Resolution Earthquake Catalogs. *Seismological Research Letters*, *92*(2A), 909-922. Retrieved from `https://doi.org/10.1785/0220200337` doi: 10.1785/0220200337

Hirata, T. (1989). A correlation between the b value and the fractal dimension of earthquakes. *Journal of Geophysical Research: Solid Earth*, *94*(B6), 7507–7514.

KNMI. (2022a). *FDSN Event Web Service.* `http://rdsa.knmi.nl/fdsnws/event/1/query?starttime=1995-01-01T00%3A00%3A00&minlatitude=53.&maxlatitude=53.7&minlongitude=6.4&maxlongitude=7.3&format=csv`, last accessed 2022-01-31.

KNMI. (2022b). *Netherlands Seismic and Acoustic Network.* Royal Netherlands Meteorological Institute (KNMI), Other/Seismic Network. (10.21944/e970fd34-23b9-3411-b366-e4f72877d2c5)

Kortekaas, M., & Jaarsma, B. (2017). Improved definition of faults in the groningen field using seismic attributes. *Netherlands Journal of Geosciences*, *96*(5), s71–s85. doi: 10.1017/njg.2017.24

Lahaie, F., & Grasso, J. R. (1999). Loading rate impact on fracturing pattern: Lessons from hydrocarbon recovery, lacq gas field, france. *Journal of Geophysical Research: Solid Earth*, *104*(B8), 17941–17954.

Langenbruch, C., & Shapiro, S. (2014). Gutenberg-richter relation originates from coulomb stress fluctuations caused by elastic rock heterogeneity. *Journal of Geophysical Research: Solid Earth*, *119*(2), 1220–1234.

Lewis, F., Butler, A., & Gilbert, L. (2011). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, *2*(2), 155–162.

Mandal, P., & Rastogi, B. (2005). Self-organized fractal seismicity and b value of aftershocks of the 2001 bhuj earthquake in kutch (india) [Journal Article]. *pure and applied geophysics*, *162*(1), 53-72.

Marzocchi, W., Spassiani, I., Stallone, A., & Taroni, M. (2019, 11). How to be fooled searching for significant variations of the b-value. *Geophysical Journal International*, *220*(3), 1845-1856. Retrieved from `https://doi.org/10.1093/gji/ggz541` doi: 10.1093/gji/ggz541

Mori, J., & Abercrombie, R. E. (1997). Depth dependence of earthquake frequency-magnitude distributions in california: Implications for rupture initiation [Journal Article]. *Journal of Geophysical Research: Solid Earth*, *102*(B7), 15081-15090.

Mukuhira, Y., Fehler, M. C., Ito, T., Asanuma, H., & Häring, M. O. (2021). Injection-induced seismicity size distribution dependent on shear stress. *Geophysical Research Letters*, *48*(8), e2020GL090934.

Nakaya, S. (2006). Spatiotemporal variation in b value within the subducting slab prior to the 2003 tokachi-oki earthquake (m 8.0), japan [Journal Article]. *Journal of Geophysical Research: Solid Earth*, *111*(B3).

NAM. (2021). *Reservoir data for public seismic hazard and risk assessment.*

Rajesh, R., & Gupta, H. K. (2021). Characterization of injection-induced seismicity at north central oklahoma, usa. *Journal of Seismology*, *25*(1), 327–337.

Rundle, J. B. (1989). Derivation of the complete gutenberg-richter magnitude-frequency relation using the principle of scale invariance [Journal Article]. *Journal of Geophysical Research: Solid Earth*, *94*(B9), 12337-12342.

Scholz, C. (1968). The frequency-magnitude relation of microfracturing in rock and its relation to earthquakes [Journal Article]. *Bulletin of the seismological society of America*, *58*(1), 399-415.

Scholz, C. H. (2015). On the stress dependence of the earthquake b value [Journal Article]. *Geophysical Research Letters*, *42*(5), 1399-1402.

Schorlemmer, D., Neri, G., Wiemer, S., & Mostaccio, A. (2003). Stability and significance tests for b-value anomalies: Example from the tyrrhenian sea. *Geophysical Research Letters*, *30*(16).

Schorlemmer, D., Wiemer, S., & Wyss, M. (2005). Variations in earthquake-size distribution across different stress regimes [Journal Article]. *Nature*, *437*(7058), 539.

Spada, M., Tormann, T., Wiemer, S., & Enescu, B. (2013). Generic dependence of the frequency-size distribution of earthquakes on depth and its relation to the strength profile of the crust [Journal Article]. *Geophysical research letters*, *40*(4), 709-714.

Stephens, M. A. (1970). Use of the kolmogorov–smirnov, cramer–von mises and related statistics without extensive tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *32*(1), 115–122.

Van Thienen-Visser, K., & Breunese, J. (2015). Induced seismicity of the groningen gas field: History and recent developments. *The Leading Edge*, *34*, 664-671. doi: 10.1190/tle34060664.1

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

Vlek, C. (2019). Rise and reduction of induced earthquakes in the groningen gas field, 1991–2018: statistical trends, social impacts, and policy change [Journal Article]. *Environmental Earth Sciences*, *78*(59), 1-14.

Vorobieva, I., Shebalin, P., & Narteau, C. (2020). Condition of occurrence of large man-made earthquakes in the zone of oil production, oklahoma. *Izvestiya, Physics of the Solid Earth*, *56*(6), 911–919.

Wiemer, S., & Benoit, J. P. (1996). Mapping the b-value anomaly at 100 km depth in the alaska and new zealand subduction zones [Journal Article]. *Geophysical Research Letters*, *23*(13), 1557-1560.

Wiemer, S., & Wyss, M. (1997). Mapping the frequency-magnitude distribution in asperities: An improved technique to calculate recurrence times? *Journal of Geophysical Research: Solid Earth*, *102*(B7), 15115–15128.

Wolfram Research, I. (2021).

Wyss, M. (1973). Towards a physical understanding of the earthquake frequency distribution [Journal Article]. *Geophysical Journal International*, *31*(4), 341-359.

Wyss, M., McNutt, S. R., & Wyss, M. (1998). Temporal and three-dimensional spatial analyses of the frequency–magnitude distribution near long valley caldera, california. *Geophysical Journal International*, *134*(2), 409–421.

Zechar, J. D., Gerstenberger, M. C., & Rhoades, D. A. (2010, 06). Likelihood-Based Tests for Evaluating Space–Rate–Magnitude Earthquake Forecasts. *Bulletin of the Seismological Society of America*, *100*(3), 1184-1195. Retrieved from `https://doi.org/10.1785/0120090192` doi: 10.1785/0120090192

List of Figures

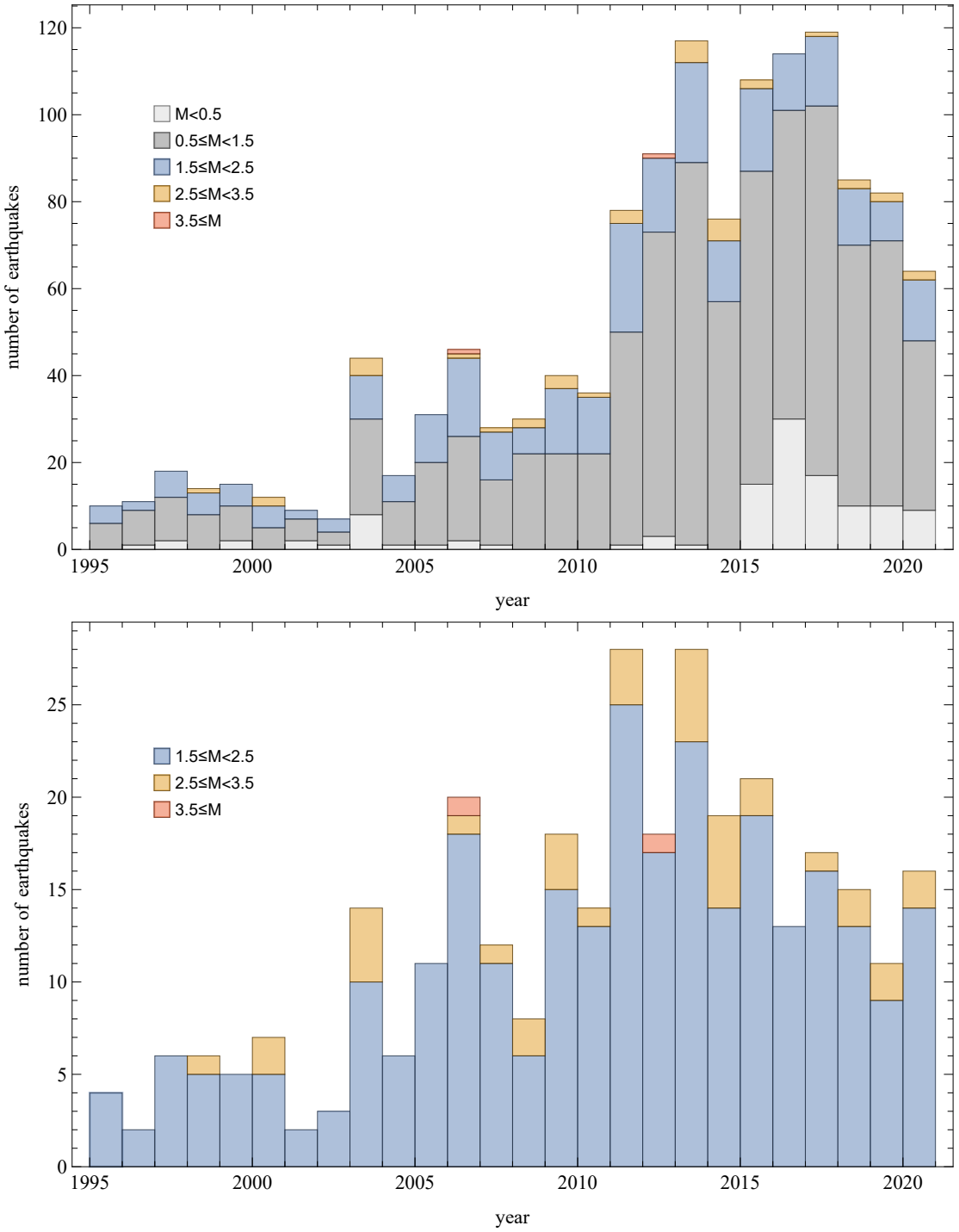

Figure 1: Number of earthquakes in the Groningen field in different magnitude categories. This Figure is created by first rounding the magnitudes to 1 decimal place and subsequently assigning the earthquakes to their magnitude categories. Only events within the outline of the Groningen gas field are included (see also Figure 2). Top view: all recorded events in said space/time window. Bottom view: the events above the minimum magnitude included in the current study (336 in total).

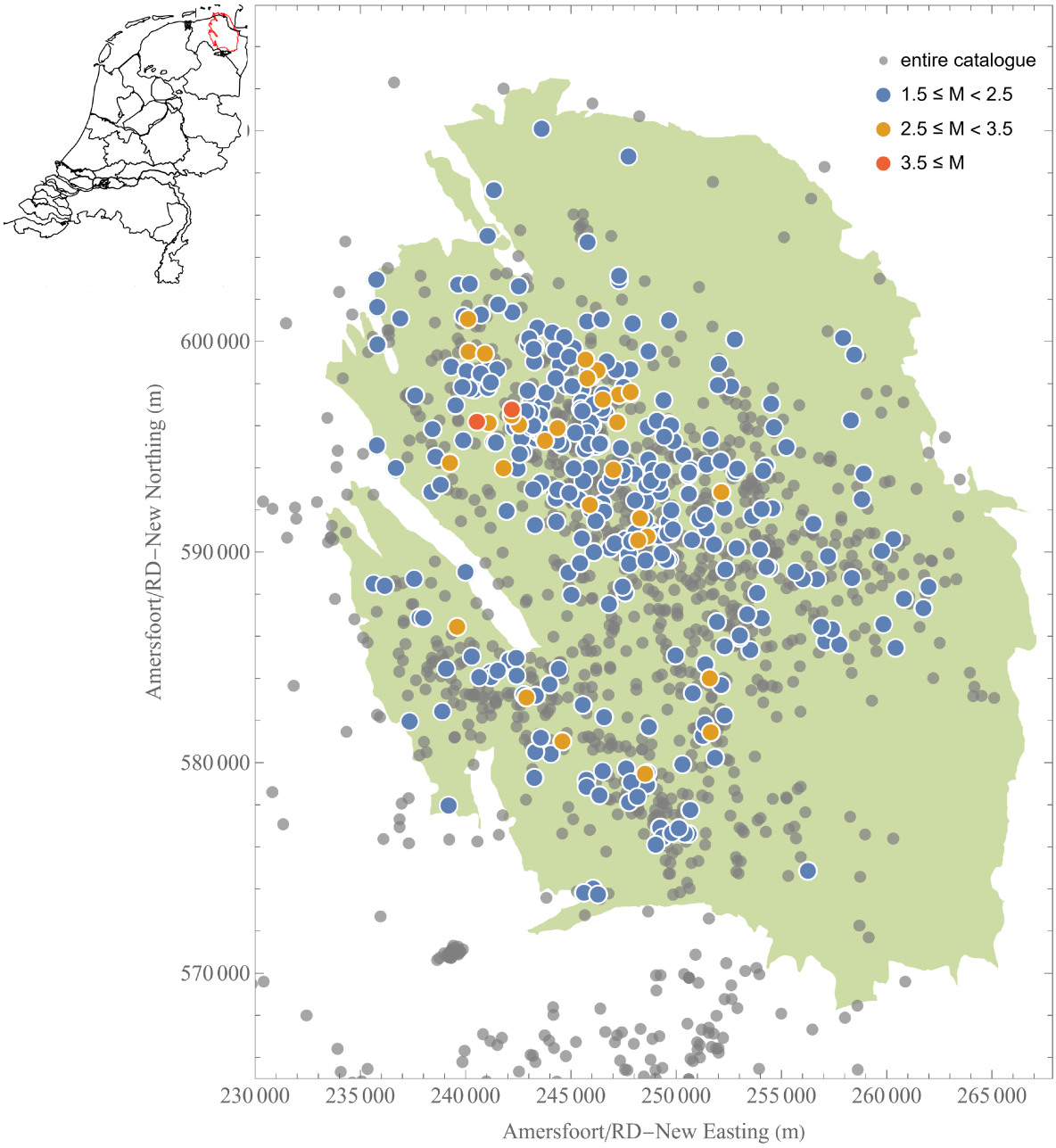

Figure 2: Map view of the Groningen gas field and its location inset (in red at the top-left corner). Locations of all recorded induced earthquakes at any time in the vicinity are shown by grey dots. The colored dots represent the earthquakes included in the current study, i.e., within the field outline and the time span from 1995-01-01 to 2022-01-01, and (1 decimal rounded) magnitudes of 1.5 and higher. The colors represent the magnitude categories, analogous to Figure 1.

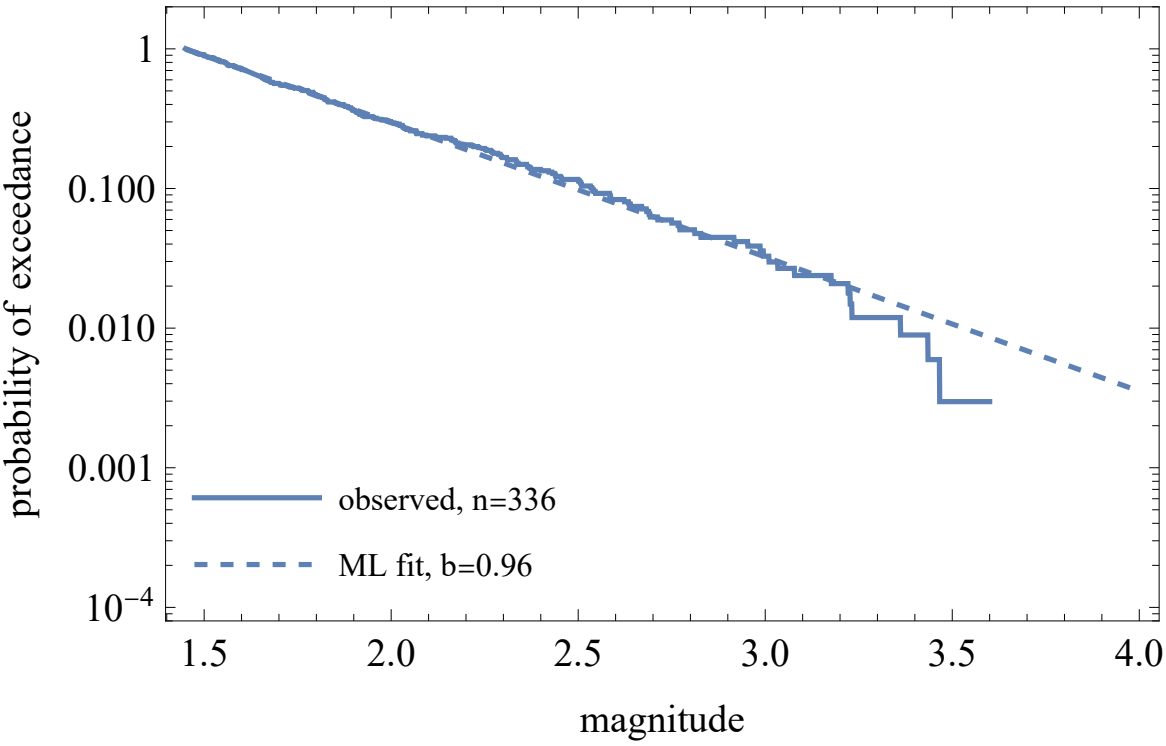

Figure 3: Empirical complementary cumulative distribution function (CCDF), or probability of exceedance per event. Also shown is the maximum-likelihood Gutenberg-Richter distribution under the assumption of a constant b-value. MLE for the constant b-value is 0.96.

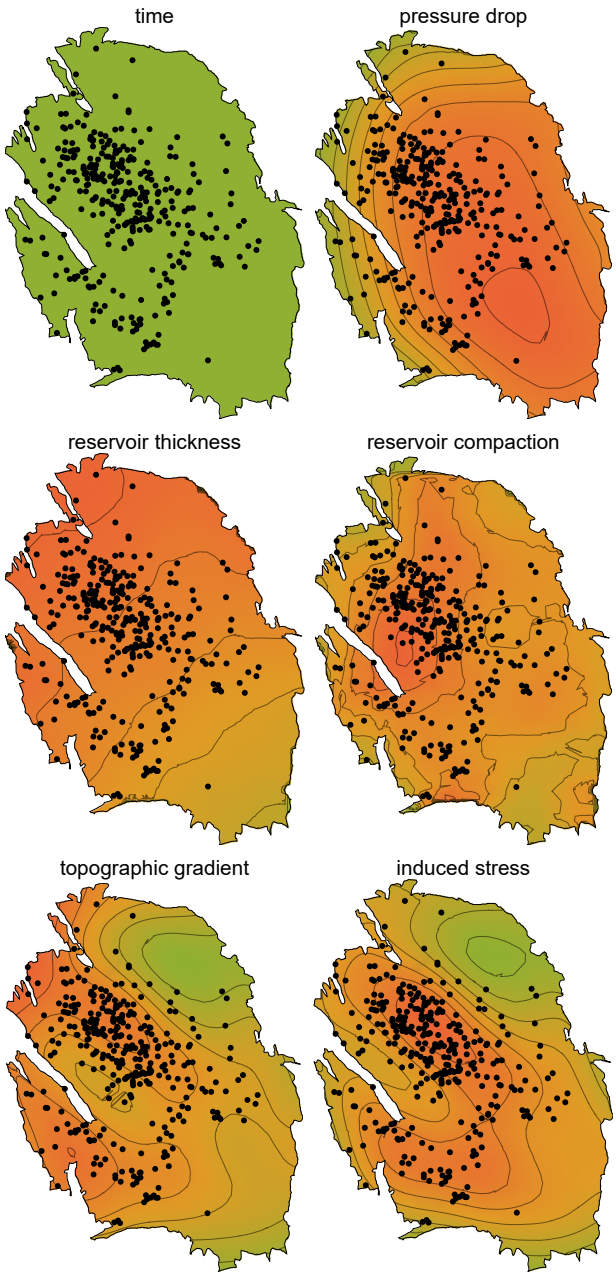

Figure 4: The six predictors for b-value variations investigated in this study represented as contour plots within the outline of the Groningen field. Each figure is individually scaled, where green colors correspond to the lowest, orange to the highest values. Representative values for the covariates are presented in Table 1. For the dynamic predictors (pressure drop, reservoir compaction and induced stress), the state at the end of the observation period (i.e., 2022-01-01) is shown.

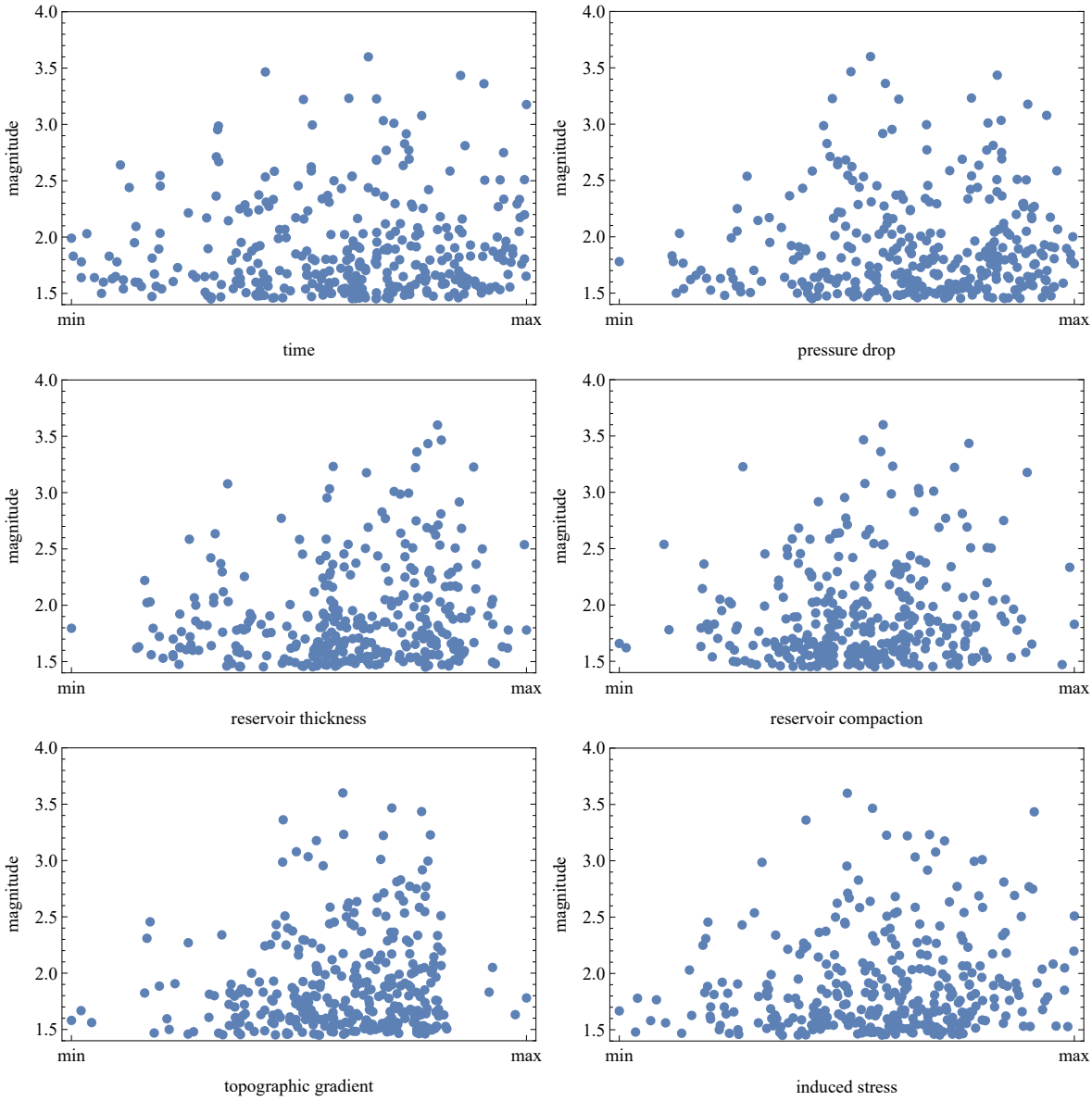

Figure 5: For each of the six prospective predictors, the magnitudes of the earthquakes in the seismic catalogue are plotted against the predictor value. Each predictor provides it own specific ordering and spacing of the catalogue.

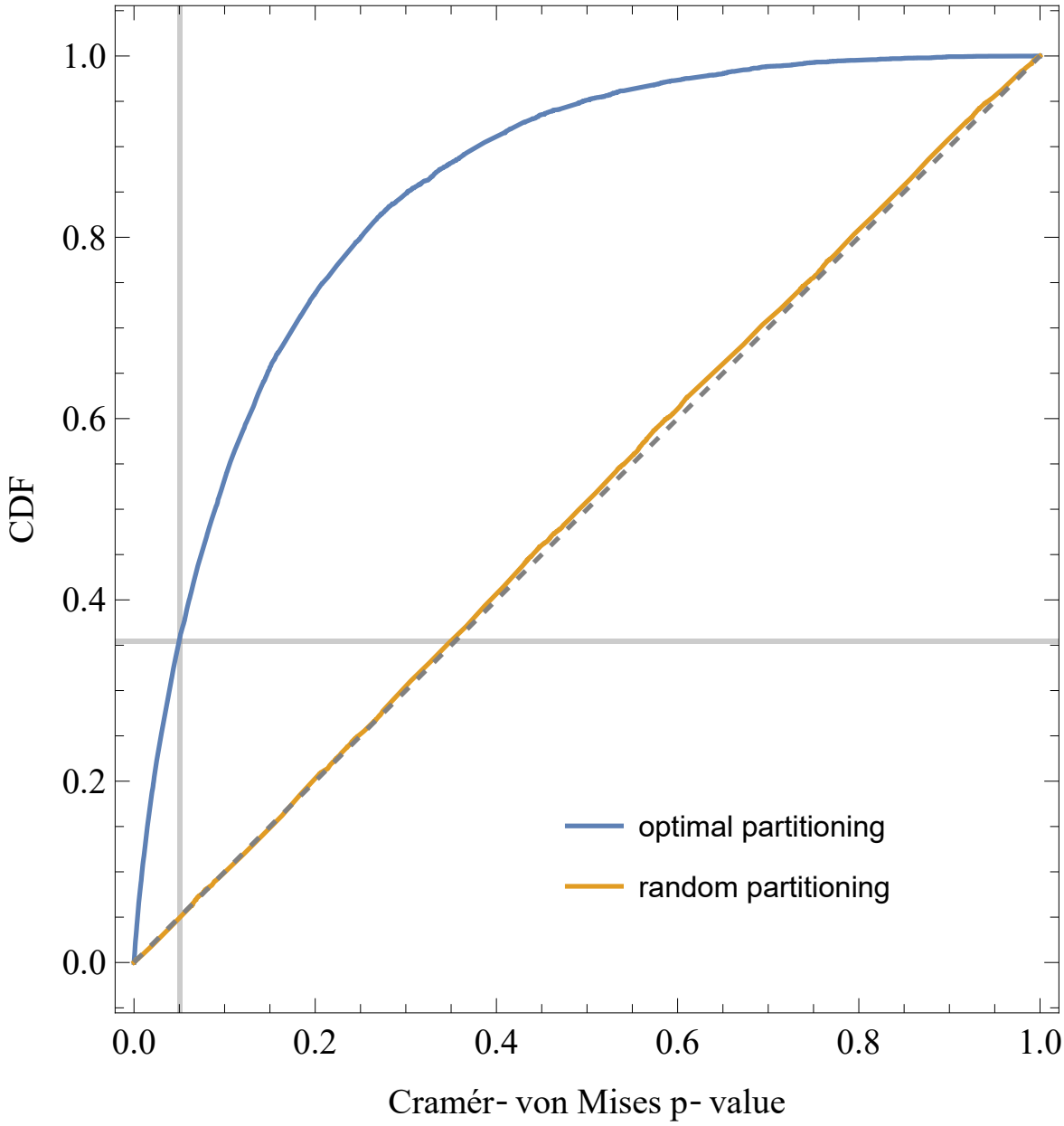

Figure 6: Results of applying the Cramér-von Mises test on two subsets of 10,000 randomly drawn catalogues ($N = 336, b = 0.96$). For each catalogue, a random split point is chosen in the catalogue, and an optimal split point is found by considering the step location of the maximum-likelihood step function. The Cramér-von Mises test is then applied, and a p-value is obtained for subsets created by the random split point and the optimal split point. The blue curve shows a CDF of p-values obtained over 10,000 catalogues for the optimal split point, while the yellow curve shows the CDF for the random split points. The Cramér-von Mises p-values for the yellow curve are distributed homogeneously between 0 and 1, while the blue curve shows an overabundance of low p-values. This indicates that the likelihood optimization corrupts the test, which should be corrected for. In fact, the blue curve provides the correction: the formal test result appears on the x-axis, while the corrected test result appears on the y-axis.

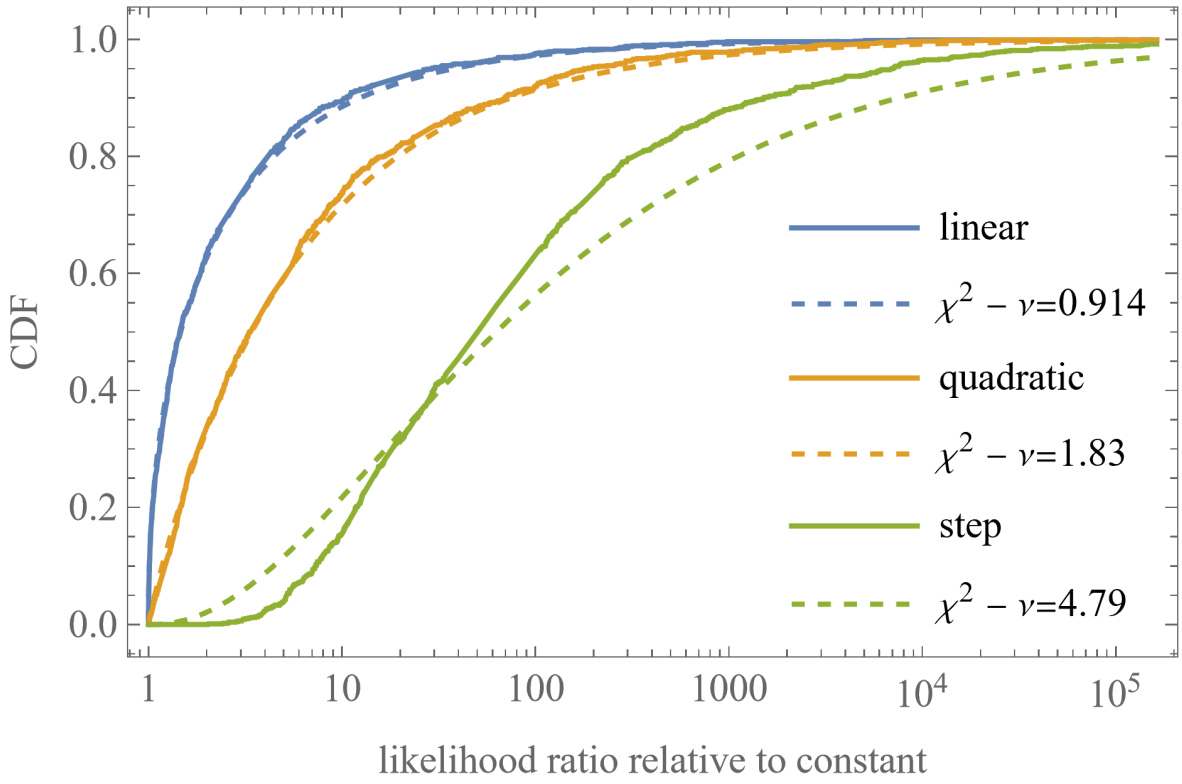

Figure 7: Likelihood ratio statistics for maximum-likelihood linear, quadratic and step b-value functions, relative to the maximum-likelihood constant b-value model. The statistics are obtained for 1,000 random reassignments of the magnitudes to the catalogue's reservoir thickness samples. Other predictors give comparable results. Both the linear and quadratic functions closely follow a maximum-likelihood chi-square distribution (dashed curves), with degrees of freedom very close to the theoretic values of 1 and 2, for functions with 1 and 2 degrees of freedom more, respectively, than the constant function. The step function, however, also has just two more parameters than the constant function, but apparently is expected to perform much better than the quadratic, and apparently is not chi-square distributed (the dashed curve is the maximum-likelihood chi-square fit to the data). As a result, the Akaike Information Criterion does not compensate adequately for the surplus degrees of freedom.

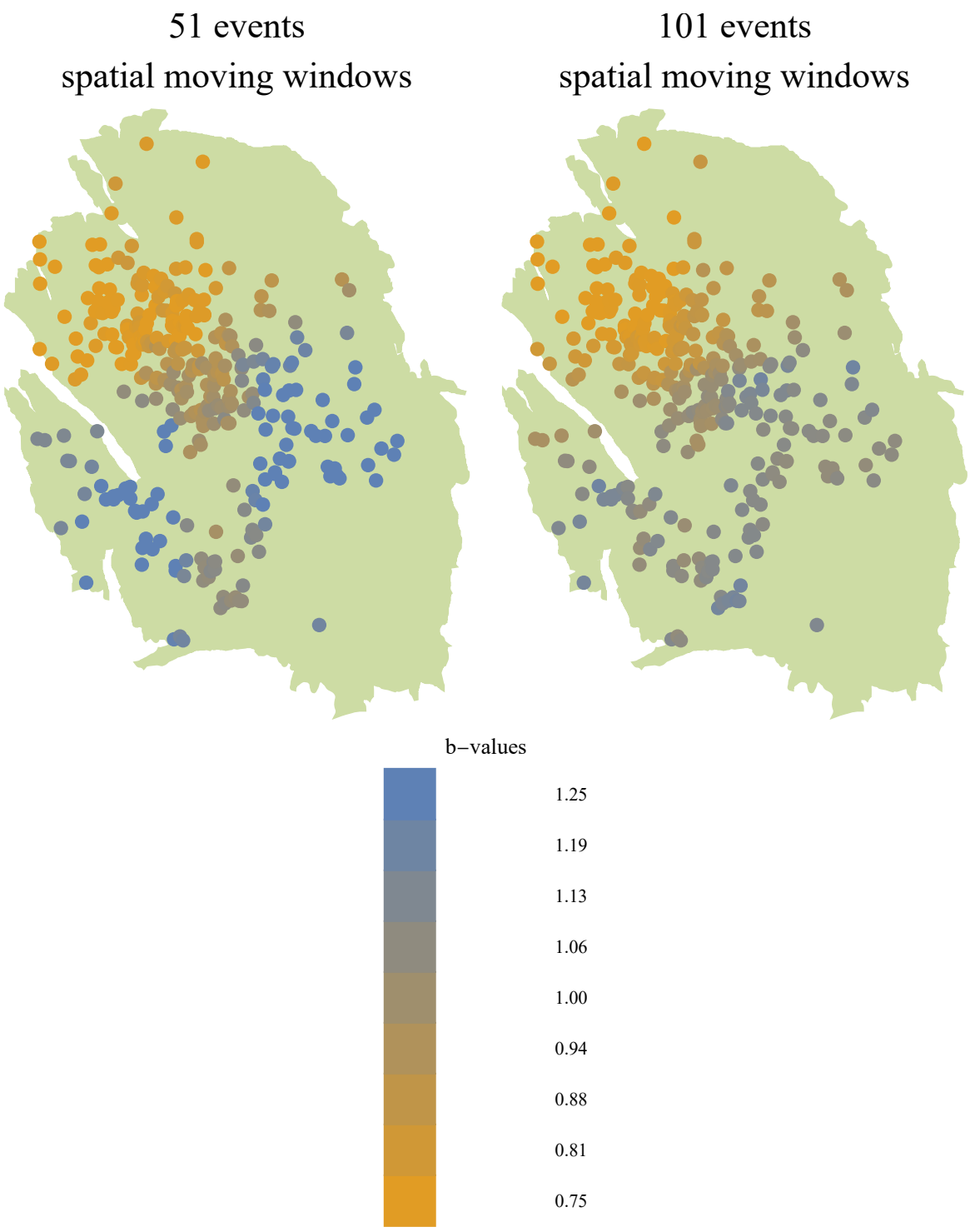

Figure 8: B-values resulting from the spatial moving window analysis. Each earthquake is assigned the MLE b-value for the sub-catalogue consisting of the event itself and its 50 (left) or 100 (right) nearest neighbors.

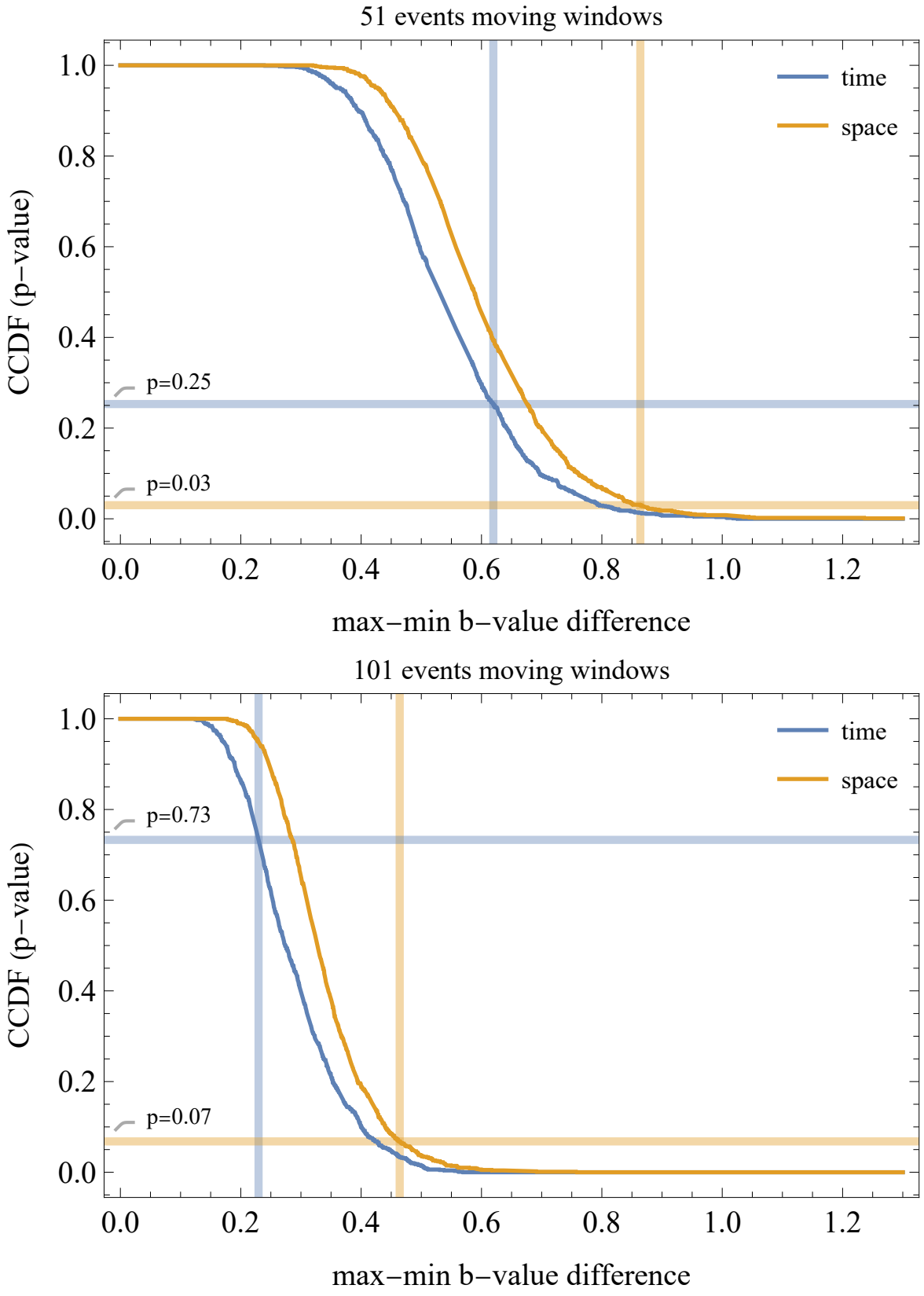

Figure 9: The graphs show empirical distributions (CCDF) for 1,000 51-event (top) and 101-event (bottom) moving window analyses in both time and space on random realizations of the null-hypothesis obtained by magnitude shuffling. The test statistic is the difference between the maximum and the minimum MLE b-value estimate in the moving window collection. Vertical bars indicate the observed values for the Groningen catalogue. The corresponding values on the vertical axis indicate the p-value.

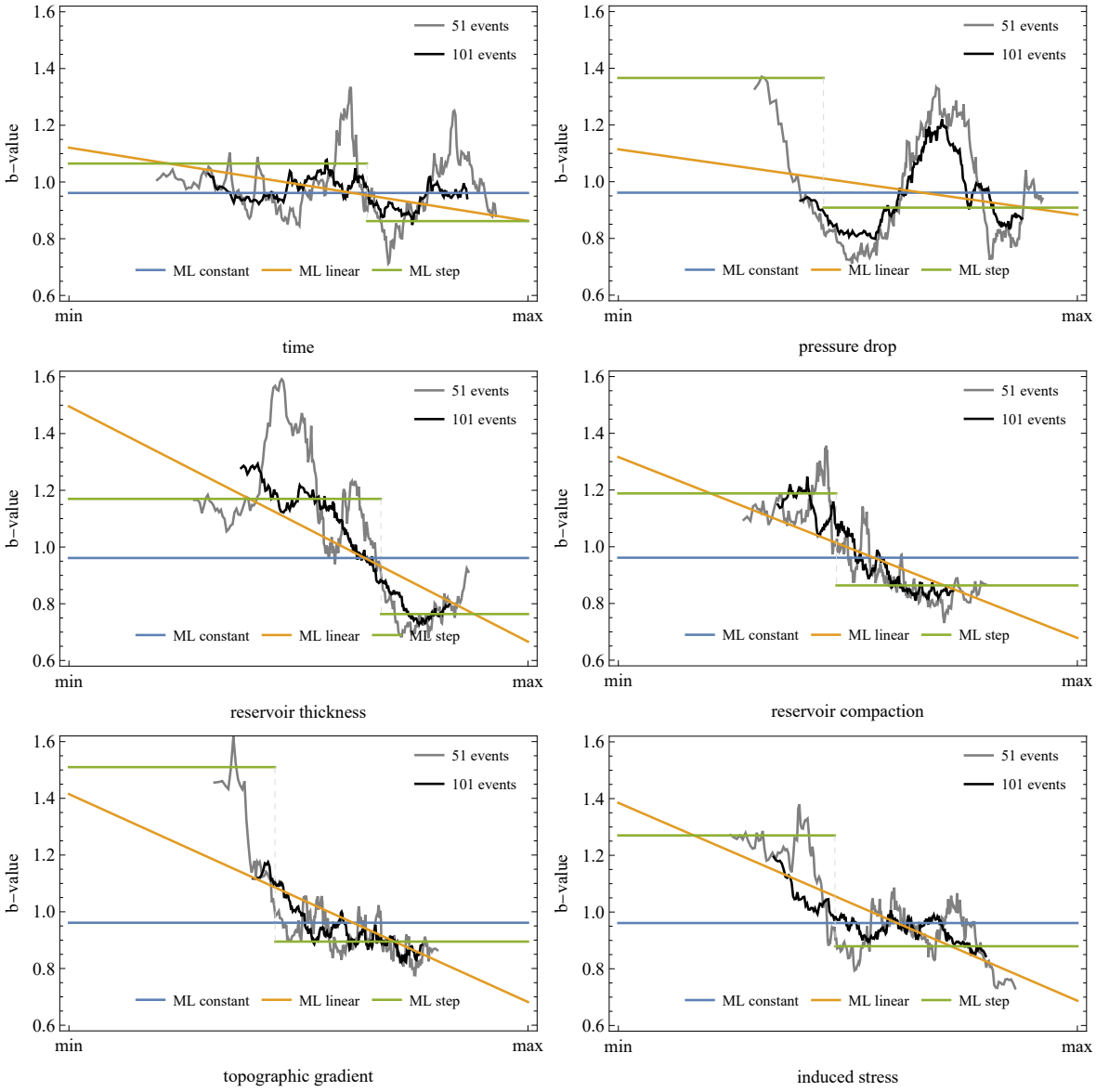

Figure 10: For each predictor, 5 lines are shown. In silver (51-event) and black (101-event) the moving windows analyses(see 2.4.2). We have chosen to plot the results for each window at the mean value of the contributing covariates. In blue, orange and green, the maximum-likelihood estimates of the constant, linear, and step models, respectively. Note that the moving-window results and the maximum-likelihood models are each independently generated from the magnitude data in Figure 5. In particular, the MLE functions are not intended to fit the moving window results.

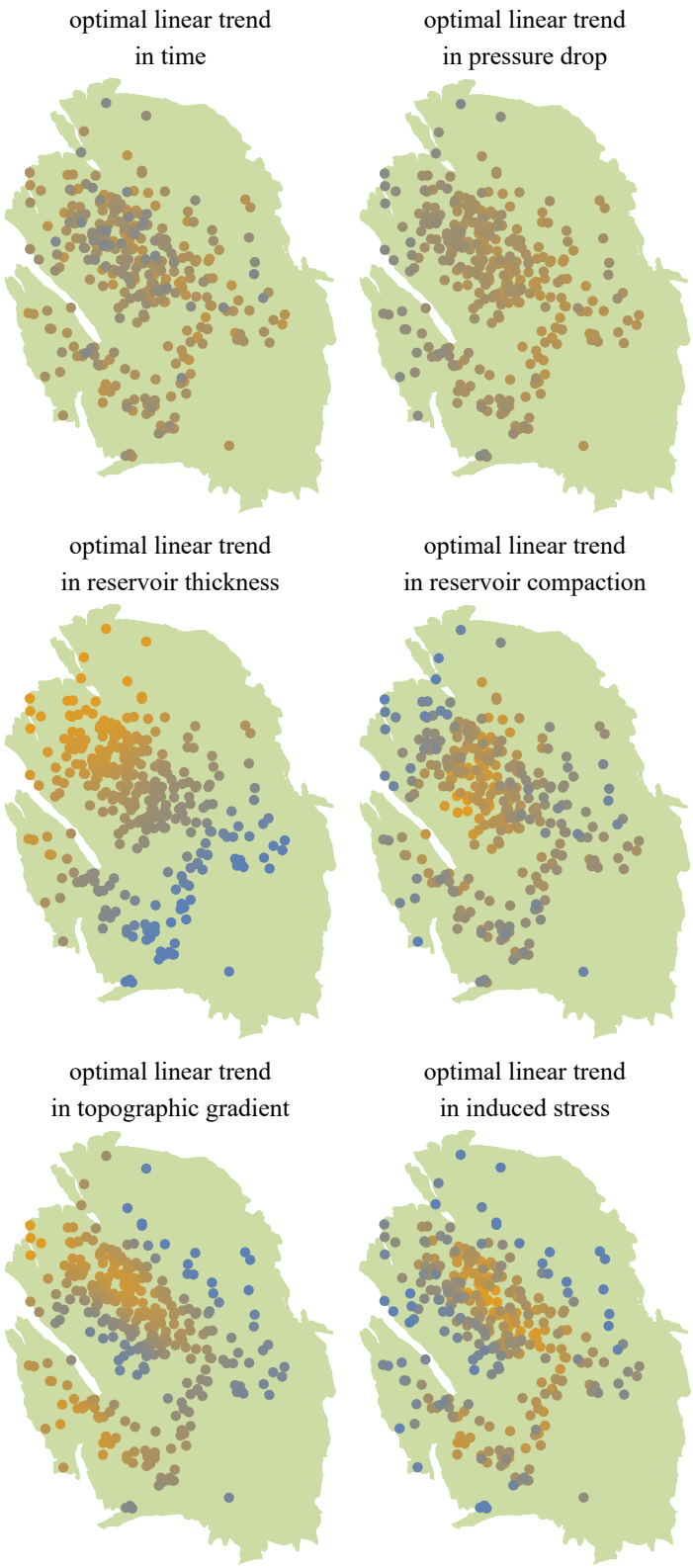

Figure 11: The MLE linear trend models for each predictor result in a b-value for each event in the catalogue. Here, we show the b-value assigned to each event in its spatial context. Yellow shades correspond to lower, blue shades to higher values of the covariate.

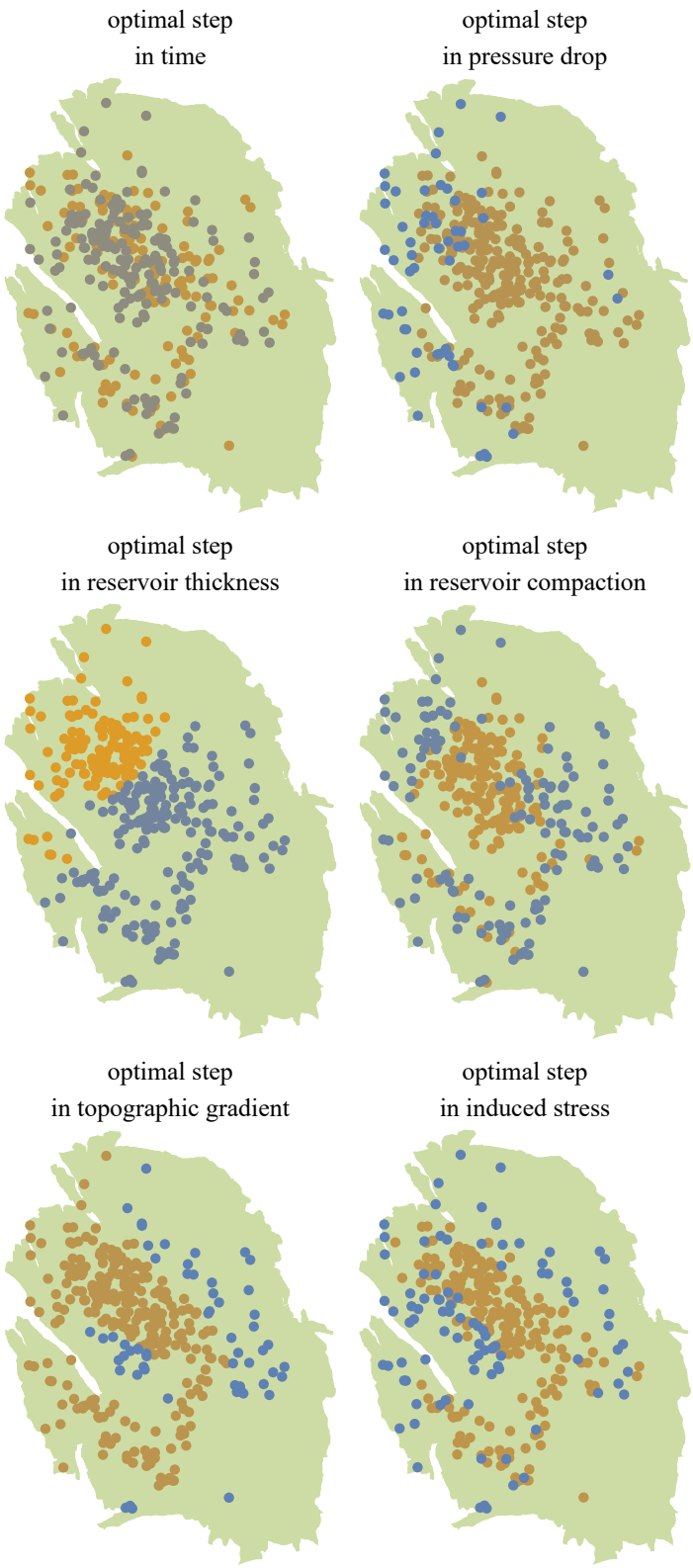

Figure 12: The MLE step-function models for each predictor result in a b-value for each event in the catalogue. Here, we show the b-value assigned to each event in its spatial context. Yellow shades correspond to lower, blue shades to higher values of the covariate.

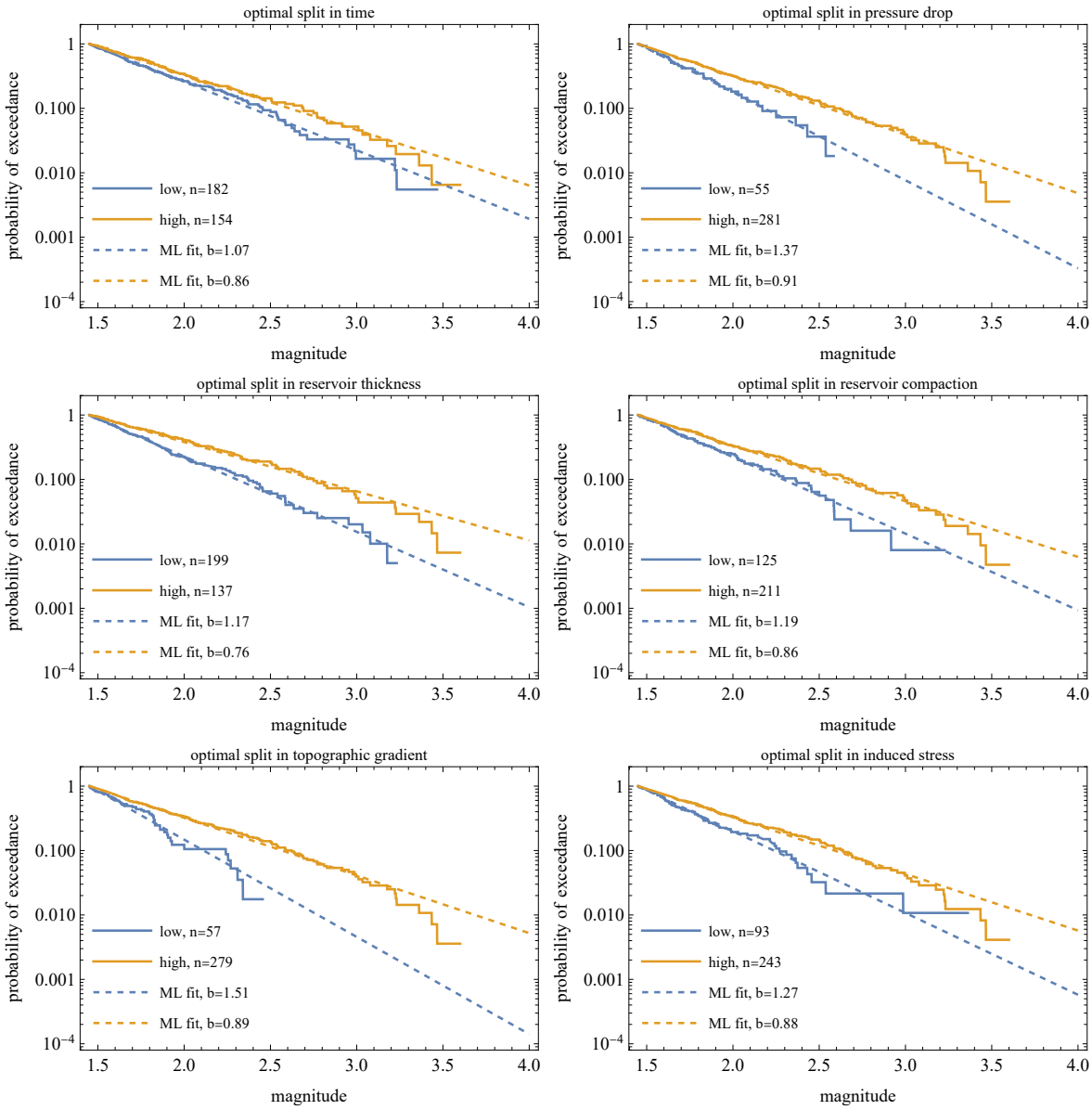

Figure 13: The MLE step-function models effectively separate the catalogue into 2 sub-catalogues, each with their own b-value: one for the low covariate values, and one for the high covariate values. This Figure shows the resulting sub-catalogues with their b-values.

List of Tables

predictor	**unit**	**min**	**max**
time	UTC	1995-01-24T10:38	2021-12-28T20:19
reservoir thickness	m	128.	310.
topographic gradient	-	0.0145	0.0442
pressure drop	bar	106.	279.
reservoir compaction	m	0.0787	0.349
induced stress	MPa	0.303	0.991

Table 1: Predictor value ranges as sampled by the catalogue.

predictor	moving window maximum likelihood b−value range		maximum likelihood linear trend model		maximum likelihood step model		
	51 events	101 events	gradient	likelihood	step size	likelihood	C−vM∗
time	22.%	73.%	23.%	24.%	84.%	54.%	37.%
reservoir thickness	0.77%	0.31%	0.1%	0.2%	28.%	0.2%	0.35%
topographic gradient	1.7%	19.%	1.5%	2.2%	16.%	1.%	8.3%
pressure drop	15.%	6.2%	38.%	38.%	26.%	15.%	33.%
reservoir compaction	21.%	4.8%	3.1%	3.2%	45.%	8.6%	20.%
induced stress	17.%	19.%	0.5%	0.6%	33.%	5.8%	6.9%

Table 2: For each of the six prospective predictors for b-value variations in the Groningen field, a total of seven statistics are compared to the distribution of results generated for the null-hypothesis, in which the predictor does not carry any information on the b-value. Realizations of the null-hypothesis are generated by random shuffling of the observed magnitudes with respect to the predictor values. The exceedance probabilities, i.e., p-values, indicate the probability that the observed statistics are the result of chance. Lower values give stronger stronger evidence for rejecting the null hypothesis that a b-value is constant. Green colors give a visual indication of the strength of the evidence.

predictor	functional form	no. parameters	relative likelihood (AIC)
−any−	constant	1	1.
time	linear	2	0.72
	quadratic	3	0.26
	step	3	0.88
reservoir thickness	linear	2	28.
	quadratic	3	10.
	step	3	250.
topographic gradient	linear	2	3.9
	quadratic	3	4.2
	step	3	43.
pressure drop	linear	2	0.55
	quadratic	3	0.53
	step	3	4.4
reservoir compaction	linear	2	3.1
	quadratic	3	1.1
	step	3	6.5
induced stress	linear	2	11.
	quadratic	3	4.5
	step	3	9.9
	tanh	3	4.9

Table 3: Relative likelihood of predictive models for b-value variations in the Groningen field consisting of simple functional forms conditioned on six possible predictor covariates. The likelihoods are calculated according to the Akaike Information Criterion and normalized relative to the likelihood of the constant model, or null hypothesis. Red and green color shades indicate worse and better performance than the null hypothesis, respectively.