UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

A generic probabilistic model for natural hazard assessment

By

Ansie Smit

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor (Mathematical Statistics)

In the Faculty of Natural and Agricultural Sciences

University of Pretoria

(June 2019)

Declaration

I, Ansie Smit, declare that the thesis, which I hereby submit for the degree Philosophiae Doctor (Mathematical Statistics) at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution

SIGNATURE:……..…….

DATE:…….

Ethics Statement

The author, whose name appears on the title page of this thesis, has obtained, for the research described in this work, the applicable research ethics approval.

The author declares that he/she has observed the ethical standard required in terms of the University of Pretoria's Code of ethics for researchers and the Policy guidelines for responsible research.

Acknowledgments

I would like to thank God for providing me with the opportunity, strength, patience and perseverance to undertake this research. Without his blessings, this would not have been possible.

I wish to express my sincere appreciation and gratitude to the following people for their constructive guidance and unfailing support, without which this thesis would not have been possible:

- My supervisor and mentor, Prof. Andrzej Kijko, for his continuous motivation, guidance, and support. Thank you for always expecting the best, while reminding me what is truly important in life.
- My co-supervisor, Prof. Alfred Stein, for his valuable input and ideas related to this thesis and publications and the writing thereof.
- My family, Arnold, Susanne, Retha, and Christa, for their endless supply of love, encouragement, interest, and cups of tea.

In addition, I should like to express my gratitude to the following people, departments, and companies for their support and insightful and helpful feedback throughout the course of my studies:

- University of Pretoria Natural Hazard Centre;
- University of Pretoria Department of Geology;
- University of Pretoria Department of Statistics and, in particular Dr I Fabris-Rotelli, Dr PJ van Staden, and Ms L Bodenstein;
- Past and current sponsors of the UP Natural Hazard Centre (MMI Holdings, Munich Reinsurance Company of Africa Limited (MRoA), and AON Benfield);
- Ms S Verryn for providing the derivation of exponential-gamma distribution of apparent event size (Eq. 3.34);

- Prof VK Gusiakov of the Novosibirsk Tsunami Laboratory of the Institute of Computational Mathematics and Mathematical Geophysics (NTL/ICMMG) SDRAS, Novosibirsk, Russia, for the tsunami dataset used in this thesis and relevant publications;
- Mr R Britz and Ms M van der Merwe from Momentum Short-term Insurance Company Limited (MSTI) for their assistance and guidance during the hail project;
- Mrs A van Heerden for the language editing required for this thesis and publications.

Finally, I should like to acknowledge the financial assistance I received:

- Research and travel were supported by the South African National Research Foundation and the South Africa Statistical Association under the SASA-NFR Grant for Vulnerable Discipline — Academic Statistics 2017. This work is based on research supported wholly or in part by the National Research Foundation of South Africa (Grant Numbers 76906, 96412, 94808 and 103724).
- University of Pretoria Natural Hazard Centre, University of Pretoria, Department of Geology.

Summary

A generic methodology for probabilistic natural hazard assessment is presented. Three area-characteristic recurrence parameters are defined by combining a Poisson process with the relevant natural-hazard-frequency–event-size power law. The distribution of the Poisson process describes the temporal characteristics present in the data and the power law describes the relationship between the frequency of events and the event sizes. The estimates for the mean rate of occurrence λ and the power law parameter b are based on empirical datasets consisting of extreme prehistoric and historical data, along with more-recent instrumental data. Likelihood functions are defined to allow for datasets to be combined and for the application of both maximum likelihood estimation (MLE) and Bayesian inference (BI). The proposed methodology accounts explicitly for aleatory and epistemic uncertainty by making provision for incomplete datasets, uncertainty associated with the observed event sizes, uncertainty associated with the parameters of the applied occurrence and event size distributions, and uncertainty associated with the occurrence of events in the dataset. These types of uncertainty are introduced in the modelling process through convolution and mixture distributions, as well as weighted likelihood functions. Existing techniques to assess the third recurrence parameter, the maximum possible event size x_{max}, are discussed briefly. The applicability of the proposed methodology is demonstrated by using a synthetic earthquake dataset, real earthquake datasets for Central Italy and the Ceres–Tulbagh region in South Africa, tsunami data for three tsunamigenic regions in the Pacific Ocean, and HAILCAST ensemble re-analysis hail data for Gauteng province, South Africa. Various combinations of the different types of assumptions, data, and uncertainty are investigated. The methodology shows the universality of the power law in assessing natural hazards. In practice, the methodology is not restricted to natural hazard assessment, but can be applied to any instance in which the frequency–event-size relationship follows a power law distribution. To illustrate this statement, financial vehicle loss information related to hail damage, obtained from a short-term insurer in South Africa, is analysed. The versatility of the modelling process provides the researcher with various options to account for incomplete data, as well as data and parameter uncertainty.

Keywords: probabilistic natural hazard assessment; incomplete and uncertain data; power law; maximum likelihood and Bayesian estimation, mixture distributions, earthquakes, tsunami, hail.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

P	This symbol denotes prehistoric data. Any references to this in the subscripts of variables and equations refer to the prehistoric datasets.
H	This symbol denotes historical data. Any references to this in the subscripts of variables and equations refer to the historical datasets.
$i = 1, \dots s$	This symbol refers to multiple sub-datasets, ranging from 1 to s. Any references to this in the subscripts of variables and equations refer to the instrumental datasets that can be subdivided.
x	Event size of the natural hazard under investigation. This can refer to any size, e.g. earthquake magnitude, tsunami intensity, area affected by wildfire, and the Volcano Explosivity Index.
x_0	The largest event size in a time interval Δt.
Δx	The positive correction factor used in the calculation of the maximum possible event size.
x_{min}	The minimum event size, above which it is certain that all the events were observed. This is also known as the level of completeness (LoC).
$x_{min}^{(i)}$	The minimum event size, above which it is certain that all the events were observed in each sub-dataset i.
$x_{max}^{(obs)}$	The maximum observed event size in the dataset.
x_{max}	The maximum possible event size.
n	The number of events observed in a specific dataset or sub-dataset.
Δt	Time interval for the Poisson process.
λ	The mean rate of occurrence, also known as the rate of occurrence. This can be measured over any time interval, usually annually.
$\lambda(x_{min})$	The rate of occurrence for the level of completeness parameter x_{min}.
$\bar{\lambda}$	The mean of the rate of occurrence λ that can fluctuate in time and space.
$\hat{\bar{\lambda}}$	The estimated mean of the rate of occurrence λ that can fluctuate in time and space.
b	The parameter representing the slope of the power law. This parameter has a physical interpretation, as it describes the relationship between small and larger event sizes.
$\bar{b}$	The mean of the power law parameter b that can fluctuate in time and space.
$\hat{\bar{b}}$	The estimated mean of the power law parameter b that can fluctuate in time and space.
β	$\beta = b \ln 10$.

$\bar{\beta}$	The mean of the β parameter that can fluctuate in time and space.
$\hat{\bar{\beta}}$	The estimated mean of the β parameter that can fluctuate in time and space
T_g	Periods of time in the dataset with missing information.
$\mathcal{I}$	Provides the background information on the relevant parameter assumptions during the modelling process.
p, q	Parameters of the two-parameter gamma distribution $GAM(p, q)$.
w	The weights associated with a type of validity index to describe the uncertainty associated with the occurrence of an event in a dataset.
L	Likelihood function.
$\boldsymbol{\theta}$	The vector consisting of the three recurrence parameters λ, β, and x_{max}.
$\pi(\cdot)$	Relevant prior distribution used in Bayesian inference.
$\boldsymbol{D}$	Variance-covariance matrix.
$\boldsymbol{I}$	Fisher information matrix.
R	Return period.
PE	Probability of exceedance representing the probability that a specific event size x or larger events will occur within a specified time interval Δt.
m	The parameter m is equivalent to x denoting the earthquake magnitude (event size). This parameter is defined in terms of moment magnitude M_W and is used only in Chapter 5.1.
i	The parameter i is equivalent to x denoting the Soloviev–Immamura tsunami intensity (event size). This parameter is used only in Chapter 5.2, and should not be confused with the sub-dataset index $i = 1, \dots s$.

Chapter 1. Introduction

Hazard is defined as the feature of a phenomenon that can cause damage.

Risk is defined as the probability of any type of loss occurring because of the hazard.

Significant hazard not necessarily results in significant risk.

1.1 Overview

Natural disasters and their effects are not new phenomena. Nature and humans have been at odds presumably since the evolution of the modern human species. Evidence of such conflict is preserved in the remnants of catastrophes, such as the eruption of Mount Vesuvius and the total destruction of Pompeii in AD 79. Proof of prehistoric (palaeo) and historical geophysical, hydrological, meteorological, and climatological disasters have been captured in the geology of the earth, as well as in the accounts from people. Over the last few decades, unprecedented global urbanisation has taken place, to the point where more than half of the global population is living currently in urban areas (IMechE, 2013). Urbanisation inevitably gives rise to the development of industries, potentially harmful, and leads to the general deterioration of the immediate and eventually global environment. Concomitant with the global migration to urban areas, the number of recorded natural disasters and their effects on society appear to have increased substantially. Since the 1960s to 1970s, the number of natural disasters (Figure 1.1) has increased three fold, with countless people affected by climatological, hydrological, and meteorological hazards (Figure 1.2). The number of occurrences of these three types of disasters shows a much sharper growth rate over time compared with biological and geophysical hazards. The total number of natural events causing fatalities of more than 100 has increased remarkably over the last 30 years (Figure 1.3). Although there is no evidence to suggest a rise in the death toll, the number of people affected, and the estimated cost of damages have increased dramatically (EM-DAT).

This apparent relationship suggests that losses from natural disasters could ultimately, by the mid-21st century, absorb the entire economic gain of a country. Such a hypothesis, rests on the extrapolation of losses based on the growth of the mean losses. In addition, as damage from natural disasters generally follows a heavy-tail power distribution, this assumption could be erroneous. Pisarenko and Rodkin (2010, 2014) provide a detailed discussion on this topic.

The intensification in the aftermath of natural hazards can be attributed to the increased concentration of people in urbanised areas, many of which are located in or near high-risk areas, such as riverbanks, forests, tectonically unstable areas, and coastal areas. Urbanisation leads to the clustering of people and accumulated wealth in a small area, which means that a natural disaster occurring today would have a much larger effect than the same event occurring 100 years ago in the same area. Barthel and Neumayer (2012) show that although no noteworthy increase can be seen in the normalised global financial insured losses over time, such an increase is noticeable in developed countries because of their accumulated wealth.

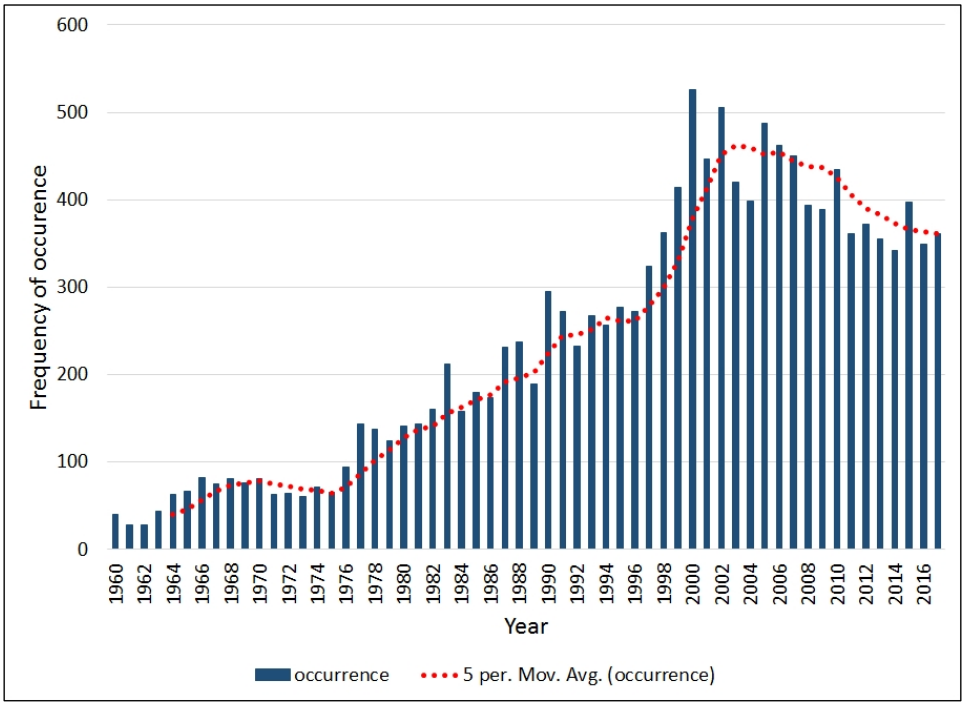


FIGURE 1.1. Increasing trend in the frequency of natural disasters from 1960 to 2016 (EM-DATA).

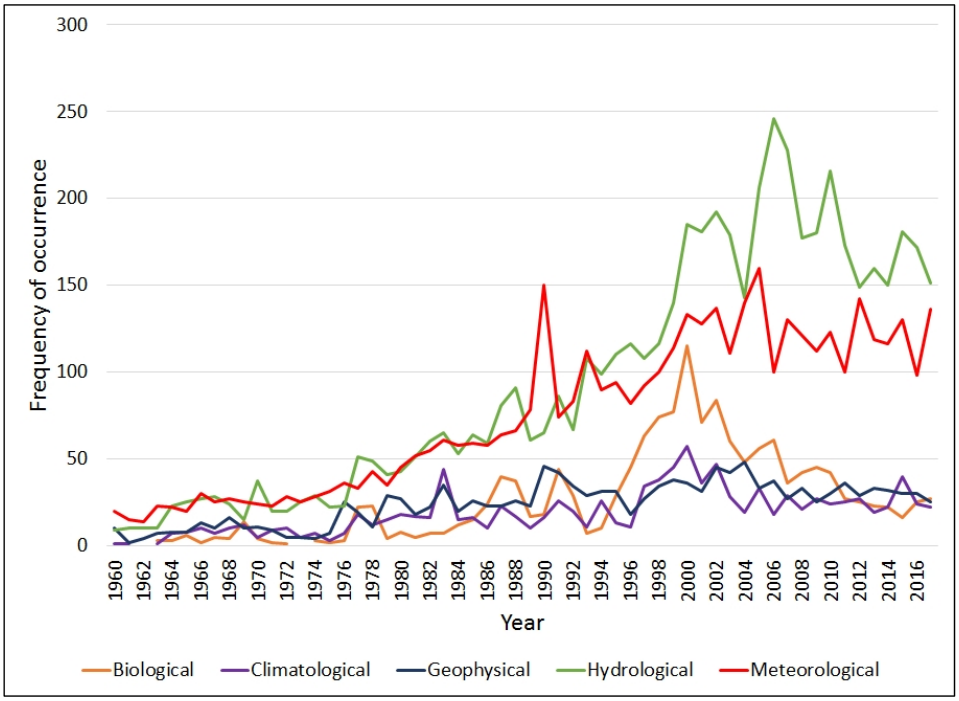


FIGURE 1.2. Increasing trend in the frequency of natural disasters per subgroup from 1960 to 2016 (EM-DATA).

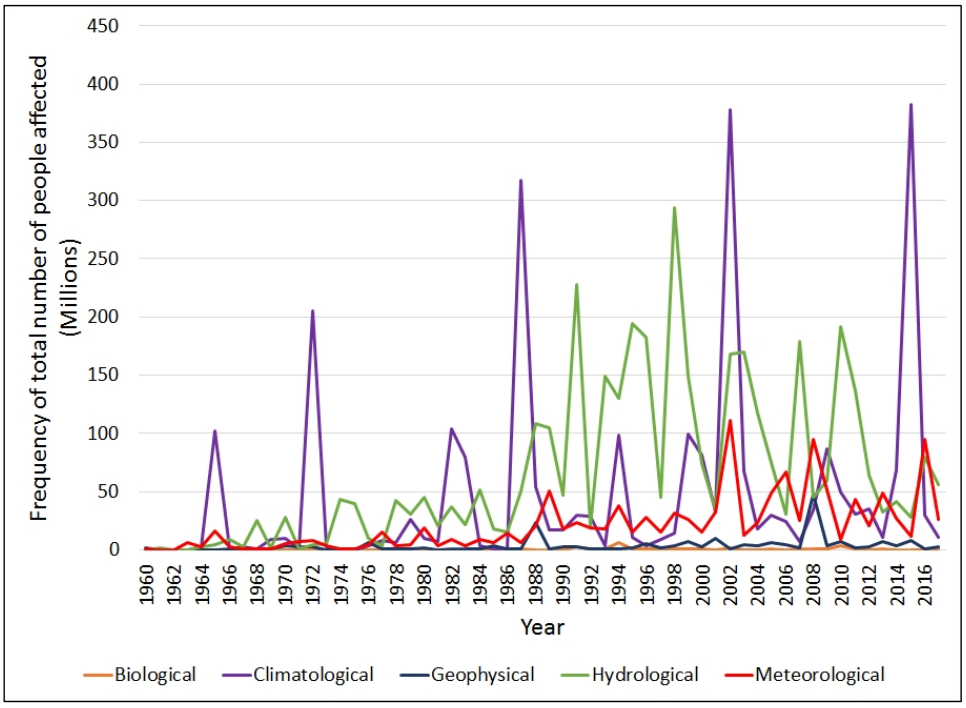


FIGURE 1.3. Frequency of the total number of people affected by natural disasters per subgroup from 1960 to 2016 (EM-DATA).

Poor and lower income groups, particularly, are most affected by natural disasters. The United Nations Office for Disaster Risk Reduction (UNISDR, 2018) reports that since the year 2000, low income countries had experienced on average of 130 fatalities per million people living in hazardous areas compared with the average of 18 fatalities per million in higher income countries. Poor and lower income groups often reside in informal settlements or slums, located in open and unbuilt areas deemed unfit for formal infrastructure development. These areas include flood plains, areas well below flood lines, and high, unstable mountain slopes prone to landslides. Houses are often built haphazardly, with inferior building plans and materials (Kohli *et al*., 2012; Kuffer *et al*., 2016).

Human activities can have a compounding effect on the hazard and risk, as in the instances of the Huang He Yellow River flood in China in 1887 and Hurricane Katrina in the southeastern United States (US) in 2005 (e.g. Brown *et al.*, 2010). In both these instances, severe storms and rainfall caused the collapse of the manmade levees, with devastating effects. On 25 January 1981, the small Karoo town of Laingsburg, South Africa, was devastated by a flash flood that resulted in over 100 fatalities and economic losses of more than ZAR 10 million. One significant contributing factor to the resulting damage was the gross underestimation of potential flood levels, resulting in infrastructure being built below the appropriate flood lines, and within a river bend where three rivers converge (Roberts and Alexander, 1982; Zawada, 1994). Furthermore, the extent of the wildfires during the summer of 1988 in Yellowstone National Park (USA) was attributable directly to the incorrect management of naturally occurring fires, which, essentially, created a giant tinderbox. The fires raged uncontrollably for months, costing nearly USD 120 million (Brown *et al.*, 2010).

Despite all the precautions in place and the research results available, we cannot ever be prepared fully for disasters. Sadly, it also takes devastating natural disasters to highlight the shortcomings in management systems, as illustrated by the Sumatra–Andaman earthquake and resulting Indian Ocean tsunami in 2004. Although comprehensive tsunami early-warning systems have been in place in the Pacific Ocean since the early 1950s, no such system existed for the Indian Ocean prior to 2004. This devastating event not only affected Indonesia but also Thailand, India, Sri-Lanka, Madagascar, and the

eastern coast of Africa. The earthquake and tsunami in Tohoku-Oki (Japan) in 2011 further illustrates our vulnerability to such disasters. As Japan is located in the 'Ring of Fire', the country experiences a high frequency of mega-earthquakes, resulting in Japan having one of the most advanced building codes and tsunami early-warning systems in the world. In 2011, however, the damage wrought by the earthquake paled in comparison with the damage from the resulting tsunami. More than 13 000 people were killed, 15 000 are still missing, and 335 000 people were left homeless. Approximately 190 000 buildings were damaged or destroyed, the most spectacular being the Fukushima Nuclear Power Reactor (Norio *et al.*, 2011). The total cost associated with this disaster amounted to approximately USD 210 billion, with insured losses amounting to USD 40 billion, making this event one of the costliest natural disasters in human history (Munich Re, Press Release 2012[1]). Furthermore, this event had a severe global economic impact and it re-ignited the international debate on the safety of nuclear energy.

In September 2015, the 2030 Agenda for Sustainable Development Goals (SDG) was adopted at a special United Nations summit. The 2030 Agenda contains 17 goals, with the ultimate aim to eradicate all forms of global hunger. The UNISDR released a reflection paper titled *Disaster Risk Reduction and Resilience in the 2030 Agenda for Sustainable Development*[2], which investigated the links between disaster risk reduction and the 2030 Agenda for SDG and the Sendai Framework for Disaster Risk Reduction 2015–2030. Direct links between disaster risk reduction and each of the 17 goals are made, highlighting the importance of including natural hazard and risk assessments to assist in reduction and mitigation strategies.

The successful modelling of natural hazards and their associated risks is important for human health, safety, and economic growth. Various industries use probabilities of exceedance, return periods, and the maximum possible event size to generate products for safeguarding society. This is done by means

[1] www.munichre.com/en/media-relations/ publications/press-releases/2012/2012-01-04-press-release/index.html; last accessed 2014/07/18.

[2] https://www.unisdr.org/files/46052_disasterriskreductioninthe2030agend.pdf; last accessed 2018/09/15.

of disaster management, and financial and engineering mitigation strategies. As natural hazard events occur frequently across national borders and affect different countries, research emphasis is moving from country-specific to regional investigations (e.g. Alfieri *et al.,* 2013). Research is focused mainly on complex hazard and risk modelling in an attempt to reduce the effect of catastrophes. These types of models are used primarily by the engineering sector and the insurance and reinsurance industries, as well as disaster management centres.

1.2 Research Statement

The statistical modelling of natural hazards and their associated risks is hampered often by a lack of complete datasets with good quality data. Such lack dictates the type of methodology applied, and care must be taken to reduce the potential hazard estimation errors. Underestimation could lead to fatalities and economic losses (Wyss *et al.*, 2012; Kousky, 2014; Howard, 2018), whereas overestimation could result in overpriced and excessive safety measures (Kossobokov *et al*., 2015).

Most statistical hazard assessment methods have been developed in countries that have extensive event datasets for different hazard types. This does not apply to developing countries, particularly in Africa, where, in many instances, instrumental data are not collected at a level that allows the effective employment of such modelling tools. Key factors influencing the completeness of event datasets include historical, socioeconomic, and demographic variations, as well as variation in the event monitoring networks. In most instances, the level of completeness is a monotonically increasing function of time, where the more-recent portion of the datasets is more complete than the older parts. Furthermore, the incorporation of additional prior information into the calculation of hazard models is necessary to constrain the results. This prior information includes additional evidence for the parameters and processes that govern the particular natural hazard, expert opinions from independent sources, as well as experiences in areas with similar but non-overlapping hazard conditions.

This thesis presents a generic probabilistic model for natural hazard assessment that is capable of assessing any type of natural hazard by utilising information from various sources and accounting for data uncertainty. The methodology is designed to account for observed data that are incomplete, and has uncertainty in recorded event sizes, uncertainty regarding the parameters of the applied distributions that describe the occurrence and physical characteristics of the hazard, as well as uncertainty in the occurrence of the events in the particular datasets. In addition, user-friendly computational tools have been developed.

The proposed methodology assesses three key area-characteristic recurrence parameters, namely λ the mean annual rate of occurrence of an event, the power law parameter b, and $x_{\max}$ the maximum possible size of an event. This assessment is done for a specific area or source for different natural hazard event datasets consisting of incomplete and uncertain data. These three key recurrence parameters are used currently in seismic hazard assessment. The methodology is an extension of the procedures by Kijko and Dessokey (1987), Kijko and Sellevoll (1989, 1992), and Kijko and Smit (2012) that provide for the incompleteness of the earthquake event datasets by accounting for the largest historical events and instrumental records with different levels of completeness (LoC), as well as for the uncertainty in earthquake event size (magnitude) determination. For the rest of this thesis, λ will be referred to as the rate of occurrence.

In addition to making provision for incompleteness and the event size uncertainty in natural hazard datasets, the methodology for natural hazard assessment considers the discrepancy between the data and the applied models describing the event occurrence and size, and accounts for the uncertainty of occurrence, i.e. the validity of events contained in the dataset. Both maximum likelihood estimation (MLE) and Bayesian inference (BI) are utilised in the estimation of λ and b.

1.3 Research Aims and Objectives

The aim of this thesis is to develop a generic probabilistic model that assesses natural hazard. Its aim is to serve as a computational tool for industries, such as civil engineering, insurance, and disaster management. The description in this thesis is focused on natural phenomena, such as earthquakes, tsunamis, hail, floods, and fires and their associated risk.

The hazard methodology applied in this thesis is an extension of the current formalisms utilised by the University of Pretoria Natural Hazard Centre, Africa (UPNHCA), developed originally for seismic hazard assessment. Four data quality characteristics that affect natural hazard modelling are considered in an attempt to utilise as much of the available data as possible, so as to reduce the epistemic and aleatory uncertainty associated with modelling. These four data quality characteristics are:

- the incompleteness of the event datasets,
- uncertainty in event size determination,
- uncertainty associated with the parameters in the applied event occurrence and size distributions,
- uncertainty of event occurrence (e.g. was the event a tsunami because of a geological disturbance, or waves because of substantial meteorological storms?).

The following objectives were identified to achieve the general aim of this thesis:

1. A literature study of hazard assessment methodologies used for natural hazards.
2. Definition and development of the mathematical formalisms for a generic probabilistic natural hazard model that can be used to determine the expected hazard in an area for different types of natural hazards, and data uncertainties.
3. Estimation of the hazard recurrence parameters using maximum likelihood estimation (MLE) and/or Bayesian inference (BI).

4. Comparison of MLE and BI estimates of recurrence parameters when applying differently constructed datasets and distributions.
5. Description of where and how the model should be adapted to incorporate the distribution of the physical process for any natural hazard under review.
6. Application of the developed mathematical formalisms to practical examples, which can include but are not limited to earthquakes, tsunamis, storms (hail), sinkholes, fires, and floods.
7. Development of user-friendly computational tools.
8. Publication of the relevant results in journals accredited with the Department of Higher Education and Training (DHET) in South Africa.
9. Integration of the published results into the final thesis manuscript for submission.

1.4 Research Outputs

Three papers have been published in field-specific journals, with the third paper currently under review for publication in an environmental statistical journal. A fourth paper is being drafted for publication in a field-specific journal.

Data, examples, results, and conclusions from publications related to the thesis were added to the thesis with minimal changes. Below is a description of each publication, with the objectives it addresses, and in which chapters and sections of the thesis the largest part of the particular publication can be found.

Articles published

1. Kijko, A., Smit, A., Sellevoll, M.A. (2016). Estimation of Earthquake Hazard Parameters from Incomplete Data Files. Part III. Incorporation of Uncertainty of Earthquake-Occurrence Model. *Bulletin of the Seismological Society of America*, 106(3). pp 1210–1222. DOI: 10.1785/0120150252. Full text available at the University of Pretoria library repository link

https://repository.up.ac.za/bitstream/handle/2263/55817/Kijko_Estimation_2016.pdf?sequence=1.

This paper on the application of the methodology in terms of earthquakes was the first published. It addresses objectives 1, 2, 3, 6, 7, 8, and 9, as set out in Section 1.3. In this paper, a substantial part of the methodology is discussed, as in Chapter 3, in terms of the Gutenberg–Richter power law for earthquake magnitudes. The example used in the publication is added with minimal amendments to Chapter 5, Section 5.1. My contribution to the paper was approximately 50% and focussed on literature review, adaptation of theory, co-writing the article, and the computation and analyses of the data.

2. Smit, A., Kijko, A., Stein, A. (2017). Probabilistic Tsunami Hazard Assessment from Incomplete and Uncertain Historical Catalogues with Application to Tsunamigenic Regions in the Pacific Ocean. *Pure and Applied Geophysics*. 174(8). pp 3065–3081. DOI: 10.1007/s00024-017-1564-4. Full text available at the University of Pretoria library repository link https://repository.up.ac.za/handle/2263/62181?show=full.

   In this paper, the methodology is presented as applied to tsunamis in the Pacific Ocean. This paper addresses objectives 1, 2, 3, 5, 6, 7, 8, and 9, as set out in Section 1.3. In the paper, a substantial part of the methodology is discussed, as in Chapter 3, in terms of the Soloviev–Imamura power law for tsunami intensities. The example used in the publication is added with minimal amendments to Chapter 5, Section 5.2. My contribution to the paper was approximately 75% and focused on the literature review, adaptation of theory, the computation and analyses of the data, interpretation of results, formulated discussion items and drafted conclusions, and co-writing the article.

3. Smit A., Stein A., Kijko A. (2019). Bayesian Inference in Natural Hazard Analysis for Incomplete and Uncertain Data. *Environmetrics*. p.e2566. DOI: 10.1002/env.2566.

This paper addresses objectives 1–9, as set out in Section 1.3. In the paper, a substantial part of the methodology is discussed, as in Chapter 3, in terms of the general power law. Parts of the text of this paper are used with minimal changes in this thesis, e.g. Section 3.3.2. The example used in the paper includes parts of the examples used in Chapter 4 and Section 5.2. My contribution to the paper was approximately 75% and focused on the literature review, adaptation of theory, the computation and analyses of the data, interpretation of results, formulated discussion items and drafted conclusions, and co-writing the article.

Article currently in preparation

4. Smit A., Kijko A., Stein, A., Dyson, L. Probabilistic Hail Hazard and Risk Analysis.

   This paper will address objectives 1–9, as set out in Section 1.3. In the paper, a substantial part of the methodology is discussed, as in Chapter 3. The example described in Section 5.3 will be used in the paper.

Conferences and Workshops

Over the last few years, I have presented the results from this thesis at several national and international conferences. I have also attended relevant workshops, as listed below.

- *All-Russian Scientific Conference in Geodynamical Processes and Natural Hazards. Lessons of Neftegorsk*, 26 to 30 May 2015, Yu. Sakhalinsk, Russia. Oral presentation titled *Common Problems in Estimation of Seismic Hazard Parameters and their Solutions* by A Kijko (presenter) and A Smit.
- *All-Russian Scientific Conference in Geodynamical Processes and Natural Hazards. Lessons of Neftegorsk*, 26 to 30 May 2015, Yu. Sakhalinsk, Russia. Oral presentation *Probabilistic Tsunami Hazard Assessment from Incomplete and Uncertain Data* by A Kijko (presenter) and A Smit.

- *Advances in Extreme Value Analysis and Application to Natural Hazard* (EVAN), 16 to 18 September 2015 in Santander, Spain. Poster presentation titled *New Procedure for Probabilistic Hazard Assessment for Incomplete and Uncertain Data* by A Smit (presenter), A Kijko, I Fabri-Rotelli, PJ van Staden.
- *57th Annual Conference of the South African Statistical Association (SASA)*, 30 November to 4 December 2015 in Pretoria, South Africa. Oral presentation titled *New Procedure for Probabilistic Hazard Assessment for Incomplete and Uncertain Data* by A Smit (presenter), A Kijko, I Fabri-Rotelli, PJ van Staden.
- *European Geosciences Union General Assembly (EGU)*, 17 to 22 April 2016, Vienna, Austria. Oral presentation titled *Probabilistic Seismic Hazard Assessment for Incomplete and Uncertain Data* by A Smit (presenter) and A Kijko.
- *58th Annual Conference of the South African Statistical Association*, 28 November to 2 December 2016, Cape Town, South Africa. Oral presentation titled *Probabilistic Tsunami Hazard Assessment for Incomplete and Uncertain Data* by A Smit (presenter), A Kijko, I Fabri-Rotelli, PJ van Staden.
- *Extreme Value Theory Conference*, Delft, the Netherlands, 26 to 30 June 2017. Oral presentation titled *Process Characteristic Extreme Value Distributions* by A Smit (presenter), A Kijko, and A Stein.
- *IAPSO-IAMAS-IAGA Joint Conference*, Cape Town, South Africa, 27 August to 1 September 2017. Oral presentation titled *Procedure for Probabilistic Tsunami Hazard Assessment from Incomplete and Uncertain Data* by A Smit (presenter), A Kijko, and A Stein.
- Attended the workshop titled *R Software Fundamentals with Applications to Statistical Analysis and Modelling*, 21 to 27 July 2017, University of the Witwatersrand, Johannesburg, South Africa.
- *59th Annual Conference of the South African Statistical Association*, 27 to 30 November 2017, Bloemfontein, South Africa. Oral presentation titled *Extreme Distributions Based on Process Characteristics* by A Smit (presenter), A Kijko, and A Stein.

- Attended summer school on *Bayesian Inference: Foundations and Application*, 17 to 27 January 2018, Betties Bay, South Africa, hosted by the University of Stellenbosch.
- University of Pretoria, Department of Statistics Seminar Series, 27 March 2018. Oral presentation titled *Generic Statistical Method of the Assessment of Hazard and Risk* by A Smit (presenter), A Kijko, and A Stein.
- *60th Annual Conference of the South African Statistical Association*, 26 to 29 November 2018, Roodepoort, South Africa. Oral presentation titled *Probabilistic Hail Hazard and Risk Analyses* by A Smit (presenter), A Kijko, and A Stein.

Awards

- Second prize for best oral presentation in the Young Statistician stream at the 58th Annual Conference of the South African Statistical Association, 28 November to 2 December 2016, Cape Town, South Africa. Oral presentation titled *Probabilistic Tsunami Hazard Assessment for Incomplete and Uncertain Data* by A Smit (presenter), A Kijko, I Fabri-Rotelli, and PJ van Staden.
- The 2018 Herbert Sichel Medal for the best statistical paper for the published article *Probabilistic Tsunami Hazard Assessment from Incomplete and Uncertain Historical Catalogues with Application to Tsunamigenic Regions in the Pacific Ocean* by A Smit, A Kijko, and A Stein (2017) in *Pure and Applied Geophysics*, 174(8). pp 3065–3081. DOI: 10.1007/s00024-017-1564-4.

1.5 Contributions to Scientific Research

The field of natural hazard assessment is vast, with countless relevant multi-disciplinary publications and books having been issued already. The aim of this thesis is to provide a focused overview and one potential framework that could be applied to various types of natural hazards and data types.

The developed methodology uses the probabilistic hazard assessment for seismic recurrence parameters, as originally initiated by Kijko and Dessokey (1987) and Kijko and Sellevoll (1989, 1992). Their methodology describes how to assess the seismic recurrence parameters λ, the rate of occurrence, the Gutenberg–Richter b-parameter (b-value), and the maximum possible earthquake magnitude (event size) x_{max}, taking into consideration extreme and instrumental data, the incompleteness of the dataset, as well as making provision for uncertainty in earthquake magnitude.

In this thesis and associated publications, the methodology is extended by incorporating several aspects into the modelling process, as illustrated in Figure 1.4.

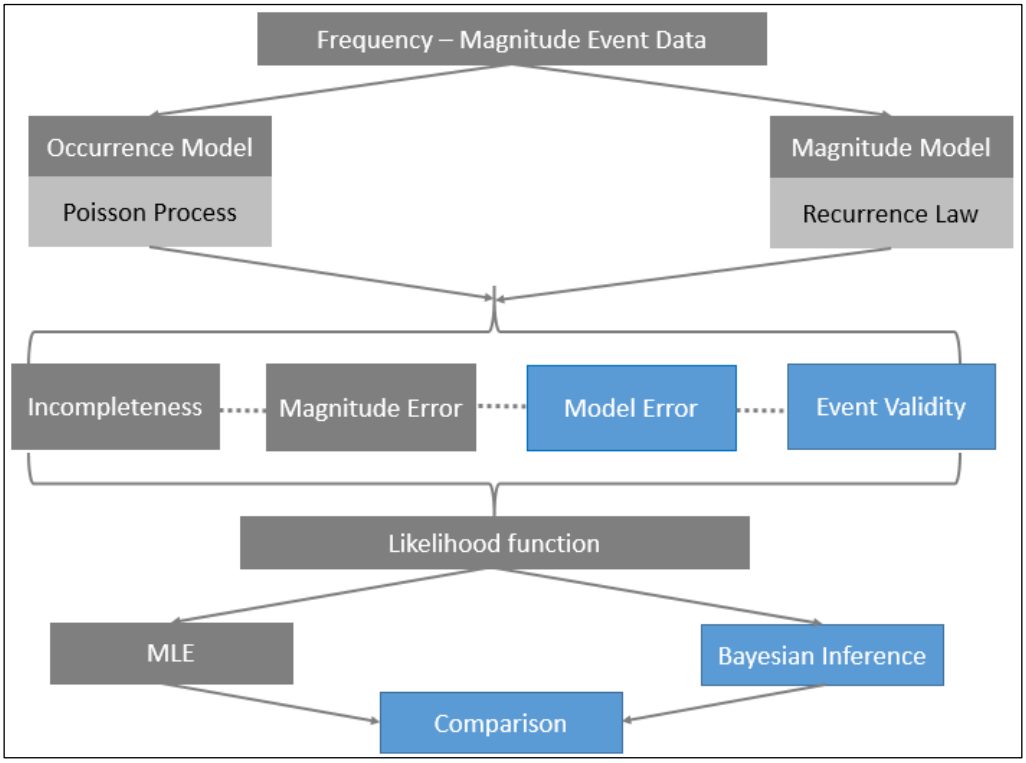


FIGURE 1.4. Schematic illustration of the applied methodology to combine prehistoric, historical, and instrumental data, as well as account for data incompleteness, uncertainty in the size of the events, uncertainty associated with the parameters in the applied occurrence and event size distributions, and uncertainty of the validity of the events in the dataset. The grey blocks represent the modelling process, as defined by Kijko and Sellevoll (1989, 1992). The blue blocks represent the contribution of this thesis to the methodology.

These aspects are:

- Uncertainty associated with the parameters of the applied occurrence and event size distributions.

  This uncertainty is incorporated to account for instances where weak dependencies can be observed between events, thereby violating the distributional requirements for the use of the Poisson distribution. For this thesis, mixture distributions with a gamma mixing distribution are used to aid with the modelling of weak dependencies or space-time fluctuations the investigated processes.

- Uncertainty in the occurrence of events in the dataset.

  To incorporate as much available information as possible into the modelling process, prehistoric and historical records should be utilised by using the additive property of likelihood functions. By introducing the weighted likelihood function and Bayesian formalism to the methodology in this thesis, the researcher has the option to include all the events and retrieve a rate of occurrence that is more realistic.

- Using prior information to estimate parameters using Bayesian inference.

  Employing Bayesian inference is new in the scope of this methodology. The synthetic data example (Section 4) illustrates the effect BI has on the estimated parameters. Additionally, an example is presented of how knowledge on hazard parameters can be used as prior information for the financial risk assessment of the same natural phenomena. It is shown in the example of hail that a direct link can be established between the hazard and risk assessments by employing a physical law distribution with the same parameters, and using the hazard assessment as a priori information for the risk assessment.

- The power law in natural hazard assessment.

  It is shown that by employing power laws to describe the physical characteristics of natural hazards and in some circumstances, their associated risks, the derived methodology is generic

and can be applied to any natural hazard dataset where events can be described in terms of a frequency–event-size relationship.

- Application of methodology to new types of hazards.

  The methodology is applied for the first time to real tsunami events, hail climatology events, and hail financial loss data. This shows that by interchanging the power law distribution relevant to the field of study, the methodology is generic. Furthermore, various aspects of the estimation process are discussed.

1.6 Overview of Thesis

The thesis is divided into seven chapters and two appendices. Chapter 1 provides an introduction and overview of the importance of natural hazard modelling, followed by the research statement, aims and objectives. The various research outputs in the form of publications and conference presentations are discussed, as well as the contribution to scientific research.

In Chapter 2, the literature review for this thesis is presented. Natural hazard modelling in the broad sense is discussed, with more-detailed and field-specific literature reviews provided for each of the practical examples in Chapter 5. In addition, potential data sources are investigated, as well as the data quality characteristics associated with natural hazards, and the application of power laws.

The methodology is introduced in Chapter 3 by means of a series of assumptions. The principal model and its variations are introduced and extended to account for the data quality characteristics of incomplete event datasets, uncertainty in event size determination, uncertainty associated with the parameters of the applied event occurrence and size distributions, and uncertainty of event occurrence. Parameter estimation, the assessment of the maximum possible event size, the calculation of the hazard

estimates in terms of return periods and probabilities of exceedance and model comparison methods are also discussed.

The methodology and the combination of the models are tested on a synthetic earthquake dataset in Chapter 4. The focus of this chapter is on how the introduction of the data incompleteness and uncertainties and the applied estimation process influence the parameter estimates, as well as the contribution of the input information. In Chapter 5, the methodology is applied to real-world examples. These examples are the earthquake datasets for the Ceres–Tulbagh region in South Africa (Section 5.1.1) and Central Italy (Section 5.1.2), tsunami data for three tsunamigenic regions in the Pacific Ocean (Section 5.2), HAILCAST ensemble re-analysis hail data for Gauteng province, South Africa (Section 5.3.2), and financial vehicle loss information related to hail, from a short-term insurer in South Africa (Section 5.3.3).

The conclusions related to each of the analyses in Chapters 4 and 5 are discussed in the respective chapters. The thesis concludes with overall conclusions and remarks presented in Chapter 6, and with ideas for future research in Chapter 7. Derivations of key equations are provided in the Appendix.

1.7 Data and Resources

Ceres–Tulbagh seismic event dataset. The instrumentally recorded events were selected mainly from the available database provided by the Council for Geoscience, Pretoria, and the International Seismological Centre in the United Kingdom (http://www.isc.ac.uk). The dataset used in the analysis spans a period of approximately 268 years, namely from 1751/01/01 to 2012/01/31.

Central Italy seismic event dataset. The zonation of Italy was done according to the ZS9 zones of MPS (2004). Extreme prehistoric data were obtained from Galli *et al*. (2008) and the historical and

instrumental data from the European–Mediterranean Earthquake Catalogue (EMEC, 2012[3]; Grünthal and Wahlström, 2012).

Tsunami data for Kuril–Kamchatka, Japan, and South America. The international tsunami database used in this study was provided by Dr VK Gusiakov of the Novosibirsk Tsunami Laboratory of the Institute of Computational Mathematics and Mathematical Geophysics (NTL/ICMMG) SDRAS, Novosibirsk, Russia (HTDB/WLD, 2013).

HAILCAST ensemble data. The data for hail occurrence in South Africa comprise ensemble data from the HAILCAST model using ERA-INTERIM reanalysis data. This information was provided by Prof. L Dyson from the Department of Geography, Geoinformatics and Meteorology at the University of Pretoria.

Hail insurance loss data. A dataset consisting of individual losses incurred from hail events was provided by Momentum Short-term Insurance Company Limited (MSTI) for the period 2007 to 2017.

The MATLAB computer software[4] and MS Excel[5] was used for the calculations. The code is available on request, from the University of Pretoria Natural Hazard Centre, for academic purposes.

[3] https://www.gfz-potsdam.de/en/section/seismic-hazard-and-stress-field/data-products-services/emec-earthquake-catalogue/, last accessed 2019/01/25.

[4] https://www.mathworks.com/products/matlab.html, last accessed 2018/09/28.

[5] https://products.office.com/en-za/excel, last accessed 2018/09/28.

Chapter 2. Literature Review

The literature relevant to this study includes several disciplines of the natural sciences, namely, mathematics, physics, and statistics. The initial development of the mathematical formalisms will rely heavily on the mathematical and statistical procedures followed in seismic hazard assessment, as discussed in Kijko and Dessokey (1987), Kijko and Sellevoll (1989, 1992), and Kijko and Graham (1998, 1999). This chapter provides a general overview of natural hazard assessment, the data sources, nature of the data, and power laws in nature. A more-detailed discussion relating to the methodology and the respective applications is presented in Chapters 3, 4, and 5.

2.1 Hazard Analyses

There are two types of natural hazard assessment methods, namely deterministic and probabilistic, which are related closely. Deterministic methods assess the effect of an event occurrence at a specified site from a single, specified source. They aim to assess the effect of a single event, usually the worst-case scenario, which provides valuable information for disaster management agencies for planning purposes. In addition, several deterministic scenarios can be aggregated to identify the worst-case credible scenario (e.g. for tsunamis Tonini *et al.*, 2011; Grilli *et al.*, 2011; Harbitz *et al.*, 2012).

Probabilistic scenarios focus on assessing the probabilistic nature of a hazard by considering all the potential sources (both locally and regionally) that can contribute to the overall hazard and risks. This type of assessment provides a comprehensive estimation of the hazard for the entire area under investigation. This, however, complicates the assessment process, as it can incorporate a vast amount and different types of information (e.g. Geist and Parsons, 2006; Power *et al.*, 2007; Thio *et al.*, 2007; Kijko, 2011; Sørensen *et al.*, 2012; Brizuela *et al.*, 2014). Large-scale research is required therefore, relevant to all the potential sources and the movement of an event, such as the propagation of seismic

and tsunami waves or the path that a severe weather system or fire will follow. Probabilistic methods are crucial in risk mitigation processes for disaster management, engineering, and insurance companies to understand the short-, medium-, and long-term characteristics of events in order to develop safeguards for society. For this reason, and the nature of the data, as discussed in the remainder of this section, the rest of the thesis focusses on probabilistic hazard assessment procedures.

The nature and quality of the data often dictate the type of probabilistic methodology that can be applied. This ranges from pure empirical methods using only observed data, computational analyses using numerical modelling of the physical process, e.g. tsunami and flood propagation modelling and climatological ensembles, as well as Bayesian methods that combine empirical data with prior information. Empirical and Bayesian models remain the preferred methods for inferences. As these methods were developed primarily for large datasets, the statistical accuracy of the results does not necessarily hold when applied to small datasets. Therefore, it is vital to collect and incorporate as much historical information as possible to stabilise and improve the accuracy of the results. Prehistoric and historical information is essential, especially in terms of return periods and the probabilities of exceedance.

Uncertainties can be introduced in the modelling process through, e.g. Bayesian techniques, mixture distributions, and convolution principles (Sections 3.2.2 and 3.2.3). Some probabilistic hazard analysis procedures provide for uncertainty in the data by separately modelling the different sources of an event and subsequently combining them. For landslides, uncertainty is accounted for by separately modelling the probability of occurrence owing to earthquakes and rainfall and afterward combining the two to obtain a single probability of occurrence estimate (Corominas and Moya, 2008; Wartman *et al.*, 2013). Probabilistic tsunami modelling functions in a similar manner, in that the probability of occurrence of various types of triggering events (e.g. earthquakes, land- and submarine landslides, and volcanic activity) is estimated and combined to assess the likelihood of a local or trans-oceanic tsunami (Geist and Parsons, 2006). This type of modelling requires extensive research and knowledge of the fundamentally different physical occurrences, making the assessment difficult and time consuming.

Pisarenko and Rodkin (2010) reviewed empirical distributions that were developed for the analysis of empirical natural hazard and risk data. These include traditional and field-specific parametric distribution functions, of which heavy-tailed extreme distributions are popular. Relevant applications are found across all subfields of historical, geophysical, hydrological, meteorological, and climatological disasters. Risk assessments in terms of the victims, fatalities, and economic losses resulting from such events are often modelled in this manner as well (e.g. Pisarenko and Rodkin, 2014).

Extreme value theory (EVT) is a widely used technique and is often used in the assessment of extreme natural hazard and risk. EVT focuses on modelling only the tail behaviour especially in cases where it is assumed that extreme events do not follow the same behaviour as small and intermediate events (e.g. Pisarenko and Sornette, 2003; De Haan and Ferreira, 2007). Attention is given also only to the largest events exceeding a specific threshold, as their occurrences are usually unforeseen, and their effect dwarfs the combined effects of the smaller events. These large events are also the most memorable, making it more likely that most, if not all, would have been recorded. Block maxima and peaks-over-threshold (POT) are techniques used to determine the threshold from which events are included in the assessments. Popular distributions in EVT include the Generalised Pareto (GPD) and the Generalized Extreme Value distributions (GEV) in the form of the Gumbel, Fréchet and Weibull distributions (e.g Pisarenko and Sornette, 2003; De Haan and Ferreira, 2007).

Other empirical analyses consist of parametric, semi-parametric, and non-parametric modelling, stochastic modelling of conditional probabilities of recurrence times (e.g. Orfanogiannaki and Papadopoulos, 2007), or modelling by means of cumulative frequency–event-size relationship laws (e.g. Gutenberg and Richter, 1942; Soloviev, 1970; and references in Table 2.2). Empirical methods are highly dependent on the completeness and quality of the dataset. Adequate historical and instrumental information is required to counter the high levels of bias that small datasets can introduce to the results. In addition, empirical distributions can be ambiguous, as any descriptors of the underlying physical characteristics of the natural hazard are removed (Pisarenko and Rodkin, 2010). Instead, a phenomenon is described only in terms of its location, shape, and the spread of the applied distribution.

Physical distribution laws are parametric distributions that retain the physical properties of the natural hazard event. Typical systems include the Gaussian (normal) law, the Boltzmann law (an exponential distribution), and the Pareto power law that can be transformed to the exponential distribution by the transformation $y = F_Y^{-1}(F(x))$ (Balakrishnan and Basu, 1995; Pisarenko and Rodkin, 2014). Power laws are used frequently to assess the relationship between the frequency of occurrence and the size of the event, with the power law parameter indicating the relationship between the large and small events. In Pisarenko and Rodkin (2010, 2014), detailed reviews are presented on the heavy-tailed power law and the use of the EVT to estimate the parameters. In Section 2.4, power laws and their application to natural hazards, as well as the proposed methodology, are discussed.

2.2 Data Sources

Natural perils are typically analysed from systematic data recorded with the help of networks of specialised measurement stations. These datasets can be augmented with data obtained from the insurance industry, as well as non-systematic data obtained from historical recordings and palaeo-environmental studies. These datasets are often incomplete in terms of their spatial and temporal context. With the growth of social media platforms, these datasets can be augmented to create datasets with finer resolutions, although these are often of lower quality. A brief overview of the three types of data sources is provided below.

2.2.1 Systematic and non-systematic data

The most reliable systematic data are provided by calibrated networks of instruments, such as seismographs and accelerographs for earthquakes, ocean buoys and tide gauges for tsunamis, stream gauges for terrestrial bodies of water, rain gauges for rainfall to investigate flood occurrences, and inclinometers for landslides. For other, more complex hazards, such as volcanoes, several tools are

employed to measure the seismicity, ground deformation, released gasses, lava flow, and pyroclastic flow and surge. Developed countries often have extensive datasets, e.g. Japan has an instrumental earthquake dataset spanning more than 1 000 years. The Japanese seismic network consists of approximately 1 000 accelerometers and 1 800 strong-motion seismograph stations spaced at 20–25 km intervals (Furumura *et al.*, 2011). This is in stark contrast with developing countries, where event datasets are dependent on information from temporary or external networks. Examples of international data sources are the International Seismological Centre[6], the United States Geological Survey Science Data Catalog[7], and the National Oceanic and Atmospheric Administration[8]. Pisarenko and Rodin (2010) provide a review of the available datasets, terminology, as well as the parameterisation and classification of events.

Non-systematic data include information that is not measured systematically over time. Prehistoric and historical data are measurements of natural perils that are typically quantified using qualitative methods during focussed studies in pre-defined areas. Prehistoric data are collected by palaeo-environmental studies and historical information from recorded accounts by eye-witnesses of the events. Accurate instrumental recordings of natural disasters represent only an extremely small part of the overall historical timeline; therefore, it is likely that the largest events are not included in the dataset. This omission could result in the underestimation of the hazard and a subsequent underestimation of the vulnerability or risk for the area under investigation. Over the years, various fields of natural hazards have started to include prehistoric and historical data into hazard assessments, e.g. floods (Baker 1987, Fernandes *et al*, 2010, Lam *et al*., 2017), earthquakes and tsunamis (McCalpin, 2009; Heidarzadeh and Kijko, 2011), and tropical cyclones (e.g Nott, 2003). Although prehistoric and historical data are not recorded with the precision of instrumental data, they do preferentially record the more extreme events.

[6] isc.ac.uk, last accessed 2018/09/25.

[7] data.usgs.gov/datacatalog, last accessed 2018/09/25.

[8] noaa.gov, last accessed 2018/09/25.

A detailed discussion of the characteristics of systematic and non-systematic information is presented in Section 2.3.

2.2.2 Insurance data

The insurance and reinsurance industry provides a second source of information on natural perils and the attendant loss of lives and property. Over the years, multi-peril natural and man-made catastrophe loss data sources have been made available online. These includes databases, such as Sigma by Swiss RE[9], NatCatSERVICE[10] by Munich Re, OasisHUB[11], and Emergency Events Database EM-DAT[12], established by the Centre for Research on the Epidemiology of Disasters (CRED) at Leuven University (Université catholique de Louvain) in Belgium.

Kron *et al.* (2012) provide a detailed discussion on the effective use of locally sourced and international data operators and identify the potential pitfalls. A typical pitfall is that losses are aggregated across perils for an individual event. For example, a small distinction may or can be made as to the exact cause of the damage. Additionally, various insurers may classify events differently, making it difficult to combine information from diverse companies in a homogeneous manner. To comprehend the effect of economic losses over time, care should be taken to normalise the values for any potential change in population and the socioeconomic status of the specific country. Financial loss information may be of help in assessing future potential risks from natural perils, particularly in the absence of specific damage information with which to build statistical loss models.

[9] institute.swissre.com/research, last accessed 2018/09/25.

[10] munichre.com/en/reinsurance/business/non-life/natcatservice, last accessed 2018/07/25.

[11] https://oasishub.co/, last accessed 2018/09/13.

[12] emdat.be/database, last accessed 2018/07/25.

2.2.3 Social media data

Social media or volunteer data consist of information that can be obtained from the vast array of digital technologies. Photos and video recordings can be valuable in collecting information on the location, size, and damage of natural phenomena (Blair and Leighton, 2012). The United States Geological Survey (USGS) website Did You Feel It? (DYFI)[13] is an interactive tool that collects information and generates maps of earthquakes based on the information provided by individuals. This information is combined with instrumental recordings and translated into scales of earthquake intensity (Wald *et al*., 2011). The most often used scales are the Modified Mercalli Intensity Scale or the European Macroseismic Scale (EMS). These scales describe the intensity of the observed ground movement (shaking). Data on the observed intensity of an event help in investigating the propagation of the seismic waves, the correct classification of observed damage, and the updating of existing damage matrices and curves.

Social media are also used in other fields. In 2013, Schnebele and Cervone (2013) combined remote sensing flood hazard assessment with data from internet searches for news, videos, and photographs. Authors such as Guan and Chen (2014) have shown how natural disasters can be spatially and temporally characterised during their evolution by using social media. Social media provide valuable information not only in terms of hazard assessment but also for disaster risk-mitigation planning.

Using social media in combination with various types of geographic information systems (GIS) is useful in data sharing for hazard and risk mitigation (Manfré *et al.*, 2012). Using the information from past events helps to understand where and how natural disasters occur in a specific area, serving as prior information in natural hazard and risk analyses.

[13] https://earthquake.usgs.gov/data/dyfi/; last accessed 2018/09/22.

2.3 Nature of Empirical Data

Geophysical, hydrological, meteorological, and climatological data can be incomplete and uncertain. The quality of the event dataset of a country or a region is dependent on its recording history. Accurate instrumental and systematic recordings of natural disasters have emerged only over the last approximately 100 years. As instruments became more sensitive and accurate, the quality and quantity of the observed data improved, with the location, time, and size of the events recorded with care. The compiled instrumental datasets of these events contain information on the level of completeness (LoC) and the measured event sizes. The event size x, an element of the random variable X, is described in terms of energy, magnitudes, intensities, or the affected areas. The LoC, also denoted x_{min}, is defined as the minimum size in the dataset for a specific time interval, assuming that all the events were observed and recorded. This type of dataset represents only a fraction of the overall historical time line. Accordingly, there is a real possibility that the observed instrumental datasets do not contain the records of large, or even the largest, observed events. This results in an underestimation of the hazard and the subsequent risks and vulnerabilities.

In an effort to include more information to supplement the instrumental datasets and reduce epistemic uncertainty, extensive research has been done on collecting reliable sources of prehistoric and historical information. These two types of datasets are often considered to contain data that are of lower quality than those of instrumental data. This can be attributed to additional levels of uncertainty relevant to the exact location, date, and size of the events recorded. The information generally contained in historical event datasets was observed typically before the advent of the instrumental age. The quality of this information, therefore, depends on whether these events were observed and accurately described according to the effect of the event on the surrounding environment. The observations in this event dataset is are not limited to extreme events, and observations of events of different sizes can be included. Prehistoric events, in contrast, are events recorded with palaeo-environmental studies. This requires the extraction and analyses of geological sediments and rocks with the aim of assessing changes in the

chemical, physical, and biological nature of sediment deposits over the last few thousand years. Such investigations are expensive, as the potential area where a natural event could have taken place must be identified, trenching must be done, and chemical and geological analyses must be conducted. One disadvantage of palaeo-studies is that the date of occurrence and the size of an event will only fall within a specific time and size interval. It addition, it is currently not possible to identify all the events that occurred in prehistoric times. The prehistoric event set, therefore, represents a temporally and spatially biased view of the occurrences of natural events.

Several fields of natural hazard assessment incorporate prehistoric and historical information. Nott (2003) has discussed the importance of including prehistoric data, using tropical cyclones, tsunami, terrestrial floods, and landslides in Australia as examples. The inclusion of this type of data helps to identify possible changes in the stationarity assumptions. Prehistoric and historical data are included more frequently in hazard assessments, particularly in relation to earthquakes, tsunamis, and floods (Cohn and Stedinger, 1987; Ogata, 1999a; IAEA TECDOC-1767, 2015; Grezio *et al*, 2017; Lam *et al*., 2017; Zöller, 2018).

Figure 2.1 provides an illustration of the typical nature of the data that are used to assess hazard parameters. Different types of uncertainty can be identified for each of the datasets. Prehistoric data are subject to uncertainty relevant to the time of occurrence and event size, as well as incompleteness relevant to the equal probability of detecting an event over a spatial grid. In both historical and instrumental event datasets, incompleteness and uncertainty occur relevant to the observed event size, with varying levels of certainty on the exact location, and the probability of an event size being observed above the LoC. Furthermore, prehistoric and historical data can contain some uncertainty in the event occurrence or validity of the classification of an event. This is observed in tsunami datasets, as it can be unclear whether the observed effects are from a tsunami or from severe storms.

Area-characteristic recurrence parameter assessment for the typical data scenario shown in Figure 2.1 can be addressed in a manner similar to that in Kijko and Dessokey (1987), Kijko and Sellevoll (1989,

1992), Kijko and Smit (2012), and Kijko *et al.* (2016). It is assumed that the dataset can be divided into several smaller sub-datasets consisting of prehistoric (P), historical (H), and multiple complete, typically instrumental datasets ($i = 1, \dots, s$), respectively. The prehistoric and historical sub-datasets typically consist of the largest events that occurred over the last few thousand years and last few hundred years, respectively. The most-recent events are usually recorded with the aid of sophisticated instruments and networks, and these can be classified as a complete dataset. It is assumed that each of these subsets is complete for event sizes exceeding a certain level of completeness (LoC) $x_{\min}^{(i)}$, during a period of time Δt_i for $i = 1, \dots, s$. This approach permits the occurrence of 'gaps' (T_g) to account for missing event records. Missing records can be attributed to e.g. event recording equipment or networks being non-operational, whereas socioeconomic and social factors could also play a role.

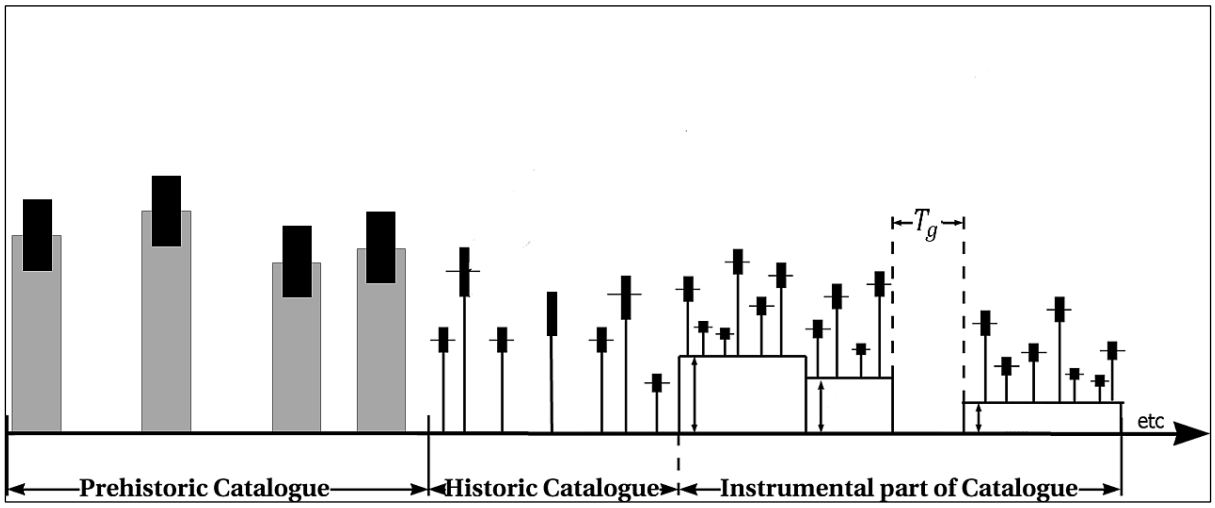


FIGURE 2.1. Illustration of a typical event dataset that can be used to obtain the three key parameters for the specified natural hazard assessment (modified after Kijko and Sellevoll, 1992).

Natural systems have different data characteristics. Hewitt (1970) reviewed some of these characteristics, including randomly and identically distributed observations that are not necessarily independent, potential spatial and temporal patterns, and observations that can show one or more

underlying physical processes that could regulate recurrence patterns. Such characteristics can differ from each other when prehistoric, historical, and instrumental datasets are compared. Table 2.1 shows the typical characteristics of prehistoric, historical, and instrumental data in terms of how the data are observed.

TABLE 2.1. Underlying assumptions of the various types of data found in natural event datasets.

	Prehistoric data	**Historical data**	**Instrumental data**
Time period represented	Represent data on events thousands of years ago.	Hundreds of years ago.	Last 100–150 years.
How event was measured	Recorded through focussed geological studies of particular events, e.g. palaeoseismology.	Human accounts	Instruments
Time of occurrence	Uncertain. No specific distribution	Certain	Certain
Location of event	Certain	Not necessarily certain	Certain
Event size error	Uncertain. Follows uniform distribution	Uncertain. Follows Gaussian or another symmetric distribution	Measured errors. Follows Gaussian or another symmetric distribution

Many of the event size measurement scales are defined in terms of the log scale, e.g. the earthquake magnitude scale, tsunami intensity scale, and the Volcano Explosivity Index. Other measurement scales represent a compromise between linear and log scales, such as the Beaufort wind velocity scale, the Saffir–Simpson damage potential scale for hurricanes or tropical cyclones, the Fuiita and Parson tornado scale, and the TORRO hail size scale. Blong (2003) provides an overview of the characteristics of various event size magnitude and intensity scales.

In Kijko and Dessokey (1987), Kijko and Sellevoll (1989, 1992), Kijko and Smit (2012), and Kijko *et al*. (2016), some of the above data quality characteristics, i.e. incompleteness, and data and event size uncertainty, are addressed with reference to the earthquake frequency–event-size (magnitude) relation by Gutenberg−Richter. Datasets that contain one or more of the three types of data (prehistoric,

historical, and instrumental) were used. Aleatory and epistemic uncertainty were considered in dealing with incomplete datasets, errors in the observed earthquake magnitude (event size), and uncertainty regarding the parameters of the applied occurrence and event size distributions. Aleatory uncertainty represents the inherent randomness of the phenomenon, whereas epistemic uncertainty represents the lack of knowledge. The probabilistic earthquake hazard in terms of earthquake magnitude is defined relevant to three recurrence parameters, namely λ, the rate of occurrence; the Gutenberg–Richter b-value (parameter); and the maximum possible earthquake magnitude x_{max}.

2.4 Power Laws in Nature

Power laws are used often to model natural systems, including the Gutenberg–Richter relation for earthquake magnitude (Gutenberg and Richter, 1942, 1956), tsunami intensity (Soloviev, 1970; Houston *et al*., 1977; Horikawa and Shuto, 1983; Burroughs and Tebbens, 2005; Geist and Parsons, 2006), landslide areas (Caccavale *et al*., 2017; Malamud *et al*., 2004), solar flare intensity and burned area for wildfires (Newman, 2005; Clauset *et al*., 2009), and air pollution (Shi and Liu, 2009). Publications such as Newman (2005), Burroughs and Tebbens (2001), and Geist and Parsons (2014) discuss the power laws found in various types of natural systems. Table 2.2 provides examples of the power laws found in nature.

TABLE 2.2. Some examples of frequency–event-size relations following power laws, as observed in nature.

Type of Hazard	Governing Frequency–Size Relation	Reference
Earthquakes	$log_{10} N(m) = a - bm$ m: seismic magnitude (a logarithmic variable), $N(M \geq m)$: cumulative number of seismic events with magnitude $m \geq LoC$; a, and b: parameters related to tectonic features.	Gutenberg and Richter (1942, 1956)
Tsunami	$n(i) = a10^{-bi}$ i: Soloviev–Imamura tsunami intensity, $n(I \geq i)$: number of tsunami events per annum, a and b: parameters related to the wave height features of a tsunami wave.	Soloviev (1970), Geist and Parsons (2006)
Floods	$Q(T) = CT^{\xi}$ $Q(T)$: maximum discharge associated with recurrence interval T, C and ξ: parameters.	Sachs *et al.* (2012)
Landslides	$N_{CL} = CA_L^{\beta}$ A_L: landslide area; N_{CL}: cumulative number of landslides with area $> A_L$; β: power law parameter, C: parameter.	Malamud *et al.* (2004), Caccavale *et al.* (2017)
Solar Flares	$dN = Ax^{-\xi}dx$ dN: number of events recorded with parameter x of interest between $[x; x + dx]$, A: parameter, ξ: power law parameter.	Newman (2005), Clauset *et al*. (2009)
Wildfires	$\frac{dN_F}{dA_F} = N_c = CA_F^{-\xi}$ N_C: cumulative number of events per year, A_F: burned area, ξ: power law parameter; C: parameter.	Newman (2005), Clauset *et al*. (2009)
Air pollution	$N = Cr^{-\xi}$ r: magnitude per daily pollution index, N: cumulative number of events $\geq r$, C: parameter, ξ: power law parameter.	Shi and Liu (2009)
Climate change variables (e.g. vapour pressure)	$N = cr^{-\lambda}$ N: cumulative number of events per time unit, r: size, λ: power law parameter, c: parameter.	Liu *et al.* (2014)
Tornadoes	$N = \frac{C}{\alpha - 1} x^{-(\alpha - 1)}$ N: cumulative number of events per time unit, x: size, $\alpha - 1$: power law parameter, C: parameter.	Machado *et al.* (2015)

In this thesis, power laws are used in the frequency–event-size analysis, equating the frequency with the respective event sizes (Burroughs and Tebbens, 2001, 2005; Newman, 2005; Clauset *et al*., 2009; Machado *et al*., 2015). Power laws are observed when the number of event occurrences $n(x)$ with event size in the interval $[x, x + dx]$ tends to be a linear line either on the log-log or on the log-linear scale. Assuming the event size x is measured on a logarithmic scale, this linear relationship is expressed as $\ln n(x) = c - kx$ and transforms to $n(x) = Ce^{-kx}$, with k representing the power law parameter and $C = \exp(c)$ a constant. One disadvantage of using the frequency–event-size relation is the division of the size variable into bins, which is not always possible to do. The cumulative frequency–event-size relation overcomes this problem and allows measurement sizes to fall on a continuous scale (Newman 2005).

The cumulative frequency–event-size relation $n_{X \geq x}$ equals

$$n_{X \geq x} = \frac{C}{k-1} 10^{-(k-1)x}, \tag{2.1}$$

and represents the number of event equal or exceeding x. By taking the logarithm, Eq. 2.1 transforms into

$$\log n_{X \geq x} = a - bx\,, \tag{2.2}$$

representing a line with $a = \log\left(\frac{C}{k-1}\right)$ and $b = k - 1$ (Newman, 2005), with a considered to be constant. The cumulative frequency−event-size power law probability distribution is also referred to as a Pareto distribution. Eq. 2.2 can be transformed into

$$n_{X \geq x} = e^{\alpha - \beta x}, \tag{2.3}$$

with $\beta = b \ln 10$ and the constant $\alpha = a \ln 10$.

As an example, we will use the Gutenberg–Richter cumulative frequency–size relation of Gutenberg and Richter (1942, 1956) to discuss the rest of the thesis. This relation shows how the cumulative

number of events greater than or equal to x relates to the size of the event x, as given in Eq. 2.2 , with parameters a and b related to the tectonic setting of the area under investigation. The event size x is measured on the logarithmic scale, thereby providing a linear relationship between the logarithm of both, namely the frequency (histogram) and cumulative frequency and earthquake event size. A typical example is provided in Figure 2.2 of the histogram for the number of earthquakes on a logarithmic scale when investigating the frequency–event-size relation. This depiction is seen also in the histograms of other types of natural hazards. The right-hand side of the histogram can be modelled by Eq. 2.2. Contrary to expectation, the left-hand side does not increase linearly as the size of the events progressively declines. This is attributable to small events not being observed.

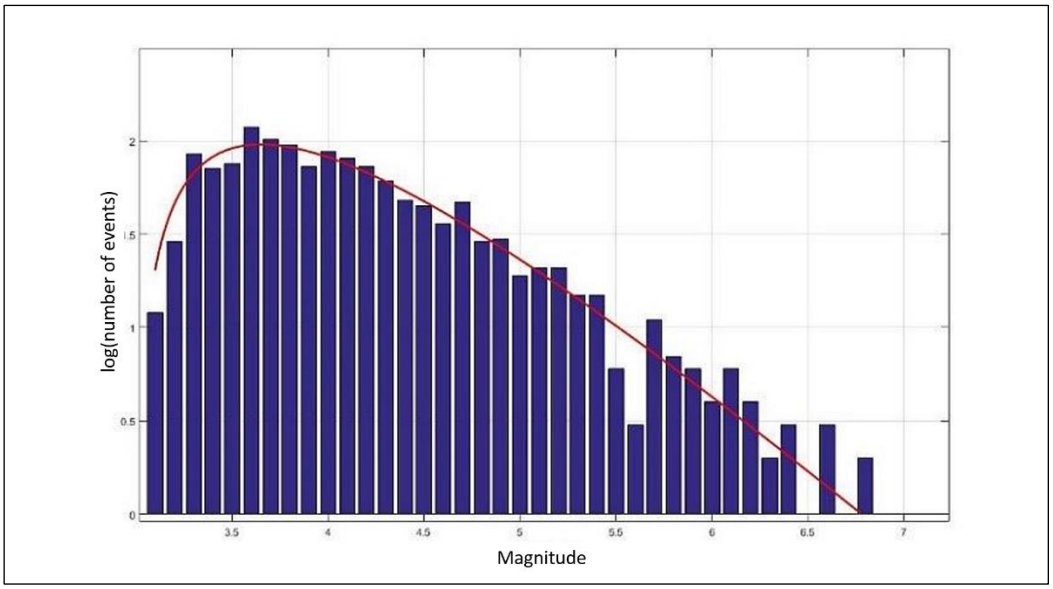


FIGURE 2.2. Histogram of apparent distribution of events in a hypothetical seismic dataset.

Pisarenko and Rodkin (2014) provide a detailed discussion on employing heavy-tail power laws in the hazard assessment of extreme events. Their reasoning, i.e. using only these types of events, stems from these events being the brunt of the associated effect on humans in terms of loss of life and financial

losses. Datasets for these events are considered complete, thereby facilitating a straightforward implementation of statistical methods.

Due to the devastating effects of a natural hazard event on society, extensive research has been devoted to understanding and modelling such events, with the aim of imparting resilience to society. Almost all the statistical models currently used require datasets that are complete in order to provide statistically sound estimates. This requirement, however, cannot be met in all instances. Accordingly, the methodology first developed by Kijko and Sellevoll (1989) and extended in Kijko and Sellevoll (1992) shows that it is possible to utilise datasets that are highly incomplete and that have uncertain event sizes.

The underlying relationship between frequency and event size provides a unique opportunity to define a single methodology that can be used in any setting where data can be described through frequency– or cumulative frequency–event-size relations following a power law. This can be done by following the same framework for earthquake hazard modelling as that described by Kijko and Sellevoll (1989, 1992). Unlike the method of Pisarenko and Rodin (2010, 2014), using the previously mentioned framework does not restrict the user to employing only extreme events but facilitates utilising as much of the data as is possible. This provides an opportunity to include all the available types of data in the analyses, namely prehistoric, historical and instrumental data, and prior independent information. Therefore, the model parameters can be assessed in a more reliable way.

Two aspects are lacking in the framework by Kijko and Sellevoll (1989, 1992). The first relates to instances of the data not following the required strict independence and stationarity of the investigated process. Space–time fluctuations of the natural process are common, and provision should be made to include them.

The second aspect relates to the uncertainty in occurrence of event or the validity of the events. For example, tsunami datasets can contain an additional parameter or validity index that shows the confidence of the authors of the dataset as to whether the evidence supports the classification of a tsunami event or only a severe storm. By introducing the weighted likelihood function to the

methodology, the researcher has the option to include all the events and retrieve a rate of occurrence that is more realistic.

The framework by Kijko and Sellevoll (1989, 1992) estimates parameters using maximum likelihood estimation (MLE). To use prior information, there is a need for Bayesian inference (BI). Prior information is arguably a valuable tool to estimate parameters from and should be useful in stabilising the results based on incomplete and uncertain data.

2.5 Summary

In conclusion, Chapter 2 provides a broad overview of the different types of data and statistical methodologies typically used in natural hazard analyses. Various data sources exists and include systematics and non-systematic data, insurance data and social media data. Each of these data types can add information necessary to better model the underlying behaviour of the hazard under investigation. The nature of typical empirical natural hazard datasets are also evaluate, listing their respective advantages and disadvantages. These data characteristics can typically be modelled by sub-setting the data according to certain parameters. Power laws are used to derive empirical probability density functions (PDF) and cumulative probability density functions (CDF). This is done to ensure the modelling process remains true to the physical characteristics of the natural phenomena under investigation. From the empirical distributions, model-based PDFs and CDFs are defined and used to build likelihood functions for each of these sub-datasets, MLE and BI can be applied in the estimation of the recurrence parameters (Chapter 3).

Chapter 3. Methodology

"All *models are wrong, but some are useful." George Box*

In this chapter, a methodology is described that can be used to assess the potential natural hazard associated with some physical properties of geophysical, hydrological, meteorological, or climatological phenomena. For this purpose, the hazard will be described using two types of distribution processes that characterise natural events. These are the temporal, frequency occurrence distribution, and the event size distribution. For both processes, a probability distribution is defined, and parameter estimates are derived. In this way, the associated probabilities of exceedance and the return periods for different event sizes can be defined.

The methodology applied comprise of estimating the hazard recurrence parameters based on different types of input data and using power laws or their transformation. Estimation of the rate of occurrence λ, the power law parameter b, and the maximum possible event size x_{max} are based on information provided by prehistoric, historical and instrumental data types. The parameter λ will be referred to as rate of occurrence for the rest of the thesis.

Several assumptions are made to define the principal and extended methodology. Section 3.1 defines the methodology as applied in Kijko and Dessokey (1987) and Kijko and Sellevoll (1989). Section 3.2 introduces various types of uncertainty, thereby extending the principal model. Section 3.3 discusses parameter estimation using maximum likelihood estimation (MLE) and Bayesian inference (BI). One method for the estimation of the maximum possible event size x_{max} is provided in Section 3.4, the hazard estimates in Section 3.5, and the model comparisons in Section 3.6.

3.1 Principal Model

The methodology is first defined in terms of a principal model for a single, complete event dataset. For the principal model, the occurrence (Section 3.1.1) and event-size distributions (Section 3.1.2) are defined to describe the temporal and physical nature of the data. Next, two possible adjustments to the event-size distribution are introduced (Section 3.1.3). The adjustments focus on the potential shifting of the distribution to start at an event-size other than 0, and a potential upper truncation. Lastly, the same equations are described for extreme prehistorical and historical data in Section 3.1.4.

Assumption 1. *The random variable X refers to the size of the observed events. The sample space for event sizes is defined in terms of N ($j = 1, \dots N$), the random number of possible independent and identically distributed (iid) event sizes, $X = \{X_1, X_2, \dots X_N\}$.*

3.1.1 Occurrence distribution

Assumption 2*. The occurrence of observed events in a specific area is a stationary Poisson process with a rate of occurrence parameter $\lambda > 0$. For a random number of N events, within an time interval Δt, the probability of observing n events equals*

$$P_N(n) = \frac{e^{-\lambda \Delta t}(\lambda \Delta t)^n}{n!}, \qquad n = 0,1,2 \dots. \tag{3.1}$$

References to instances where this assumption was used can be found in, e.g. Cornell (1968), Lomnitz (1973), Gardner and Knopoff (1974), Geist and Parsons (2006), Corominas and Moya (2008), and Rougier *et al.* (2018).

The Poisson parameter $\lambda > 0$ describes the rate of event occurrences. It is also known as the rate of exceedance. An estimate of the parameter λ is obtained by dividing the number of events by the

specified time interval, i.e. $\lambda = n/\Delta t$. The violation of the requirement that events have to be independent is discussed in Section 3.2.

Following Assumption 2, the inter-arrival times of events are defined by assuming that Δt is a random, time to the first arrival, variable. The probable of first arrival can be considered as the probability of no events being observed in a specific time interval Δt, i.e.

$$Pr(\Delta T > \Delta t) = 1 - F_{\Delta T}(\Delta t), \tag{3.2}$$

where $F_{\Delta T}(\Delta t)$ is the cumulative distribution function (CDF) of Δt, defined as $P_N(n)$ with the parameter $n = 0$. Therefore

$$F_{\Delta T}(\Delta t) = 1 - e^{-\lambda \Delta t} \qquad \Delta t \geq 0, \tag{3.3}$$

with the probability distribution function (PDF) defined as

$$f_{\Delta T}(\Delta T) = \frac{dF_{\Delta T}(\Delta T)}{d\Delta T} = \lambda e^{-\lambda \Delta t} \quad \Delta t \geq 0. \tag{3.4}$$

The memoryless property often associated with the Poisson distribution translates from the memoryless property for the inter-arrival times, which is an exponential distribution. In other words, the failure for an event to occur does not affect the inter-arrival time (Balakrishnan and Basu, 1995; Bain and Engelhardt, 1992).

3.1.2 Event size distribution

Assumption 3. *The frequency–event-size relation of random variable X follows a linear equation, defined in terms of N (random) possible iid events* $X = (X_1, X_2, \ldots X_N)$, *which are related to the frequency of occurrence, as defined by the log-transformed power law in Eq. 2.2 and the exponential distribution of Eq. 2.3. The two parameters a and b are a function of the area under investigation.*

Equation 2.2 can be viewed as the log-transformed cumulative histogram of the number of event sizes $x_j \geq x$. Following Eq. 2.3 the CDF and PDF for the event size equal

$$F_X(x) = 1 - e^{-\beta x} \quad x > 0, \tag{3.5a}$$

$$f_X(x) = \beta e^{-\beta x} \quad x > 0, \tag{3.5b}$$

with the parameter $\beta = b \ln(10)$. The derivations are available in the Appendix.

3.1.3 Adjustments to event size distribution

3.1.3.1 Shifted event size distribution

Different levels of completeness (LoC) can be observed in datasets. For instrumental data, LoC is a function of the accuracy of instrumentation and the coverage of an observational network. Prehistoric and historical datasets are dependent on whether the event was felt and how well its effects were recorded. The LoC of an instrumental dataset is dependent on the quality of the instruments, which decreases with the increasing ability of instruments to detect smaller events. Nevertheless, there are still instances of smaller size events being undetected, which is particularly relevant when measuring instruments that are faulty or not yet able to detect small events. Small recorded events are often not of interest, as they seldom cause noteworthy damage or loss. These events are sometimes ignored during hazard analyses, depending on the aim of the project. This results in a typical histogram, as in Figure 3.1, showing a decrease in the number of small event sizes compared with that expected. The apparent distribution, therefore, varies from the expected true distribution. By introducing the LoCs, the methodology makes provision for deviations from the power law at small event sizes, i.e. the left tail of the distribution.

Assumption 4. *The event dataset is considered complete for event sizes equal to and exceeding a certain event size, also known as the level of completeness (LoC) or* x_{min}.

The introduction of parameter x_{min} results in the starting point of the distribution being shifted, as illustrated in Figure 3.1. This shift is reflected in the rate of occurrence (Cosentino *et al*., 1977; Youngs and Coppersmith, 1985; McGuire and Arabasz, 1990; Benjamin and Cornell, 2014) by redefining the Poisson distribution in Eq. 3.1 as the probability of observing N events for event sizes $x_j \geq x_{min}$ $(j = 1, \dots N)$ in the time interval Δt.

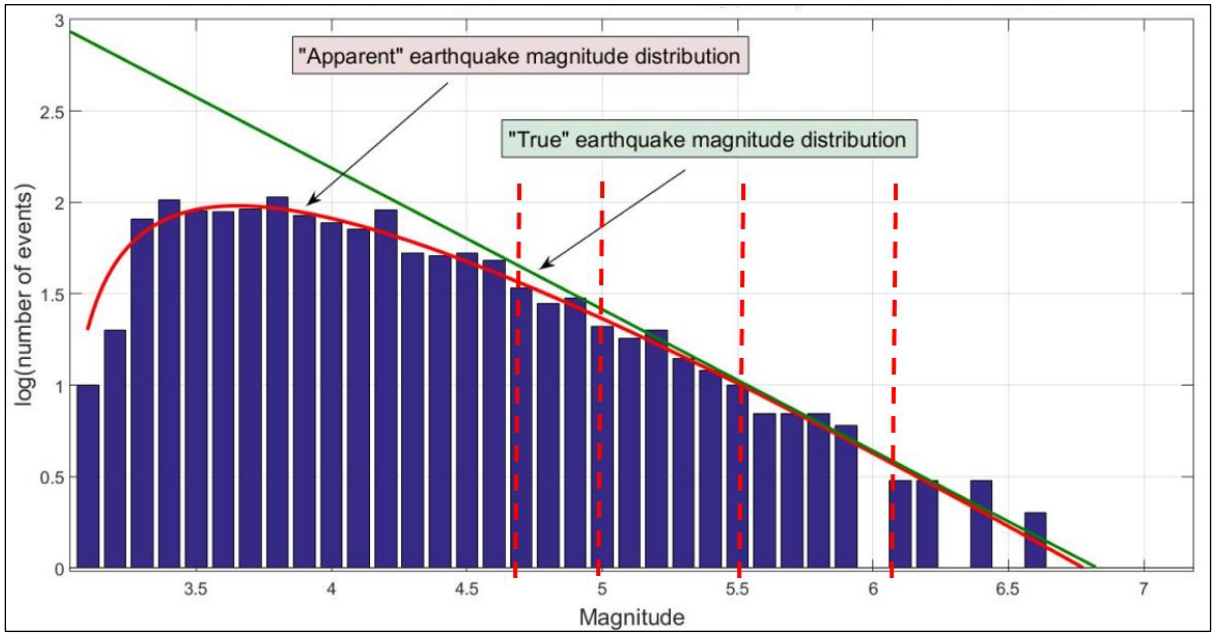


Figure 3.1. Histogram of the apparent distribution of events in a hypothetical seismic dataset. The dashed, vertical red lines represent examples of potential shifts in the event size distribution according to the chosen level of completeness x_{min}. Various methods exist to assess this parameter, which can lead to different results (modified from Kijko and Smit, 2017).

Parameter x_{min} has a substantial effect on the final estimates, as seen in Section 3.3. Several different methodologies exist to assess this parameter, with no clear preferred one. These include goodness-of-fit analysis, maximum curvature, b-value stability estimation method, entire-magnitude-range method, median based assessment of segment slope, and assessment of x_{min} by the method of moments. These methods and their performance are discussed in detail in, e.g. De Witt (2013), and Mignan and Woessner (2012).

For the purpose of this thesis, the visual cumulative method by Mulargia and Tinti (1985) was applied for the estimation of x_{min}. This method aims to identify periods in time where the event dataset is complete for a specified event size. All events are divided into their respective event-size intervals and represented graphically as the cumulative number of events over time. The time period where the cumulative trend evens out to approximate a straight line is considered the cut-off date. The LoC for the period is the lowest event size for that time period. This method helps identify both the sub-datasets and the respective LoC to be used in this thesis (Section 3.2).

Assumption 5. *The frequency–event-size distribution is shifted from zero to x_{min}. The respective conditional CDF and PDF of the shifted event size distribution equal*

$$F_X(x|x_{min}) = \begin{cases} 0 & x < x_{min} \\ 1 - e^{-\beta(x-x_{min})} & x \geq x_{min} \end{cases}, \tag{3.6a}$$

$$f_X(x|x_{min}) = \begin{cases} 0 & x < x_{min} \\ \beta e^{-\beta(x-x_{\min})} & x \geq x_{min} \end{cases}. \tag{3.6b}$$

The derivations are available in the Appendix.

3.1.3.2 Truncated from the top event size distribution

Natural hazards are driven by forces of nature adhering to certain physical constraints. These constraints are unique to the different types of hazards. As events behave in a specific manner, given the input stimulus, some assumptions regarding natural events can be made. One of these is the assumption that an upper limit or maximum event size x_{max} exists. The upper limit depends on a series of factors, including the type of hazard and the physical properties of the region. For example, the potential earthquake magnitude of a specific fault line is related to the fault line geometry. This physical law indicates that an upper limit x_{max} exists (Wells and Coppersmith, 1994; Gibowicz and Kijko, 1994; Pisarenko *et al.*, 1996; Sahakian *et al*., 2017). Similarly, the physical processes that govern events, such as tsunamis (Burroughs and Tebbens, 2001), landslides (e.g. Guzzetti *et al.*, 2008), floods (e.g. Fernandes *et al.*, 2010), and fires (e.g. Cumming, 2001) are subject to upper limits. Its parameter depends on the type of hazard and the physical properties of the region.

The maximum possible event size or the upper limit of the distribution is an important parameter in any hazard modelling process. It represents the worst-case scenario, for which society has to make provision. Assessing the upper limit of the power laws can be problematic if only a few large events are available for analyses. In such instance, the inclusion of the largest prehistoric and historical information plays a crucial role in constraining the estimation of the upper limit of the event size distribution.

Assumption 6. *An upper limit for event size x_{max} exists for the area under investigation.*

Adjusting the shifted event size distribution (Eq. 3.6) to include a truncated upper bound, as shown in Figure 3.2, leads to the shifted-truncated event size distribution. In this instance, the power law is defined between x_{min} and x_{max}. The distribution is normalised, such that the conditional CDF and PDF are defined, respectively, as (Cosentino *et al*., 1977; Youngs and Coppersmith, 1985; McGuire and Arabasz, 1990; Benjamin and Cornell, 2014)

$$F_X(x|x_{min},x_{max}) = \begin{cases} 0 & x < x_{min} \\ \dfrac{e^{-\beta x_{min}} - e^{-\beta x}}{e^{-\beta x_{min}} - e^{-\beta x_{max}}} & x_{\min} \le x \le x_{\max}, \\ 1 & x > x_{max} \end{cases} \quad (3.7a)$$

$$f_X(x|x_{min},x_{max}) = \begin{cases} \dfrac{\beta e^{-\beta(x-x_{min})}}{1 - e^{-\beta(x_{max}-x_{min})}} & x_{min} \le x \le x_{max} \\ 0 & x < x_{min}; x > x_{max} \end{cases}. \quad (3.7b)$$

The derivations are available in the Appendix.

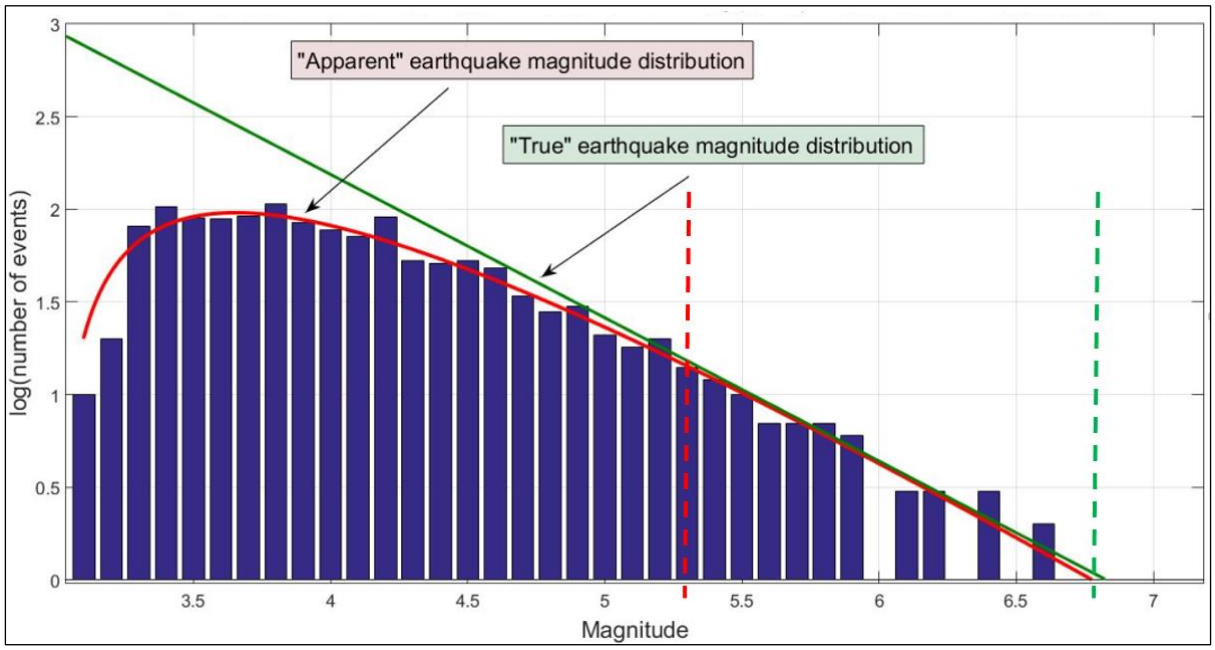


FIGURE 3.2. Histogram of the apparent distribution of events in a hypothetical seismic dataset. The dashed, vertical red line represents the LoC x_{min} and the green line the upper limit x_{max}. The power law defined between these two limits will be evaluated (modified from Kijko and Smit, 2017).

3.1.4 Inclusion of prehistoric and historical data

By adjusting Assumptions 1 and 6 to adhere to the characteristics of datasets that contain only extreme events, similar frequency–event-size distributions are defined, respectively, for prehistoric and

historical datasets (Figure 3.3). Let x_{min} represent the LoC of the entire dataset that includes the extreme prehistoric, historical, and instrumental event datasets. The values $x_H \geq x_{min}$ and $x_P \geq x_{min}$ represent the smallest observed event sizes in the historical and the prehistoric datasets, respectively. For the probabilistic model, both datasets are considered to be governed by extreme distributions.

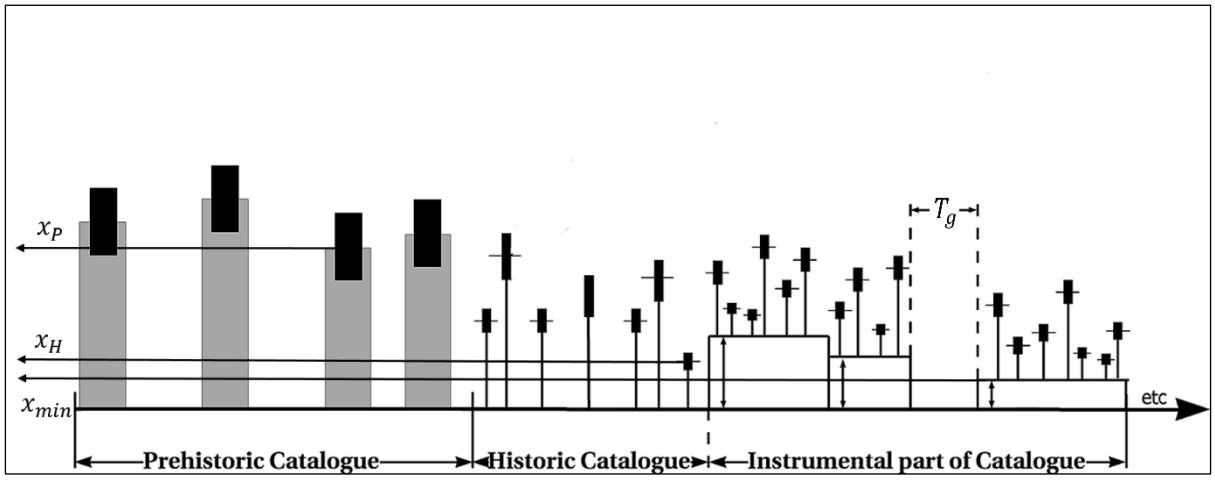


FIGURE 3.3. An event dataset that can be used to obtain three key recurrence parameters for natural hazard assessment. The values x_P and x_H represent the smallest observed event sizes in the respective prehistoric and historical datasets. The parameter x_{min} represents the smallest level of completeness (LoC) of the entire dataset (modified after Kijko and Sellevoll, 1992).

Compared with complete datasets, where all the events that are above the LoC ($x_j \geq x_{min}$) are included in the analyses, only the largest event within a specific time interval is used when analysing extreme prehistoric and historical datasets. It is possible, therefore, that within a specific time interval in the prehistoric and historical datasets, several events could be ignored in favour of the largest event. An extreme event size distribution defined for the prehistoric dataset is based the parameters of the frequency–event-size recurrence relation (power law) and, therefore, retains the connection with the event size distribution of the instrumental dataset, as defined in Sections 3.1.2 and 3.1.3. A similar distribution can be defined for historical datasets by modifying the specified time interval.

Assumption 7. *Let N_P be the random number of events in an extreme prehistoric dataset. The random event size variable $X = \{X_1, X_2, \dots X_{N_P}\}$ are independently, identically distributed (iid) with $x_j \geq x_{min}$ for $j = 1, \dots n$. The occurrence of the events is assumed to be a stationary Poisson process within time interval Δt_P. The probability to observe n_P events in the time interval Δt_P equals*

$$P_{N_P}(n_P) = \frac{e^{-\lambda_P \Delta t_P}(\lambda_P \Delta t_P)^{n_P}}{n_P!}, \qquad n_P = 0,1,2 \dots \tag{3.8}$$

where λ_P represents the rate of occurrence. The Poisson process for historical data is defined in a similar way.

Identity 1. *The rate of occurrence for event size x equals (e.g. Benjamin and Cornell, 2014)*

$$\lambda(x) = \lambda(x_{min})[1 - F_X(x)],$$

with $\lambda(x_{min})$ is the rate of occurrence for x_{min}, and $F_X(x)$ the applied CDF used to calculate $\lambda(x_{min})$.

This parameter helps to assess the rate of occurrence for any event size when at least one other rate of occurrence is known (Campbell, 1982, Kijko and Sellevoll, 1989; 1992).

Assumption 8*. Let x_0 be the largest event size in the time interval Δt_P. For iid event sizes, the CDF and PDF (Coles, 2001) equal*

$$F_X^{MAX}(x_0) = [F_X(x_0)]^{n_P}, \tag{3.9a}$$

$$f_X^{MAX}(x_0) = n_P[F_X(x_0)]^{n_P - 1} f_X(x_0). \tag{3.9b}$$

The derivations are available in the Appendix.

The shifted extreme event size distribution is defined by substituting Eq. 3.6a into Eq. 3.9a and Eq. 3.6b into Eq. 3.9b.

$$F_X^{MAX}(x_0|x_{min}) = \left[1 - e^{-\beta(x_0 - x_{min})}\right]^{n_P} \quad x_0 \geq x_{min}, \tag{3.10a}$$

$$f_X^{MAX}(x_0|x_{min}) = n_P\left[1 - e^{-\beta(x_0 - x_{min})}\right]^{n_P - 1}\beta e^{-\beta(x_0 - x_{min})} \quad x_0 \geq x_{min}. \tag{3.10b}$$

Similarly, the shifted-truncated extreme event size distribution, truncated at the maximum possible event size x_{max}, can be described in terms of the conditional CDF and PDF by substituting Eq. 3.7a into Eq. 3.9a and Eq. 3.7b into Eq. 3.9b.

$$F_X^{MAX}(x_0|x_{min}, x_{max}) = \begin{cases} 0 & x_0 < x_{min} \\ \left[\dfrac{e^{-\beta x_{min}} - e^{-\beta x_0}}{e^{-\beta x_{min}} - e^{-\beta x_{max}}}\right]^{n_P} & x_{min} \leq x_0 \leq x_{max}, \\ 1 & x_0 > x_{max} \end{cases} \tag{3.11a}$$

$$f_X^{MAX}(x_0|x_{min}, x_{max}) =$$

$$\begin{cases} \left[\dfrac{e^{-\beta x_{min}} - e^{-\beta x_0}}{e^{-\beta x_{min}} - e^{-\beta x_{max}}}\right]^{n_P - 1} \dfrac{n_P \beta e^{-\beta(x_0 - x_{min})}}{1 - e^{-\beta(x_{max} - x_{min})}} & x_{min} \leq x_0 \leq x_{max} \\ 0 & x_0 < x_{min}; x_0 > x_{max} \end{cases}. \tag{3.11b}$$

Assumption 9. *Let x_0 be the largest event size in a specified time interval Δt_P, and n_P is the random number of events with a probability distribution $P_{N_P}(n_P)$. The probability that the maximum event size in the dataset does not exceed x_0 equals*

$$P(\max(X) \leq x_0|\Delta t) = P_{N_P}(n_P)[F_X(x_0)]^{n_P}. \tag{3.12}$$

From Assumption 9 (Epstein and Lomnitz, 1966) and the theorem of total probability (Cramér, 1961), follows the conditional CDF for extreme events x_0 in a time interval Δt_P

$$F_X^{MAX}(x_0|\Delta t_P) = \sum_{n_P=0}^{\infty} P_{N_P}(n_P)[F_X(x_0)]^{n_P}, \tag{3.13}$$

(derivation available in the Appendix), such that the subsequent generic conditional CDF and PDF reduce to

$$F_X^{MAX}(x_0|\Delta t_P) = e^{-\lambda_P \Delta t_P [1-F_X(x_0)]}, \tag{3.14a}$$

$$f_X^{MAX}(x_0|\Delta t_P) = \lambda_P \Delta t_P f_X(x_0) e^{-\lambda_P \Delta t_P (1-F_X(x_0))}. \tag{3.14b}$$

The derivations are available in the Appendix.

The extreme event size distributions are defined by substituting the relevant equations into Eqs 3.14a and 3.14b. Substituting Eqs 3.5a and 3.5b into Eqs 3.14a and 3.14b equal

$$F_X^{MAX}(x_0|\Delta t_P) = \exp(-\lambda_P \Delta t_P e^{-\beta x_0}) \quad x_0 > 0, \tag{3.15a}$$

$$f_X^{MAX}(x_0|\Delta t_P) = \lambda_P \Delta t_P \beta \exp(-\beta x_0) \exp\left[-\lambda_P \Delta t_P e^{[-\beta x_0]}\right] \quad x_0 > 0. \tag{3.15b}$$

Equation 3.15a is the Gumbel I extreme distribution (Gumbel, 1962), with distribution parameters defined as $\alpha > 0$ and u with $x_0 \in (-\infty, +\infty)$ and

$$F_{MAX}^{\mathrm{I}}(x_0) = \exp(-\exp(-\alpha(x_0 - u))),$$

with $\alpha = \beta = b\, ln(10)$ and $-\lambda_P \Delta t_P = \exp(\alpha u)$ (Epstein and Lomnitz, 1966).

Similarly, the conditional CDF and PDF of the extreme shifted event size distribution are defined by substituting Eqs 3.6a and 3.6b into Eqs 3.14a and 3.14b to equate to

$$F_X^{MAX}(x_0|x_{min}, \Delta t_P) = \exp(-\lambda_P \Delta t_P \mathrm{e}^{-\beta(x_0 - x_{min})}) \quad x_0 \geq x_{min}, \tag{3.16a}$$

$$f_X^{MAX}(x_0|x_{min},\Delta t_P) = \lambda_P \Delta t_P \beta \exp\left(-\beta(x_0 - x_{min})\right) \exp\left(-\lambda_P \Delta t_P e^{-\beta(x_0 - x_{min})}\right) \quad x_0 \geq x_{\min}, \tag{3.16b}$$

and the conditional extreme shifted-truncated event size CDF and PDF are defined by substituting Eqs 3.7a and 3.7b into Eqs 3.14a and 3.14b to equate to

$$F_X^{MAX}(x_0|\Delta t_P, x_{min}, x_{max}) = \exp\left(-\lambda_P \Delta t_P \left(\frac{e^{-\beta x_0} - e^{-\beta x_{max}}}{e^{-\beta x_{min}} - e^{-\beta x_{max}}}\right)\right) \quad x_{min} \leq x_0 \leq x_{max}, \tag{3.17a}$$

$$f_X^{MAX}(x_0|\Delta t_P, x_{min}, x_{max}) = \frac{\lambda_P \Delta t_P \beta \left[e^{-\beta(x_0 - x_{min})}\right]}{1 - e^{-\beta(x_{max} - x_{min})}} \exp\left(-\lambda_P \Delta t_P \left(\frac{e^{-\beta x_0} - e^{-\beta x_{max}}}{e^{-\beta x_{min}} - e^{-\beta x_{max}}}\right)\right) \quad x_{min} \leq x_0 \leq x_{max}. \tag{3.17b}$$

Equation 3.17a can also be considered a Gumbel Type III equation (Benjamin and Cornell, 2014), where $F_Y^{\text{MAX}}(y) = \exp(-\left(\frac{w-y}{w-u}\right))^k$, such that $w = -e^{-\beta x_{max}}$, $y = -e^{-\beta x_0}$, $u = -e^{-\beta x_{min}}$, and $k = 1$.

3.2 Extension of the Principal Model

The assumptions for the principal model require events to behave in a strict manner. In reality, the recorded information is seldom complete and not all events have been captured. In many instances, records are incomplete and uncertain, and are available only for a limited number of years. In regard to natural phenomena, certain countries are unable to measure and record events of all sizes owing to the lack of appropriate infrastructure. In certain instances, the occurrence of large events could destroy measuring equipment, such as river flow measurement stations, and much time could pass before the

instruments are replaced. Therefore, the recorded datasets inherently have different sub-datasets, with varying levels of completeness.

As the quality and quantity of measurement instrumentation have improved over the years, so has the accuracy with which events are recorded. Currently, the location and size of natural hazards can be identified with a small error, in most cases. Older events, however, are recorded with substantial errors, particularly concerning the size of the event. Often, event sizes are estimated based on qualitative descriptions. In such instances, uncertainty could be present in the event size estimation, which could result in the severe under- or overestimation of the actual hazard and the subsequent risk.

Investigations into natural hazard datasets often reveal spatial and temporal dependencies in the data. These dependencies can manifest through cycles, oscillations, or otherwise related fluctuations, and violate the assumptions of independence and stationarity required in the modelling process (Section 3.1). A common way of dealing with this problem is to decluster the datasets by removing smaller related events and keeping only the largest event in the cluster. In relation to earthquakes, this implies the removal of fore- and aftershocks. The disadvantage of declustering is that not only are the data points removed to better fit a model but also it could lead to underestimation of the hazard. Ideally, any methodology must provide for these types of dependencies.

The methodology described in Kijko *et al*. (2016) and Smit *et al*. (2017) accounts for the above aleatory and epistemic uncertainty by making provision for incompleteness in the dataset, uncertainty in the event size determination, and uncertainty associated with the parameters in the applied occurrence and event size distributions. The methodology is focused primarily on empirical data but, as shown later, is flexible enough to utilise any additional information in an attempt to improve the parameter estimation.

The aim of this section is to update the principal model in Section 3.1 to account for four common data quality problems when working with real-world data, namely:

- incompleteness of the datasets

- uncertainty in the event size
- uncertainty associated with the parameters of the applied occurrence and event size distributions
- uncertainty in the event occurrence (validity).

3.2.1 Incompleteness

The principal model of Section 3.1 can be applied to the datasets of the form illustrated in Figure 3.3 in three ways. Either the prehistoric and historical data are discarded or a high, single LoC x_{min}, that holds for all the different types of datasets, is selected. Both options facilitate the application of established methodologies. Both these options (discarding prehistorical and historical data or selecting a high, single LoC x_{min}) have the disadvantage that it reduces the number of events for model calibration. This could affect the estimation of both model parameters, the power law parameter, i.e. the slope of Eq. 2.2, and the estimated rate of occurrence and the return periods of events. A third option is to subset the data further according to different LoCs.

Several authors, such as Molchan *et al.* (1970), Weichert (1980), Kijko and Sellevoll (1989, 1992), Rosenblueth (1986), Rosenblueth and Ordaz (1987), and Kijko and Smit (2012) have investigated the possibility of sub-setting data according to a type of LoC. These authors divided the data according to time or according to event size levels over time. For this thesis, the data are subdivided according to the format of Figure 3.3, following Kijko and Sellevoll (1989, 1992), Kijko and Smit (2012), Kijko *et al.* (2016), and Smit *et al.* (2017**,** 2019).

Assumption 10. *A typical dataset related to natural hazard assessment consists of prehistoric, historical, and instrumental data or a combination thereof. It is assumed that the prehistoric (P) and historical (H) datasets are independent, and that the instrumental dataset (I) can be divided into s*

subsets. Each of the s instrumental sub-datasets has a LoC $x_{min}^{(i)}$ for $i = 1, \dots, s$. *Each sub-dataset contains a random number of* N_i *events, within a time interval* Δt.

Assumption 11. *The model parameters are estimated utilising all the available data types by defining likelihood functions for prehistoric (P), historic (H) and instrumental datasets* $i = 1, \dots, s$.

By structuring the dataset following Assumptions 11 and 12 and applying the additive property of likelihood functions (Rao, 1973), a single likelihood function L_{Total} for the entire dataset can be defined. Following this formula allows the user to construct likelihood functions around data gaps T_g.

$$L_{Total}(\lambda, \beta | \boldsymbol{\mathcal{I}}_P, \boldsymbol{\mathcal{I}}_H, \boldsymbol{\mathcal{I}}_I) = L_P(\lambda, \beta | \boldsymbol{\mathcal{I}}_P)\, L_H(\lambda, \beta | \boldsymbol{\mathcal{I}}_H)\, L_I(\lambda, \beta | \boldsymbol{\mathcal{I}}_I), \tag{3.18}$$

where L_P and L_H denote the likelihood functions based on the prehistoric and historical parts of the dataset, and L_I is the likelihood function based on the instrumental sub-datasets. The background information matrices for the three types of datasets are defined as $\boldsymbol{\mathcal{I}}_P = (\boldsymbol{x_P}, \boldsymbol{\Delta t_P})$, $\boldsymbol{\mathcal{I}}_H = (\boldsymbol{x_H}, \boldsymbol{\Delta t_H})$, and $\boldsymbol{\mathcal{I}}_I = (\boldsymbol{n_I}, \boldsymbol{\Delta t_I}, \boldsymbol{X_I})$. For the prehistoric dataset, $\boldsymbol{x}_P$ and $\boldsymbol{\Delta t}_P$ are the n_P vectors of the prehistoric event sizes x_{Pj} that occurred within time intervals Δt_{Pj}, where $j = 1, \dots, n_P$. The vectors $\boldsymbol{x_H}$ and $\boldsymbol{\Delta t_H}$ are defined similarly for the historic dataset. For the instrumental dataset, the vectors are defined as $\boldsymbol{n_I} = [n_1, n_2, .. n_s]$ and $\boldsymbol{\Delta t_I} = [\Delta t_1, \Delta t_2, .. \Delta t_s]$, and the matrix $\boldsymbol{X_I} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots \boldsymbol{x_s})$ represents the vectors of events x_{ij} with $i = 1, \dots s$, the number of instrumental sub-datasets, and $j = 1, \dots n_i$, the number of events in subset i.

The likelihood function for the prehistoric dataset $L_P(\lambda, \beta | \boldsymbol{\mathcal{I}}_P)$ is constructed, following the chosen extreme conditional PDF $f_X^{MAX}(x_0 | \Delta t_P)$, from Eqs 3.15b, 3.16b or 3.17b, in N_P consecutive time intervals $\boldsymbol{\Delta t_P} = \{\Delta t_{P1}, \Delta t_{P2}, \dots, \Delta t_{Pn_P}\}$ equals

$$L_P(\lambda, \beta|\boldsymbol{\mathcal{I}_P}) = \prod_{j=1}^{n_P} f_X^{MAX}(x_{0Pj}|\Delta t_P). \quad (3.19a)$$

The historical likelihood function $L_H(\lambda, \beta|\boldsymbol{\mathcal{I}_H})$ is defined in a similar manner.

The likelihood function for the instrumental datasets $L_I(\lambda, \beta|\boldsymbol{\mathcal{I}_I})$ is a combination of the likelihood functions, following from the defined frequency and event size distributions. In the instance of activity rate λ being independent of the power law parameter β, then

$$L_i(\lambda, \beta|\boldsymbol{\mathcal{I}}) = \prod_{i=1}^{s} L_{\lambda i}(\lambda|n_i, \Delta t_i) \times L_{\beta i}(\beta|\boldsymbol{x}_i), \quad (3.19b)$$

for $i = 1,..s$. The conditional PDFs for the occurrence and event size distributions chosen by the researcher are substituted into Eqs 3.18 and 3.19a and b. In the instance of a single dataset or only instrumental data, parameter estimation is relatively effortless (see Sections 3.3.1.1 and 3.3.1.3). In the instance of prehistoric and historical data being combined with instrumental data, the function becomes non-linear, and can be solved only with numerical methods. Other examples of constructing a total likelihood function to combine various types of data are employed in earthquake, tsunami, and extreme flood analyses (e.g. Stedinger and Cohn 1986; Cohn and Stedinger, 1987; Fernandes *et al.*, 2010; Kijko *et al*., 2016; Lam *et al*., 2017, Smit *et al*., 2017; Cloete *et al*., 2018).

Further extensions of the likelihood functions are explored in Section 3.2.4. Section 3.3 explores the estimation of the likelihood functions using MLE and Bayesian inference (BI).

3.2.2 Parameter uncertainty

The standard and most simplistic assumption about the occurrences of natural hazards is that they follow a Poisson distribution (Assumption 2). This assumes that the occurrence of events of natural phenomena is random and mutually independent. In other words, the distribution has stationary and independent

increments, with the rate of occurrence depending on the number of events in the specified time interval. The distribution, therefore, depends only on the size of the time interval and not its start or end-points. In addition, the distributions of each time increment are independent of every other time increment. This means that in a specific area, the rate of occurrence (λ) and the power law parameter b, remain constant over time.

This assumption can be true in most cases, and results in a mathematical formalism that is easy to apply. By assuming that the probability of event occurrence in a small time interval Δt is negligible, the probability of observing at least one event within time interval Δt is

$$P(at\ least\ one\ event\ in\ time\ \Delta t) = 1 - e^{-\lambda \Delta t},$$

which for $\lambda \Delta t < 0.1$ can be approximated as $\lambda \Delta t$ (Baker, 2013).

The assumption that recorded events follow a stationary, homogeneous Poisson process is not always true, particularly where prehistoric and historical data are incorporated into the analyses (Nott, 2003). Temporal, spatial, or spatial–temporal dependencies result in a potentially non-homogeneous effect (Cornell and Winterstein, 1988; Benito *et al*., 2004). These dependencies can manifest through cycles (periodic variations), short-term oscillations (seasonal variations), and/or purely random fluctuations **(**Daykin *et al*., 1993; Cunningham *et al*., 2012). These dependencies can be attributed to various factors, including the state of stress in rock in the instance of seismic event occurrence (Simpson and Richards, 1981; Gibowicz and Kijko, 1994; Sharma *et al*., 2013; Scholz, 2015) and, by inference, tsunamis (e.g. Ogata and Abe, 1991, Mora *et al*., 2000, Karakaisis *et al*., 2002, Talbi and Yamazaki, 2009, Sharma *et al*., 2013; Scholz, 2015). Rainfall, deforestation, and earthquakes influence the occurrence of landslides (Tatard *et al.*, 2010), whereas weather phenomena, such as El Niño Southern Oscillation (ENSO) affect meteorological and hydrological occurrences (Khaliq *et al*., 2006; Egüen *et al*., 2016). Therefore, an assumption that the stationary Poisson distribution can be used to describe the temporal occurrence of events is not necessarily true and neither is the assumption of independent, identically distributed event sizes.

These levels of uncertainty of potential non-homogeneous effects in the data are often referred to as statistical uncertainty or aleatory uncertainty, where the variability in the data is random and cannot necessarily be reduced by additional information. The inherent randomness in the data is the unpredictable physical properties of the data. Epistemic uncertainty refers to the lack of understanding of the process owing to insufficient data (McGuire and Toro, 1986). In an interesting article, Der Kiureghian and Ditlevsen (2009) discuss the difference between aleatory and epistemic uncertainties, how uncertainty manifests, and how to deal with it. Anderson *et al.* (2000) describe how the primary form of uncertainty in the data is epistemic in nature but is confused sometimes with aleatory uncertainty. Additionally, the authors state that the decision on whether epistemic or aleatory uncertainty is applicable depends on the context and the application of information in the 'modelling universe'. It is possible, therefore, for a certain uncertainty to migrate from aleatory to epistemic based on how the 'modelling universe' changes.

In the case of natural hazard assessment, uncertainties in the data can prove difficult to identify and/or remove. Often, they are dealt with by using complex modelling of dependence structures, as well as simulations. This situation is attributable to most natural hazard event datasets not only being incomplete or not long enough but also having certain elements of uncertainty related to location and size. A wide range of opinions on spatial and temporal fluctuations of, e.g. seismogenic processes and, by inference, tsunami occurrences are discussed i.e. by Ogata and Abe (1991), Mora *et al.* (2000), Karakaisis *et al.* (2002), Talbi and Yamazaki (2009), Parsons and Geist (2008), Tatard *et al.* (2010), Witt *et al*. (2010).

Cornell and Winterstein (1988) showed that a deviation of up to 30% could be observed in the Poisson process of an earthquake dataset before the assumption of a stationary Poisson process no longer holds and negatively affect the results. Rhoades (1996) indicated that if the dataset spanned a sufficient time period (Δt), the effects of short-term clustering in an earthquake dataset because of fore- and aftershocks would not have a noteworthy influence on the results.

This conclusion can be extended to other types of natural phenomena. Nevertheless, it is important that this type of uncertainty in the data is accounted for, if possible, to ensure that the perceived hazard is not under- or overestimated.

3.2.2.1 Mixture distributions

One way of providing for weak dependencies and non-homogeneity in the data (excluding trends), is to assume that the parameters of the distributions are random. In the non-stationary Poisson process, the constant mean of the Poisson process ($\lambda\Delta t$) is assumed a random variable and is incorporated using mixture or Bayesian distributions (e.g. Pisarenko *et al*., 1996; Pisarenko and Lyubushin, 1997; Ogata, 1999b; Rotondi and Varini, 2007; Vicini *et al.*, 2013). This method is used also in economics, finance, and insurance (McDonald and Butler, 1987; Daykin *et al*., 1993; Cunningham *et al*., 2012). Both mixture and Bayesian distributions assume that a parameter of a distribution, in this instance λ and β, is distributed randomly. Classic Bayesian theory requires prior information to be independent regarding the data (Von der Linden *et al*., 2014). Unfortunately, in natural hazard assessments, the Bayes' rule is frequently used incorrectly by incorporating *a priori* information that is dependent on the data (Fernandes *et al*., 2010; Yadav *et al*., 2013). With mixture distributions, the uncertainty in the data follows directly from the observations by assuming that the relevant parameters are stochastic variables and subjected to change (Daykin *et al*., 1993; Cunningham *et al*., 2012), imparting a semi-parametric feel to the modelling process. This technique is used also to model different latent structures in a dataset (Marin *et al*., 2005). The term Bayesian is used often to refer to a mixture or compound distributions for continuous variables in field-specific literature, e.g. Mortgat and Shah (1979), Campbell (1982, 1983), Parvez (2007), and Yazdani and Kowsari (2013). The mixing distribution can be viewed as *a priori* information and the mixture distribution as the predictive posterior distribution. However, whereas the resulting posterior distributions are the same (for the distributions defined in this thesis), the requirement of application of the Bayesian and mixture distributions are fundamentally different.

Alternative methods to consider model parameter uncertainty are described in Mortgat and Shah (1979) who used a combination of the Bernoulli and beta distributions, whereas Dong *et al.* (1984) and Stavrakakis and Tselentis (1987) used a combination of uniform and multinomial distributions. Bender and Perkins (1993) and Rhoades *et al.* (1994) provide reviews of various alternative ways to deal with various uncertainties. Furthermore, Shi and Bolt (1982) and Guttorop and Hopkins (1986) addressed the question of fluctuations in the hazard parameters by applying an approximate variance stabilising transformation to the power law parameter b, which is assumed to be a slowly varying random variable.

Mixture distributions assume that the occurrence distributions are defined for the random variable X by the original CDF $\mathcal{F}_A$ for the parameter θ_i (for $j = 1, \ldots x$), from the parameter space Θ, which is also a random variable described by the mixing distribution $\mathcal{F}_B$. The mixture distribution for Θ is denoted as $\mathcal{F}_A \bigwedge_\Theta \mathcal{F}_B$. In an instance of the mixing distribution $\mathcal{F}_B$ having a discrete origin with probabilities p_j, the mixture distribution is also discrete, known as a countable mixture. Assuming that the CDF of $\mathcal{F}_A$ takes the form $F_X(x;\theta)$, dependent on θ, the CDF of the countable mixture Y, in the discrete instance, takes the form (Campbell, 1982, 1983; Johnson *et al*., 2005)

$$F_Y(y) = \sum_\theta F_X(x;\theta) p_\theta(\theta),$$

with the moments of Y defined as

$$E[Y^n] = m_Y^{(n)} = \sum_\theta m_X^{(n)}(\theta) p_\theta(\theta).$$

The continuous instance or the infinite mixture distribution is defined in terms of the conditional CDF $F_X(x;\theta)$, which is dependent on the random variable θ, with a PDF form $h_\theta(\theta)$. The resultant mixture takes the form

$$F_Y(y) = \int_{\theta} F_X(x;\theta) h_\theta(\theta)\, d\theta \quad \theta \in \Theta, \tag{3.20}$$

with the moments of Y defined as

$$E[Y^n] = m_Y^{(n)} = \int_{-\infty}^{\infty} m_X^{(n)}(\theta) h_\theta(\theta) d\theta, \tag{3.21}$$

such that the mean and variance reduce to

$$m_Y = \int_{-\infty}^{\infty} m_X(\theta) h_\theta(\theta) d\theta, \tag{3.22a}$$

$$\sigma_Y^2 = m_Y^{(2)} - \left(m_Y^{(1)}\right)^2 = \int_{-\infty}^{\infty} m_X^{(2)}(\theta) h_\theta(\theta) d\theta - \left(\int_{-\infty}^{\infty} m_X^{(1)}(\theta) h_\theta(\theta) d\theta \right)^2. \tag{3.22b}$$

The distribution of Y can be described as a weighted average of the CDF of X for each realisation of θ.

This thesis accounts for the randomness of the recurrence parameters with the help of the two-parameter gamma mixing density function. This function is a convenient choice in many research fields, as it is flexible enough to take on different forms through the shape (p) and scale (q) parameters of the gamma distribution. Examples of probabilistic models using gamma distribution are found in engineering (Hamada *et al*., 2008), seismology (Benjamin, 1968; Campbell, 1982, 1983; Benjamin and Cornell, 2014), transport (Lord and Park, 2008), and the insurance and risk industry (Klugman *et al*., 2008).

The choice of the gamma distribution is not random. Similar to the Poisson process, it is infinitely divisable, a resultant condition from iid random variables. The gamma distribution is also the maximum entropy distribution for the Pearson Type III distributions (Pearson 1894, 1895). This means that the distribution commits only to the known information, and limits the prior information required. In addition, it adheres to the characteristic of systems governed by physical laws that always tend to move toward maximum entropy. As Jaynes (1957) said, the resulting maximum entropy distribution, "...*is the*

least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information". Compared with principal distributions, mixture distributions have the additional benefit that the unconditional variance of the mixture distribution is larger because of the assumed uncertainty in the mixing distribution (Klugman *et al*., 2008).

Assuming a gamma mixing distribution $\sim GAM(p,q)$ that equals

$$h_Z(z) = \frac{z^{(q-1)}p^q}{\Gamma(\mathrm{q})}\exp(-pz) \quad z,p,q>0, \tag{3.23a}$$

with the gamma function defined as

$$\Gamma(q) = \int_0^\infty z^{(q-1)}\exp(-z)\,dz \quad q>0, \tag{3.23b}$$

such that the mean, variance, and coefficient of variance equal

$$\begin{array}{cc} Mean: & \mu_z = q/p \\ Variance: & \sigma_z^2 = q/p^2. \\ Coefficient\ of\ variance & v_z = \sigma_z/\mu_z \end{array} \tag{3.23cc}$$

The gamma function is a logical choice, as it is a conjugate prior for both the Poisson and exponential distributions. In terms of Bayesian statistics, $h_Z(z)$ can be seen as *a priori* distribution and the mixture distribution $F_Y(y)$ from Eq. 3.20, as the predictive posterior distribution with hyper-parameters p and q.

3.2.2.2 Parameter uncertainty in occurrence distribution

Assumption 2 is replaced by Assumption 12 by incorporating the two-parameter gamma distribution to account for uncertainty in the temporal occurrence.

Assumption 12. *The occurrence of observed events is a non-homogeneous Poisson process.*

The probability to observe a number of n events in the specific area under investigation, for the time interval Δt, for a non-homogeneous Poisson distribution equals

$$P_N\left(n|\bar{\lambda}, \Delta t, v_\lambda\right) = \int_0^\infty P_N(n; \Delta t) f_\Lambda(\lambda) d\lambda,$$

resulting in the Poisson-gamma distribution, with gamma hyper-parameters p_λ and q_λ. The probability is then defined as

$$P_N\left(n|\bar{\lambda}, \Delta t, v_\lambda\right) = \frac{\Gamma(n + q_\lambda)}{n!\,\Gamma(q_\lambda)} \left(\frac{p_\lambda}{\Delta t + p_\lambda}\right)^{q_\lambda} \left(\frac{\Delta t}{\Delta t + p_\lambda}\right)^{n}. \tag{3.24}$$

where v_λ is the coefficient of variance for $\bar{\lambda}$. The derivation is available in the Appendix.

Equation 3.24 is equivalent to the negative binomial distribution. Interestingly, the resultant Bayesian formulation with the gamma distribution as *a priori* information also results in the negative binomial distribution (Karlis and Xekalaki, 2005). The mean, variance, and coefficient of variance for λ are defined in terms of the parameters p_λ and q_λ, and are given in Eq. 3.25.

$$E(\lambda) = \bar{\lambda} = \frac{q_\lambda}{p_\lambda}, \tag{3.25a}$$

$$var(\lambda) = \sigma_\lambda^2 = \frac{q_\lambda}{p_\lambda^2}, \tag{3.25b}$$

$$v_\lambda = \sigma_\lambda / \mu_\lambda. \tag{3.25c}$$

Alternatively, $P_N\left(n|\bar{\lambda}, \Delta t, v_\lambda\right)$ can be rewritten as

$$P_N\left(n|\bar{\lambda},\Delta t,v_\lambda\right) \equiv P_N\left(n|\bar{\lambda},\Delta t,q_\lambda\right) = \frac{\Gamma(n+q_\lambda)}{n!\,\Gamma(q_\lambda)}\left(\frac{q_\lambda}{\bar{\lambda}\Delta t+q_\lambda}\right)^{q_\lambda}\left(\frac{\Delta t\bar{\lambda}}{\Delta t\bar{\lambda}+q_\lambda}\right)^{n}. \quad (3.26)$$

The derivation is available in the Appendix**.**

3.2.2.3 Parameter uncertainty in event size distribution

Similarly, the two-parameter gamma distribution is used to account for uncertainty in the parameters of the event size distribution, with Assumption 13 replacing Assumption 3. This is applied to the conditional CDFs and PDFs of the shifted-truncated event size distributions (Eqs 3.7a and b), and the extreme event size distributions (Eqs 3.17a and b). The same process can be followed for the unaltered and shifted distributions. The exponential-gamma function is also known as the Lomax or Pareto II function (Arnold, 2014).

For the rest of this thesis, the conditional CDFs and PDFs will be referred to only as CDF and PDF.

Assumption 13. *The frequency–event-size relation of random variable X is defined in terms of N (random) possible events* $X = (X_1, X_2, \ldots X_N)$ *that account for the possibility that the N events are random, weakly dependent, and identically distributed.*

The normalised, shifted-truncated exponential event size distribution of Eq. 3.7a is altered to allow for a varying β parameter. The CDF and PDF of the shifted-truncated exponential-gamma event size distribution equal

$$F_X\left(x|x_{min},x_{max},\bar{\beta},v_\beta\right) = C_\beta\left[1-\left(\frac{q_\beta}{\bar{\beta}(x-x_{min})+q_\beta}\right)^{q_\beta}\right] \quad x_{min}\le x\le x_{max}, \quad (3.27a)$$

$$f_X\left(x|x_{min},x_{max},\bar{\beta},v_\beta\right)=C_\beta\bar{\beta}\left[1+\frac{\bar{\beta}(x-x_{min})}{q_\beta}\right]^{-(q_\beta+1)}\quad x_{min}\le x\le x_{max},\tag{3.27b}$$

and

$$C_\beta=\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_{max}-x_{min})+q_\beta}\right)^{q_\beta}\right]^{-1}.\tag{3.27c}$$

The parameters p_β and q_β represent the parameters of the gamma function, such that $p_\beta=\bar{\beta}/\sigma_\beta^2$ and $q_\beta\equiv v_\beta^{-2}=\bar{\beta}^2/\sigma_\beta^2$, and $\bar{\beta}$ and σ_β^2 denote the mean and variance of β, derived in a similar manner as Eqs 3.25a and b. The derivations of Eqs 3.27a to c are available in the Appendix.

A similar extension is conducted for the extreme shifted-truncated distributions by substituting Eqs 3.27 a and b into the generalised extreme value distribution formulas defined in Eqs 3.14 a and b. The CDF and PDF of the shifted-truncated extreme exponential-gamma distribution equal

$$F_X^{MAX}\left(x_0\middle|\Delta t_P,\bar{\beta},\bar{\lambda}_P,x_{min},x_{max},\boldsymbol{v_P}\right)=$$
$$\exp\left(-\bar{\lambda}_P\Delta t_P\left[1-C_\beta\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_0-x_{min})+q_\beta}\right)^{q_\beta}\right]\right]\right)\quad x_{min}\le x_0\le x_{max},\tag{3.28a}$$

$$f_X^{MAX}\left(x_0\middle|\Delta t_P,\bar{\beta},\bar{\lambda}_P,x_{min},x_{max},\boldsymbol{v_P}\right)=$$
$$\bar{\lambda}_P\Delta t_P\bar{\beta}C_\beta\left[1+\frac{\bar{\beta}(x_0-x_{min})}{q_\beta}\right]^{-(q_\beta+1)}$$
$$\times\left\{\exp\left[-\bar{\lambda}_P\Delta t_P\left(1-C_\beta\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_0-x_{min})+q_\beta}\right)^{q_\beta}\right]\right)\right]\right\}\quad x_{min}\le x_0\le x_{max},\tag{3.28b}$$

where $\boldsymbol{v}=(v_\lambda,v_\beta)$ is the coefficient of variance of λ and β, and C_β, as defined in Eq. 3.27c.

3.2.3 Uncertainty in determination of event size

Uncertainty in event size measurement refers to errors that occur because of uncalibrated instruments, or when making use of qualitative assessments of the effect of an event on the environment. One way of dealing with the uncertainty of the event size is to assume that the observed size of an event (apparent event size) x consists of the true size y and an error ε, such that $x = y + \varepsilon$. The uncertainty in size measurement in this thesis is addressed by using hard-bound and soft-bound models utilising the convolution theorem.

Assumption 14. *The random variable X refers to the apparent size of events, such that* $X = Y \pm E$*, where* Y *is the random true event size and* E *is the random, stochastically independent error. The CDF for X equals*

$$F_X(x) = P(Y + \mathrm{E} \leq x),$$

and the PDF of X is

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

Following that the two random variables Y and E are independent (Tinti and Mulargia, 1985), with the respective PDFs of $f_Y(y)$ and $g_{\mathrm{E}}(\varepsilon)$. The CDF of X is evaluated using the convolution theorem ($(f * g)(x)$) and the Leibniz differentiation rule of an integral as

$$F_X(x) = \int_{-\infty}^{\infty} f_Y(y) \left[\int_{-\infty}^{x-y} g_{\mathrm{E}}(\varepsilon) d\varepsilon \right] dy,$$

$$f_X(x) = (f * g)(x) = \int_{-\infty}^{\infty} f_Y(x - \varepsilon) g_E(\varepsilon) d\varepsilon = \int_{-\infty}^{\infty} f_Y(\varepsilon) g_E(x - \varepsilon) d\varepsilon.$$

3.2.3.1 Hard-bound model

Hard-bound models assume that the event size uncertainty is described as a uniform distribution $UNIF(-\delta, \delta)$. This is typically assumed in the instances of prehistoric data, where only a range for the event size can be provided. For the hard-bound models $\delta = \frac{1}{2}(x_U - x_L)$, where x_L for the lower bound and x_U for the upper bound of the apparent event size x.

The hard-bound PDF of $f_X(x)$ follows from the convolution and equals

$$f_X(x|\delta) = \frac{1}{2\delta}\int_{-\delta}^{+\delta} f_Y(x-\varepsilon)d\varepsilon.$$

For the interval $x \geq 0$, the support of the event size distribution, the equation updates to

$$f_X(x|\delta) = \frac{1}{2\delta}[F_Y(x+\delta) - F_Y(x-\delta)],$$

such that

$$f_X(x|\delta) = \frac{1}{2\delta}[F_Y(x_U) - F_Y(x_L)] \quad x_L \leq x \leq x_U. \tag{3.29}$$

Using the general formula of Eq. 3.29, the chosen CDF and PDF can be updated to make provision for event size uncertainty. The updated CDF and PDF, when applying the shifted event size distributions (e.g. Eqs 3.6 and 3.16), with or without making provision for parameter uncertainty, reduce to

$$F_X(x|x_{min}, \delta) = \frac{1}{2\delta}\int f_X(x|x_{min}, \delta)dx, \tag{3.30a}$$

$$f_X(x|x_{min},\delta) =$$

$$\frac{1}{2\delta}\begin{cases} F_Y(x+\delta|x_{min},\delta) - F_Y(x_{min}-\delta|x_{min},\delta) & x_{min}-\delta \le x \le x_U \\ F_Y(x+\delta|x_{min},\delta) - F_Y(x-\delta|x_{min},\delta) & x_L \le x \le x_U;\ x_L > x_{min} \end{cases}. \quad (3.30b)$$

In practice, the cut-off of the lower part of the event size range is not as sharp as depicted in these equations. The transition between $x < x_{min}$ and $x \ge x_{min}$ occurs more gradually, with $x < x_{min}$ (i.e. $x_{min} - \delta$) still being recorded but not representing a complete set of event sizes. The equations are, therefore, normalised to start from x_{min}, where the researcher is sure that all the events have been recorded. This is done by dividing the PDF for $x_{min} \le x$ by the corresponding normalising coefficient $1 - F_Y(x_{min}|x_{min},\delta)$.

$$F_X(x|x_{min},\delta) = \frac{1}{2\delta}\int f_X(x|x_{min},\delta)dx, \quad (3.31a)$$

$$f_X(x|x_{min},\delta) =$$

$$\frac{1}{2\delta}\begin{cases} 0 & x < x_{min} \\ \dfrac{F_Y(x+\delta|x_{min},\delta) - F_Y(x-\delta|x_{min},\delta)}{1 - F_Y(x_{min}|x_{min},\delta)} & x_L \le x \le x_U \text{ and } x \ge x_{min} \end{cases}. \quad (3.31b)$$

For the shifted-truncated distribution as defined in Eqs 3.7, 3.17, 3.27 or 3.28, Eq. 3.31 yields

$$F_X(x|x_{min},x_{max},\delta) = \frac{1}{2\delta}\int f_X(x|x_{min},x_{max},\delta)dx, \quad (3.32a)$$

$$f_X(x|x_{min},x_{max},\delta) =$$

$$\frac{1}{2\delta}\begin{cases} 0 & x < x_{min} \\ \left[\dfrac{F_Y(x+\delta|x_{min},x_{max}) - F_Y(x-\delta|x_{min},x_{max})}{1 - F_Y(x_{min}|x_{min},x_{max},\delta)}\right] & x_L \le x \le x_U \text{ and } x_{min} \le x < x_{max}-\delta \\ \left[\dfrac{1 - F_Y(x-\delta|x_{min},x_{max},\delta)}{1 - F_Y(x_{min}|x_{min},x_{max},\delta)}\right] & x_{max}-\delta \le x \le x_{max}+\delta \end{cases}. \quad (3.32b)$$

The relevant CDFs, with or without making provision for parameter uncertainty, are substituted into Eq. 3.32b to obtain the hard-bound event size distributions for the shifted-truncated distributions. The CDFs for Eq. 3.32b are obtained by solving the integral with numerical methods.

3.2.3.2 Soft-bound model

A second method to introduce the event size error E is to assume that the boundaries of the error follow a continuous distribution, with a support for real numbers $\mathbb{R}$. The Gaussian distribution is used typically, as seen in Tinti and Mulargia (1985), who first introduced the soft-bound model for earthquake magnitudes by defining the error as $\varepsilon \sim N(0, \sigma_x^2)$. Let

$$g_{\sigma_x}(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\varepsilon}{\sigma_x}\right)^2}.$$

The apparent new distribution for event size equals

$$F_X(x|x_{min}, x_{max}) = \int_{x_{min}}^{x_{max}} \left[f_Y(y) \int_{-\infty}^{x-y} g_{\sigma_x}(\varepsilon) d\varepsilon \right] dy. \tag{3.33}$$

The CDF of the soft-bound model for the shifted-truncated exponential-gamma distribution (Eq. 3.28a) was derived by Ms. S. Verryn (personal communication, 2011) and is provided in the Appendix and Kijko *et al.* (2016) as

$$F_X\left(x|x_{min}, x_{max}, \bar{\beta}, v_\beta\right) = \frac{C_\beta \bar{\beta} q_\beta^{q_\beta+1}}{2\sigma_x}\{A + B\} \tag{3.34}$$

$$A = \frac{(r_1 + r_2\alpha)^{-q_\beta}}{r_2 q_\beta}\Bigg|_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}},$$

$$B = \left(\frac{2}{\pi}\right)^{1/2} \sum_{h=0}^{\infty} \frac{(-1)^h}{2^h h!\,(2h+1)} \frac{1}{b^{2h+2}}$$

$$\times \sum_{j=0}^{2h+1} \frac{(2h+1)!\,(-r_1)^j (r_1 + r_2\alpha)^{2h+1-q_\beta-j}}{(2h+1-j)!\,j!\,(2h+1-q_\beta-j)} \Bigg|_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}},$$

in which $C_\beta = \left[1 - \left(\frac{q_\beta}{q_\beta + \bar{\beta}(x_{max} - x_{min})}\right)^{q_\beta}\right]^{-1}$ (Eq. 3.27c), $r_1 = q_\beta + \bar{\beta}(x - x_{min})$, $r_2 = \bar{\beta}\sigma_x$, $\alpha = q_\beta + \bar{\beta}(x - x_{min})$ and $b = -\bar{\beta}\sigma_X$. The derivations are available in the Appendix. The relevant PDF can be derived by differentiating Eq. 3.34 but, owing to its complexity, it is solved numerically.

3.2.4 Uncertainty of event occurrence

In some natural phenomena, the underlying, mechanical trigger of the event is sometimes questionable. This is particularly relevant to prehistoric and historical tsunami occurrences, as it can be difficult to distinguish between the eroded environmental effects of a tsunami and severe storm surges and floods. Earthquake datasets struggle to distinguish between triggered and induced events, and landslides can be caused by extreme rainfall or earthquakes. Some datasets have an additional variable that expresses the validity or quality of the event. For example, the Genesis and Impact of Tsunamis on European Coasts (GITEC) catalogue criteria (Tinti and Maramai 1996; Tinti *et al.*, 2001) is a tsunami validity index for each observation ranging from 0, the event is considered extremely improbable, to 4, the event is considered a definite tsunami with probability near 1.

The presence of uncertain and questionable event data with respect to the applied model can affect the likelihood function, leading to an erroneous assessment of the recurrence parameters. To consider the uncertainty associated with an event, the standard likelihood function is replaced with the weighted likelihood (WL) function defined in Section 3.2.1. The weighted maximum likelihood estimation (WMLE) is used widely in many branches of the sciences, particularly physics (e.g. Lyons, 1989). The

formalism of the WMLE extends Wald's general principle of the maximum likelihood parameter estimation (Hu and Zidek, 2002; Markatou *et al.,* 1998; Wang, 1992, 2001; Wang *et al.,* 2004). In addition, the procedure preserves at least the first-order asymptotic properties of the classic likelihood function, leading to estimators with the usual asymptotic behaviour (Markatou *et al.,* 1998; Wang, 2006).

The weighted likelihood function in the context of the above methodology is defined by introducing $w_j \equiv w(x_j)$ as the known weight of observation x_j. For example, the likelihood function $L_{\beta i}(\bar{\beta}|\boldsymbol{x}_i)$, in the instance of all the uncertainty types being accounted for, equals

$$L_{\beta i}(\bar{\lambda}, \bar{\beta}|\boldsymbol{\mathcal{I}_I}) \equiv \prod_{j=1}^{n_i} f_X^{\mathrm{i}}(x_j|\bar{\beta}, \boldsymbol{\mathcal{I}_I})^{w_j}. \tag{3.35}$$

The weights range from [0, 1], where 1 is equivalent to 100% assurance of the validity of an event. The effect of questionable events in the assumed model is reduced. The background information matrix for Eq. 3.35 is defined as $\boldsymbol{\mathcal{I}}_I = (\boldsymbol{x}_i, \boldsymbol{v_i}, x_{min}^{(i)}, x_{max})$ with $\boldsymbol{v_i} = (v_{\lambda i}, v_{\beta i})$.

The validity index in event datasets provides an opportunity to calculate the distribution of the most probable number of events in the dataset. For instance, if the researcher would like to distinguish between the number of tsunamis and storm surges, higher weights can be assigned to events that have evidence that a tsunami occurred. The probable number of tsunami events would, therefore, be equal or less than the total number of events in the dataset.

Following this description, and under the assumption that $w_j < 1$, the probability that n out of N events in the dataset occurred or equivalently the most probable number of events, takes the form of a binomial sum (Benjamin and Cornell, 2014)

$$P(\ n \text{ out of } N \text{ events in dataset occurred})$$

$$= \left[\prod_{k_1=1}^{n} w_{k_1} \cdot \prod_{k_2=n+1}^{N} (1 - w_{k_2}) \right]$$

$$+ \left[(1 - w_{n+1}) \prod_{k_1=2}^{n+1} (w_{k_1}) \cdot \prod_{k_2=n+2}^{N} (1 - w_{k_2}) \right] \tag{3.36}$$

$$+ \cdots \left[\prod_{k_1=(N-n)+1}^{N} w_{k_1} \cdot \prod_{k_2=1}^{(N-n)} (1 - w_{k_2}) \right],$$

where $j = 1, \ldots, N$. The probability of all N events in the dataset having a validity index $w_j = 1$ equals

$$P(N \text{ events in dataset}) = \prod_{j=1}^{N} w_j, \tag{3.37}$$

for N independent events in the dataset. The average and standard deviation of the most probable number of events are

$$E(n \text{ out of } N \text{ eventsoccurred }) = \sum_{j=1}^{N} w_j, \tag{3.38a}$$

$$SD(n \text{ out of } N \text{ possible events occurred}) = \sqrt{\sum_{j=1}^{N} w_j (1 - w_j)}. \tag{3.38b}$$

The model parameters can be estimated in several ways. If the likelihood function of the data is known, the obvious choice is the WMLE. If any additional independent information about the model parameters is available, BI can replace WMLE.

3.3 Recurrence Parameter Estimation

Several techniques can be applied to assess the model recurrence parameters λ and β. These include the least squares method (LS), method of moments (MM), maximum likelihood estimation (MLE), and Bayesian inference (BI). The LS method is probably the least applicable, as application of this procedure is valid only when the model error follows a Gaussian distribution. Furthermore, the LS procedure cannot deal adequately with sudden and extreme shifts to the 'ecosystem' of the phenomena under investigation (Holmes *et al.*, 2008).

The estimation process in this thesis is dictated by the use of likelihood functions to incorporate various types of data, and both MLE and BI are discussed. The different variations of the event size distributions can be applied to the likelihood functions. For illustration purposes, the shifted-truncated exponential-gamma distributions are used.

The likelihood functions for the s instrumental sub-datasets ($i = 1, \ldots s$), as applied to Poisson and shifted-truncated mixture distributions (Eqs 3.26 and 3.27a and b), with unknown model parameters $\bar{\lambda}$ and $\bar{\beta}$, follows the definition in Eq. 3.19b

$$L_{\lambda i}\left(\bar{\lambda}\middle|\boldsymbol{\mathcal{I}}_{\lambda i}\right) = const_i \left(\frac{1}{\bar{\lambda}_i \Delta t_i + q_\lambda}\right)^{q_\lambda} \left(\frac{\bar{\lambda}_i \Delta t_i}{\bar{\lambda}_i \Delta t_i + q_\lambda}\right)^{n_i}, \tag{3.39a}$$

with

$$const_i = \bar{\lambda}(x_{min}) \left[1 - F_X\left(x|x_{min}^{(i)}, x_{max}, \bar{\beta}, v_\beta\right)\right] \frac{q_\lambda^{q_\lambda} \Gamma(n_i + q_\lambda)}{n_i! \, \Gamma(q_\lambda)},$$

and

$$L_{\beta i}\left(\bar{\beta}\middle|\boldsymbol{\mathcal{I}}_{\beta i}\right) = \left[C_\beta \bar{\beta}\right]^{n_i} \prod_{j=1}^{n_i} \left[1 + \frac{\bar{\beta}}{q_\beta}\left(x_{ij} - x_{min}^{(i)}\right)\right]^{-(q_\beta + 1)}, \tag{3.39b}$$

with, $\boldsymbol{\mathcal{J}}_{\lambda i} = (n_i, \Delta t_i, v_\lambda,)$, $\boldsymbol{\mathcal{J}}_{\beta i} = (\boldsymbol{x}_i, x_{min}^{(i)}, x_{max}, v_\beta,)$, $v_\lambda = \sigma_\lambda/\mu_\lambda$, $v_\beta = \sigma_\beta/\mu_\beta$, and C_β is as defined in Eq. 3.27c. The derivations are available in the Appendix.

The likelihood function for the shifted-truncated extreme prehistoric data taking into account parameter uncertainty is defined by substituting Eq. 3.28b into Eq. 3.19a, which equals

$$L_P(\bar{\lambda}, \bar{\beta}|\boldsymbol{\mathcal{J}}_P) = \prod_{j=1}^{n_P} \left[\begin{array}{c} \bar{\lambda}_P \Delta t_P \bar{\beta} C_\beta \left[1 + \dfrac{\bar{\beta}(x_0 - x_{min})}{q_\beta}\right]^{-(q_\beta+1)} \\ \times \left\{ \exp\left[-\bar{\lambda}_P \Delta t_P \left(1 - C_\beta \left[1 - \left(\dfrac{q_\beta}{\bar{\beta}(x_0 - x_{min}) + q_\beta} \right)^{q_\beta} \right] \right) \right] \right\} \end{array} \right], \qquad (3.40)$$

with the $\boldsymbol{\mathcal{J}}_P = (\boldsymbol{x}_P, \boldsymbol{\Delta t}_P, x_{min}, x_{max}, \boldsymbol{v}_P)$, and $\boldsymbol{v_P} = (v_\lambda, v_\beta)$. The likelihood function for historical data is derived in a similar manner.

3.3.1 Maximum likelihood estimation

The maximum of the likelihood function is obtained by solving the system of two equations $\frac{\partial \ell}{\partial \bar{\lambda}} = 0$ and $\frac{\partial \ell}{\partial \bar{\beta}} = 0$, where $\ell = ln[L(\boldsymbol{\theta}|\boldsymbol{\mathcal{J}})]$ with $\boldsymbol{\theta} = \left(\bar{\lambda}, \bar{\beta}\right)^T$. For large samples sizes, the MLE procedure has a number of desirable properties. Not only does it provide numerically precise estimates but also it facilitates easy model comparisons with the goodness-of-fit tests, such as the likelihood ratio test or Akaike information criterion (AIC). In addition, it has asymptotic properties under certain smoothness constraints, which include the estimator being asymptotically consistent and independent for the existence of a unique maximum for the likelihood function. Furthermore, the estimator will be asymptotically normally distributed for large sample sizes, with variances determined by the inverse of the Fisher information matrix $\boldsymbol{I}$. In other words, $\widehat{\boldsymbol{\theta}}_{MLE} \sim N(\boldsymbol{\theta}, \left[\boldsymbol{I}\left(\hat{\theta}_{MLE}\right)\right]^{-1}]$. The other asymptotic properties of the MLE estimator include efficiency, thereby achieving the Cramér–Rao lower bound as the sample size of the dataset approaches infinity, and invariance, it being a minimax estimator (e.g. Davison, 2003; Chave, 2017).

3.3.1.1 Instrumental data with a single level of completeness

Following the shifted event size distribution model for a single complete dataset, under the assumptions of iid and stationary datasets, Aki (1965) and Utsu (1965) derived estimates for earthquake recurrence parameters using both MLE and MM. Surprisingly, both estimators for β are defined by exactly the same expression of

$$\hat{\beta} = \frac{1}{\bar{x} - x_{min}}, \tag{3.41}$$

where $\bar{x} = \sum x_j / n$ and x_j is the event-sizes in the dataset. Following the central limit theorem (e.g. Eadie *et al.*, 1971), if n is sufficiently large, the parameter $\hat{\beta}$ in Eq. 3.41 asymptotically approaches the Gaussian distribution, i.e. $\hat{\beta} \sim N(\bar{\beta}, \sigma_{\beta}^2)$, with an estimated standard deviation equal to

$$\hat{\sigma}_{\beta} = -\left(\frac{\partial^2 \ln L}{\partial \beta^2}\right)^{-\frac{1}{2}} = \hat{\beta}/\sqrt{n}. \tag{3.42a}$$

A slightly different, but more-accurate estimate of the standard deviation is (Zhang and Song, 1981)

$$\hat{\sigma}_{\beta} = \frac{\hat{\beta} n}{\left[(n-1)(n-2)^{\frac{1}{2}}\right]}, \tag{3.42b}$$

which, for large n is equivalent to the MLE variance $\hat{\beta}/\sqrt{n}$. The rate of occurrence in the time period Δt equals

$$\hat{\lambda} = \frac{n}{\Delta t}. \tag{3.42c}$$

The equations above reflect the simplest case when the uncertainty in event-size is ignored.

The estimators for the long-term rate of occurrence λ, for $x_j \geq x_{min}$, for the case where the uncertainty in event size is not ignored, is defined by Rhoades (1996). If Assumption 2 regarding the Poisson process holds, the number of events in the dataset exceeding a given event size is

$$\hat{\lambda}(x) = \frac{E[\#X > x]}{\Delta t}, \tag{3.43a}$$

where $E[\#X > x]$ is the expected value of the number of successes in n binomial trials, such that

$$\hat{\lambda}(x) = \frac{\sum_{j=1}^{n}\left[1 - F_X(x_j)\right]}{\Delta t}, \tag{3.43b}$$

with a variance of

$$var\left(\hat{\lambda}(x)\right) = \frac{1}{\Delta t}\sum_{j=1}^{n}\left[1 - F_X(x_j)^2\right]. \tag{3.43c}$$

The same equations can be applied in instances where $F_X(x_j)$ considers the uncertainty of both parameters and event sizes.

3.3.1.2 Instrumental data with varying levels of completeness

Maximisation of the instrumental part of the likelihood function Eq. 3.18 for an instance where no parameter or event size uncertainty is assumed provides the generalised Aki–Utsu $\hat{\beta}$-parameter estimator (Kijko and Smit, 2012)

$$\hat{\bar{\beta}} = \left(\frac{r_1}{\hat{\beta}_1} + \frac{r_2}{\hat{\beta}_2} + \cdots + \frac{r_s}{\hat{\beta}_s}\right)^{-1}, \tag{3.44}$$

where $i = 1, \ldots, s$, $r_i = n_i/n$; $n = \sum_i^s n_i$ for $\forall\, x_{ij} \geq x_{min}^{(i)}$. In Eq. 3.44, the individual parameters $\hat{\beta}_\iota$ are the Aki–Utsu estimators calculated for each individual sub-dataset using the Aki–Utsu estimate of Eq. 3.41. Equation 3.44 is generic and can be applied to various functional forms of the likelihood

functions, under condition that the likelihood function can be expressed as a multiplication of $L_i(\lambda)L_i(\beta)$.

The functional form, where provision is made for incompleteness, parameter uncertainty, uncertainty in the event size determination, and uncertainty in event occurrence yields a non-linear likelihood function. The maximisation of such likelihood functions can be obtained only by the application of numerical methods and iterative procedures.

3.3.1.3 Only extreme data

For instances where either only extreme prehistoric or historical information is available, and no additional uncertainty is accounted for, the respective likelihood function is simplified considerably. The solution for the maximum likelihood estimates for extreme data is discussed in Kijko and Dessokey, (1987) and Kijko and Sellevoll (1989), and is done by maximising and solving the system of likelihood equations (e.g. Eq. 3.40). The maximum likelihood parameter estimates for the stationary shifted-truncated distribution (Eq. 3.11b) are defined (Kijko and Dessokey, 1987) as

$$\begin{cases} \dfrac{1}{\hat{\lambda}} = \dfrac{\bar{x}_P - \overline{\Delta t_P \boldsymbol{a}}}{A_2 - A_1} \\ \dfrac{1}{\hat{\beta}} = \dfrac{\overline{\Delta t_P x_P \boldsymbol{a}} - \overline{\Delta t}_P A_2 x_{max}}{\overline{\Delta t_P \boldsymbol{a}} - \overline{\Delta t}_P A_2} \end{cases}, \tag{3.45}$$

with

$$\begin{aligned} \bar{x}_P &= \sum_{j=1}^{n_P} x_{0j} \Big/ n_P, \\ \overline{\Delta t}_P &= \sum_{j=1}^{n_P} \Delta t_{Pj} \Big/ n_P, \end{aligned} \tag{3.46}$$

$$\overline{\Delta t_P \boldsymbol{a}} = \sum_{j=1}^{n_P} t_{Pj} e^{-\beta x_{0j}} \Big/ n_P,$$

$$\overline{\Delta t_P x_P \boldsymbol{a}} = \sum_{j=1}^{n_P} t_{Pj} x_{0j} e^{-\beta x_{0j}} \Big/ n_P,$$

$$A_1 = e^{-\beta x_{min}},$$

$$A_2 = e^{-\beta x_{max}},$$

where $\boldsymbol{a}$ is a column vector, with elements $a(x_{0j}) = e^{-\beta x_{0j}}$ and $j = 1, \dots, n_P$.

In the instance where $x_{max} \to +\infty$ and Δt_P is a constant, parameter estimates reduce to the maximum likelihood estimates of the first Gumbel distribution (Kimball, 1946). The inclusion of the different types of uncertainties reduces the model to a non-linear equation, the solution of which requires the application of iterative techniques.

3.3.2 Bayesian inference

The accuracy of the maximum likelihood estimates obtained by the maximisation of likelihood (Eq. 3.18) depends upon the quality of the observed dataset. Small datasets often do not yield reliable estimates for natural hazards, as they provide only a limited view of the characteristics of the physical process. By including prior information in the estimation process, the hazard estimates are improved and stabilised.

Any information describing the natural process under review can be used as *a priori* information. A few typical examples of this is seen in seismology, which includes recurrence parameters for geologically similar areas or other geophysical information (Campbell, 1982, 1983; Dong *et al.*, 1984; Papoulia *et al*., 2001; Nomura *et al*., 2011; Yazdani and Kowsari, 2013). Silva *et al*. (2017 and references therein) provide ideas for what can and is being used as prior information in extreme frequency flood analyses.

Patskoski and Sankarasubramanian (2018 and references therein) discuss various potential prior information that can be used in time-series-based hydrological studies, such as tree rings, observed annual maximum events, and sea surface temperature. These include, among others, the probable maximum flood discharge (Fernandes *et al.*, 2010), expert judgement (Viglione *et al.*, 2013), and climate covariates (Sun *et al*., 2014). Other geology-related hazards, such as tsunamis and landslides, can include prior information, such as seismotectonics (Geist and Uri, 2012), rainfall thresholds (Berti *et al.*, 2012), and expert opinions (Aleotti and Chowdhury, 1999; Yazdani and Kowsari, 2013). Uninformative priors are popular also among different types of hazards (Lyubushin and Parvez, 2010; Yadav *et al*., 2013; Cooley *et al*., 2007).

Following the Bayesian rule, the posterior distribution $z(\boldsymbol{\psi}|\boldsymbol{\mathcal{I}})$ for the likelihood function of L_{Total} is constructed (Eq. 3.47), with *a priori* probability defined as $\pi(\boldsymbol{\psi})$, in which $\boldsymbol{\psi} = \left(\bar{\lambda}, \bar{\beta}, x_{max}\right)^T$, the parameters $\bar{\lambda}$ and $\bar{\beta}$ as defined in Eqs 3.39a, b and 3.40, and $\boldsymbol{\mathcal{I}}$ is the background information that denotes all the assumptions related to the particular investigation. In seismology, BI is performed typically directly on the shifted-truncated distribution, with the gamma distribution used often as a prior to account for parameter uncertainty. This distribution is a conjugate prior to both the Poisson and the exponential distributions, resulting in closed-from expressions. In this thesis, the parameter uncertainty is accounted for explicitly by using continuous mixture (compound) distributions. Additional information, independent of the dataset, is incorporated using prior functions and BI. These priors could be taken as uninformative, with $\pi(\boldsymbol{\psi}) = const$, when little or no information is available. Alternatively, following the law of large numbers and the central limit theorem, the choice of prior $\pi_\beta(\bar{\beta})$ for $\bar{\beta}$ is the Gaussian distribution, with $\bar{\beta} \sim N\left(\mu_\beta, \sigma_{\bar{\beta}}^2\right)$ (Aki, 1965; Shi and Bolt, 1982; Kijko and Graham, 1999). The mean rate of occurrence is a parameter that is specific to the region and not always known. An uninformative prior $\pi_\lambda(\bar{\lambda})$ for $\bar{\lambda}$, in the form of a uniform distribution $UNIF[\lambda_A, \lambda_B]$ (Dong *et al.*, 1984), is assigned. Similarly, an uninformative prior is assumed for the estimation of the maximum possible event size for the range $UNIF[x_{max}^A, x_{max}^B]$. In Eq. 3.47 it is assumed that the prior information for $\bar{\beta}$ and $\bar{\lambda}$ is independent. Alternatively, a dependent prior $\pi(\boldsymbol{\psi})$ can be assumed, as in Pisarenko *et*

al. (1996) that implemented uniform priors for$\bar{\lambda}$, $\bar{b}$, and x_{max} that are constant on a parallelepiped (also see Pisarenko and Lyubushin, 1997; Lyubushin *et al.*, 2002; Lyubushin and Parvez, 2010). Assuming that the prior information for the three recurrence parameters is independent, the joint prior distribution $\pi(\boldsymbol{\psi})$ is defined as $\left[\pi_{\lambda}(\bar{\lambda})\pi_{\beta}(\bar{\beta})\pi_{x_{max}}(x_{max})\right]$. The posterior distribution $z(\boldsymbol{\psi}|\boldsymbol{\mathcal{I}})$ and its estimated mean $\hat{\mu}(\boldsymbol{\psi}|\mathcal{I})$ and variance $\hat{\sigma}^2(\boldsymbol{\psi}|\mathcal{I})$ are evaluated numerically, from the following formula

$$z(\boldsymbol{\psi}|\boldsymbol{\mathcal{I}}) = \frac{\left(\frac{1}{x^B_{max}-x^A_{max}}\right)\left(\frac{1}{\lambda_B-\lambda_A}\right)\exp\left(-\frac{(\bar{\beta}-\mu_\beta)^2}{\sigma^2_{\beta}}\right)L_{Total}(\boldsymbol{\psi}|\boldsymbol{\mathcal{I}})}{\int\int\left(\frac{1}{x^B_{max}-x^A_{max}}\right)\left(\frac{1}{\lambda_B-\lambda_A}\right)\exp\left(-\frac{(\bar{\beta}-\mu_\beta)^2}{\sigma^2_{\beta}}\right)L_{Total}(\boldsymbol{\psi}|\boldsymbol{\mathcal{I}})d\bar{\beta}\,\bar{\lambda}}. \quad (3.47)$$

A more detailed description of Bayesian inference can be found in the references such as Davison (2003) and Von der Linden et al. (2014).

The likelihood functions that define the combination of prehistoric, historical, and instrumental data are non-linear and require a numerical estimation technique to estimate the parameters $\hat{\bar{\lambda}}$ and $\hat{\bar{\beta}}$. Various techniques for the maximisation of functions can be employed, as described in Chapter 10 of Press *et al*. (1986). Methods for maximisation in a multi-dimensional instance include the downhill simplex method by Nelder and Mead (1965), as well as the direction-set methods. For the purpose of this thesis, the downhill simplex method is applied by solving the vector parameter $\boldsymbol{\psi} = \left(\bar{\lambda}, \bar{\beta}, x_{max}\right)^T$ simultaneously. The estimation of parameter x_{max} is discussed in Section 3.4 below.

3.4 Maximum Possible Event Size

The third parameter required in this methodology is the area-characteristic maximum possible event size x_{max}. This parameter is defined as the upper limit or largest possible event size that could occur in a specified region compared with $x_{\max}^{(obs)}$, the maximum observed event size in a dataset. This upper

limit is introduced by truncating the event size distribution from the right. Currently, two schools of thought consider either a sharp or a soft truncation (Hamilton, 1967; Page, 1968; Cosentino *et al*., 1977; Stein and Hanks, 1998; Main and Burton, 1984, Kagan, 1991, 2002a, b). Sharp truncation of a distribution refers to the probability of $x_j > x_{max}$ being equal to zero ($P(X_j > x_{\max}) = 0$). Several sharp truncation methods are discussed in Kijko (2004), Kijko and Singh (2011), Beirlant *et al.* (2017), and Vermeulen and Kijko (2017). Soft truncation refers to the assumption that the prevailing distribution decays at a faster rate than the assumed power law after point x_{max} (Sornette and Sornette, 1999; Pisarenko and Sornette, 2003). The finiteness of event sizes in terms of the dissipative nature of physical dynamic systems is discussed by several authors (e.g. Kagan, 1991, 1997, 2002 a, b; Main, 1996; Sornette and Sornette, 1999; Vere-Jones *et al*., 2001; Kagan and Schoenberg, 2001; Pisarenko and Sornette, 2003; Bird and Kagan, 2004; Kagan, 2010; Pisarenko *et al*., 2008, 2014). The authors provide arguments for several soft truncations or potential decay, including the exponential and gamma distribution.

The evaluation methods for x_{max} can be classified into deterministic and probabilistic methods. Deterministic procedures focus on the empirical relationships between event sizes and the physical process of a specific natural hazard. In seismology, the physical processes that influence the size of an event include the location, size, and type of seismogenic region and the tectonic fault length (Wells and Coppersmith, 1994; Anderson *et al*., 1996), rate of seismic moment release (Papastamatiou, 1980; Anderson and Luco, 1983; WGCEP, 1995; Stein and Hanks, 1998; Field *et al*., 1999; McGarr, 1984), and rupture processes (Ward, 1997). For landslides, the size of an event is determined by local and regional geological, hydrological, soil and vegetation characteristics, and meteorological conditions (e.g. Guzzetti *et al*., 2008). As another example, factors such as the abundance of fuel, prevailing meteorological conditions, and fire suppression efforts determine the fire event size (e.g. Cumming, 2001). Clearly, a large component of uncertainty is introduced in the deterministic assessment of the upper limit x_{max}, as the nature of the underlying physical processes of natural hazards is not always easy to formalise to construct an adequate mathematical model.

The probabilistic estimation procedures focus on assessing x_{max} based solely on the statistical analysis of observed event history in the area of interest. These methods consist of the extrapolation of the event size relationship. The first procedures for the estimation of the end-points of the distributions were probably done by Robson and Whitlock (1964), Woodroofe (1972, 1974), Weiss and Wolfowitz (1973), and Hall (1982). Examples of its application in seismology can be seen in Dargahi-Noubary (1983), Kijko and Sellevoll (1989, 1992), Pisarenko (1991), and Pisarenko *et al*. (1996). The probabilistic methods to estimate x_{max} can be classified into parametric, non-parametric estimators, the fit of a CDF to the observed event sizes, as well as EVT methods. If the frequency–event-size model is known and relatively simple, parametric procedures are used. These procedures include a deterministic procedure (Wheeler, 2009), the Tate–Pisarenko (Tate, 1959; Pisarenko *et al*., 1996; Kendall and Stuart, 1967), and the Kijko–Sellevoll methods (Cooke, 1979, Kijko, 2004). If some model anomalies are observed, such as a bi-, multi-modal, or non-linear frequency–event-size models or characteristic events, non-parametric assessment methods can be applied. Examples of non-parametric procedures are a Gaussian kernel (Parzen, 1962; Kijko *et al*., 2001), order statistics (Cooke, 1979), the few largest events method, as described by Cooke (1980), and the Robson–Whitlock and Robson–Whitlock–Cooke procedures (Quenouille, 1956; Robson and Whitlock, 1964; Cooke 1979). If the functional form of the CDF $F_X(x)$ is known, L_1- and L_2- regression analyses can be performed, particularly when large outliers are observed or when the distribution of the residuals of the CDF follows a Gaussian distribution. Authors such as Nuttli (1981) and Fröhlich (1998) investigated extrapolation methods where the frequency–event-size relation (in these instances the frequency–magnitude Gutenberg–Richter relation) are truncated at a pre-specified annual probability of exceedance. Particular interesting methods in EVT that could be applied are the methods by Alves and Neves (2014) and Alves *et al*. (2017), linear sums of extreme values, as defined by Cooke (1980), and an EVT standard maximum likelihood estimator (Coles, 2001, De Haan and Feirreira, 2007). Beirlant *et al.* (2017) estimated upper limits for the truncated generalised Pareto distribution (GPD) and the truncated Pareto distribution. The EVT methods follow a similar underlying structure to the method proposed by Kijko (2004) namely, the maximum estimated event size is equal to the observed maximum event size plus something extra.

One of the regularity conditions needed for the existence of the asymptotic properties of the maximum likelihood function is that the parameter space must be compact and the likelihood function must be twice continuously differentiable (LeCam, 1970; Davison, 2003). As the maximum possible event size x_{max} is one of the main parameters in the proposed likelihood function for the shifted-truncated distributions (e.g. Eq. 3.47), the MLE estimate of x_{max} using $L(\boldsymbol{\psi})$ will result in the conditions of regularity being violated (Cheng and Traylor, 1995; LeCam, 1970; Eadie *et al*., 1971; Davison 2003). Intuitively, if $x_{max}^{(obs)}$ is used in the evaluation range of the integral, the likelihood function will reach its maximum at $x_{max}^{(obs)}$ and not at the required possible estimated maximum event size $\hat{x}_{max}$. Several generic procedures can provide solutions for x_{max}. These procedures differ with respect to the underlying assumptions about the statistical model and/or the information available about natural hazard events. For the purpose of this thesis, the Kijko–Sellevoll methodology defined by Kijko (2004) and Kijko and Singh (2011) is applied.

Assumption 15. *A generic equation for maximum event size, is of the form*

$$\hat{x}_{max} = x_{max}^{(obs)} + \Delta x, \tag{3.48}$$

with Δx defined as a positive correction factor.

Further, it is assumed that event sizes X_j $(j = 1, \dots N)$ are iid random values, such that $X_j \geq x_{min}$ (Assumption 4), and can be described by the PDF $f_X(x|x_{min}, x_{max})$ and CDF $F_X(x|x_{min}, x_{max})$. The approximate variance of $\hat{x}_{max}$ equals (Kijko and Graham, 1998; Kijko 2004)

$$var(\hat{x}_{max}) \cong \sigma_X^2 + \Delta x^2, \tag{3.49}$$

with $\sigma_X^2 = \sigma^2\left(x_{max}^{(obs)}\right)$ the variance the maximum observed event size. The confidence intervals can be derived by assuming that z_α is defined such that

$$\left[\hat{x}_{max} \pm z_{\alpha}\sigma_{\hat{x}_{max}}\right], \tag{3.50}$$

where $x_{max}^{(obs)} = MAX(X_j)$, the maximum observed event size in the dataset. The correction factor Δx is a function of the applied model and systematically decreases as the time span of the dataset increases (Kijko *et al*., 2017). Following Cooke (1979), the correction factor Δx is of the form

$$\Delta x = \int_{x_{min}}^{x_{max}} \begin{cases} 0 & x < x_{min} \\ [F_X(x|x_{min}, x_{max})]^n & x_{min} \le x \le x_{max}. \\ 1 & x > x_{max} \end{cases} \tag{3.51}$$

Therefore, the estimator $\hat{x}_{max}$ in Eq. 3.48 takes the form

$$\hat{x}_{max} = x_{max}^{(obs)} + \int_{x_{min}}^{\hat{x}_{max}} [F_X(x|x_{min}, x_{max})]^n dx \quad x_{min} \le x \le x_{max}, \tag{3.52}$$

with the term $\hat{x}_{max}$ appearing on both sides of the equation. To estimate $\hat{x}_{max}$, an iterative procedure is applied by using $x_{max}^{(obs)}$ as a starting point. According to Cramér (1961), for large n ($n > 10$), $[F_X(x|x_{min}, x_{max})]^n \cong e^{-n[1-F_X(x|x_{min}, x_{max})]}$. This approximation provides a useful tool to obtain semi-analytical solutions for the $\hat{x}_{max}$ (Kijko, 2004).

Using Eq. 3.52, the upper limit $\hat{x}_{max}$ can be assessed with reasonable accuracy when a large amount of data is available. When the available dataset is too small or the area under investigation is not prone to the particular type of natural phenomenon, the assessment will be unreliable. The above procedure for assessing the upper limit of the event size distribution can be applied only when the event size distribution is known or can be approximated with good accuracy. The smaller the event set the larger will be the likelihood that $\hat{x}_{max}$ is underestimated (e.g. Chinnery, 1979; Bender, 1988). Therefore, improved accuracy of $\hat{x}_{max}$ estimates can be achieved by including prehistoric and historical information and any other independent information using the Bayesian formalism (e.g. Coppersmith, 1994; Kijko, 2012).

Other methods to increase the robustness of the maximum possible event size is to remove the bias in the estimation process of $\hat{x}_{max}$ (Lasocki and Urban, 2011) or to apply Bayesian statistics by introducing additional information sources (e.g. Cornell, 1994). The methodology described by Cornell (1994) introduces additional bias, which is addressed in Kijko (2012).

3.5 Hazard Estimates

Natural hazard estimates are derived in terms of probabilities of exceedance (PE) and return periods, known also as the recurrence interval R (e.g. Baker, 2013). Probabilities of exceedance represent the probability that a specific event size or larger events will occur within a specific time interval ($P(X_j \geq x_j|\Delta t)$). These probabilities are expressed usually in terms of an annual value PE_1 when investigating a period of one year, or the general PE_m when investigating a period of m years. Return periods are the inverse of the annual probability of exceedance $1/PE_1$, i.e. an annual probability of exceedance of 10% gives a return period of 10 years.

Hazard estimates should not be confused with prediction estimates. Prediction aims to identify a small and specific area and time window, where the likelihood of an event occurring is high. For this purpose, certain short-term triggers are monitored continuously. The aim of prediction is to serve as a form of early-warning system. Hazard estimates utilise past observations over decades to establish certain trends in the recurrence of events. The aim is to provide the engineering and insurance industries and disaster management centres with the necessary tools to develop infrastructure and financial instruments to alleviate the effects of natural phenomena when they do occur.

The probabilities of exceedance and return periods are usually defined per industry. In seismology, it is customary to express official seismic hazard maps in terms of a 10% probability of exceedance in 50 years. In insurance, the question often asked is what the 1–in–200 year loss event would be. This

value is required generally for solvency capital requirements and represents the amount of capital required to meet obligations to clients.

The 1–in–200 year event is calculated (e.g. Baker, 2013) to be equivalent to an annual probability of exceedance $PE_1 = 1/R$ of 0.5%, whereas the probability of exceedance in m years is calculated as the complement of PE_1, i.e. the probability of no exceedance PE_0 with

$$PE_0 = (1 - PE_1)^m. \tag{3.53}$$

Therefore, the probability of exceedance within m years is

$$PE_m = 1 - (1 - PE_1)^m, \tag{3.54}$$

so that the 10% probability of exceedance in 50 years is

$$0.1 = 1 - (1 - PE_1)^{50}$$

$$PE_1 = 0.002105$$

$$R = \frac{1}{PE_1} = 475\ years.$$

3.6 Model Comparison

In Section 3.1 a principal model was defined where the frequency–event-size distribution closely follows a power law. Section 3.2 describes four types of uncertainty that are often associated with the hazard modelling of natural phenomena and how the principal model should be adapted to account for these uncertainties. This could yield up to 24 potential different models that can be evaluated with MLE and BI, as shown in Table 3.1.

TABLE 3.1. List of model variations that can be evaluated using maximum likelihood estimation (MLE) and/or Bayesian inference (BI).

Model Acronym	Model Description	Number
SDS	Single event dataset containing only instrumental data, with one level of completeness x_{min}	(M1)
SDS_MAG	Single event dataset containing only instrumental data, with one level of completeness x_{min}. Event sizes are uncertain.	(M2)
SDS_MOD	Single event dataset containing only instrumental data, with one level of completeness x_{min}. The parameters of the applied distributions are uncertain.	(M3)
SDS_MAG_MOD	Single event dataset containing only instrumental data, with one level of completeness x_{min}. Event sizes and the parameters of the applied distributions are uncertain.	(M4)
SDS_OCC	Single event dataset containing only instrumental data, with one level of completeness x_{min}. The uncertainty of event occurrence is introduced.	(M5)
SDS_MAG_OCC	Single event dataset containing only instrumental data, with one level of completeness x_{min}. Event sizes are uncertain. The uncertainty of event occurrence is introduced.	(M6)
SDS_MOD_OCC	Single event dataset containing only instrumental data, with one level of completeness x_{min}. The parameters of the applied distributions are uncertain. The uncertainty of event occurrence is introduced.	(M7)
SDS_MAG_MOD_OCC	Single event dataset containing only instrumental data, with one level of completeness x_{min}. Event sizes and the parameters of the applied distributions are uncertain. The uncertainty of event occurrence is introduced.	(M8)
EXT	Event dataset containing only extreme events.	(M9)
EXT _MAG	Event dataset containing only extreme events. Event sizes are uncertain.	(M10)
EXT _MOD	Event dataset containing only extreme events. The parameters of the applied distributions are uncertain.	(M11)
EXT _MAG_MOD	Event dataset containing only extreme events. Event sizes and the parameters of the applied distributions are uncertain.	(M12)
EXT _OCC	Event dataset containing only extreme events. The uncertainty of event occurrence is introduced.	(M13)
EXT _MAG_OCC	Event dataset containing only extreme events. Event sizes are uncertain. The uncertainty of event occurrence is introduced.	(M14)

EXT _MOD_OCC	Event dataset containing only extreme events. The parameters of the applied distributions are uncertain. The uncertainty of event occurrence is introduced.	(M15)
EXT _MAG_MOD_OCC	Event dataset containing only extreme events. Event sizes and the parameters of the applied distributions are uncertain. The uncertainty of event occurrence is introduced.	(M16)
LOC	Event dataset containing the entire dataset with different levels of completeness $x_{min}^{(i)}$.	(M17)
LOC_MAG	Event dataset containing the entire dataset with different levels of completeness $x_{min}^{(i)}$. Event sizes are uncertain.	(M18)
LOC_MOD	Event dataset containing the entire dataset with different levels of completeness $x_{min}^{(i)}$. The parameters of the applied distributions are uncertain.	(M19)
LOC_MAG_MOD	Event dataset containing the entire dataset with different levels of completeness $x_{min}^{(i)}$. The event sizes and the parameters of the applied distributions are uncertain.	(M20)
LOC_OCC	Event dataset containing the entire dataset with different levels of completeness $x_{min}^{(i)}$. The uncertainty of event occurrence is introduced.	(M21)
LOC_MAG_OCC	Event dataset containing the entire dataset with different levels of completeness $x_{min}^{(i)}$. The event sizes are uncertain. The uncertainty of event occurrence is introduced.	(M22)
LOC_MOD_OCC	Event dataset containing the entire dataset with different levels of completeness $x_{min}^{(i)}$. The parameters of the applied distributions are uncertain. The uncertainty of event occurrence is introduced.	(M23)
LOC_MAG_MOD_OCC	Event dataset containing the entire dataset with different levels of completeness $x_{min}^{(i)}$. The event sizes and the parameters of the applied distributions are assumed uncertain. The uncertainty of event occurrence is introduced.	(M24)

In any modelling process, the question is how well the chosen distribution fits the data. A secondary question is how different distributions compare with each other. Several techniques exist to test how well a distribution follows the data, e.g. Kolmogorov–Smirnov, Anderson–Darling tests, and likelihood-ratios. Methods to discriminate between distributions include the Akaike information criterion (AIC), consistent Akaike information criteria (CAIC), correct AIC (AICc), Bayesian information criterion (BIC), Deviance Information Criteria (DIC), and the Hannan–Quinn criterion (e.g.

Davison, 2003). These criteria are dependent on the underlying log likelihood function. Another popular method to investigate fitted models is sensitivity analyses through variance- and entropy analyses (e.g. Liu *et al*., 2004; Schorlemmer *et al.*, 2007; Lam *et al.*, 2017).

Several authors, e.g. Clauset *et al.* (2009) applied the Kolmogorov–Smirnov test and likelihood ratios to power laws derived from empirical data. Deluca and Coral (2013) specifically investigated the use of the Kolmogorov–Smirnov test on non-truncated and truncated power law distributions. The authors also introduced an alternative way of using the test by generating multiple synthetic samples based on the estimates from the power law. The ensemble of the p-values was used subsequently to determine the appropriateness of the model.

In addition, graphical tools such as residual analyses are useful, as they can help to visually pinpoint where a model appears to depart from the observations or where one model appears to outperform another in terms of agreement with the data. Alternative methods to compare models are the Fisher Information Criteria, sensitivity analyses in terms of how recurrence parameters and hazard estimates react with the various modelling options, and D-optimality that evaluates and minimises the confidence ellipsoid around the estimated parameters using the covariance matrix (e.g. Fedorov, 2013).

All the above-mentioned methods of investigating goodness of fit require datasets that are complete, independent modelling parameters, and/or that the distributions are applied to the same underlying dataset. These methods can be used to compare models for a single, complete dataset with one x_{min} (SDS, models M1–M8) or to compare models containing only extreme data (EXT, models M9–M16). Once the data are divided into sub-datasets with different levels of completeness, or instrumental data are combined with prehistoric and historical data, the dataset is no longer single or complete (models M17–M24). In such instances, the recurrence model parameters $\hat{\hat{\lambda}}$ and $\hat{\hat{\beta}}$ are also dependent on each other, as seen in Kijko and Sellevoll (1989). For the above stated reasons it is difficult to compare all 24 possible models with standard goodness-of-fit techniques to determine which combination of input data and types of uncertainty would best describe the natural phenomena under investigation.

To investigate the contribution of observed information in a model, the log of the likelihood function is used to determine the Hessian information matrix (matrix of second derivatives) for $\boldsymbol{\theta} = \left(\bar{\lambda}, \bar{\beta}\right)^T$

$$J(\boldsymbol{\theta}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\boldsymbol{2}}}, \tag{3.55a}$$

if it exists. This provides details on the relative quantity of information that each sub-dataset contributes to the parameter estimate, under the assumption that the model is correct and that the true value of the parameter is $\boldsymbol{\theta}$. In the instance where the log likelihood function is the sum of m components, the information matrix equals

$$J(\boldsymbol{\theta}) = -\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\boldsymbol{2}}} = \sum_{j=1}^{m} -\frac{\partial^2 \log f(y_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\boldsymbol{2}}}. \tag{3.55b}$$

The Fisher information matrix is defined by taking the expectation of the information matrix $J(\boldsymbol{\theta})$

$$I(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\boldsymbol{2}}}\right], \tag{3.56}$$

with $I\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)$ representing the observed Fisher information matrix evaluated at the MLE estimate. For large sample sizes, the Fisher information matrix is both finite and positive definite. The inverse of the matrix subsequently serves as the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}$ in the asymptote**.** A variance-covariance matrix $\mathbf{D}(\hat{\bar{\lambda}}, \hat{\bar{\beta}})$ of the estimated $\hat{\bar{\lambda}}$ and $\hat{\bar{\beta}}$ is calculated according to the formula by Edwards (1972)

$$\mathbf{D}\left(\hat{\bar{\lambda}}, \hat{\bar{\beta}}\right) = -\begin{bmatrix} \dfrac{\partial^2 \ell}{\partial \bar{\lambda}^2} & \dfrac{\partial^2 \ell}{\partial \bar{\lambda} \partial \bar{\beta}} \\ \dfrac{\partial^2 \ell}{\partial \bar{\beta} \partial \bar{\lambda}} & \dfrac{\partial^2 \ell}{\partial \bar{\beta}^2} \end{bmatrix}^{-1}, \tag{3.57}$$

where derivatives are calculated at the point $\bar{\lambda} = \hat{\bar{\lambda}}$ and $\bar{\beta} = \hat{\bar{\beta}}$. The Fisher information matrix and subsequent covariance-matrix $\mathbf{D}(\hat{\bar{\lambda}}, \hat{\bar{\beta}})$ are used to calculate the percentage contribution of each individual dataset or prior information to the estimated parameters.

The choice of model comparison is, in this instance, a function of what is considered reasonable given the input data, choice of distributions, and types of uncertainty. Each aspect of the modelling process should be evaluated on the grounds that it 'makes sense' in the context of the physical process and available data.

3.7 Summary

Chapter 3 introduced the statistical methodology of the generic model through a series of steps and assumptions. Each step is described in the respective sections. Section 3.1 provides a description of the principal model on the assumption that events are independent and identically distributed and can consist of prehistoric, historic and instrumental records. The shifted- and shifted-truncated variations on the principal model are introduced to account for small events not observed and a potential upper limit in the distribution. In Section 3.2, the principal model and its variations are extended to account for the data characteristics of incomplete event datasets, uncertainty in event size determination, uncertainty of the applied event occurrence models, and uncertainty of event occurrence. Parameter estimation is discussed in Section 3.3, with Section 3.4 concentrating on the assessment of the area-characteristic maximum possible event size. Section 3.5 focusses on the calculation of the hazard estimates in terms of return periods and probabilities of exceedance. Model comparisons methods and their potential pitfalls as it relates to the methodology in Chapter 3 are briefly discussed in Section 3.6. Figure 1.4 in Section 1.5 provides a schematic illustration of the described methodology in this Chapter.

Chapter 4. Synthetic Data Example

The behaviour of the proposed methodology is evaluated using an earthquake dataset synthesised by using the Monte Carlo simulation to mimic typical prehistoric, historical, and instrumentally recorded earthquake events. The prehistoric data were generated using Eq. 3.28b substituted into Eq. 3.32b, as well as the historical data with the historical derivations of these equations. The individual instrumental datasets of the different levels of completeness were generated using Eq. 3.34 and visual cumulative method (Mulargia and Tinti, 1985). Each sub-dataset has different and decreasing LoC $x_{min}^{(i)}$ over time, and different event size errors. The apparent earthquake event sizes, in terms of moment magnitude M_W were generated for β = 2.302 (b = 1), and a mean rate of occurrence for a LoC equal to 4, i.e. $\bar{\lambda}(x_{min}$=4) = 10 using Eq. 3.26. The chosen values for the parameters are observed typically in areas exhibiting tectonic seismicity. The choice of $\bar{\lambda}(x_{min} = 4) = 10$ is random. Parameter variation of 25% was introduced, as well as uniform event size errors for prehistoric and historical data and Gaussian errors for instrumental data. Table 4.1 provides the input for the generation of the data. The evaluated time periods were chosen to simulate the way earthquake data are observed usually. Each sub-dataset, therefore, has different and decreasing LoC $x_{min}^{(i)}$ and event size errors, as expected in an earthquake dataset.

The recurrence parameter estimates $\hat{\bar{\lambda}}$ and $\hat{\bar{\beta}}$ are calculated for 24 possible scenarios using both maximum likelihood estimation (MLE) and Bayesian inference (BI) to show the effect of each of the uncertainty steps included in the analysis (Table 4.2). This can be regarded also as an analysis of the sensitivity of the parameters to changes in the underlying data and the functional models applied. The independent priors $\pi(\bar{\beta})$ and $\pi(\bar{\lambda})$ are restricted, respectively, to $\bar{b}\sim N(1.0, 0.01)$ for $\bar{\beta}(= \bar{b}\ln 10)$, and an uninformative prior for $\bar{\lambda}$. The choices for the priors follow the true global Gutenberg–Richter b-parameter estimate for tectonic-active areas (El-Isa and Eaton, 2014) for $\pi(\bar{\beta})$, and the assumption of a lack of knowledge of the rate of occurrence in the region in question for $\pi(\bar{\lambda})$. To assess the effect

of including the various epistemic and aleatory uncertainties on the recurrence parameter estimates $\hat{\bar{\lambda}}$ and $\hat{\bar{\beta}}$, the parameter x_{max} was set to 9.65 across all scenarios investigated. An analysis of the various datasets indicated a percentage variation of approximately 25%, and was included in the modelling process whenever MOD was included.

TABLE 4.1. Synthetic earthquake magnitude dataset for the shifted-truncated exponential-gamma distribution with $b = 1$ and $\lambda(x_{min} = 4) = 10$.

Input	**Prehistoric (P)**	**Historical (H)**	**Instrumental 1 ($c1$)**	**Instrumental 2 ($c2$)**
Year start	100 000 BC	1500-01-01	1970-01-01	2001-01-01
Year end	1 AD	1969-12-31	2000-12-31	2017-12-31
Time periods Δt	5 000 years	[50–2.5] years	Annual	Annual
Level of completeness $x_{min}^{(i)}$	7.0	6.0	5.0	4.0
Magnitude error	0.5	0.5	0.3	0.1
Number of events n_i	20	20	42	161
Maximum observed magnitude x_{max}^{obs}	9.53	8.84	7.78	6.5

All results for parameter $\hat{\bar{\beta}}$ are discussed in terms of $\hat{\bar{b}}$ ($\hat{\bar{\beta}} = \hat{\bar{b}} \ln 10$), the power law exponent. This parameter is often used in the subsequent hazard assessments, e.g. calculation of peak ground acceleration (PGA) in seismology. The proposed methodology described in Chapter 3 was applied to a synthetic earthquake dataset. In testing the effects of the various types of uncertainty and estimation procedures, 48 model variations were investigated. The recurrence parameters $\hat{\bar{\lambda}}$ and $\hat{\bar{b}}$ were estimated using MLE and BI, while keeping the maximum possible event size x_{max} constant. The results for parameters $\hat{\bar{\lambda}}$ and $\hat{\bar{b}}$ for each scenario are provided in Table 4.2. From this table, it is clear that the estimated $\hat{\bar{b}}$-parameter depends on the type of input information and the type of uncertainty taken into consideration. This table shows the percentage contribution of each type of dataset to the estimates of the two recurrence parameters. Figures 4.1 and 4.2 compares the estimates $\hat{\bar{b}}$ and $\hat{\bar{\lambda}}$ for the modelled scenarios. Figures 4.3 to 4.5 show to which extent the estimates $\hat{\bar{\lambda}}$ and $\hat{\bar{b}}$ rely on the respective input information. Figure 4.6 shows the comparison of the output for the estimated return periods.

TABLE 4.2. Output of the estimated earthquake recurrence parameters $\hat{\bar{\lambda}}$ and $\hat{\bar{b}}$ according to the mixture-convolution occurrence and event size distributions for the shifted-truncated distribution with x_{min} and x_{max}. Return periods (R) are provided for event magnitude 8.0 M_W. The assumed true Gutenberg–Richter $\bar{b}$-parameter is assumed equal to 1.0 and the true mean rate of occurrence for event size 6.0 M_W is $\bar{\lambda}(x = 6.0) = 0.1$.

Scenario	**Estimation Method**	**Estimated Mean Rate of Occurrence $\hat{\bar{\lambda}}$ for $x = 6.0$**	**Estimated $\hat{\bar{b}}$-parameter**	**Return Period (R) in Years for $x = 8.0$**	**Percentage Contribution ($\hat{\bar{\lambda}}$)**				**Percentage Contribution ($\hat{\bar{b}}$)**			
					PH	**H**	**Inst**	**Prior**	**PH**	**H**	**Inst**	**Prior**
SDS	MLE	0.27 ± 0.03	0.70 ± 0.09	97.0	0	0	100	0	0	0	100	0
	BI	0.19 ± 0.02	0.84 ± 0.07	259	0	0	100	0	0	0	65.5	34.4
SDS_MAG	MLE	0.20 ± 0.03	0.83 ± 0.11	234.0	0	0	100	0	0	0	100	0
	BI	0.16 ±0.02	0.92 ±0.08	451.0	0	0	100	0	0	0	62.6	37.4
SDS_MOD	MLE	0.28 ± 0.08	0.72 ± 0.10	64.5	0	0	100	0	0	0	100	0
	BI	0.21 ± 0.06	0.87 ± 0.08	134.0	0	0	100	0	0	0	63.5	36.5
SDS_MAG_MOD	MLE	0.20 ± 0.06	0.88 ± 0.12	140.0	0	0	100	0	0	0	100	0
	BI	0.17 ± 0.05	0.95 ± 0.08	202.0	0	0	100	0	0	0	60.8	39.2
SDS_OCC	MLE	0.13 ±0.02	0.74 ±0.13	255	0	0	100	0	0	0	100	0
	BI	0.09 ± 0.02	0.91 ± 0.08	820.0	0	0	100	0	0	0	58.3	41.7
SDS_MAG_OCC	MLE	0.13 ± 0.02	0.74 ± 0.13	255.0	0	0	100	0	0	0	100	0
	BI	0.09 ± 0.02	0.91 ± 0.08	822.0	0	0	100	0	0	0	58.3	41.7
SDS_MOD_OCC	MLE	0.13 ± 0.04	0.77 ± 0.14	161.0	0	0	100	0	0	0	100	0
	BI	0.09 ± 0.03	0.93 ± 0.09	356.0	0	0	100	0	0	0	57.1	42.9
SDS_MAG_MOD_OCC	MLE	0.13 ± 0.04	0.77 ± 0.14	161.0	0	0	100	0	0	0	100	0
	BI	0.09 ± 0.03	0.93 ± 0.09	356.0	0	0	100	0	0	0	57.1	42.9

Scenario	Estimation Method	Estimated Mean Rate of Occurrence $\hat{\bar{\lambda}}$ for $x = 6.0$	Estimated $\hat{\bar{b}}$-parameter	Return Period (R) in Years for $x = 8.0$	Percentage Contribution ($\hat{\bar{\lambda}}$)				Percentage Contribution ($\hat{\bar{b}}$)			
					PH	H	Inst	Prior	PH	H	Inst	Prior
EXT	MLE	0.54 ± 0.13	1.41 ± 0.07	1220.0	24.7	75.3	0	0	86.6	13.4	0	0
	BI	0.38 ± 0.09	1.28 ± 0.06	954.0	23.1	76.9	0	0	70.2	18.2	0	11.5
EXT _MAG	MLE	0.25 ± 0.05	1.31 ± 0.07	1690.0	27	73	0	0	92.9	7.1	0	0
	BI	0.20 ± 0.04	1.21 ± 0.05	1330.0	25.3	74.7	0	0	69	17.3	0	13.7
EXT _MOD	MLE	0.47 ± 0.08	1.50 ± 0.0	685.0	21	79	0	0	50	50	0	0
	BI	0.32 ± 0.08	1.33 ± 0.07	596.0	16.9	83.1	0	0	64	20.8	0	15.2
EXT_MAG_MOD	MLE	0.26 ± 0.04	1.48 ± 0.0	1160.0	27.6	72.4	0	0	50	50	0	0
	BI	0.19 ± 0.04	1.28 ± 0.06	844.0	22.2	77.8	0	0	62.8	20.3	0	16.9
EXT _OCC	MLE	0.49 ± 0.16	1.43 ± 0.1	1450.0	23.4	76.6	0	0	85.2	14.8	0	0
	BI	0.28 ± 0.09	1.21 ± 0.07	968.0	20.4	79.6	0	0	61.5	21.7	0	16.8
EXT_MAG_OCC	MLE	0.22 ± 0.06	1.31 ± 0.1	1980.0	26.1	73.9	0	0	92.2	7.8	0	0
	BI	0.16 ± 0.04	1.17 ± 0.07	1340.0	23	77	0	0	59.3	21.7	0	19
EXT_MOD_OCC	MLE	0.41 ± 0.1	1.50 ± 0.00	783.0	20.2	79.8	0	0	50	50	0	0
	BI	0.22 ± 0.07	1.24 ± 0.08	629.0	13.7	86.3	0	0	55.8	24.1	0	20.2
EXT_MAG_MOD_OCC	MLE	0.23 ± 0.05	1.49 ± 0.0	1360.0	27	73	0	0	50	50	0	0
	BI	0.14 ± 0.04	1.20 ± 0.07	865.0	19.1	80.9	0	0	54.2	24.1	0	21.7
LOC	MLE	0.11 ± 0.01	0.97 ± 0.02	798.0	3.5	8.8	87.7	0	79.9	8.3	11.8	0
	BI	0.11 ± 0.01	0.97 ± 0.02	802.0	3.5	8.8	87.7	0	73.6	10.6	13.3	2.8

Scenario	Estimation Method	Estimated Mean Rate of Occurrence $\hat{\lambda}$ for $x = 6.0$	Estimated $\hat{b}$-parameter	Return Period (R) in Years for $x = 8.0$	Percentage Contribution ($\hat{\lambda}$)				Percentage Contribution ($\hat{b}$)			
					PH	H	Inst	Prior	PH	H	Inst	Prior
LOC_MAG	MLE	0.09 ± 0.01	1.00 ± 0.02	1110.0	4.1	8.7	87.2	0	75.8	11.9	12.3	0
	BI	0.09 ± 0.01	1.00 ± 0.02	1110.0	4.1	8.7	87.2	0	68.9	13.7	14.1	3.5
LOC_MOD	MLE	0.1 ± 0.02	1.09 ± 0.04	355.0	0.7	32.5	66.8	0	79.5	8.1	12.3	0
	BI	0.10 ± 0.02	1.08 ± 0.03	347.0	0.1	31.9	68	0	68.3	11.6	14.9	5.2
LOC_MAG_MOD	MLE	0.08 ± 0.02	1.16 ± 0.04	479.0	7.6	34.1	58.3	0	77.6	11	11.4	0
	BI	0.08 ± 0.01	1.14 ± 0.04	456.0	6.2	32.9	60.9	0	65	14.3	14.6	6.1
LOC_OCC	MLE	0.07 ± 0.01	0.92 ± 0.03	1030.0	3.4	8	88.6	0	77.5	7.3	15.2	0
	BI	0.07 ± 0.01	0.93 ± 0.03	1080.0	3.5	8	88.5	0	65.2	11.4	17.7	5.7
LOC_MAG_OCC	MLE	0.06 ± 0.01	0.94 ± 0.03	1270.0	3.9	7.9	88.2	0	75.1	9.6	15.3	0
	BI	0.06 ± 0.01	0.94 ± 0.03	1330.0	4	7.9	88.1	0	62	13.5	18	6.4
LOC_MOD_OCC	MLE	0.07 ± 0.02	1.07 ± 0.05	483.0	3.4	25.4	71.3	0	77.2	7	15.7	0
	BI	0.07 ± 0.01	1.05 ± 0.04	466.0	2.6	24.9	72.5	0	58.6	13.5	18.9	9
LOC_MAG_MOD_OCC	MLE	0.06 ± 0.01	1.08 ± 0.05	582.0	5.9	24.5	69.6	0	75.9	9.1	15	0
	BI	0.06 ± 0.01	1.06 ± 0.04	555.0	5	24	71.1	0	56.6	15.1	18.7	9.6

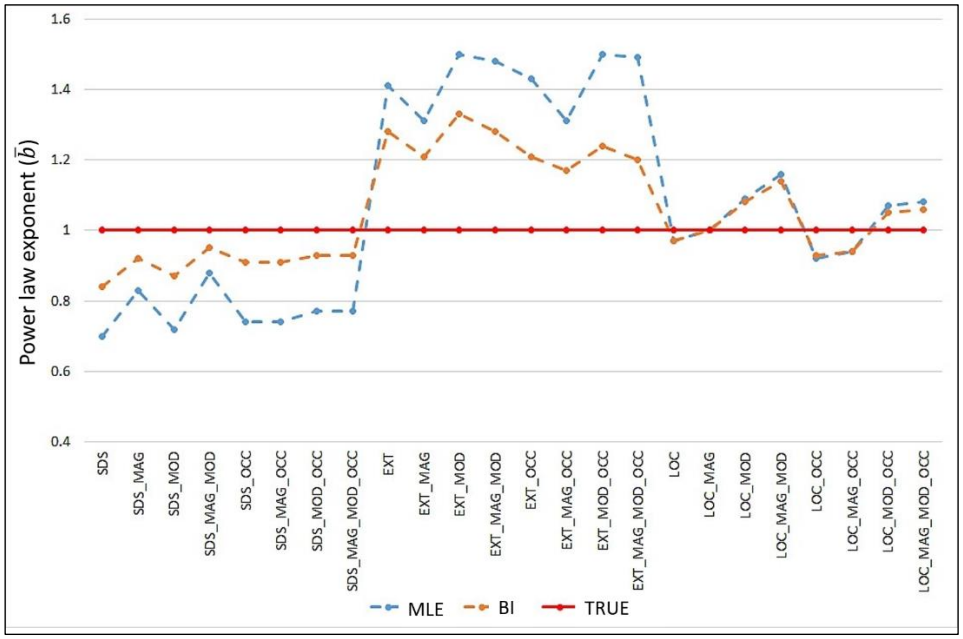


FIGURE 4.1. Estimated $\hat{\bar{b}}$-parameters per model, as defined in Table 3.1. The output from MLE (in blue) is compared with the output from BI (orange line), with a Gaussian prior $\pi(\bar{b}) \sim N(1.0, 0.01)$.

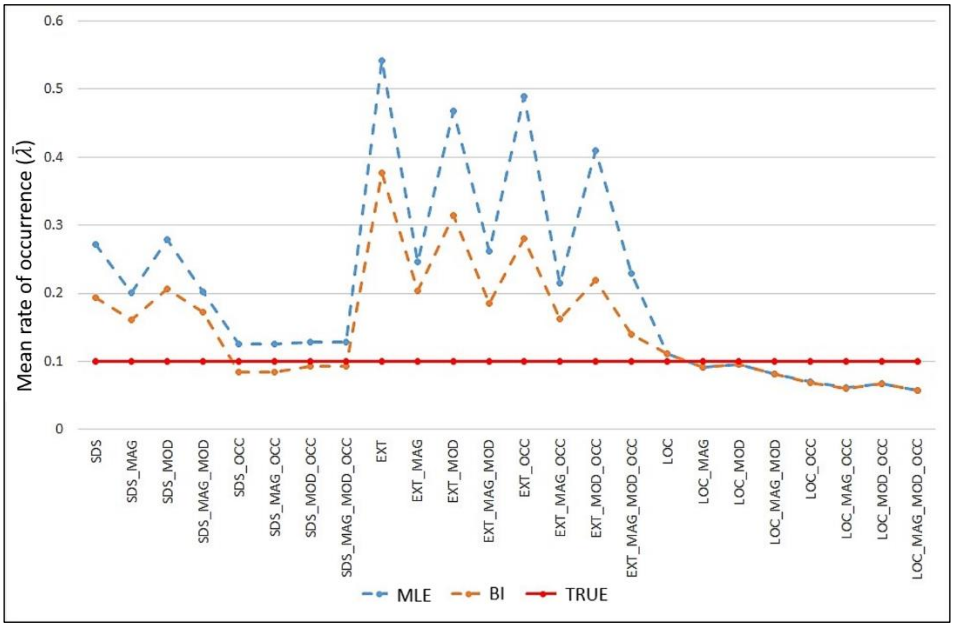


FIGURE 4.2. Estimated $\hat{\bar{\lambda}}$-parameter per model, for an earthquake of magnitude 6.0 M_W, as defined in Table 3.1. The output from MLE (in blue) is compared with the output from BI (orange line), with a uniform prior $\pi(\bar{\lambda})$.

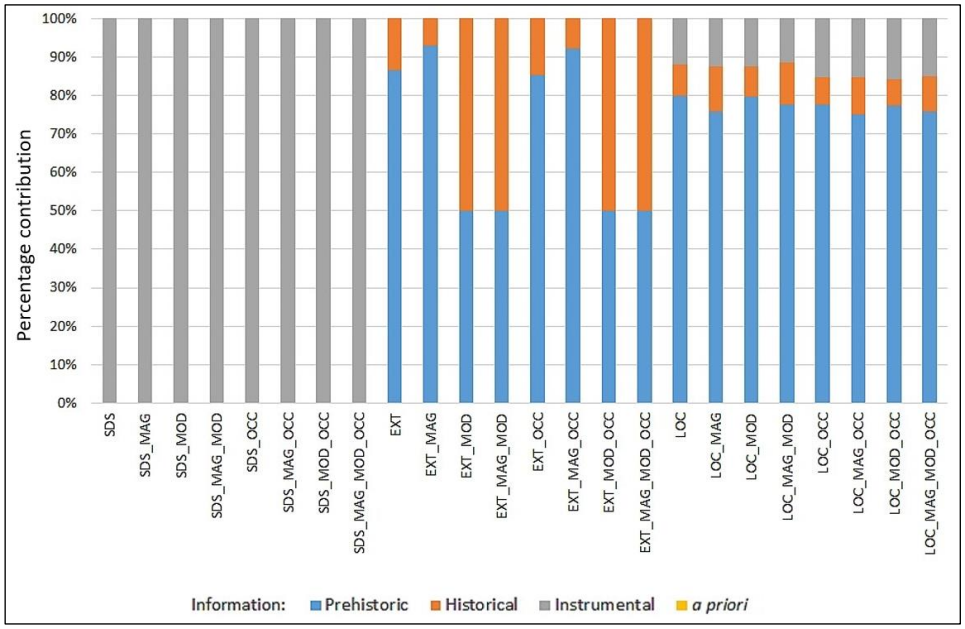


FIGURE 4.3. Percentage contribution per model (as defined in Table 3.1) of the various datasets to the Gutenberg–Richter $\hat{\bar{b}}$-parameter estimates using MLE.

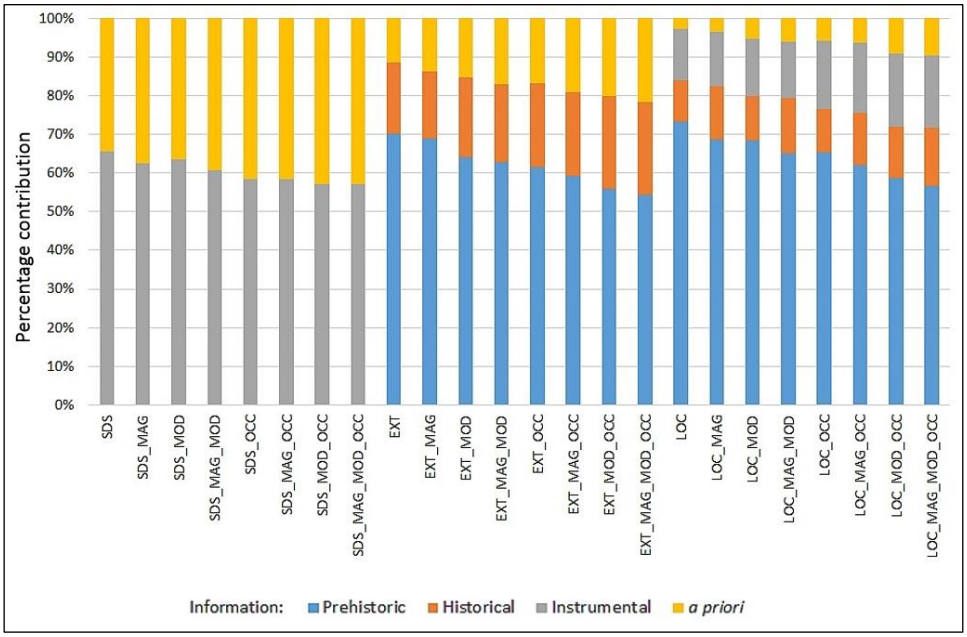


FIGURE 4.4. Percentage contribution per model (as defined in Table 3.1) of the various datasets to the Gutenberg–Richter $\hat{\bar{b}}$ -parameter estimates using BI, with a Gaussian prior $\pi(\bar{b}) \sim N(1.0, 0.01)$.

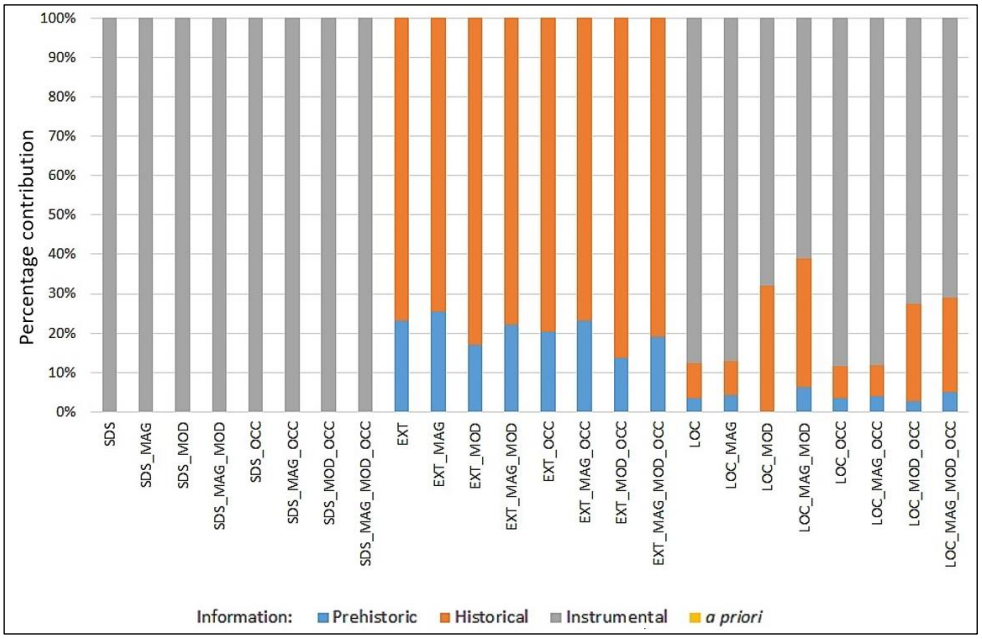


Figure 4.5. Percentage contribution of each sub-dataset to the MLE (blue) and BI (orange) of the mean rate of occurrence $\hat{\bar{\lambda}}$ per scenario. The uniform prior $\pi(\bar{\lambda})$ has no effect on the estimates.

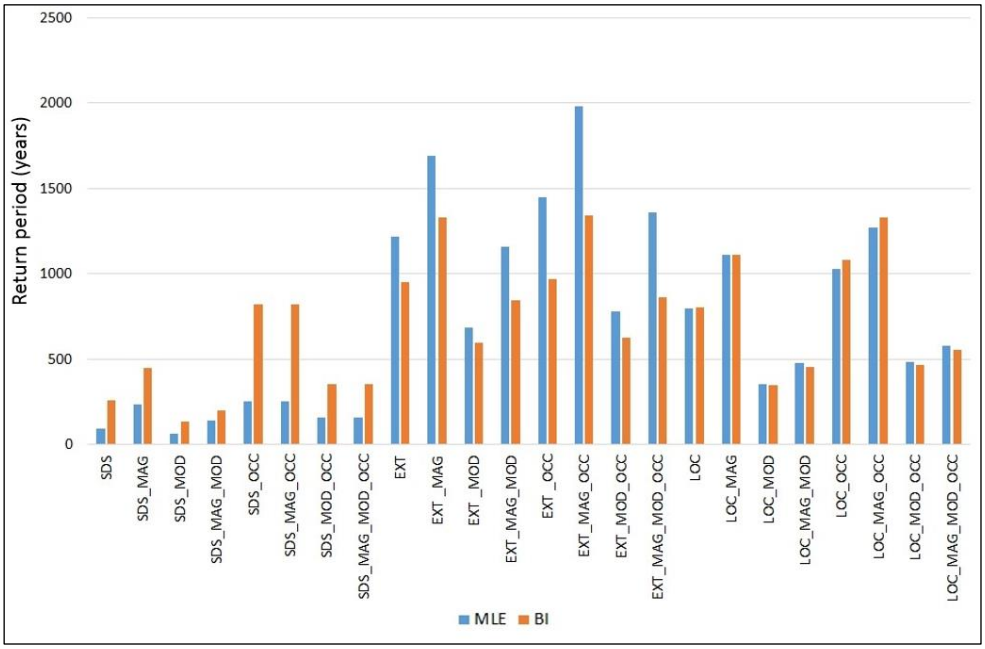


FIGURE 4.6. Comparison of return periods for each sub-dataset, using maximum likelihood estimation (MLE) in (blue) and Bayesian inference (BI) in (orange) for a synthetic earthquake event size $x = 8.0\ M_W$.

Figure 4.1 compares the MLE and BI estimation results for the $\hat{\bar{b}}$-parameter. With the introduction of the various combinations of parameter and earthquake magnitude uncertainty, the historical and prehistoric datasets start playing a role. The estimated $\hat{\bar{b}}$-parameters for the single datasets (SDS) are underestimated compared with the estimates from the extreme-only datasets (EXT). The combined datasets (LOC) yield estimates closer to the true value of $\bar{b} = 1$. With the inclusion of event size uncertainties (MAG) and parameter (MOD) uncertainties, the validity index (OCC) and prior information yield superior results compared with the MLE estimation for the SDS. In addition, the prior information stabilises the results from the EXT dataset, but the results from the MLE and BI for the LoC datasets are close. Event size uncertainty has a stronger effect on the estimates than does parameter uncertainty. The combination of the two types of uncertainties provides the best results using the SDS dataset but the worst results using the EXT dataset. The effect of the combination of uncertainty types on extreme data is carried over to the LoC datasets. The combination of event size uncertainty, parameter uncertainty, and validity index affect the SDS and extreme data in diverse ways, as well. This is carried over into the LOC datasets. From this simulation, it is clear that the choices made when building the model are crucial and the selection should be done with care.

Figure 4.2 compares the MLE and BI estimates for rate of occurrence $\hat{\bar{\lambda}}$ of earthquake event size $x \geq 6.0\ M_w$, with the true parameter equal to $\bar{\lambda}(x_{min} = 6.0) = 0.1$. For the SDS models, overestimation of the true value 0.1 occurs. The introduction of a validity index has the most influence on the parameters and reduces the rate of occurrence for the SDS to being closer to the true value. Utilising only the extreme dataset (EXT) has a much larger overestimation effect than that observed in the SDS or the combined datasets with varying levels of completeness (the LOC models M17–M24 in Table 3.1). Here, the introduction of the weights has little to no effect. This is attributable possibly to the small likelihood of a large event being misclassified. For the extreme dataset, the event size uncertainty has the most effect on the estimates of the rate of occurrence. The models LOC, LOC_MAG, and LOC_MODEL have the closest estimates to the true value of 0.1. The introduction of the validity index results in a slight underestimation of the rate of occurrence, as does the introduction of the various

combinations of uncertainty. Overestimation of the rates of occurrence for SDS and EXT datasets results in the underestimation of return periods and, therefore, the overestimation of hazard. The slight underestimation of $\hat{\bar{\lambda}}$ for the LOC datasets results in an overestimation of the return periods and the underestimation of the associated hazard. The introduction of prior information for the $\bar{b}$-parameter has a slight effect on the estimates for $\hat{\bar{\lambda}}$ for SDS and the EXT datasets. This effect is small for the LOC datasets. The relationship can change as the assumptions about the assumed uncertainties change.

From this simulation, it appears that the introduction of model and event size uncertainties provides estimates quite close to the true assumed values. Conversely, the influence of the type of uncertainty introduced manifests more clearly in the return periods. The inclusion of the validity of events generates higher return periods, signifying a lower hazard. Uncertainty in size also increases the return period, but the inclusion of parameter uncertainty reduces the return period (i.e. increased hazard). The introduction of the validity of events (OCC models) practically halved the mean rate of occurrence, which could have a severe effect on hazard classifications.

Figures 4.3 to 4.5 show the contribution of the various datasets to the MLE and BI estimates $\hat{\bar{\lambda}}$ and $\hat{\bar{b}}$. The contribution of each dataset to the parameters depends on the type of uncertainty included and the estimation process used. For this synthetic earthquake example, prehistoric and historical data play a larger role in the estimation of $\hat{\bar{b}}$ compared with the more-recent instrumental data playing a larger role in the estimation process for $\hat{\bar{\lambda}}$. The introduction of parameter uncertainty (MOD models) in the MLE results for the EXT dataset yields a contribution that is more even for prehistoric and historical data (Figure 4.3). The introduction of various types of uncertainty does not have a large effect on the percentage contribution of the combined LOC datasets. The contribution of prior information for BI (Figure 4.4) is the largest for the $\hat{\bar{b}}$ estimates based on SDS data and the smallest for the LOC datasets. The more uncertainty is introduced the larger is the role of the prior information in the estimation process.

As a flat prior is assumed, there is no difference in the percentage contribution of the information for $\hat{\bar{\lambda}}$ between the MLE and BI estimation procedures, shown in Figure 4.5. As the two recurrence parameters were solved simultaneously, the assumed prior for $\bar{b}$ had an indirect effect on $\hat{\bar{\lambda}}$. In this example, the $\hat{\bar{b}}$-parameter estimate relies heavily on the more-recent instrumental information, than on historical or complete information. The introduction of parameter uncertainty in the LOC dataset plays the main role in which dataset the estimation process would rely on the most. In such an instance, the process also relies on the historical information to help stabilise and improve the accuracy of the results.

Figure 4.6 shows a comparison of the return periods for $x = 8.0\ M_W$ for the MLE and BI estimates for the 48 models. The aim of this figure is to inspect the effects that different datasets, models, and estimation processes have on the return periods for large event sizes. The influence of a priori information on the return periods for the SDS dataset is larger than is that for the extreme or combined datasets. This is particularly true in instances where only the validity of the event and the validity of the event combined with event size error are used. The inclusion of prior information yields higher return periods compared with the MLE estimation process when applied to the SDS data and models, i.e. lower hazard. It has the opposite effect on the extreme dataset and models, providing higher hazard levels than those of the MLE estimates. The estimates from MLE and BI for the combined dataset LOC are quite close. When applying the extreme dataset, the return periods for large event sizes are larger compared with the return periods for the SDS dataset; however, combining the datasets seems to balance this out. This means that return periods are yielded that are larger compared with using a single dataset with a high level of completeness, and return periods are lower than when only extreme data are applied. This should provide a more realistic estimate of the hazard for the area under investigation. Interestingly, lower return periods (i.e. higher hazard) are observed with the inclusion of only parameter uncertainty and the combination of parameter and event size uncertainty, regardless of the type of dataset used. The same pattern is observed when the validity index is included. The type of dataset used, and the uncertainty included in the modelling process, therefore, has a notable effect on the hazard estimates.

Chapter 5. Real-World Applications

The methodology described in Chapter 3 is applied to various datasets relating to the occurrence of natural hazards to illustrate how the proposed procedure can be adjusted to suit a particular dataset. Four examples are presented in the form of seismic hazard for the Ceres–Tulbagh region in South Africa, seismic hazard in Central Italy, tsunami hazard for regions in the Pacific Ocean, and hail hazard for Gauteng province, South Africa. The hail hazard example is extended to hail risk by applying the methodology to insured losses and incorporating the hail hazard estimates as *a priori* information. For each dataset, the input conditions and estimation process differ.

5.1 Earthquakes

The seismic hazard for the Ceres–Tulbagh region in South Africa is investigated using historical and instrumental data (Section 5.1.1). Estimates were derived using MLE[14]. The seismic hazard for Central Italy (Section 5.1.2) illustrates the inclusion of prehistoric, historical, and instrumental data[15].

The types of uncertainty considered in Section 5.1 are the incompleteness of data, event size (magnitude) uncertainty, and uncertainty of the parameters in the applied distributions (M12 in Table 3.1). The shifted-truncated event size distributions for instrumental and historical data were used. The validity of occurrence was not considered.

[14] The description of the analyses, tables, and discussion is provided below as it was published in the Bulletin of the Seismological Society of America (Kijko, Smit and Sellevoll, 2016). Minor amendments were made to the text, as seen in the references to 'the authors', 'this paper' instead of this thesis, 'magnitude' instead of event size, and 'catalogue' instead of event dataset. In the paper, the acronyms KS I, KS II, and KS III refer to the publications by Kijko and Sellevoll (1989), Kijko and Sellevoll (1992), and Kijko, Smit and Sellevoll (2016), respectively.

[15] The results were presented at the Extreme Value Theory Conference, 26–30 June 2017 by Smit, Kijko, and Stein.

5.1.1 Historical and instrumental data

For the purpose of illustration, the proposed methodology is applied to the area that had experienced the strongest and most devastating earthquake in contemporary South African history. The 29 September 1969 M_W 6.3 Ceres–Tulbagh event occurred approximately 100 km from the major metropolitan city of Cape Town (Green and Bloch, 1971; Green and McGarr, 1972; Kijko *et al*., 2003; Krüger and Scherbaum, 2014). Several buildings in the area suffered serious structural damage, which varied from the almost total destruction of old and poorly constructed buildings to large cracks appearing in better-constructed infrastructure. Twelve people were killed and many more injured. This event resulted in an insured loss of USD 7.4 million at the time of occurrence; however, the uninsured loss was approximately 3.5 times higher (Davies and Kijko, 2003).

Reported seismicity in the vicinity of Cape Town dates back to 1620. The seismicity here is typical of an intraplate region and is characterised, compared with world standards, by low-level activity. The earthquakes are distributed randomly in space and time. Probably, the most common practice in seismic-hazard analysis is that the maximum possible earthquake magnitude is estimated from the magnitude–fault-length relationships (Wells and Coppersmith, 1994; Leonard, 2010; Stirling *et al*., 2013). The correlation between most of the observed earthquakes in South Africa and the surface expression of major geological features is generally not clear (Fernandez and Guzman, 1979 a,b; Brandt *et al*., 2005). The estimated maximum possible earthquake magnitude is, therefore, calculated by utilising the observed seismic-event catalogue.

The seismic-event catalogue for South Africa, especially the historical catalogue, is highly incomplete because large parts of the country are sparsely populated, and the detection capabilities of the seismic network are not uniform (Saunders *et al.*, 2008). The seismic-event dataset used in this study was compiled from several sources. After critical analysis of each of the data sources, the main contribution to pre-instrumentally recorded seismicity was that of Brandt *et al.* (2002). The instrumentally recorded events were selected mainly from the available catalogues provided by the Council for Geoscience,

Pretoria, South Africa, and the International Seismological Centre in Edinburgh, Scotland. For illustrative purposes, the dataset used in the present analysis spans the period 1 January 1751 to 31 January 2012 (see Section 1.7). The events were selected from within a circle, with a radius of 300 km from the anticipated epicentre, which is the Ceres–Tulbagh earthquake (33.28′ S, 19.70 ′ E) (Kijko *et al*., 2003). The seismic-event dataset was divided into an incomplete historical part, consisting of only the largest events (Table 5.1), and the complete instrumental part. It was assumed that earthquake magnitudes for the incomplete part of the dataset were determined with a standard error equal to 0.3 magnitude units. The complete part of the dataset was divided further into three sub-datasets, each with different levels of completeness and assumed magnitude standard errors of 0.3, 0.2, and 0.1 (Table 5.2). The sub-datasets and their respective LoC were obtain using the visual cumulative method (Mulargia and Tinti, 1985). Based on several tests of the seismicity of the selected area, it was assumed that the uncertainty of the earthquake-model parameters was equal to 25%. This implied that the space–time variation of parameters λ and β, relative to their mean values, was equal to 0.25. Figure 5.1a depicts the mean return periods, and Figure 5.1b depicts the probability that a given magnitude would be exceeded at least once in any year for the selected area, as estimated by the proposed procedure (KSS III, Kijko *et al.*, 2016). Each graph also provides the calculated standard error for the calculated values.

Three models were tested, namely KS I (magnitude and model uncertainties were ignored), KS II (model uncertainties were ignored), and KSS III, where both magnitude and model uncertainties were accounted for. KS 1 refers to LOC (M17) in Table 3.1, KS II to LOC_MAG (M18), and KSS III to LOC_MAG_MOD (M20). Table 5.3 provides the relevant estimated recurrence parameters for λ and b for all the models using the same datasets. The Kijko–Sellevoll–Bayes' method was used to estimate $\widehat{m}_{max}$, as described in Eq. 3.52 of Section 3.4. The estimated parameters are close in value, with the largest difference observed in the estimated maximum possible magnitude $\widehat{m}_{max}$. Although KS II estimates $\widehat{m}_{max}$ as 7.8, it also has the largest associated error. Again, the return periods best illustrate the effect of uncertainty included in the data. For the same earthquake magnitude M_w 6.3 observed in

1969, the hazard return periods are estimated as 391 years for KS I, 492 years for KS II, and 353 years for KSS III. The model accounting for magnitude error only (KS II) provided the lowest hazard-estimate, and the model accounting for both magnitude and model uncertainty provided the highest hazard. The 139-year difference between these two estimates can make a marked difference in how civil engineering industries design earthquake-proof structures in this region.

Table 5.1. The largest earthquakes that occurred within a 300 km radius of the Ceres–Tulbagh earthquake epicentre. The catalogue starts on 1751/01/01 and ends on 1970/12/31. The standard error of earthquake magnitude determination, when applied, was assumed 0.3.

Event	Date	Magnitude	Event	Date	Magnitude
1	1766/07/14	4.3	15	1950/09/30	5.5
2	1809/12/04	6.3	16	1951/06/13	4.7
3	1811/06/02	5.7	17	1952/01/28	5.4
4	1819/04/14	4.3	18	1953/02/26	4.4
5	1835/11/11	4.3	19	1957/09/30	4.2
6	1857/08/14	5.0	20	1960/08/29	4.8
7	1899/09/13	5.0	21	1963/08/27	5.0
8	1902/05/28	4.3	22	1964/02/21	4.3
9	1908/12/30	4.0	23	1965/09/28	4.3
10	1911/07/06	4.0	24	1966/03/01	4.3
11	1921/10/09	5.0	25	1967/06/16	4.3
12	1926/08/11	4.0	26	1969/09/29	6.3
13	1940/10/13	4.3	27	1970/04/14	5.7
14	1941/10/23	4.3			

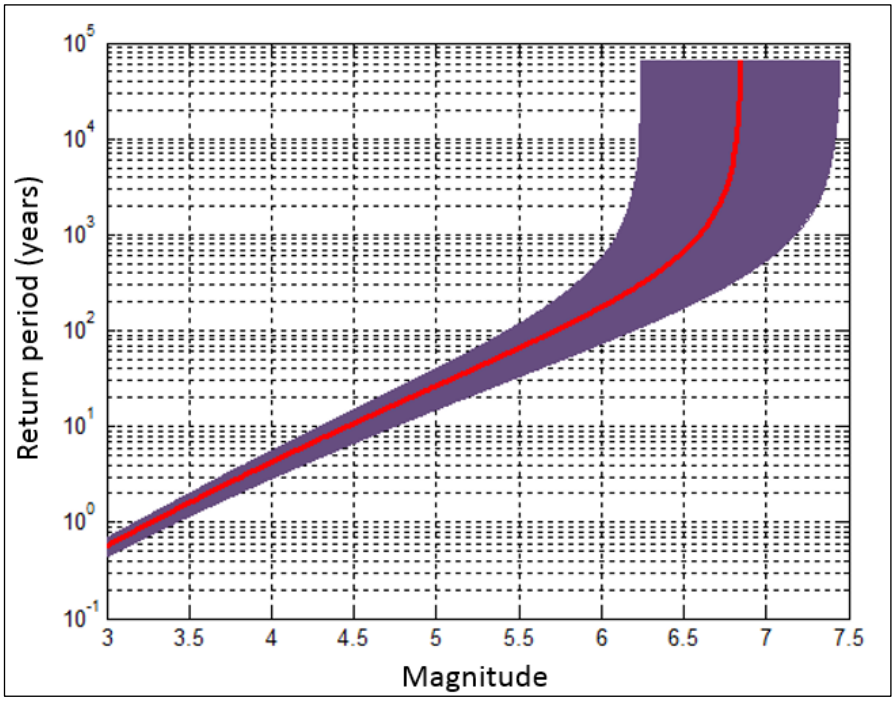
TABLE 5.2. Summary of the complete parts of the catalogues.

Sub-catalogue	Start Date	End Date	Level of Completeness (M_w)	Number of Events	Standard Error of Earthquake Magnitude Determination
1	1971/01/01	1990/12/31	4.0	7	0.3
2	1991/01/01	1995/12/31	3.5	2	0.2
3	1996/01/01	2013/01/31	3.0	29	0.1

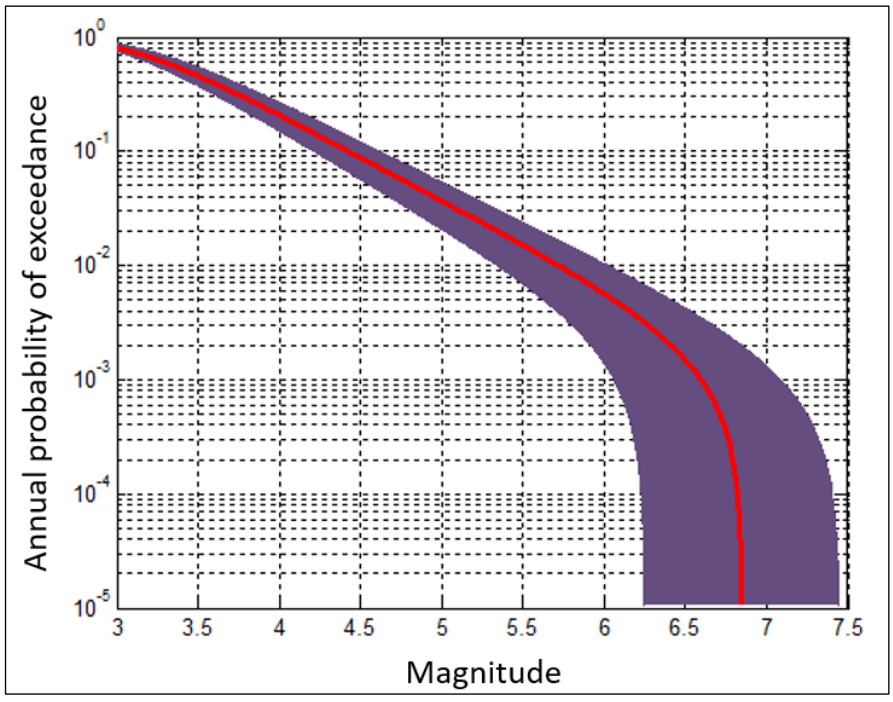
TABLE 5.3. Estimated recurrence parameters for the Ceres–Tulbagh earthquake epicentre. The catalogue starts on 1751/01/01 and ends on 1970/12/31.

Recurrence parameter estimates	KS I	KS II	KS III
Gutenberg–Richter b-parameter	$\hat{b} = 0.8 \pm 0.06$	$\hat{b} = 0.9 \pm 0.06$	$\hat{\bar{b}} = 0.9 \pm 0.07$
Rate of occurrence λ (for $x_{min} = 3.0$)	$\hat{\lambda} = 1.8 \pm 0.3$	$\hat{\lambda} = 1.7 \pm 0.3$	$\hat{\bar{\lambda}} = 1.8 \pm 0.4$
Maximum possible magnitude $\hat{m}_{max}$	7.0 ± 0.74	7.8 ± 1.5	6.9 ± 0.60
Return period for $m = 6.3$	391 years	492 years	353 years
Probability of exceedance $m = 6.3$			
1 year	0%	0%	0%
50 years	10%	10%	13%
100 years	20%	18%	24%
250 years	50%	40%	50%

To capture the effect of incorporated uncertainties in both the magnitude determination and the selected earthquake model, two comparisons were made. In the first comparison (Figure. 5.2), the differences were investigated between the estimates of the mean return periods according to KS I (magnitude and model uncertainties were ignored) and two respective alternative instances. These alternative instances are (1) only the uncertainty of the earthquake magnitude is taken into account (KS II), and (2) only the uncertainty of the earthquake-model parameters is considered. The purpose of such a comparison was to isolate and capture the effects of the two uncertainties. The comparison in Figure 5.2 of the classic instance (KS I), with the instance in which only magnitude uncertainties are considered, confirms the acknowledged fact that accounting for magnitude errors leads to an increase in return periods or, equivalently, to a decrease in seismic hazard (e.g. Tinti and Mulargia, 1985; Rhoades and Dowrick, 2000; McGuire, 2004). The uncertainty of the earthquake-model has an opposite effect, namely it leads to a decrease in the return periods or, equivalently, an increase in seismic hazard.



(a)



(b)

FIGURE 5.1. a) The estimated mean return periods with the level of confidence per magnitude, and b) the probability that a given magnitude will be exceeded at least once in any year for KSS III (Kijko *et al*., 2016). Each graph also provides the calculated standard error for the calculated values.

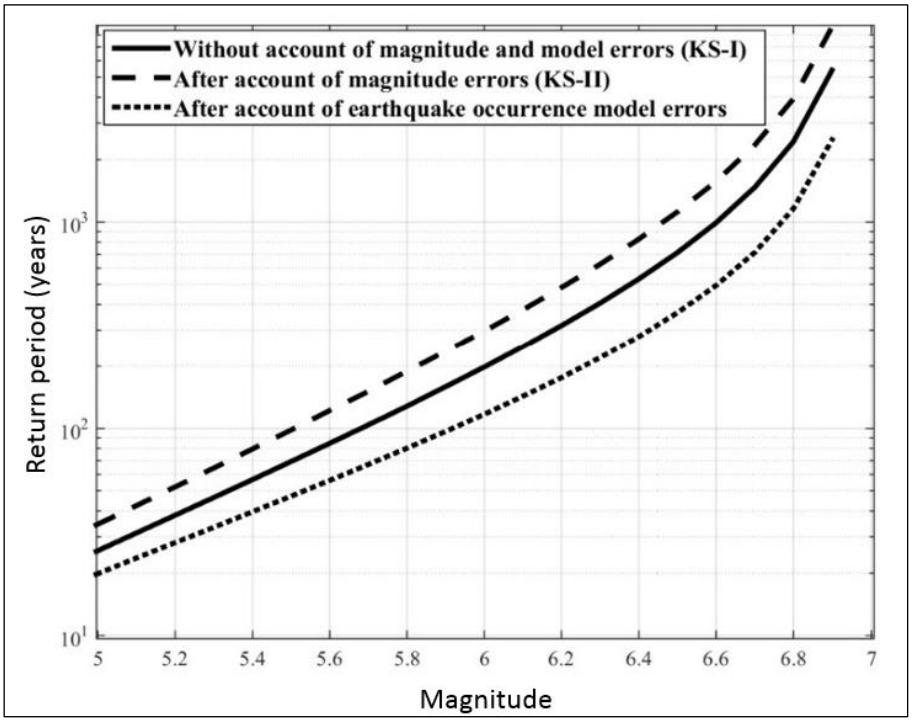


FIGURE 5.2. Comparison between the individual effects of the magnitude uncertainties and the event occurrence model uncertainties on the mean return periods. The solid line shows KS I, when both the uncertainties are ignored. The dashed line captures the effect of magnitude uncertainties implemented in KS II, whereas the dotted line shows the effect of accounting for the parameter uncertainties as in KSS III.

The second comparison (Figure 5.3) shows two estimates of the mean return periods according to KS I, when both magnitude and model uncertainties are ignored; as well as the same estimates as those obtained by KSS III, when both the magnitude and the event occurrence model uncertainties are considered. The increase in return periods, as a result of magnitude uncertainty, and the decrease in return periods, as an effect of the uncertainty of the earthquake occurrence model, average out to an overall decrease in return periods. Consequently, the calculated hazard levels increase, as the effect of introducing model uncertainty is stronger than is the effect of magnitude uncertainty.

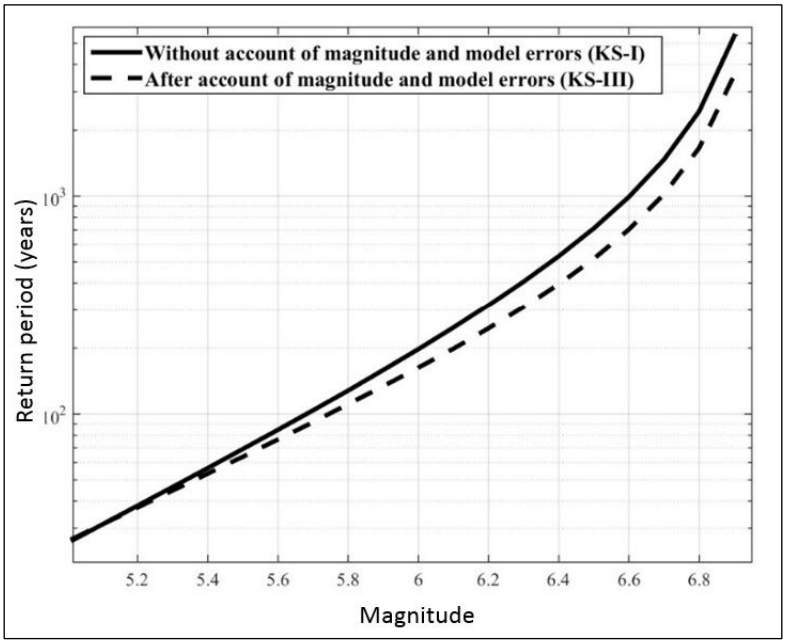


FIGURE 5.3. Comparison of mean return periods calculated without accounting for magnitude and parameter uncertainties as per KS I (solid line), with return periods calculated taking into account the magnitude and parameter uncertainties as per KSS III (dashed line). The dashed line is a balance of the two opposite effects of the modelled uncertainties.

Several additional aspects should be investigated in the future. These include (1) a more comprehensive investigation of the universal applicability of the above results to seismic hazard modelling in areas with different levels of seismic activity, (2) the sensitivity of the hazard parameters to different degrees of model uncertainty, (3) the effect of different distribution models of magnitude uncertainty, and (4) the relationship between magnitude uncertainty and uncertainty in the earthquake occurrence and event size models.

5.1.2 Prehistoric, historical and instrumental Data

To test the methodology on prehistoric, historical, and instrumental data, seismic events in Central Italy were used according to the ZS9 zones of the Working Group MPS (2004). Extreme prehistoric data

were obtained from Galli *et al*. (2008) and the historical and instrumental data from the European-Mediterranean Earthquake Catalogue (EMEC, 2012[16]; Grünthal and Wahlström, 2012). Figure 5.4 shows the earthquake events used in the dataset, with Table 5.4 showing the input information. Here, the expected spatial bias of the prehistoric events in the dataset (red dots) is clearly visible.

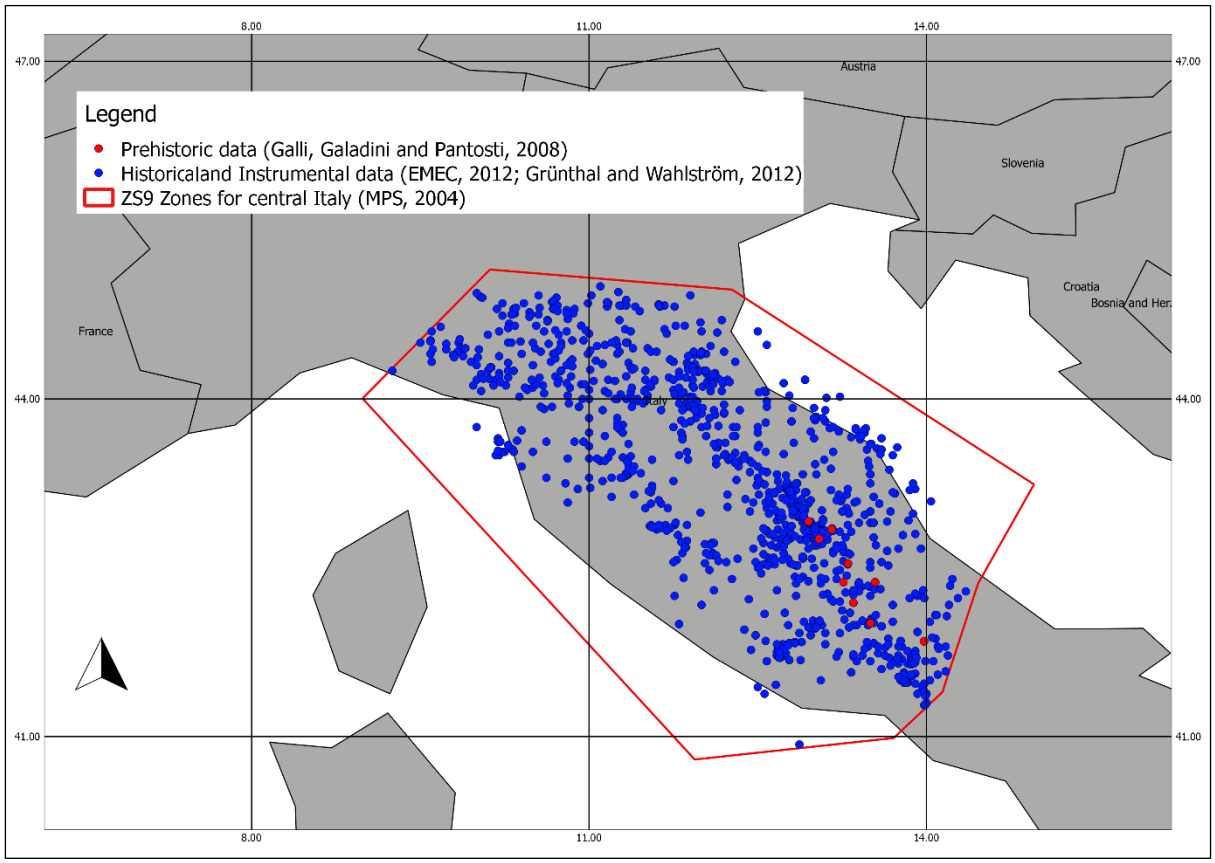


FIGURE 5.4. Earthquake events in Central Italy. Blue dots are the historical and instrumental events and the red dots the extreme prehistoric events. The red line denotes the boundaries of Central Italy ZS9 zone.

[16] https://www.gfz-potsdam.de/en/section/seismic-hazard-and-stress-field/data-products-services/emec-earthquake-catalogue/, last accessed 2019/01/25.

Five sub-datasets were created with levels of completeness $x_{min}^{(i)}$ ranging from 4.0 to 5.1. Uniform event size errors were assumed for the prehistoric and historical datasets, with a Gaussian error for the instrumental sub-datasets. Parameter variation was accounted for at a level of 25%. The maximum possible event size $\hat{x}_{max}$ was estimated using the Kijko–Sellevoll–Bayes' method, as discussed in Eq. 3.52 of Section 3.4.

TABLE 5.4. Input information for the assessment of the recurrence parameters for the Central Italy earthquake event dataset.

	Extreme Prehistoric	**Historical**	**Instr. 1**	**Instr. 2**	**Instr. 3**
Year start	18000 BC	1000 AD	1625	1891	1991
Year end	999 AD	1624	1890	1990	2006
LoC $x_{min}^{(i)}$		5.1	4.3	4.3	4.0
Number of events n_i	22	25	139	306	93
Magnitude standard error	0.5	0.5	0.5	0.3	0.1
Observed maximum magnitude x_{max}^{obs}	7.0	7.2	7.0	7.1	5.8

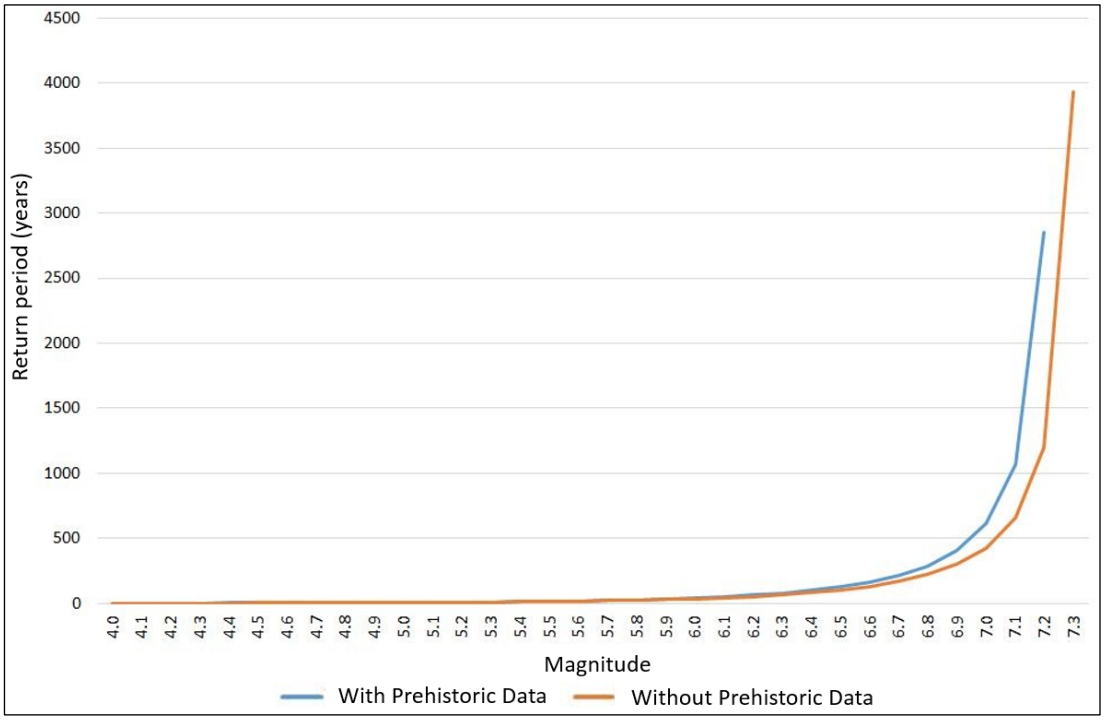
The Gutenberg–Richter frequency–event-size shifted-truncated power law and MLE estimation are employed as they best address the availability and the quality of the data in this example The recurrence parameters are estimated with (example 1) and without (example 2) the prehistoric data (Table 5.5). Both examples made used the LOC_MAG_MOD model (M20 in Table 3.1), with example 2 dropping the prehistorical likelihood function from Eq. 3.18. Good $\hat{\hat{b}}$-estimates are obtained in both instances that are close to the theoretical Gutenberg–Richter b-parameter of 1 for tectonic-active areas. The estimate of the mean rate of occurrence $\hat{\hat{\lambda}}$ for the level of completeness $x_{min} = 4.0$ is not affected; however, the mean return periods of large event sizes are affected. For example, the return periods for event size $x = 7.0$, for the estimate that considers prehistoric data (WPD), differ by 186 years from the estimate

ignoring the prehistoric data (WOPD), exhibiting a larger return period, i.e. smaller hazard. In this example, the extreme prehistoric dataset contributed 28% of the total information to the $\hat{\hat{b}}$-parameter estimate. In a subsequent test, however, this contribution dropped dramatically to almost 6% as the event size error increased. The instrumental datasets contributed the most to the rate of occurrence of the area. Where prehistoric data are not considered, the historical and instrumental data contribute equally to the $\hat{\hat{b}}$ parameter, but the contribution to $\hat{\hat{\lambda}}$ from the instrumental data increases.

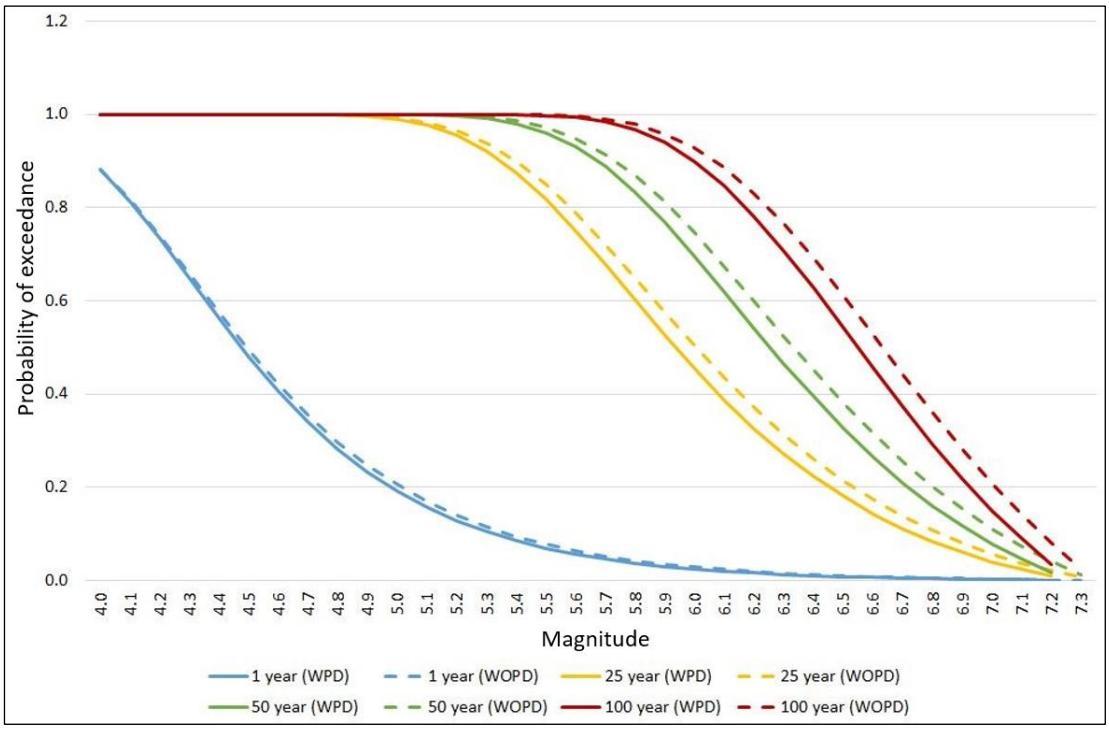
TABLE 5.5. Seismic recurrence parameters for the earthquake event dataset of Central Italy, taking into account complete, historical, and extreme prehistoric data. WPD takes into consideration prehistoric data and WOPD excludes prehistoric data from the estimation process.

	Type of Dataset	**With or Without Prehistoric Data**	**Estimated Mean Rate of Occurrence $\hat{\hat{\lambda}}(x_{min} = 4)$**	**Estimated $\hat{\hat{b}}$-parameter**	**Maximum Possible Magnitude (x_{max})**	**Return Period for $x = 7.0$**
Estimation		WPD	2.3 ± 0.3	1.11 ± 0.1	7.3 ± 0.5	611 years
		WOPD	2.3 ± 0.3	1.07 ± 0.05	7.4 ± 0.5	425 years
Percentage contribution	Extreme prehistoric data	WPD	6.2%	28.8%		
		WOPD				
	Historical data	WPD	17.8%	34.4%		
		WOPD	18.1%	48.0%		
	Instrumental data	WPD	76.0%	36.8%		
		WOPD	81.9%	52.0%		

Figure 5.5 provides a comparison of the return periods and probabilities of exceedance with prehistoric data (WPD) and without prehistoric data (WOPD). The return periods for larger event sizes are larger (i.e. lower hazard) when prehistoric data are included, and the same is true for the respective probabilities of exceedance for 1, 25, 50, and 100 years.



(a)



(b)

FIGURE 5.5. a) Return period and b) probabilities of exceedance for 1, 25, 50, and 100 years for earthquake event sizes between 4 and 8 for the earthquake event dataset of Central Italy, taking into account complete, historical, and extreme prehistoric data.

5.1.3 Discussion

The methodology proposed in Chapter 3 was applied to seismic event datasets for the Ceres–Tulbagh area in the Western Cape Province, South Africa, as well as to Central Italy. In this section, only a selected number of combinations were tested. For the Ceres–Tulbagh example, MLE was applied to a combined dataset assuming no event size or parameter uncertainty, to a combined dataset with only event size uncertainty, and to a combined dataset with event size and parameter uncertainty. The example confirms the statements by Tinti and Mulargia (1985), Rhoades and Dowrick (2000), and McGuire (2004) that the inclusion of event size uncertainty yields longer return periods and, therefore, a decrease in seismic hazard. The inclusion of parameter uncertainty has the opposite effect, leading to an increase in seismic hazard.

The investigation into the seismic hazard for Central Italy focused on the effect of prehistoric data on the estimated recurrence parameters and hazard estimates using MLE. In this example, there was no marked difference in the estimates of the mean rate of occurrence $\hat{\hat{\lambda}}$ or the estimates of the Gutenberg–Richter frequency–event-size $\hat{\hat{b}}$-parameter. The effect was, however, notable in the hazard estimates for large event sizes. By excluding the prehistoric data, smaller return periods and probabilities of exceedance were obtained, increasing the hazard. This could potentially lead to the overestimation of the hazard and associated risk for the investigated area.

Proposed future research in the above regard is the inclusion of seismic prior information through BI, determining the effect for different distribution models in terms of event size uncertainty, and whether there would be a difference in the behaviour of the models when investigating areas with different levels of seismic activity (e.g. tectonic vs mining-related/-induced events).

5.2 Tsunami

The tsunami hazard was investigated for three tsunamigenic regions in the Pacific Ocean, namely Japan, Kuril–Kamchatka, and the western coast of South America. Estimates were derived using historical and instrumental data and MLE[17]. In subsequent analyses, the validity of occurrence and prior information were introduced for the assessment of tsunami hazard for Japan[18].

5.2.1 Introduction

Since 2004, four devastating tsunamis following major earthquakes have caused unprecedented destruction and loss of life. The events in Sumatra (2004), Japan (2011), and Chile (2010, 2015) have re-emphasised the need for comprehensive and accurate hazard assessment that considers both earthquakes and tsunamis. Earthquake hazard assessment procedures, both probabilistic and deterministic, are well defined and applied; however, until quite recently, the same could not be said about tsunami hazard assessment.

Tsunamis can be generated by earthquakes, volcanogenic processes, submarine landslides, and meteorites. Among these, earthquakes with frequent occurrences generate the most destructive tsunamis, whereas the other processes have a much smaller probability of occurrence. Therefore, most current research is focussed primarily on assessing the tsunami hazard generated by earthquakes.

In essence, the tsunami hazard for a specified area can be assessed in two ways, namely by deterministic and probabilistic procedures. Deterministic methodologies focus on a single event, usually the worst-

[17] The introduction, description of the analyses, tables, and discussion are provided below, as published in Pure and Applied Geophysics, Smit, Kijko and Stein (2017). The paper considered incompleteness of data, event size (tsunami intensity) uncertainty, and parameter (model) uncertainty. Minor amendments were made to the text, as seen in references made to 'the authors', 'this paper' or 'this study' instead of this thesis, 'tsunami intensity instead of event size, and 'catalogue' instead of event dataset.

[18] The results were published in Smit, Stein and Kijko (2019).

case scenario, by modelling the resultant wave run-up height (or tsunami intensity). Wave run-up height refers to the far up (inland) the wave extended. Little or no emphasis is placed on the probability of such an event occurring (e.g. Tinti and Armigliato, 2003; Hébert *et al*., 2005; Paulatto *et al*., 2007; Lorito *et al*., 2008; Shaw *et al*., 2008; Wijetunge, 2014). Such an approach provides valuable information for disaster management agencies charged with response planning to minimise the potential effects. Recently, research has focused on assessing the deterministic tsunami hazard with the help of the worst-case credible tsunami scenario method (Tonini *et al*., 2011; Grilli *et al*., 2011; Harbitz *et al*., 2012). This involves the development of an aggregated scenario based on several independently assessed deterministic scenarios. Probabilistic tsunami hazard assessment (PTHA), in turn, considers all possible earthquake scenarios (both locally and regionally) that could generate tsunamis of various intensities. Accordingly, this approach provides a comprehensive estimate of the total hazard that confronts a specific area (e.g. Geist and Parsons, 2006; Power *et al*., 2007; Thio *et al*., 2007, Sørensen *et al*., 2012; Brizuela *et al*., 2014). Determining the probability of occurrences and return periods is crucial for the risk mitigation processes of disaster management, engineering, and insurance companies. The disadvantage of PTHA through probabilistically assessing the occurrence of potential sources is that it requires extensive research and modelling to identify all the potential sources such as earthquakes, and to assess the hazard and the resultant wave propagation. Moreover, these models have to consider aleatory and epistemic uncertainties to prevent the underestimation of the tsunami hazard. Aleatory uncertainty refers to the natural stochastic nature, which is characteristic of the physical system, whereas epistemic uncertainty refers to the incompleteness of the information in the system.

PTHA can be conducted using three distinct procedures. In the first procedure, PTHA comprises of computational analyses, which take into account the known and available information on the earthquake source parameters, as well as the propagation of waves, the recurrence rate, and the underlying uncertainties. This procedure generally follows the same route as standard probabilistic seismic hazard analysis (PSHA), as developed by Cornell (1968) and McGuire (1976). Examples of the application of this procedure are available in Lin and Tung (1982), Rikitake and Aida (1988), Ward and Asphaug

(2000), Downes and Stirling (2001), Ward (2001, 2003), Geist (2005), Geist and Parsons (2006), Liu *et al*. (2007), Burbidge *et al*. (2008), González *et al*. (2009), Power *et al*. (2007), Sørensen *et al*. (2012), and Brizuela *et al*. (2014). The steps followed in computational PTHA are the identification and characterisation of tsunamigenic source recurrences parameters. Such characterisation includes the estimation of the coastline-characteristic maximum possible earthquake magnitude, the creation of tsunamigenic scenario earthquake event catalogues based on identified sources, numerical propagation modelling, and the estimation of potential inundation. The results of all possible scenario events are subsequently combined to create tsunami hazard curves and maps. The techniques used to provide explicit provision for epistemic uncertainty in the PTHA include the use of a logic-tree approach, similar to that used for PSHA (Petersen *et al*., 2002; Annaka *et al*., 2007).

A second procedure is based on Bayesian inferences (Grezio *et al*., 2010, 2012; Yadav *et al*., 2013; Tatsumi *et al*., 2014; Anita *et al*., 2015). This procedure allows the incorporation of all independent prior information pertaining to the tsunami hazard for a specified region. In this way, the knowledge of the physical nature of the process is combined with the likelihood of the process, based on historical event information.

The third procedure focuses entirely on the empirical analysis of historical tsunami event records (Burroughs and Tebbens, 2005; Tinti *et al*., 2005; Burbidge *et al*., 2008). Accordingly, the assessment of the hazard can proceed without prior knowledge of the type or location of the process that triggered the tsunami. If sufficient data are available from tsunami catalogues, site-specific hazard curves can be developed or, alternatively, be used as regional *a priori* information relating to far-field tsunami-generating sources (Geist and Parson, 2006). Using this procedure, the PTHA for an area can be assessed by means of stochastic modelling of the conditional probabilities of tsunami recurrence times (Orfanogiannaki and Papadopoulos, 2007) or by means of frequency–magnitude relationships (Soloviev, 1970; Rikitake and Aida, 1988; Tinti, 1993). Empirical methods, however, are highly dependent on the completeness and quality of the tsunami catalogue.

The major problem with the empirical modelling of tsunami intensity data is that it requires adequate historical and instrumental information for the tsunami run-up heights or intensity values. This problem is discussed in various studies, e.g. Burroughs and Tebbens (2005) and Anita *et al*. (2015). Moreover, small datasets can introduce high levels of uncertainty and bias to the results. The proposed procedure of the current example aims to solve this problem by applying methodology specifically designed to assess recurrence parameters for highly incomplete and uncertain datasets.

The methodology proposed in this study is focused on the empirical assessment of the tsunami recurrence parameters for probabilistic tsunami hazard assessment, based on a methodology similar to that described by Kijko *et al*. (2016) for earthquakes (Chapter 3). These recurrence parameters are the mean tsunami activity rate (rate of occurrence) λ, the Soloviev frequency–intensity power law b-parameter, and the coastline-characteristic, maximum possible tsunami intensity $i_{\max}$. The three tsunami recurrence parameters are derived from tsunami catalogues by taking into consideration the incompleteness of the catalogue, uncertainty in the tsunami intensity determination, as well as the uncertainty associated with the applied tsunami occurrence and intensity models. The uncertainty in the parameters of the tsunami models is introduced by assuming that both the tsunami rate of occurrence λ and the b-parameter are random variables, each described by the secondary mixing distribution, such as the gamma distribution. This approach results in the extension of the classic frequency–intensity Soloviev–Imamura relation and the Poisson distribution of the number of tsunamis with their mixture distribution counterparts (Benjamin, 1968; Campbell, 1982, 1983). This procedure is adapted for application to a tsunami event catalogue to assess the probabilities of exceedance for a specified tsunami intensity i, and the relevant return periods.

Similar to earthquakes, the number of tsunami events observed is related to the intensity (size) of the events. Studies by Soloviev (1970), Houston *et al*. (1977, including references therein), Horikawa and Shuto (1983), Burroughs and Tebbens, (2005), and Geist and Parsons (2006) indicate that tsunami intensity can be described by a power law distribution in the format

$$n(i) = a10^{-bi}, \tag{5.1}$$

where i is the Soloviev–Imamura tsunami intensity, $n(i)$ is the number of tsunami events per annum, and a and b are defined as constants. Defining $\beta = b\ln(10)$ and $\alpha = \ln a$, Eq. 5.1 can be rewritten as

$$n(i) = e^{\alpha-\beta i}. \tag{5.2}$$

Empirical studies show that the tsunami intensity i follows the same power law distribution as the earthquake magnitude m in the instance where the applied scale of tsunami strength is one of the scales, as introduced by Sieberg (1927), Soloviev (1970), Ambraseys (1962), Papadopoulos and Imamura (2001), and Geist and Parsons (2006). Equation 5.2 is, therefore, equivalent to Eq. 2.2, and the classic Gutenberg–Richter frequency–magnitude relation $\log n(m) = a - bm$ (Gutenberg and Richter, 1942; 1956), with m indicating the earthquake magnitude. The power law defined in Eq. 5.1 is not the only model used to describe the distribution of tsunami intensity i. A comprehensive review and discussion of alternative tsunami size distributions were presented by Burroughs and Tebbens (2005) and Geist and Parsons (2006). The Soloviev (1970) scale is used in the current study as it is the standard intensity scale used in most tsunami event datasets.

Regardless of the applied tsunami intensity scale, the intensity i is linked with the average tsunami run-up height h along a coastline. For example, when the intensity i is expressed in the units of the Soloviev scale (Soloviev, 1970), the value of i is calculated from the average tsunami run-up height h (in metres) along a coastline, as

$$i = \log_2\left(\sqrt{2h}\right). \tag{5.3}$$

It is assumed that the tsunami intensities i occurring along a stretch of coastline are continuous, independent, random values distributed according to the power law (Eq. 5.1). Following, for example, Burroughs and Tebbens (2001, 2005), the PDF and the CDF of tsunami intensity take the same functional form as the CDF for earthquake magnitude (Page, 1968; Cosentino *et al*., 1977). The PDF

and CDF are, therefore, defined for intensity i, for a minimum threshold parameter i_{min}, and is truncated from the top by the coastline-characteristic maximum possible tsunami intensity $i_{\max}$, such that Eq. 3.11 (a and b) can be re-written by replacing event size x with the tsunami intensity i.

The tsunami occurrence model is defined based on the assumption that the temporal distribution of the tsunami events observed along a stretch of coastline can be modelled by a Poisson process (e.g. Geist and Parsons, 2006). This corresponds with a similar assumption for the temporal distributions of tsunami-generating processes, such as earthquakes (Cornell 1968; Lomnitz 1973; Gardner and Knopoff, 1974; Cao and Gao, 2002), submarine landslides (Urlaub *et al*., 2013), volcanogenic processes (Dzierma and Wehrmann, 2010), and meteorites (Ward and Asphaug, 2000). Approximately 75% of the global recorded historical tsunamis are attributed to earthquakes (Gusiakov, 2009). The Poisson-gamma distribution, as defined in Eq. 3.24, is applied to determine the mean rate of occurrence. The historic data were assessed using the historical derivation of Eq. 3.28b, substituted into Eq. 3.32b. The individual instrumental datasets of the different LoC were assessed using Eq. 3.34.

5.2.2 Results

The proposed methodology is applied to three tsunamigenic regions in the Pacific Ocean, namely Japan (JAP), Kuril–Kamchatka (K-K), and the western coast of South America (SAM). The international tsunami database used in this study (see Section 1.7) hosts global historical and instrumental tsunami observations for the period 47 BC to 2015, with the main intensity measurement being the Soloviev–Imamura scale, ranging from -5 to 5. The boundaries for the identified tsunamigenic regions were used as provided in the NTL/ICMMG database. These boundaries, depicted in Figure 5.6, follow the zoning used in the NGDC map by Lockridge and Smith (1984) and Gusiakov (2005).

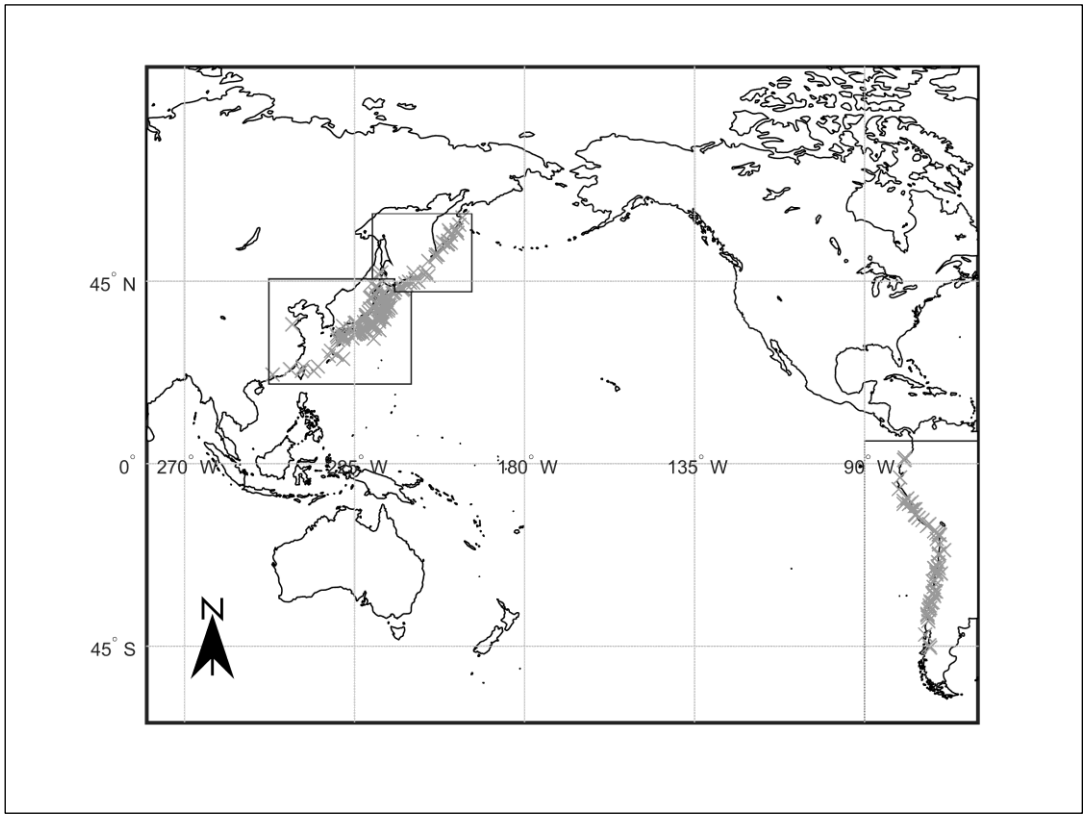


FIGURE 5.6. Positions of the three investigated tsunamigenic regions in the Pacific Ocean (modified after Gusiakov, 2005), being the source locations of the tsunamis for Japan (JAP), Kuril–Kamchatka (K-K), and South America (SAM).

The GITEC catalogue criteria (Tinti and Maramai, 1996; Tinti *et al*., 2001) were used to ensure that the data employed represented the best available information. This criterion assigns a validity index to each observation in the database, such that index value 0 indicates that the event is considered extremely improbable (probability near 0%), 1 is considered improbable (probability approximately 25%), 2 as questionable (probability approximately 50%), 3 as probable (probability approximately 75%), and 4 as a definite tsunami (probability near 100%). The final catalogues for the three identified regions contain only tsunami events with a validity index of 3 or higher. Events with no coordinates (latitude or longitude) or no intensity values were excluded. The final catalogues for Japan, Kuril–Kamchatka, and South America contain 214, 73, and 79 events, respectively.

In the first step, the historical and complete sub-catalogues were identified for each of the three regions. Each sub-catalogue and level of completeness were identified with the help of the visual cumulative method (Mulargia and Tinti, 1985). Provision was made for intensity errors of 0.5 and 0.1 units, respectively, for the historical and instrumental complete catalogues in each of the regions. Table 5.6 provides the input parameters, as well as the estimated recurrence parameters based on the proposed procedure for probabilistic tsunami hazard assessment.

TABLE 5.6. Probabilistic tsunami hazard assessment input parameters and estimated recurrence parameters. The parameter $\boldsymbol{i_{min}}$ represents the level of completeness (LoC) and SE represents the standard error (assumed uncertainty) in the intensity estimation.

		Japan (JAP)	**Kuril–Kamchatka (K-K)**	**South America (SAM)**
Original number of events in identified zones		214	73	79
Dataset information				
Historical	Period	684 AD to 1960	1737–1959	1513–1959
	Number of events	79	6	8
	LoC (i_{min}^{H})	1.00	3.0	3.50
	Intensity SE	0.5	0.5	0.5
Complete	Period	1961–2011	1960–2009	1960–2015
	Number of events	40	21	8
	LoC ($i_{min}^{(i)}$)	-2.00	-0.5	2.0
	Intensity SE	0.1	0.1	0.1
Observed maximum intensity		4.2 ± 0.1	4.0 ± 0.5	4.1 ± 0.1
Estimated recurrence parameters				
Mean rate of occurrence ($\hat{\bar{\lambda}}$)		1.5 ± 0.4	0.3 ± 0.1	0.2 ± 0.1
Frequency–intensity $\hat{\bar{b}}$-parameter		0.4 ± 0.04	0.2 ± 0.1	0.5 ± 0.2
Coastline-characteristic maximum intensity $\hat{\imath}_{max}$		4.3 ± 0.2	4.2 ± 0.5	4.2 ± 0.2

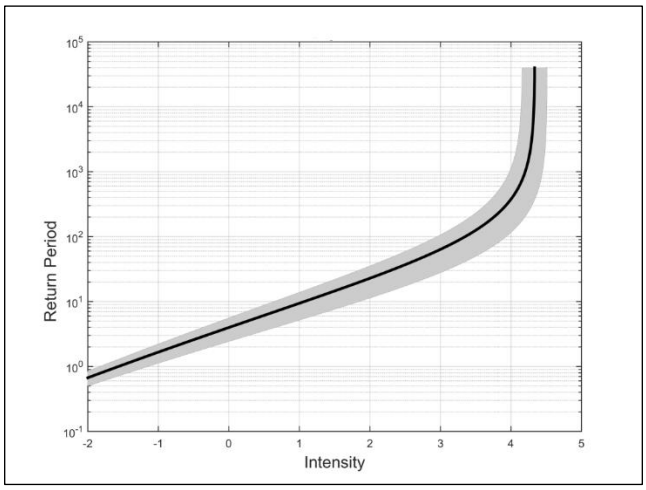
It was assumed that the uncertainty for the tsunami-model parameters was equal to an adjustable, arbitrary value of 25%, meaning that the space–time variation in the estimated recurrence parameters of $\hat{\bar{\lambda}}$ and $\hat{\bar{\beta}}$

(or equivalently the $\hat{\hat{b}}$-parameter) was 0.25. This 25% uncertainty follows from extensive investigations into the analogous parameters of earthquake occurrence models based on tectonic datasets across the globe. Figure 5.7 depicts the estimated mean return period, Figure 5.8 the annual probability of exceedance, and Figure 5.9 the probabilities of exceedance for 5, 10, and 25 years for various tsunami intensities. Figures 5.7 and 5.8 also provide the one standard deviation confidence intervals for the mean return periods and probabilities of exceedance. The values are calculated using the mean and standard deviations from the variance–covariance matrix. From a physical point of view, these confidence intervals reflect the uncertainty of the input data, as well as the uncertainty associated with the parameters in the applied distributions. Table 5.7 provides examples of the return period and probabilities of exceedance for each of the regions for $i \geq 1.5$ and 2.5 and for the time periods 1, 10, and 25 years.

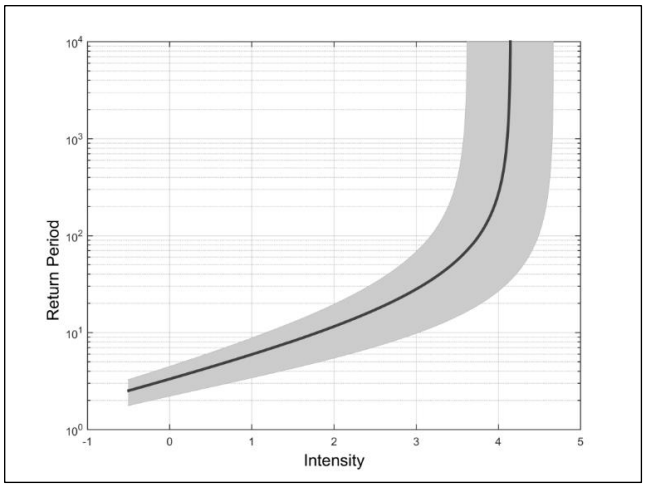
TABLE 5.7. Return period and probabilities of exceedance for Japan, Kuril–Kamchatka, and South America for intensities $i \geq 1.5$, 2.0, and 2.5 and for time periods 1, 10, and 25 years.

Tsunami Intensity	Recurrence variable	JAP	K-K	SAM
$i \geq 1.5$	Return Period (yrs)	15	8	-
	Prob. Exceedance (%): 1 year	7	11	-
	Prob. Exceedance (%): 10 years	49	69	-
	Prob. Exceedance (%): 25 years	80	94	-
$i \geq 2.0$	Return Period (yrs)	23	12	6
	Prob. Exceedance (%): 1 year	4	8	15
	Prob. Exceedance (%): 10 years	35	57	79
	Prob. Exceedance (%): 25 years	65	87	97
$i \geq 2.5$	Return Period (yrs)	37	17	12
	Prob. Exceedance (%): 1 year	3	6	8
	Prob. Exceedance (%): 10 years	24	43	55
	Prob. Exceedance (%): 25 years	48	75	86

(a)



(b)



(c)

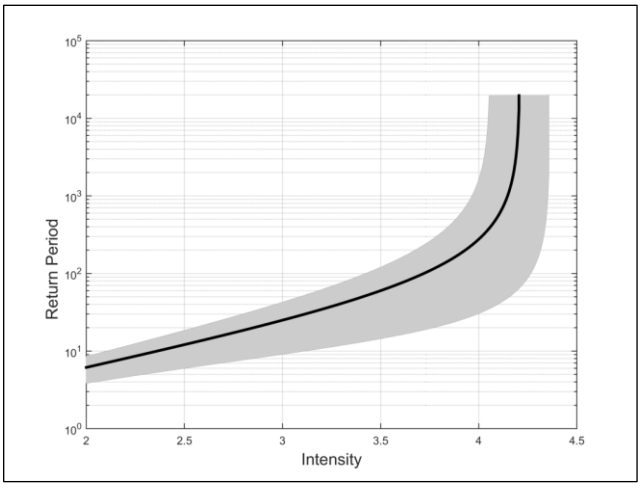
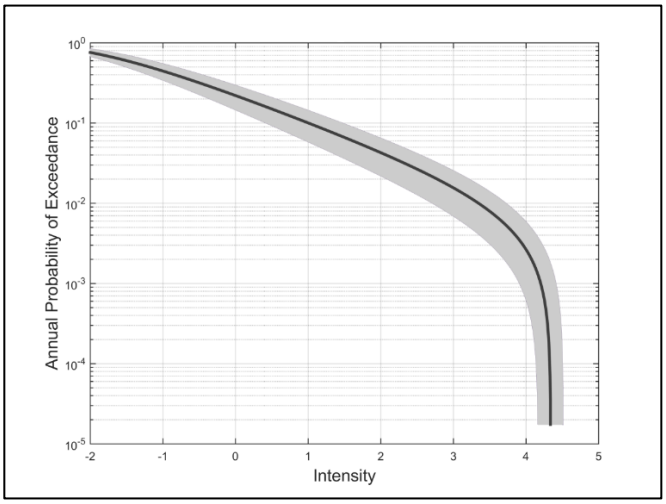
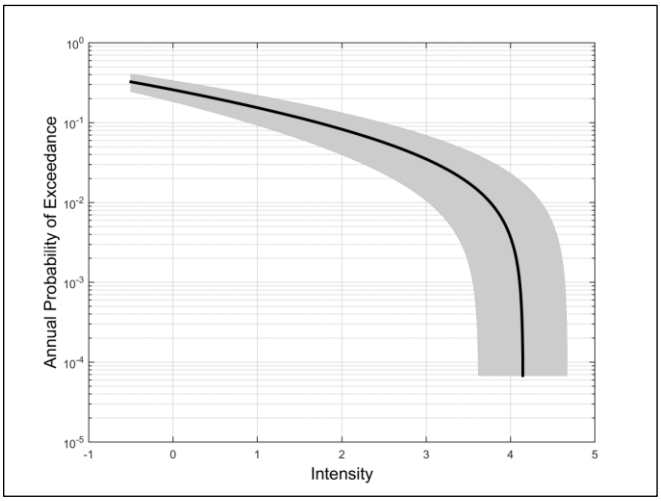


FIGURE 5.7. The annual mean return period (in years) and its associated uncertainty for various tsunami intensities for a) Japan, b) Kuril–Kamchatka, and c) South America. The shaded area represents the one standard deviation confidence interval for the calculated values.

(a)



(b)



(c)

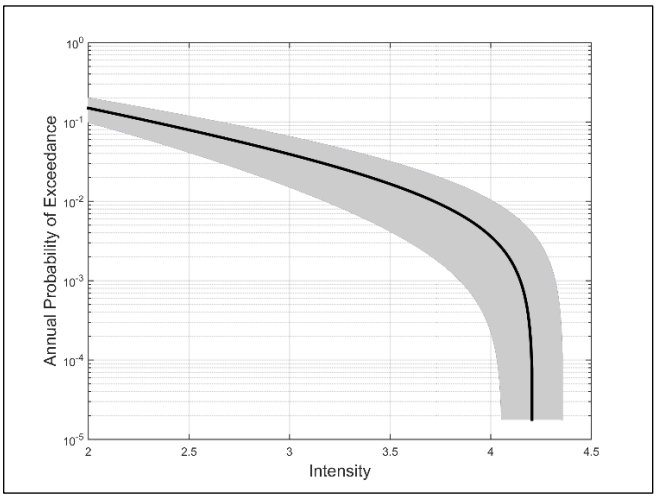
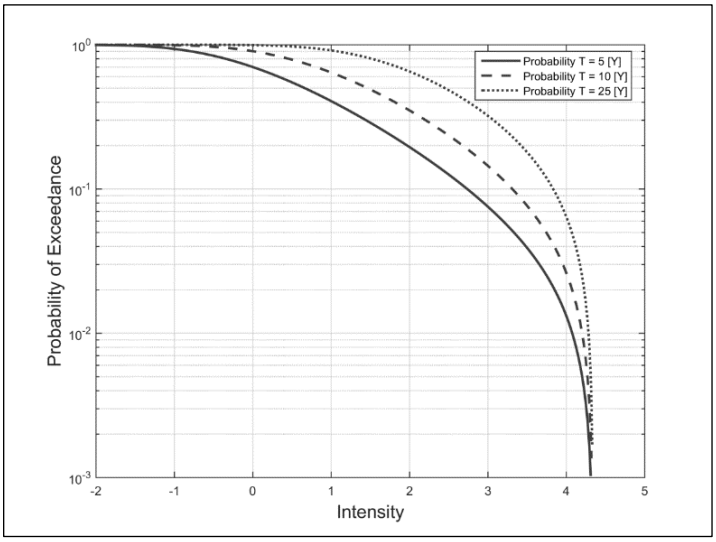
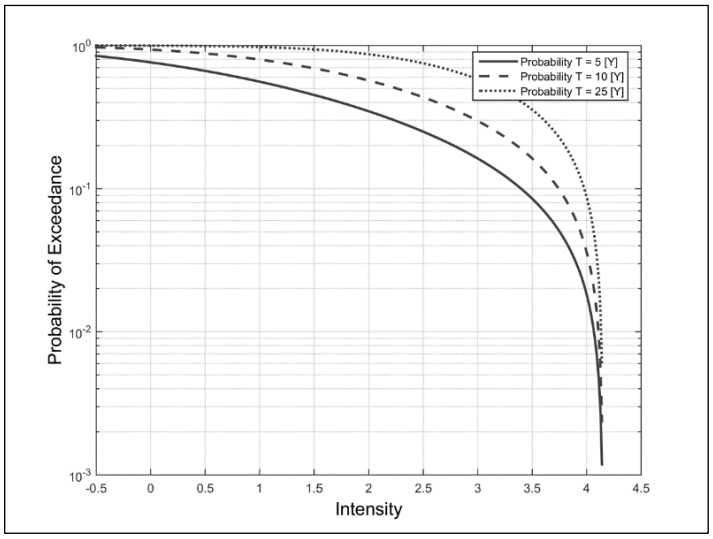


FIGURE 5.8. The annual probability of exceedance and its associated uncertainty for various tsunami intensities for a) Japan, b) Kuril–Kamchatka, and c) South America. The shaded area represents the one standard deviation confidence interval for the calculated values.

(a)



(b)



(c)

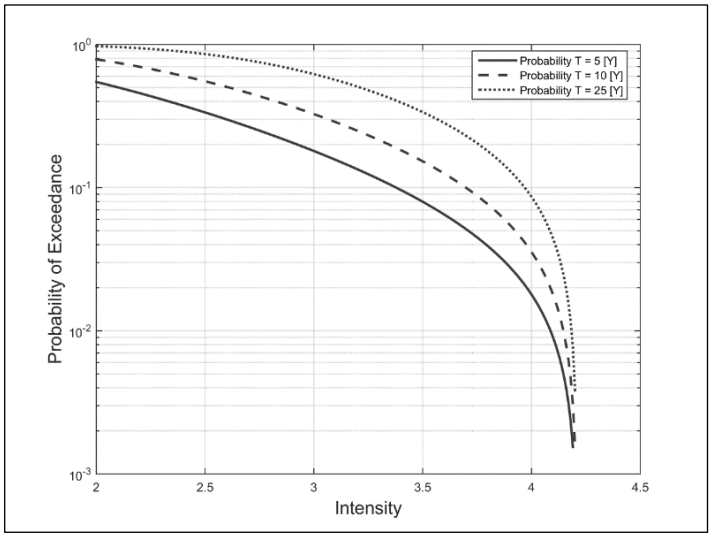


FIGURE 5.9. The probability of exceedance for 5, 10, and 25 years for various tsunami intensities for a) Japan, b) Kuril–Kamchatka, and c) South America.

The results obtained were compared with the b-parameter estimates obtained by Orfanogiannaki and Papadopoulos (2007). These authors investigated the regions of Japan, Kuril–Kamchatka, and South America by developing stochastic approaches to utilise observed earthquakes in order to assess the potential for tsunami generation. They investigated the probability of tsunami occurrence, firstly, by using conditional probabilities for tsunami occurrence as a function of time and, secondly, by obtaining the product of the conditional probabilities with the ratio of the number of earthquakes generating tsunamis and the total number of earthquakes by applying the total probability theorem. Orfanogiannaki and Papadopoulos (2007) utilised the database, as described in Gusiakov (2001), for the time periods 1900 to 2000. Their $\hat{b}$-parameters, obtained with the help of linear regression for the three regions of interest, are 0.34 for Japan, 0.56 for Kuril–Kamchatka, and 0.34 for South America. The associated estimated mean rates of occurrence $\hat{\hat{\lambda}}$ for the three regions are 1.02 for Japan, 2.26 for Kuril–Kamchatka, and 1.64 for South America. The results obtained for Japan are comparable to those obtained by Orfanogiannaki and Papadopoulos (2007), but not to those for Kuril–Kamchatka and South America. In addition, Orfanogiannaki and Papadopoulos (2007) and Koravos *et al*. (2015) investigated the tsunami hazard for these areas by calculating the conditional probabilities of occurrence for specified tsunami intensities. These values are not comparable to the probabilities of exceedance calculated in the present study.

A question that immediately arises is what the difference will be between the results of the PTHA using the mixture distributions in comparison with the results provided by the same model when no uncertainty in the parameters of the models is taken into consideration. As a test, the data for Japan were applied to the defined mixture distributions (Poisson-gamma and exponential-gamma), as well as to the principal distributions of Poisson and exponential. Japan was chosen, as it is the dataset with the most observations to illustrate the effect of the validity index. The results are provided in Table 5.8. A noteworthy increase can be seen in the estimated parameters and probabilities of exceedance, as well as a substantial decrease in the return periods. The estimated hazard, therefore, increased when uncertainty regarding the parameters of the applied distributions was taken into consideration.

TABLE 5.8. Comparison of the model for Japan with and without the application of mixture distributions to account for uncertainty in the parameters of the applied distributions.

Estimates	**With mixture distributions (WMD)**	**Without mixture distribution (WOMD)**	**Percentage increase (WMD-WOMD)/WOMD**
Rate of occurrence	$\hat{\bar{\lambda}} = 1.5 \pm 0.4$	$\hat{\lambda} = 0.9 \pm 0.1$	64%
Frequency–intensity b-parameter	$\hat{b} = 0.4 \pm 0.04$	$\hat{b} = 0.34 \pm 0.02$	18%
Coastline-characteristic maximum intensity i_{max}	4.35 ± 0.17	4.35 ± 0.18	0%
Return Periods			
$i \geq 1.5$	15 yrs	19 yrs	-24%
$i \geq 2.0$	23 yrs	30 yrs	-25%
$i \geq 2.5$	37 yrs	50 yrs	-25%
Prob. of Exceedance: 1 year			
$i \geq 1.5$	7%	5%	24%
$i \geq 2.0$	4%	3%	24%
$i \geq 2.5$	3%	2%	25%
Prob.of Exceedance: 10 years			
$i \geq 1.5$	49%	40%	17%
$i \geq 2.0$	35%	28%	20%
$i \geq 2.5$	24%	18%	22%
Prob. of Exceedance: 25 years			
$i \geq 1.5$	80%	73%	10%
$i \geq 2.0$	65%	56%	14%
$i \geq 2.5$	48%	40%	18%

To illustrate the effects of including the uncertainty of an event occurrence by using a validity index, as well as implementing Bayesian inference as the estimation method, these were incorporated into the assessment of the tsunami recurrence parameters for the Japan tsunami dataset. In the previous example (Smit *et al*., 2017), all the events with a validity index less than 3 were ignored. For comparison purposes, the dataset preparation followed the same process, with the exception that all events were included, regardless of the validity index. Table 5.9 shows the input information for the analysis used in this example, where the validity index and BI were used to estimate the recurrence parameters.

Table 5.10 provides a comparison between the estimated recurrence parameters using MLE estimation (Smit *et al*., 2017) that considered model uncertainty and tsunami intensity uncertainty, with the inclusion of the validity index, as well as the Bayesian estimates with the validity parameters. The maximum possible event size was assumed known as $i_{max} = 4.3 \pm 0.2$ (Table 5.6), as estimated in Smit *et al.* (2017).

TABLE 5.9. Probabilistic tsunami hazard assessment input parameters for Japan.

	Smit *et al*. (2017) datasets. Excluding events with validity index ≤ 2.	**New datasets. Including all events.**
Historical Period	684 AD to 1960	684 AD to 1960
Number of events	79	118
LoC (i_{min}^{H})	1.0	1.0
Intensity SE	0.5	0.5
Complete Period	1961–2011	1961–2011
Number of events	40	43
LoC ($i_{min}^{(i)}$)	-2.0	-2.0
Intensity SE	0.1	0.1
Observed maximum intensity	4.2±0.1	4.2±0.1
Percentage variation	25	25
a priori information for $\hat{\hat{b}}$-parameter mean	None	0.34
a priori information for $\hat{\hat{b}}$-parameter SE	None	0.034
Coastline-characteristic intensity (i_{max})	4.3±0.2	4.3±0.2

The introduction of the validity index allows the use of more events in the estimation process compared with Smit *et al.* (2017), which ignored events with a small index. This, combined with the introduction of prior information, has a decreasing effect on the estimated return periods, which increases the hazard. In addition, both the validity index and the prior information have a decreasing effect on the expected probabilities of exceedance for the tsunami intensity sizes compared with the MLE estimates in Smit *et al*. (2017).

TABLE 5.10. Return periods and probabilities of exceedance for tsunami intensities $i \geq 1.5$, 2.0, and 2.5, and percentage contribution of datasets to the estimates. Results are provided for the time periods 1, 10, and 25 years.

Estimated recurrence parameters	**Smit *et al*. (2017)**	**MLE**	**Bayesian inference**
Model accounts for the following uncertainties:			
Model parameter uncertainties	Yes	Yes	Yes
Uncertainty in intensity size measurement	Yes	Yes	Yes
Uncertainty in the occurrence of events	No	Yes	Yes
Estimation technique	MLE	MLE	Bayesian inference
Mean rate of occurrence ($\hat{\hat{\lambda}}$)	1.5 ± 0.4	2.4 ± 0.8	1.8 ± 0.4
Frequency–intensity $\hat{\hat{b}}$-parameter	0.4 ± 0.04	0.4 ± 0.05	0.4 ± 0.03
$\boldsymbol{i \geq 1.5}$			
Return period	15	11	10
Prob. exceedance 1 year	7	0	0
Prob. exceedance 10 years	49	1	1
Prob. exceedance 25 years	80	1	1
$\boldsymbol{i \geq 2.0}$			
Return period	23	17	16
Prob. exceedance 1 year	4	0	0
Prob. exceedance 10 years	35	0	1
Prob. exceedance 25 years	65	1	1
$\boldsymbol{i \geq 2.5}$			
Return period	37	29	26
Prob. exceedance 1 year	3	0	0
Prob. exceedance 10 years	24	0	0
Prob. exceedance 25 years	48	1	1
% contribution			
Historical data to $\hat{\hat{b}}$-parameter	92.9	94.3	73.1
Complete data to $\hat{\hat{b}}$-parameter	7.1	5.7	5.3
a priori information to $\hat{\hat{b}}$-parameter	0	0	21.6
Historical data to $\hat{\hat{\lambda}}$	98.5	94.9	98.5
Complete data to $\hat{\hat{\lambda}}$	1.5	5.1	1.5
a priori information to $\hat{\hat{\lambda}}$	0	0	0

The prehistoric, historical, and instrumental data play different roles, as observed for the simulated dataset. The historical dataset spans more than 1 000 years compared with 50 years for the assumed instrumental dataset. In this example, both estimates $\hat{\hat{\lambda}}$ and $\hat{\hat{\beta}}$ rely heavily on the historical dataset (Figure 5.10) that is substantially larger than the instrumental dataset. In this instance, introducing the validity index and prior information not only increased the hazard estimates with short return periods but also decreased the probabilities of exceedance of the different tsunami intensities.

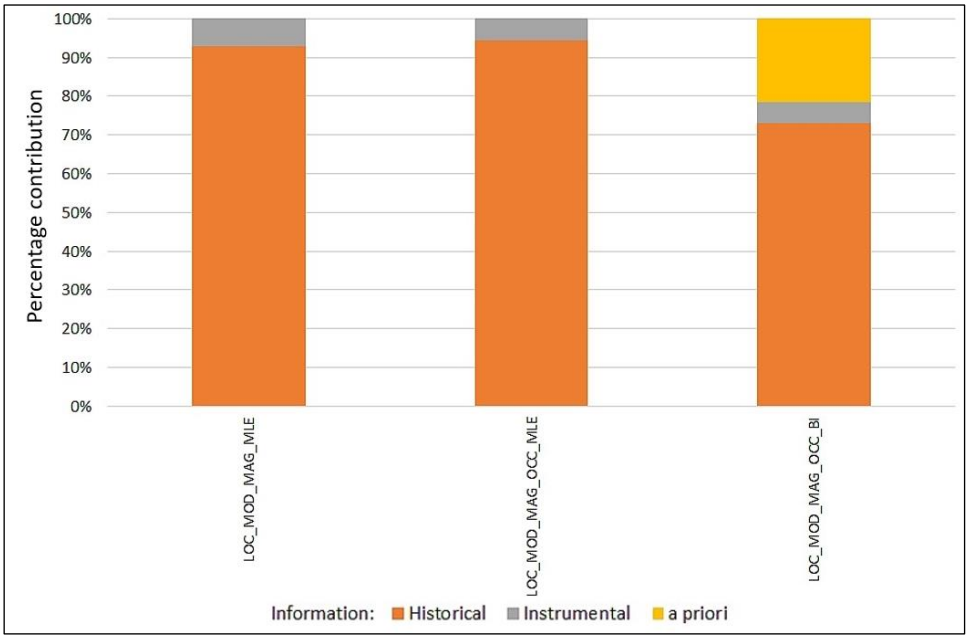


FIGURE 5.10. Percentage contribution of each sub-dataset to MLE and BI estimates of the $\hat{\hat{b}}$-parameter taking into consideration the validity index associated with the Japan tsunami dataset.

5.2.3 Discussion

Considerable effort has been devoted over many years to assessing the tsunami hazard at vulnerable coastal areas, yet, large-scale uncertainty remains associated with the applied methodologies and research results. This uncertainty is attributable in part to the different generating mechanisms, ranging from near- and far-field earthquakes to submarine slumps and volcanoes. In addition, the bathymetry

of the affected local coastline influences the intensity and height of the tsunami wave, either increasing or reducing the force of the incoming wave. Ultimately, the small number of events in any tsunami catalogue increases the number of assumptions that have to be made in tsunami hazard assessment process.

A methodology is presented to assess the coastline-characteristic tsunami hazard based on empirical catalogues. The methodology, originally defined in terms of probabilistic earthquake hazard assessment, is highly flexible, allowing for the transition to PTHA. It considers only the observed tsunami intensity and makes no assumptions about the source of the tsunami. The formalisms focus on assessing the tsunami recurrence parameters, the tsunami rate of occurrence λ, and the Soloviev–Imamura frequency–intensity power law b-parameter based on incomplete catalogues, the uncertainty in the tsunami intensity determination, as well as the uncertainty associated with the parameters of the applied tsunami occurrence and intensity models. The procedure explicitly allows for both aleatory and epistemic uncertainty by introducing mixture distributions, where the recurrence parameters are assumed to be random variables that can fluctuate. For this purpose, the gamma distribution was introduced as the mixing distribution in both the tsunami occurrence and the intensity distributions. The ability to view the recurrence parameters as random variables allows the user to define models that closely echo reality. As seen in Table 5.8, the exclusion of parameter uncertainty has a noteworthy effect on the estimated hazard, indicating that the simpler models can lead to the underestimation of the tsunami hazard for a region.

Furthermore, the recurrence parameters $\hat{\bar{\lambda}}$ and $\hat{\bar{\beta}}$ for Japan were estimated using the validity index and BI, while keeping the maximum possible tsunami intensity constant to that estimated in Smit *et al.* (2017). In the synthetic earthquake example (Chapter 4), the rate of occurrence prefers more-recent information compared with the power law parameter that is partial to extreme events to dictate its slope. More emphasis is placed on prior information, as more variation is explicitly accounted for. The opposite is observed when evaluating the tsunami dataset for Japan. Here, both $\hat{\bar{\lambda}}$ and $\hat{\bar{\beta}}$ rely heavily on the historical dataset that is substantially larger than the instrumental dataset. In this instance, the

introduction of the validity index and prior information not only increased the hazard estimates with short return periods, but also decreased the probabilities of exceedance of different tsunami intensities. The examples used historical and complete catalogues, but prehistoric tsunami information can also be included in the analysis.

It is safe to assume that the other parameters associated with tsunamis, such as the direction of the wave propagation and the local bathymetry, would affect the hazard estimates, resulting in locally adjusted site effects. The accuracy of the estimated recurrence parameters can be enhanced by incorporating these site effects by means of Bayesian statistics, similar to the procedure followed by Grezio *et al.* (2010, 2017) and Yadav *et al*. (2013). A posterior probability is created by combining physical knowledge of the tsunami process, as *a priori* probability, with the likelihood function $L_{Total}(\boldsymbol{\psi})$.

Several factors have to be considered when this proposed methodology is used. Firstly, the formalisms are dependent on the choice of LoC. This concern can be addressed potentially through the application of formalisms similar to those by Kijko and Smit (2017), who provide two new equations for the estimation of seismic recurrence parameters without using LoC. In addition, the uncertainty associated with the intensity error influences the derived recurrence parameters, as does the choice of uncertainty associated with the occurrence model. The choice of the frequency–intensity distribution could also affect the results, as these formalisms do not consider the local bathymetry conditions. Care has to be taken to ensure that a model is presented that is as objective as possible and, in addition, can accommodate the various uncertainties without severely over- or underestimating the hazard.

5.3 Hail

This example examines both hail hazard and hail risk by using the defined methodology. Hail hazard is determined by using hail size estimates derived from ensemble HAILCAST modelling run on ERA-ITERM re-analysis data (Dyson *et al*., unpublished). Hail risk is analysed using insured losses based on observed insurance claims from Momentum Short-Term Insurance Company (MSTI) in South Africa. This demonstrates how both the hazard and the risk can be described by the same recurrence parameters, and how the hazard results can serve as *a priori* information for risk. These next sections provide background information on both hail hazard and risk, assessments of the hail hazard for Gauteng province, South Africa, and the associated hail risk for vehicles.

5.3.1 Introduction

Hail is one of the largest single-event natural hazards and can cause extensive damage. Global cumulative insured losses often exceed USD 1 billion (Changnon, 2008; Gunturi and Tippet, 2017). Insured losses cover the damage to assets of residential and commercial infrastructure, agriculture, and vehicles. In some countries, as is the case in Australia, hailstorms contribute to the most frequent and largest natural catastrophe losses for the insurance industry (Eichner, 2015). Meteorologists, geographers, economists, and the engineering, construction, and agriculture sectors are among the industries that can benefit from understanding hail frequency and characteristics (Punge and Kunz, 2016). Each of these industries continuously tries to understand the underlying phenomena but also to design and plan for their respective economic safety and profitability. Agriculture is typically the most susceptible to hail, even with smaller size events, as entire crops can be wiped out during a single storm and loss of livestock can occur. Severe convective storms can cause notable economic damage to property and, in rare instances, severe injuries and even the loss of human life. The economic losses can be crippling, e.g. as with the events of 8–10 June 2014 over the northern parts of France, and the central and western parts of Belgium, with the insured losses amounting to 2.3 billion Euro. During the

same year, an event was recorded in Brisbane, Australia, that caused damage of AUD 1.35 billion. References to these and other insured losses related to hailstorms can be found in publications such as Munich Re (Eichner, 2015) and Punge and Kunz (2016).

Hailstorms are high-energy, convective weather systems mostly associated with thunderstorms, although the reverse is not necessarily true. Ice crystals form because of the freezing of moist vapour (rain) caught in strong updrafts during the storm. As the storm updrafts continue, the ice crystals accumulate more ice, with irregular-shaped hailstones being formed (Punge and Kunz, 2016). Once these stones are heavy enough to overcome the updrafts, the hailstones will start dropping to Earth. In many instances, these hailstones will melt before reaching the ground, becoming raindrops again. As convective storms are continuously moving, an elongated footpath of hail is formed. Changnon *et al.* (2009) and references therein provide a review of the formation of hail and hailstorms and their occurrence in the USA, whereas Punge and Kunz (2016) present a review of hail in Europe.

Several factors can affect the type of damage that a hailstorm can cause. These include the frequency and size of the stones (diameter), kinetic energy, ground wind speed, the amount of hail per measured area (volume), the shape of the hailstones, and the physical properties of the source of impact (Webb *et al.*, 2009; Changnon *et al*., 2009; Punge *et al.*, 2014, and references therein). A combination of these factors determines the intensity of the storm. The volume of the hailstones and the ground wind speed during the storm are difficult to measure and vary greatly between locations. Detailed studies are required on these two factors, and they are rarely included in hail hazard analyses.

The closely correlated factors of size and kinetic energy are strong indicators of the type of damage that can be expected (e.g. Leigh, 1998; Punge *et al.*, 2014). The TORRO (the Tornado and Storm Research Organization) Hail Intensity Scale[19] introduced by Webb *et al*. (1986) and revised by Sioutas *et al.* (2009), provides a good indication of the type of damage that can be associated with various hail sizes

[19] http://www.torro.org.uk/hscale.php; last accessed 2018/721.

and kinetic energy. In addition, the TORRO scale can be modified slightly to reflect the effect of hail on the building materials of a particular country. Table 5.11, as presented in Webb *et al.* (2009, Table 1), serves as an example of this scale that was developed to reflect hail damage observed in Britain.

TABLE 5.11. TORRO Hail Intensity Scale (Webb *et al*., 2009).

Level	Intensity Category	Typical Hail Diameter (mm)[a]	Probable Kinetic Energy (J-m^2)	Typical Damage Effects
H0	Hard hail	5–9 (pea)	0–20	No noticeable damage
H1	Potentially damaging	10–15 (moth ball)	>20	Slight general damage to fruit, crops
H2	Significantly damaging	16–20 (marble, grape)	>100	Significant general damage to fruit, crops, vegetation
H3	Severe	21–30 (walnut)	>300[b]	Severe general damage to fruit, crops, damage to glass and plastic structures, paint and wood scored
H4	Severe	31–40 (pigeon egg> squash ball)	>500[c]	Widespread glass damage, vehicle bodywork damage
H5	Destructive	41–50 (golf ball>pullet egg)	>800	Extensive destruction of glass, damage to tiled roofs, significant risk of injuries
H6	Destructive	51–60 (chicken egg)		Bodywork of grounded aircraft dented, brick walls pitted
H7	Destructive	61–75 (tennis ball>cricket ball)		Severe roof damage, risk of serious injuries
H8	Destructive	76–90 (large orange> soft ball)		Severe damage to aircraft bodywork
H9	Super hailstorms	91–110 (grapefruit)		Extreme structural damage. Risk of severe or fatal injuries to person caught in the open.
H10	Super hailstorms	>110 (melon)		Extreme structural damage. Risk of severe or fatal injuries to persons caught in the open.

[a]Approximate range (typical maximum size in bold), since other factors (e.g. number and density of hailstones, hail fall speed, and surface wind speeds) affect severity. For non-spheroidal hailstones, the diameter refers to the mean of the coordinates.

[b]Vinet (2001) noted that kinetic energies > 300 J-m^2 were associated with total crop losses in France

[c]Fraile *et al*. (2003) indicated that "very severe" hailfalls were those with kinetic energy >500 J-m^2 and/or maximum hail size > 30 mm in diameter.

More-specific intensity scales can be developed, e.g. an intensity–susceptibility index for various crops (e.g. Changnon *et al*., 2009). This is a varying index, which describes the effect on crops from hail impact, based on the point in time in the lifespan of the plant, as well as the point in time of the hail season (Changnon, 1967; Changnon and Stout, 1967; Olivier and Van Rensburg, 1992). These indices are dependent on the type of crops and hail climatology of the area under investigation.

Hail hazard is expressed often as the number of incidences, or hail frequency, in a specific area over a predefined time period. This includes reviews of spatial variability in annual and seasonal patterns. Hail frequencies can be based on point or area relationships. Point relationships focus only on the point of measurement, whereas area relationships consider the entire affected area. This is done by using hailstrips, a narrow area with high winds and high amount of hailstones; hailstreaks, an area with continuous spatial and temporal coverage of hailstones; and hailswathes, an area where two or more hailstreaks are observed within a certain distance and time (Changnon *et al.*, 2009).

Over the years, several sources for hail observations have been used in an effort to establish hail hazard or climatologies and their links to insurance claims (e.g. Leigh and Kuhnel, 2001; Sander *et al*., 2013; Changnon, 2008; Brown *et al*., 2015; Allen *et al*., 2017). These include using insurance data, national weather spotter services, specialised field campaigns (e.g. Changnon *et al*., 2009, p 12), hailpad observations, and observations from weather stations (Allen *et al*., 2017 and references therein). Hailpads[20] are useful in identifying the location, hail size, and kinetic energy at the moment of impact. Establishing such a network, however, is expensive and problems could be encountered, such as hailpad saturation, identification of the hail shapes, drag, and spatial incompleteness. Using insurance data to supplement hail observational data can be complicated. The focus of the industry is often on the overall cumulative losses of a given portfolio in a specific period of time and less on the loss from a single hailstorm. This can lead to fewer events being counted than are observed in meteorological terms. Moreover, these datasets are subject to low spatial and temporal resolution, incompleteness, and

[20] https://definedterm.com/hailpad/286788, last accessed 2019-05-26.

heterogeneous observational systems (Punge and Kunz, 2016). Furthermore, reporting bias exists, as small hail occurrences can go unreported owing to lack of equipment or low population density at the location of the event (Allen and Tippett, 2015; Allen *et al*., 2015).

The lack of reliable observational data has led to the development of alternative methods to estimate hail climatologies, assisting researchers in understanding the characteristics of hail for future predictions. These methods rely on using proxy data from remote sensing instruments of numerical weather prediction systems. Weather radars are popular because of their high spatial and temporal resolution (e.g. Punge and Kunz, 2016), and can give some indication of hail events. Numerical weather prediction systems are used to build hail climatologies based on atmospheric conditions. Since the early years of the current century, climatologists have employed parameters from re-analysis data combined with observed hail events to develop area-specific hail climatologies (e.g. Brooks *et al*., 2003, 2007; Allen and Karoly, 2014; Dyson *et al.*, unpublished).

Applied methodologies for assessing hail climatologies focus mainly on hail frequency analyses (e.g. Punge and Kunz, 2016) that characterise an area according to the number of hail event days in a month or a year. Several authors have attempted to link hail occurrences and size with other environmental factors by using synoptic composites and station proximity analyses. Other methods include models, such as HAILCAST, Regional Climate models (RCM), or Poisson regression (e.g. Allen *et al.*, 2015). Kunz and Puskeiler (2010) combined 3D/2D radar, lightning, radio sounding, and insurance data to identify hailstorm tracks and they evaluated the hazard using EVT and generalised Pareto distributions to assess the return periods. By applying the Gumbel I distribution to observations from US hail reports, Allen *et al.* (2017) developed spatial hazard maps for the continental US, consisting of hail event size return intervals for grid sizes of $1^\circ \times 1^\circ$. The applied dataset is homogenised for the entire US and only complete information, with hail sizes larger or equal to 25 mm (1 inch) was used. Smaller event sizes that could have a detrimental effect on agricultural crops and vegetation were ignored. The events used were for the period 1979 to 2013, and historical information between 1955 and 1978 was ignored because of spatial and temporal incompleteness.

Distributions used in modelling the maximum hail size per storm appear to differ based on the measurement techniques. These include the exponential and power law probability distribution functions, inverse Rayleigh, and the gamma distribution (Punge and Kunz, 2016). Large uncertainty is associated with choosing the type of distribution to use because of the various types of data uncertainty, whether observational or derived from empirical ensemble models.

5.3.2 Hail hazard

South Africa has both summer and winter rainfall regions. In the austral summer rainfall season (October–March), convective thunderstorms are an almost daily occurrence, usually a single event lasting less than one hour over any particular location (Visser and Van Heerden, 2000). Hail events are observed typically for the months September to December (Pienaar *et al*., 2015) over the Highveld of South Africa. This region includes the densely populated Gauteng province. Most of the severe and damaging events occur in the late afternoons or early evenings. On 1 November 1985, a severe hailstorm struck Pretoria CBD and surrounding areas. Collapsed roofs, damage to vehicles, and the windows of buildings were observed, with a total estimated financial loss of ZAR 400 million (unadjusted) (Terblanche, 1985; Caellum, 1991; Grobler, 2001). Recent severe hail events in Gauteng include the hailstorms in 2012 and 2013 that wreaked havoc on motorists and infrastructure, with financial losses in excess of ZAR 2 billion ($\pm$ USD 160 million). The loss resulting from the hailstorm on 27 November 2013 was double that suffered in Edenvale (Greater Johannesburg region) from the hailstorm of 2012 (ZAR 1 million). The hailstorm events in 2013 were severe, causing power disruptions and extensive damage to buildings, houses, and vehicles (AON, 2014; PWC 2014), with the event of 28 November 2013 described as '*the single worst insurance event in South Africa's history'* (PWC, 2014). Upon further investigation, Pienaar *et al.* (2015) found significant circulation pattern anomalies in the atmospheric properties observed during the 2012 and 2013 hail seasons compared with the period 1979–2011. Such anomalies could potentially serve as warnings of higher frequencies of damaging hail events during a particular season.

Similar to the rest of the world, South Africa has a lack of reliable hail observations. The South African Weather Service (SAWS) that maintains official hail information, has a limited number of manned stations able to record hailfall at or close to the station. Moreover, event sizes are only recorded when the diameter is larger than 5 mm. Such lack of observations obviously makes the development of a hail climatology particularly difficult. The format of the currently observed hail climatological data precludes the successful application of the methodology described in Chapter 3.

In an attempt to build a hail climatology for the period 1979–2017 over South Africa, (Dyson *et al*., unpublished) used the HAILCAST model first developed for South Africa by Poolman (1992) and later updated by Brimelow *et al.* (2002) for Canada. The HAILCAST model is a one-dimensional coupled model using a steady-state cloud model and a time-dependent hail-growth model. HAILCAST conducted ensembles of frequency and maximum hail size diameters using ERA-ITERM pseudo–proximity soundings over South Africa at a resolution of 0.75°.

The hail hazard is analysed using four grid points 339, 339, 367, and 368 that covers Gauteng. These grid points were provided by the HAILCAST model. The grid points were not combined due to topographical changes in Gauteng that can influence the recurrence parameters. The hail size used for analysis consists of the mean hail size over the ensemble data for each day. All events of size less than 1 cm were discarded, as it is difficult to distinguish between hail and ice rain for such small sizes. At this stage, the raw ensemble data were not available to assess the standard deviation of the mean hail size predicted by the model. The ensemble data for a single grid point was analysed and yielded a mean standard deviation of $\sigma_x = 0.1$ cm for hail size events. It was, therefore, assumed that the event size uncertainty could be modelled using the Gaussian distribution, with a mean equal to the observed event size and standard deviation of 0.1 cm. Analysing the log of the cumulative frequency of event sizes yielded a percentage variation in the data of approximately 5% at each grid point. The shifted-truncated distribution was applied that takes into account a single dataset, with parameter and event size uncertainty (M4 in Table 3.1).

Table 5.12 provides the input parameters for the four grid points, as depicted in Figure 5.11. For this example, it was assumed that there was no prior information available and MLE was used to estimate the recurrence parameters. The Kijko–Sellevoll–Bayes method was used to estimate the maximum possible hail size. The estimated parameters are provided in Table 5.13. Figure 5.12 provides the cumulative frequency–event-size behaviour, return period, and probabilities of exceedance for 5, 10, and 20 years for each of the four grid points.

A discussion of the results is available in Section 5.3.4. In the next section, the results of the probabilistic hail hazard assessment will be used as *a priori* information in the assessment of the probabilistic hail risk for the same grid points in Gauteng.

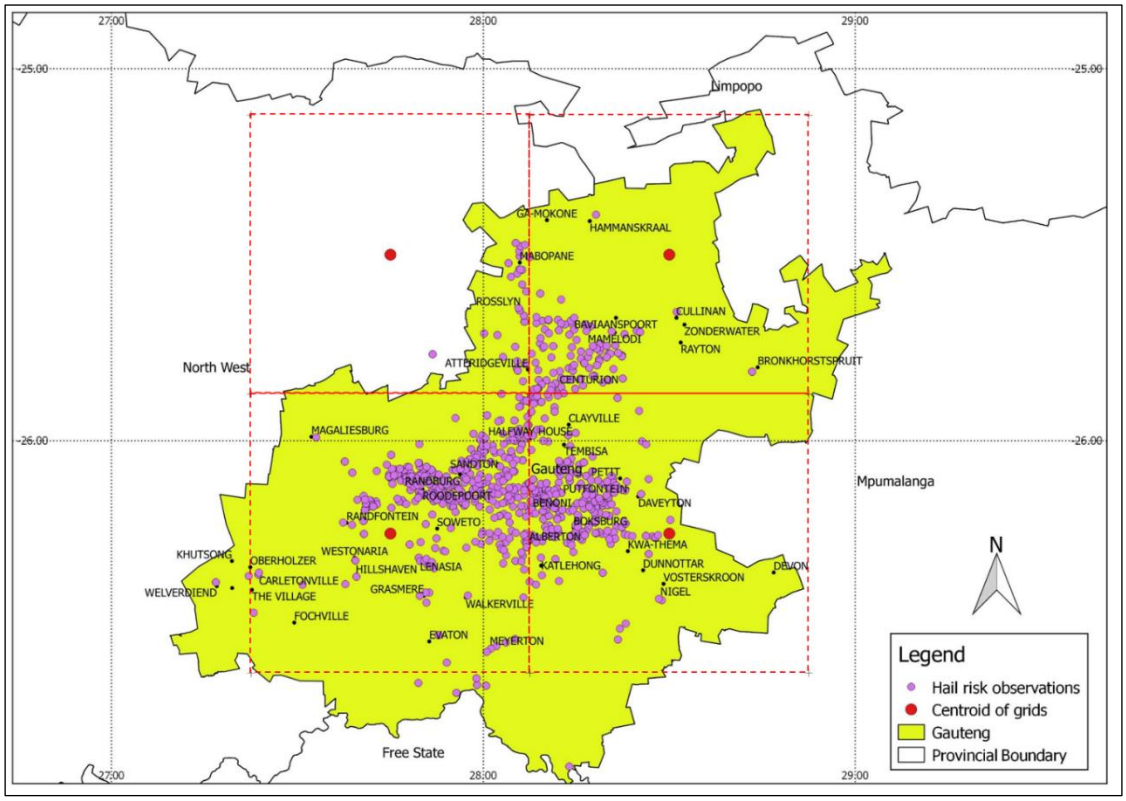


FIGURE 5.11. Map of Gauteng with the centroids of the grid points used in the HAILCAST model. The hail hazard is determined for each of these grid points. Grid 338 is the bottom left quadrant, grid 339 the bottom right, grid 367 the top left quadrant, and grid 368 the top right. The locations of incurred insurance losses at suburb level are shown in purple.

TABLE 5.12. Probabilistic hail hazard assessment input parameters for Gauteng, South Africa, based on ensemble HAILCAST data for the period 1979/06/10 to 2016/12/31.

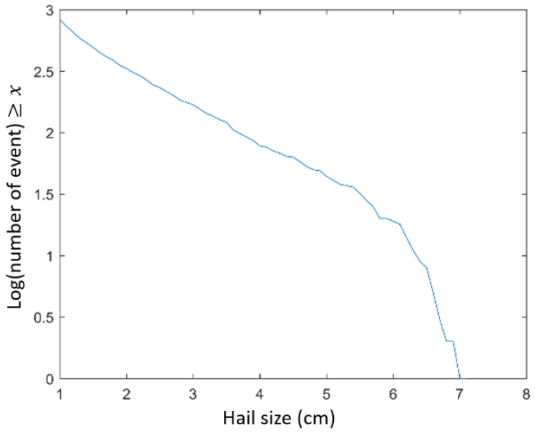
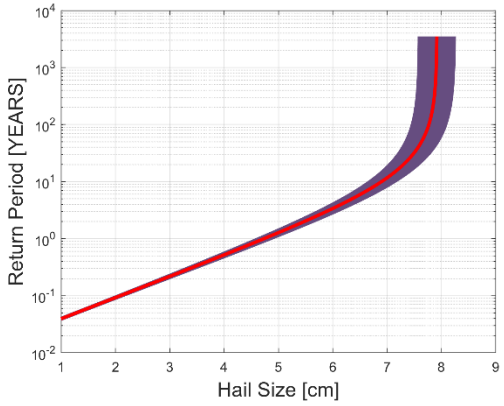
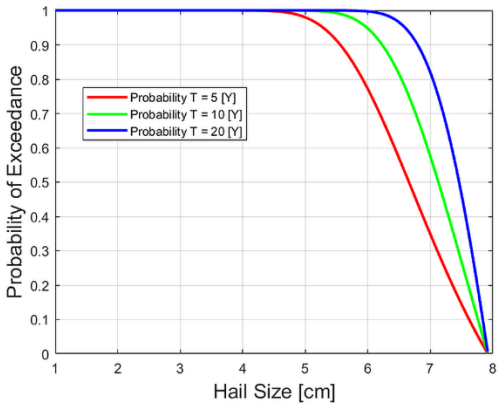
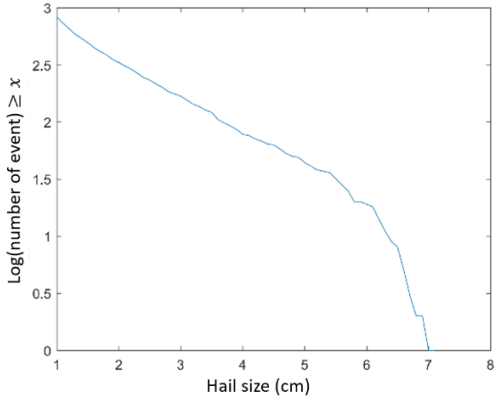
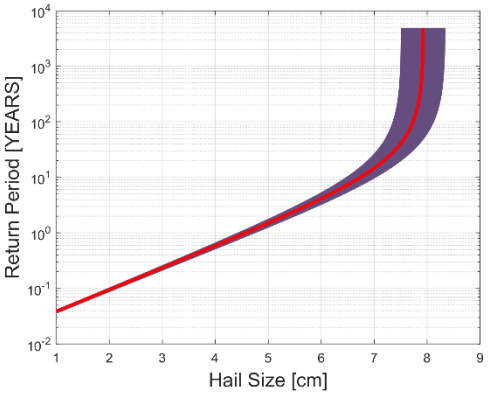
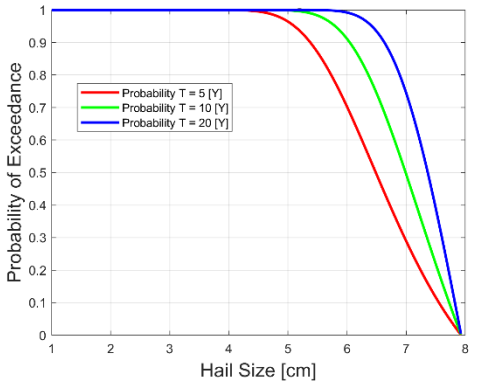
	Grid 338	**Grid 339**	**Grid 367**	**Grid 368**
GPS centre points	[-26.25, 27.75]	[-26.25, 28.5]	[-25.5, 27.75]	[-25.5, 28.5]
Number of events	951	976	556	525
LoC (in cm)	1.0	1.0	1.0	1.0
Hail size SE* (in cm)	0.1	0.1	0.1	0.1
Observed maximum hail size (in cm)	7.59±0.1	7.53±0.1	6.72±0.1	6.3±0.1
Percentage variation	5	5	5	5

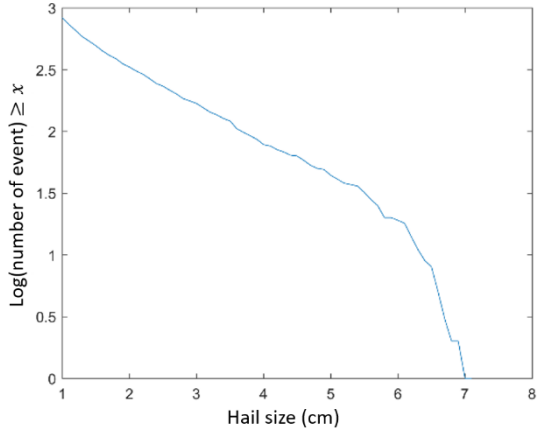
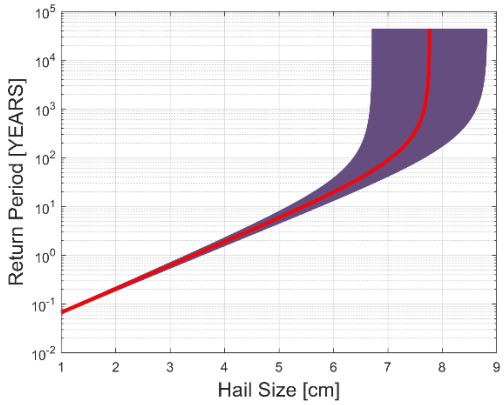
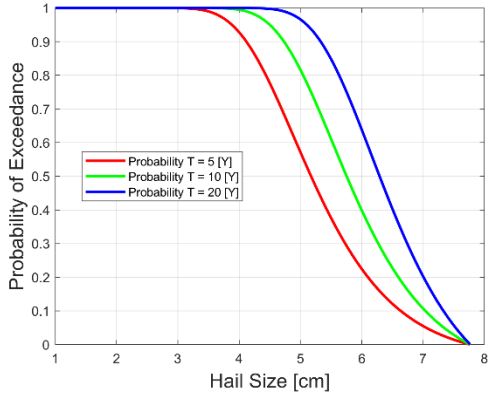
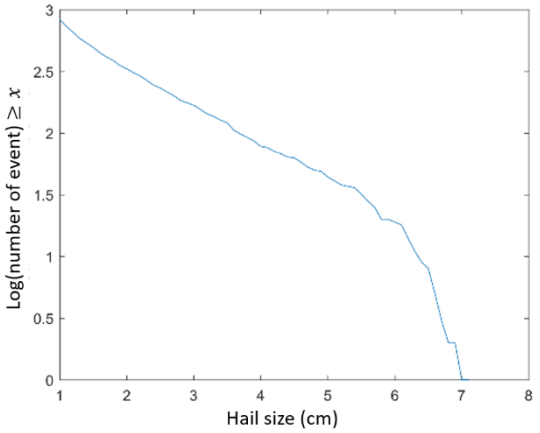
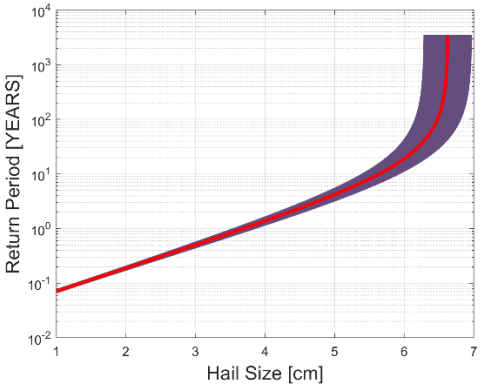
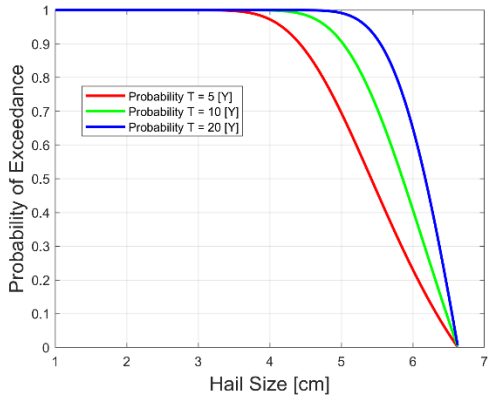
* SE represents the standard error.

TABLE 5.13. Estimated probabilistic hail hazard recurrence parameters for Gauteng, South Africa, based on ensemble HAILCAST data for the period 1979/06/10 to 2016/12/31.

Estimated parameters	**Grid 338**	**Grid 339**	**Grid 367**	**Grid 368**
Mean rate of occurrence $\hat{\hat{\lambda}}(x_{min} = 1.0$ cm)	25.3 ± 1.5	26.0 ± 1.5	14.8 ± 1.0	14.0 ± 1.0
Power law parameter $\hat{\hat{b}}$	0.37 ± 0.01	0.39 ± 0.01	0.48 ± 0.02	0.42 ± 0.02
Estimated maximum possible hail size $\hat{x}_{max}$	7.92 ± 0.34	7.93 ± 0.41	7.77 ± 1.05	6.63 ± 0.34

FIGURE 5.12. Graphical representation of the hail hazard estimates by cumulative frequency–event-size behaviour, return period, and probabilities of exceedance for 5, 10, and 20 years for each of the four grid points over Gauteng, South Africa.

Grid	Cumulative frequency–event-size	Return Period	Probability of Exceedance
338			
339			

Grid	Cumulative frequency–event-size	Return Period	Probability of Exceedance
367			
368			

5.3.3 Hail risk

A secondary dataset, consisting of individual hail-related losses for the period 2007 to 2017 was obtained from a South African insurance company. The information provided by the insurer contains the approximate geospatial coordinates, at the suburb level, of the location of the insured entity during the hailstorm, the occurrence date, the incurred loss, as well as the liable excess of the insured.

A Bayesian vehicle hail-risk model was developed for Gauteng province, incorporating the information generated in Section 5.3.2. The incurred loss amounts were adjusted with the yearly average consumer price index (CPI), with a baseline of December 2016. The CPI index was obtained from Statistics South Africa (StatsSA)[21]. The loss information per individual vehicle was summarised as a daily loss amount. The methodology defined in Chapter 3 was applied to assess the empirical annual distribution of observed losses. This facilitated the assessment of the required recurrence parameters, namely the mean rate of occurrence (expected loss) $\bar{\lambda}$, the power law parameter $\bar{b}$, and the maximum possible hail losses $\hat{x}_{max}$.

The four zones identified for the hail hazard (Figure 5.11) were not used, as there was not enough loss information to provide reliable results per zone. The weighted average and standard deviation of the $\bar{b}$-parameter, as calculated for the hail hazard (Section 5.3.2), was derived as 0.4 and 0.04, respectively. These values were used as *a priori* information for the estimation of the $\hat{\bar{b}}$-parameter for hail risk.

The insured population can typically be divided into two groups according to their 'attitude toward insurance' or the likelihood of instituting a claim after a loss has occurred. It is observed often that the larger a loss is the less is the difference in customer behaviour; however, with small losses, this difference is rather evident. One way that insurers induce clients to participate in risk management, is offering reward programmes and/or excesses (contribution to claims), which are designed to persuade

[21] http://www.statssa.gov.za/?page_id=1854&PPN=P0141, last accessed 2019/09/15.

clients to think twice before pursuing a claim. It is, therefore, assumed that not all the possible damage from hailstorms is reported to insurance companies, particularly when the losses are relatively small. Accordingly, a single LoC of ZAR 0.1 million was introduced. The maximum observed annual loss was ZAR 7.3 million in 2012. Because of the large range of the financial losses over a short time period, an event size error of ZAR 0.5 million was assumed. This value was obtained using the visual cumulative method (Mulargia and Tinti, 1985). Little variation was observed in the data over the time period, and a total of 10% parameter variation was included. All the claims related to the events in the dataset were paid out by the insurer; therefore, each event was assigned a validity index of 1. The possibility of false claims and fraud was not considered. The input parameters for the assessment of the risk are provided in Table 5.14 and the results of the risk assessment in Table 5.15. The results of the MLE are also provided for comparison purposes. A comparison of estimated mean rate of occurrence of hail risk for vehicles is provided in Figure 5.13. The return period and probabilities of exceedance for 5, 10, and 20 years for the hail risk estimates are provided in Figure 5.14.

TABLE 5.14. Probabilistic hail risk assessment input parameters for Gauteng, South Africa, based on financial loss information for vehicles from MSTI for the period 2007/01/01 to 2017/12/31.

Input	**Gauteng**
Number of events	28
LoC	R0.1 million
Hail size SE	R0.5 million
Observed maximum hail size	R7.3 million
% variation	10%
Mean of *a priori* information for $\hat{\hat{b}}$	0.4
Standard deviation of *a priori* information for $\hat{\hat{b}}$	0.04

TABLE 5.15. Estimated probabilistic hail risk recurrence parameters for Gauteng, South Africa, based on financial loss information for vehicles from MSTI for the period 2007/01/01 to 2017/12/31.

Estimated parameters	MLE	BI
Mean rate of occurrence $\hat{\hat{\lambda}}(x_{min} = \text{ZAR 0.1 million})$	2.2 ± 0.5	2.2 ± 0.5
Power law parameter $\hat{\hat{b}}$	0.55 ± 0.12	0.42 ± 0.04
Estimated maximum possible hail size $\hat{x}_{max}$	9.9 ± 2.6	8.2 ± 0.9
Percentage contribution to $\hat{\hat{b}}$-parameter		
Hail hazard *a priori* information	0	46.5
Financial loss information from insurer	100	53.5

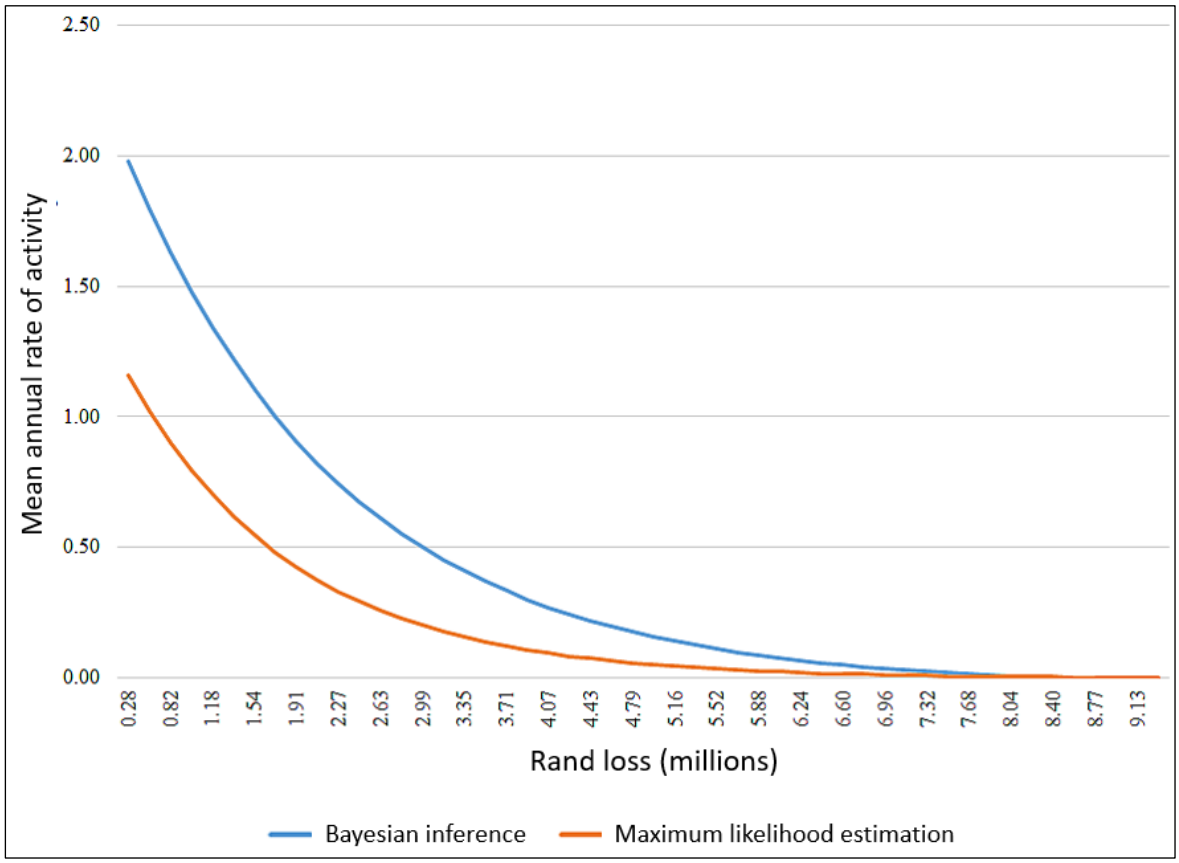
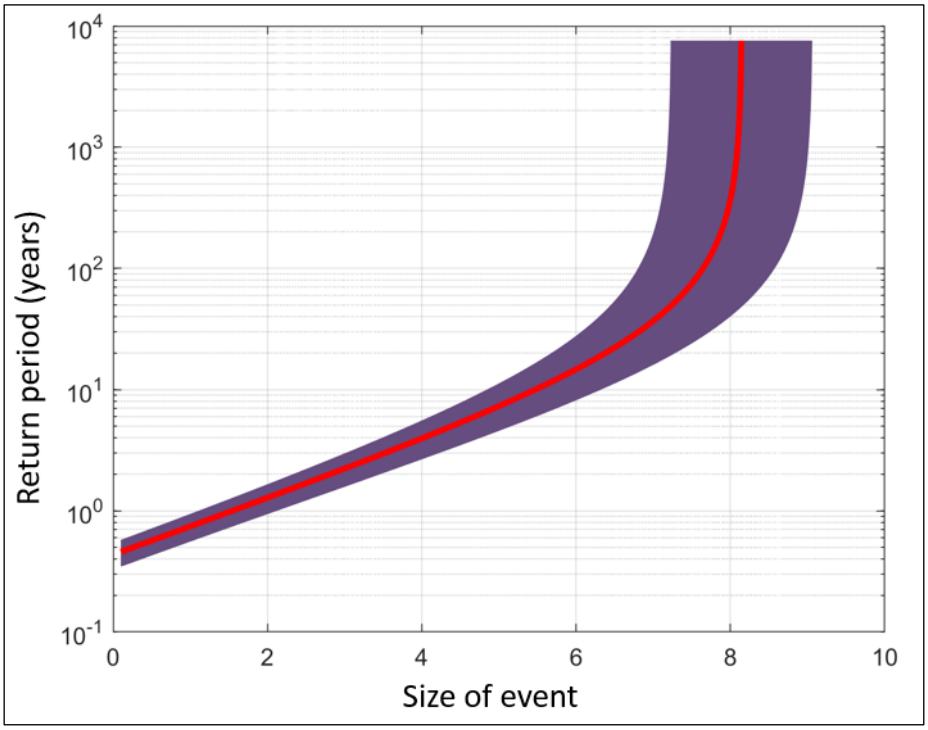
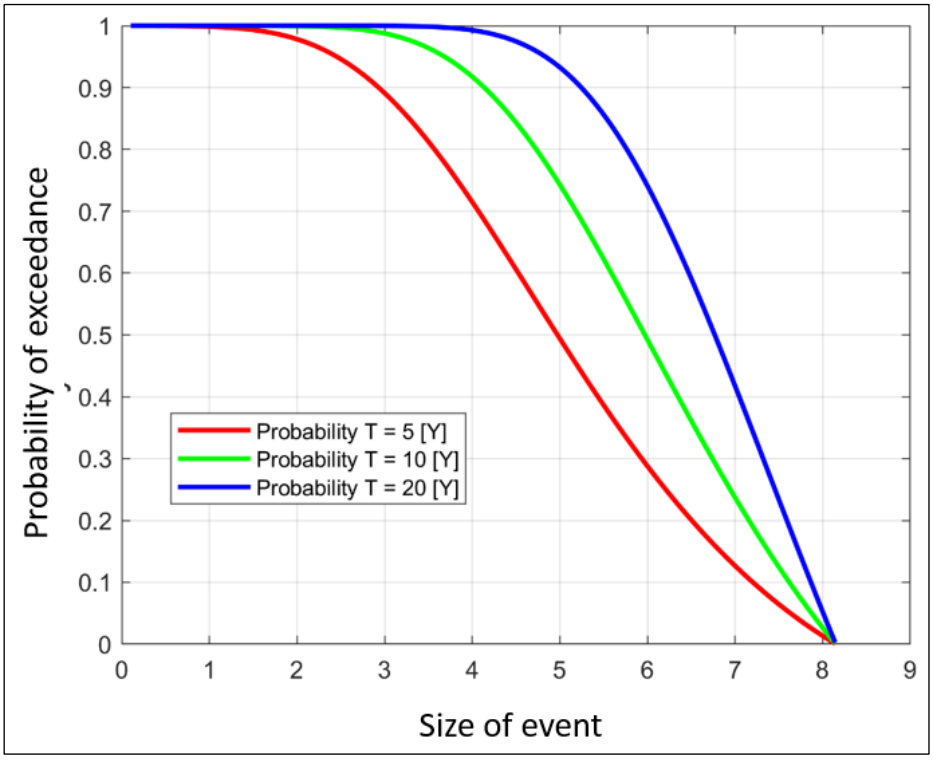


FIGURE 5.13. Comparison of estimated mean rate of occurrence of hail risk for vehicles in Gauteng by MLE and BI.



(a)



(b)

FIGURE 5.14. Graphical representation of the hail risk estimates by (a) return period and (b) probabilities of exceedance for 5, 10, and 20 years for Gauteng, South Africa.

5.3.4 Discussion

Sections 5.3.2 and 5.3.3 demonstrate how both hail hazard and hail risk can be assessed using the generic methodology defined in Chapter 3.

Four grids points in the HAILCAST ensemble dataset were used to assess the hail hazard for Gauteng. The data in all four grids points demonstrate a power law behaviour when investigating the cumulative frequency–event-size relationship (Figure 5.12). As the ensemble dataset can also be viewed as synthetic data, is the dataset is considered a complete event set. This single dataset, therefore, contributed 100% to the assessment of both estimates $\hat{\hat{\lambda}}$ and $\hat{\hat{b}}$. Minimal variation can be seen in the power law behaviour, except for large event sizes. The sharp decline in the cumulative frequency for large events supports the introduction of an upper limit in the modelling process.

The hail hazard estimates were generated using MLE under the assumption that no prior information was available. The estimated recurrence parameters have a relatively small standard deviation, with the estimated maximum possible event sizes close to the maximum observed. Sensitivity analyses were performed by varying the percentage variation in the model parameters, as well as the event size errors. Accounting for uncertainty in the recurrence parameters did not change the estimated annual rate of exceedance for event sizes in the instance where the event size error did not change. On the other hand, the event size error had the largest effect on the recurrence parameters. The middle range of event sizes was the most sensitive, resulting in the underestimation of the mean rate of occurrence $\hat{\hat{\lambda}}$.

The estimates $\hat{\hat{\lambda}}$ and $\hat{\hat{b}}$ (Table 5.13) indicate that behaviour of the recurrence parameters for 338 and 339 are similar, and the behaviour for 367 and 368 are similar. There is, however, a difference between the southern and northern grid points, most likely attributable to topographical changes. The return periods and probabilities of exceedance for the four grid points are very similar (Figure 5.13), with the grid point over Pretoria exhibiting slightly lower return periods for higher event sizes.

A small sample of the incurred vehicle hail losses for a South African insurance company was used to demonstrate how the proposed methodology could be used to assess financial risk. Only incurred losses $x \geq$ ZAR 0.1 million were analysed using MLE and BI. Through Bayesian inference, the hail hazard b-parameter was used as a priori information in the assessment of the b-parameter of hail risk. The power law parameter estimate $\hat{\hat{b}}$ from the hail hazard assessment was used as *a priori* information to constrain the risk assessment results. The introduction of the prior information reduced the estimate value of the risk power law parameter, as well as its associated standard error (Table 5.15). Because of the small number of events used in the analyses, the hail hazard prior contributed 46.5% of the information used in the estimation process. This resulted in a reduction in the $\hat{\hat{b}}$-parameter compared with the MLE risk estimates. The estimated maximum possible hail loss $\hat{x}_{max}$ and standard error were also reduced.

Figure 5.13 illustrates the estimated mean rate of occurrence $\hat{\hat{\lambda}}$ for vehicle hail risk in Gauteng when applying BI and MLE. The estimated values for MLE were lower compared with the Bayesian estimates, thereby yielding a higher estimated return period. Figure 5.14 provides the return periods and probabilities of exceedance for 5, 10, and 20 years. The return periods for MLE are higher compared with those of BI. For instance, the maximum observed loss x_{max}^{obs} = ZAR 7.3 million yields a return period of 78 years for MLE and 56 years for BI. The hail risk estimates using BI are, therefore, higher but with a lower maximum possible event size.

Future research could include analyses of the hail hazard over smaller grid sizes to account for spatial and temporal dependencies related to the topography. Additional investigations can be conducted to obtain more historical information on severe hailstorms or the application of other event-size distributions such as the generalized Pareto distribution. The loss results provided are applicable only to the portfolio of the particular insurer and are based on a small dataset. Additional hail loss information across the market will be required to build a representative hail risk model for South Africa. It is also important to note that these risk estimates are dependent on external factors, such as vehicle cost and

inflation. Regular adjustments should be made to ensure that the estimates remain in agreement with the economic realities.

Chapter 6. Remarks and Conclusions

"No mathematical tricks can replace data." Andrzej Kijko

This thesis presents a generic and flexible methodology for the probabilistic modelling of any type of natural hazard based on empirical data, aleatory and epistemic uncertainty, and prior information. The methodology combines a non-homogeneous Poisson process with the relevant natural hazard power law describing the relationship between the cumulative frequency and the measurement sizes of the events. Explicit provision is made for highly incomplete and uncertain data, the inclusion of prehistoric, historical, and prior information to stabilise and improve the accuracy of the results, uncertainties in size measurement, uncertainty in the parameters of the applied distributions when events are non-homogeneous or weakly dependent, as well as a means to provide for the uncertainty (validity) of the occurrences in the dataset. Both maximum likelihood estimation (MLE) and Bayesian inference (BI) are used to estimate the recurrence parameters λ and b (β). In addition, an existing methodology to assess the maximum possible event size x_{max} is discussed. The presented work is a natural expansion of the work by Kijko and Dessokey (1987) and Kijko and Sellevoll (1989; 1992) that developed efficient techniques for the optimal assessment of the area-characteristic seismic hazard parameters by considering incomplete datasets and uncertainty in earthquake magnitudes. The new proposed methodology builds on this foundation by describing how any physical distribution power law observed in nature could be utilised in this framework, how to account for parameter uncertainty, i.e. weak dependencies in the data, and how to account for uncertainty related to the validity of the events in the dataset.

The methodology is discussed in terms of a power law that describes the underlying physical characteristics of natural hazards. Extreme prehistoric and historical distributions are defined using the same power law as defined for the instrumental datasets that are more complete. As the precision of measurement instrumentation generally increases over time, increasingly smaller events are included in

such datasets. Therefore, to account for the varying levels of completeness between the prehistoric, historical, and instrumental datasets, sub-datasets are created, allowing for the employment of the additive property of likelihood functions to combine the datasets. In addition, it allows for the estimation of single parameter estimates for λ and b by using MLE and BI.

In instances where the data are perfectly independent and identically distributed, the probabilistic hazard assessment of these parameters is relatively easy using the analytical formulas. Natural event datasets, however, often exhibit patterns and some dependencies between the observed events. Furthermore, the event size measurements can be questionable. In some instances, particularly in prehistoric and historical datasets, the validity of an event can be questioned.

To address the concerns listed above, three types of additional uncertainty are introduced into the modelling framework to account for the knowledge gaps that occur often in natural hazard datasets. The convolution theorem is used to introduce event size errors and mixture distributions to allow for deviations from the strict requirements of the Poisson process and exponential distribution. These methods, and employing the likelihood functionality, facilitate the direct introduction of data characteristics particular to each of the prehistoric, historical, and instrumental datasets. Using the weighted likelihood function to account for the validity of an event is an effective tool to ensure that the rates of occurrence and return periods are not overestimated.

To test the versatility of the methodology, it was applied to a synthetic earthquake dataset, two real earthquake datasets, one from South Africa and the other from Central Italy, a tsunami dataset for tsunamigenic regions in the Pacific Ocean, ensemble hail event data, and vehicle hail insurance loss data from South Africa. The results for each specific example are discussed in depth in the respective subsections of Chapters 4 and 5. Therefore, only general observations from the thesis will be discussed in points 1 to 14 listed below.

1. Physical distribution power laws are used often in natural hazard studies to describe the relationship between the cumulative frequency and the event size. The proposed methodology

employs this characteristic to define a framework for probabilistic natural hazard assessment that not only allows the researcher to choose from different power law equations but also to account for different types of data uncertainty.

2. Defining the hazard in terms of power laws and likelihood functions allows different types of data to be combined, as well as to account for potential time gaps in the datasets.
3. The parameter estimates for b ($\beta = b \ln 10$) depend on the level of completeness parameter x_{min}. Unfortunately, no single method exists to assess x_{min} accurately and it remains a subjective parameter. The choice of x_{min} can have a noteworthy effect on the estimates. This aspect is not discussed in this thesis; however, refer to De Witt (2013) and Kijko and Smit (2017).
4. Prehistoric and historical information can be unreliable or not available. If the researcher feels that too much emphasis is being placed on these datasets, the methodology allows the user to reduce their contribution by using a validity index or, alternatively, to continue with only instrumental data.
5. The inclusion of different types of data uncertainty should be done prudently and with a thorough understanding of the data and the underlying physical mechanics that is the driving force of the specific natural hazard.
6. The type of uncertainty included in the analyses must be a function of the data. Consideration of the event size uncertainty usually has an opposite effect on the estimates compared with parameter uncertainty. By accounting for both, the combined effect will always be less than the largest effect of the two separately.
7. Utilising only a single event dataset with a medium to high level of completeness, or only extreme data, can result in under- or overestimation of the recurrence parameters and subsequent hazard estimates. Considering small and large events in the analyses affects the hazard estimates, especially in instances where only a single dataset or only extreme events are used. Including prior information can help to stabilise and improve the accuracy of the results.

8. The creation of sub-datasets allows the researcher to use more observations. In the synthetic example in Chapter 4, sub-setting the datasets leads to results that are more stable compared with the single complete and extreme datasets. Using prior information for sub-datasets yields BI estimates similar to MLE estimates.
9. The estimation process for the rate of occurrence λ relies more on the most-recent instrumental data, except where the historical or prehistoric datasets contain substantially more events over a longer period of time (as seen in Section 5.2). In comparison, the power law parameter b-estimates always rely more on the provided prehistoric and historical information.
10. The percentage contribution that prior information makes to the estimation of both the recurrence parameters is a function of the type and amount of uncertainty included in the modelling process.
11. The estimates for the recurrence parameters λ and b are obtained simultaneously. This results in dependencies between the parameters when multiple datasets are used. If only prior information is assumed for one parameter, it will affect the second parameter indirectly without making an explicit contribution.
12. The inclusion of the validity index as an indicator of the uncertainty associated with occurrence affects only the mean rate of occurrence and the return periods. In some instances, it could even provide a rate of occurrence estimate equivalent to half of that when the index is not included. This is a salient feature, as it translates directly to the hazard estimates.
13. Traditional model comparison techniques require using a single, complete dataset and independent model parameters. It is, therefore, difficult to compare the effectiveness and fit of the models between a single dataset (SDS), an extreme dataset (EXT), or multiple datasets with varying levels of completeness (LOC) using standard techniques.
14. As is demonstrated in Section 5.3.3, the methodology can also be applied for natural hazard risk assessment. In this section, the financial risk recurrence parameters and risk estimates are based on an insurance loss dataset. This reiterates the versatility of the methodology discussed in the thesis.

Chapter 7. Future Research

As stated in Section 1.5, natural hazard and risk assessment is a wide research field. The methodology discussed in this thesis provides a generic framework that can be used to build more complex models. Three general themes that can be investigated are:

1. **Data and parameter dependencies**. Future research into data and parameter dependencies can focus on accounting for strong spatial- and/or temporal dependencies in the data, e.g. using max-stable processes or copulas when modelling extreme events; removing the subjective level of completeness parameter x_{min} from the model; and obtaining an improved understanding of the dependencies between the recurrence parameters when using multiple sub-datasets in the modelling process.
2. **Parameter estimation and model comparisons**. Investigations can focus on alternative extreme, heavy-tail, semi-parametric, or non-parametric distributions applicable within the modelling framework; potential bi-model behaviour in the frequency–event-size relationship; the effect of different types of numerical methods on parameter estimates; developing non-central confidence intervals, where a penalty is built into the confidence intervals for either over- or underestimation (Chave, 2017); and developing a model comparison methodology that is not dependent on the same underlying dataset to be used, or on the independence of parameters.
3. **New applications.** The generic nature of the methodology can be tested on non-natural event datasets, including other financial risks and vulnerabilities governed by a frequency–event-size relationship. This opens up the possibility to extend this methodology to a multi-hazard and risk framework that would include the assessment of hazards and/or risks that occur simultaneously or as the result of a previous event, as well as their potential interrelated effects[22].

[22] http://www.interactinghazards.com/defining-multi-hazard, last accessed 2018/09/28.

References

Aki, K. (1965). Maximum likelihood estimate of b in the formula log N=a-bM and its confidence limits. *Bulletin of the Earthquake Research Institute-University of Tokyo*, 43, pp. 237–239.

Aleotti, P. & Chowdhury, R. (1999). Landslide hazard assessment summary review and new perspectives. *Bulletin of Engineering Geology and the Environment*, 58(1), pp. 21–44. https://link.springer.com/content/pdf/10.1007/s100640050066.pdf.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J. & Pappenberger, F. (2013). GloFAS-global ensemble streamflow forecasting and flood early warning. *Hydrology and Earth System Sciences*, 17(3), 1161 pp. DOI: 10.5194/hess-17-1161-2013.

Allen, J.T. & Karoly, D.J. (2014). A climatology of Australian severe thunderstorm environments 1979–2011: inter-annual variability and ENSO influence. *International Journal of Climatology*, 34(1), pp. 81–97. DOI: 10.1002/joc.3667.

Allen, J.T. & Tippett, M.K. (2015). The characteristics of United States hail reports: 1955–2014. *E-Journal of Severe Storms Meteorology*, 10(3). pp. 1-31. ejssm.org/ojs/index.php/ejssm/article/view/149/104.

Allen, J.T., Tippett, M.K., Kaheil, Y., Sobel, A.H., Lepore, C., Nong, S. & Muehlbauer, A. (2017). An Extreme Value Model for US Hail Size. *Monthly Weather Review*, 145(11), pp. 4501–4519. DOI: 10.1175/MWR-D-17-0119.1.

Allen, J.T., Tippett, M.K. & Sobel, A.H. (2015). An empirical model relating US monthly hail occurrence to large-scale meteorological environment. *Journal of Advances in Modeling Earth Systems*, 7(1), pp. 226–243. DOI: 10.1002/2014MS000397.

Alves, I.F. & Neves, C. (2014). Estimation of the finite right endpoint in the Gumbel domain. *Statistica Sinica*, pp.1811–1835. jstor.org/stable/24310971.

Alves, I.F., Neves, C. & Rosário, P. (2017). A general estimator for the right endpoint with an application to supercentenarian women's records. *Extremes*, 20(1), pp. 199–237. DOI: 10.1007/s10687-016-0260-6.

Ambraseys, N. (1962). Data for the investigation of the seismic sea-waves in the Eastern Mediterranean. *Bulletin of the Seismological Society of America*, 52, pp. 895–913.

Anderson, J.G., Brune, J.N., Anooshehpoor, R. & Ni, S.D. (2000). New ground motion data and concepts in seismic hazard analysis. *Current Science*, pp. 1278–1290. jstor.org/stable/24105280.

Anderson, J.G. & Luco, J.E. (1983). Consequences of slip rate constraints on earthquake occurrence relations. *Bulletin of the Seismological Society of America*, 73(2), pp. 471–496.

Anderson, J.G., Wesnousky, S.G. & Stirling, M.W. (1996). Earthquake size as a function of fault slip rate. *Bulletin of the Seismological Society of America*, 86(3), pp. 683–690.

Anita, G., Tonini, R., Sandri, L., Pierdominici, S., & Selva, J. (2015). A Methodology for a comprehensive probabilistic tsunami hazard assessment: Multiple sources and short-term interactions. *Journal of Marine Science and Engineering*, 3(1), pp. 23–51. DOI: 10.3390/jmse3010023.

Annaka, T., Satake, K., Sakakiyama, T., Yanagisawa, K. & Shuto N. (2007). Logic-tree approach for probabilistic tsunami hazard analysis and its applications to the Japanese coasts. *Pure and Applied Geophysics*, 164, pp. 577–592. DOI: 10.1007/978-3-7643-8364-0_17.

AON (2014). News release of 9 April 2014. Hail damage catastrophe modelling by Mandy Barrett https://www.fanews.co.za/article/short-term-insurance/15/general/1217/hail-damage-catastrophe-modelling/16493, last accessed 2018/10/05.

Arnold, B.C. (2014). Pareto distribution. Wiley StatsRef: Statistics Reference Online, pp. 1–10. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat01100.pub2.

Bain, L.J. & Engelhardt, M. (1992). *Introduction to probability and mathematical statistics*. Second Edition. Duxbury Press.

Baker, V.R. (1987). Paleoflood hydrology and extraordinary flood events. *Journal of Hydrology*, 96, pp. 79–99. DOI: 10.1016/0022-1694(87)90145-4.

Baker, J.W. (2013). An introduction to probabilistic seismic hazard analysis. White paper version, 2(1), 79 pp. https://scits.stanford.edu/sites/default/files/baker_2013_intro_psha_v2.pdf.

Balakrishnan, N. & Basu, A.P. (1995). *The exponential distribution: Theory, methods and applications*. CRC Press, Taylor and Francis Group.

Barthel F. & Neumayer E. (2012). A trend analysis of normalized insured damage from natural disasters. *Climate Change*, 113, pp. 215–237. DOI: 10.1007/s10584-011-0331-2.

Beirlant, J., Kijko, A., Reynkens, T. & Einmahl, J.H. (2017). Estimating the maximum possible earthquake magnitude using extreme value methodology: the Groningen case. *Natural Hazards*, pp. 1–23. DOI: 10.1007/s11069-017-3162-2.

Bender, B. (1988). Reliability of estimates of maximum earthquake magnitudes based on observed maxima. *Seismological Research Letters*, 59(1), pp. 1-15.

Bender, B.K. & Perkins, D.M. (1993). Treatment of parameter uncertainty and variability for a single seismic hazard map. *Earthquake Spectra*, 9(2), pp. 165–195. DOI: 10.1193/1.1585711.

Benito, G., Lang, M., Barriendos, M., Llasat, M.C., Francés, F., Ouarda, T., Thorndycraft, V., Enzel, Y., Bardossy, A., Coeur, D. & Bobée, B. (2004). Use of systematic, palaeoflood and historical data for the improvement of flood risk estimation. Review of scientific methods. *Natural Hazards*, 31(3), pp 623–643. DOI: 10.1023/B:NHAZ.0000024895.48463.eb.

Benjamin, J.R. (1968). Probabilistic models for seismic forces design, *Journal of the Structural Division – Civil Engineering Database – ASCE*, 94, (ST5), pp. 1175–1196.

Benjamin, J. R., & Cornell, C. A. (2014). *Probability, statistics, and decision for civil engineers*. Courier Corporation.

Berti, M., Martina, M. L. V., Franceschini, S., Pignone, S., Simoni, A., & Pizziolo, M. (2012). Probabilistic rainfall thresholds for landslide occurrence using a Bayesian approach. *Journal of Geophysical Research: Earth Surface*, 117(F4), pp. 1–20. DOI: 10.1029/2012JF002367.

Bird, P. & Kagan, Y.Y. (2004). Plate-tectonic analysis of shallow seismicity: Apparent boundary width, beta, corner magnitude, coupled lithosphere thickness, and coupling in seven tectonic settings. *Bulletin of the Seismological Society of America*, 94(6), pp. 2380–2399. DOI: 10.1785/0120030107.

Blair, S.F. & Leighton, J.W. (2012). Creating high-resolution hail datasets using social media and post-storm ground surveys. *Electronic Journal of Operational Meteorology*, 13(3), pp. 32–45.

Blong, R. (2003). A review of damage intensity scales. *Natural Hazards*, 29(1), pp. 57–76. DOI: 10.1023/A:1022960414329.

Brandt, M.B.C., Saunders, I. & Graham, G. (2002). Frequency–magnitude relationship and estimated m_{max} from micro-seismic data, and regional- and historical data sets for the southern part of South Africa. Africa. *Geoscience Review*, 9(3), pp. 199–210.

Brandt, M.B.C., Bejaichund, M., Kgaswane, E.M., Hattingh, E. & Roblin, D.L. (2005). *Seismic history of South Africa*. Council for Geoscience, Seismological Series. 37, 32 pp.

Brimelow, J.C., Reuter, G.W. & Poolman, E.R. (2002). Modeling maximum hail size in Alberta thunderstorms. *Weather and Forecasting*, 17(5), pp.1048-1062. DOI: 10.1175/1520-0434(2002)017<1048:MMHSIA>2.0.CO;2.

Brizuela, B., Armigliato, A. & Tinti, S. (2014). Assessment of tsunami hazards for the Central American Pacific coast from southern Mexico to northern Peru. *Natural Hazards and Earth System Sciences*, 14(7), pp.1889-1903. DOI: 10.5194/nhess-14-1889-2014.

Brooks, H.E, Anderson, A.R., Riemann, K., Ebbers, I., & Flachs H. (2007). Climatological aspects of convective parameters from the NCAR/ NCEP reanalysis. *Atmospheric Research*, 83, pp. 294–305. DOI: 10.1016/j.atmosres.2005.08.005.

Brooks, H.J., Lee, J., & Craven, J. (2003). The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmospheric Research*, 67, pp. 73–94. DOI: 10.1016/S0169-8095(03)00045-0.

Brown, D., Backhouse, F. & Oliver, S. (2010). *501 Most Devastating Disasters*. Octopus Publishing Group Limited.

Brown, T.M., Pogorzelski, W.H. & Giammanco, I.M. (2015). Evaluating hail damage using property insurance claims data. *Weather, Climate, and Society*, 7(3), pp. 197-210. DOI: 10.1175/WCAS-D-15-0011.1.

Burbidge, D., Cummins, P.R., Mleczko, R. & Thio, H.K. (2008). A probabilistic tsunami hazard assessment for Western Australia. *Pure and Applied Geophysics*, 165(11-12), pp. 2059–2088. DOI: 0.1007/978-3-0346-0057-6_6.

Burroughs, S.M., & Tebbens, S.F. (2001). Upper-truncated power laws in natural systems. *Pure and Applied Geophysics*, 158(4), pp. 741–757. DOI: 10.1007/PL00001202.

Burroughs, S.M. & Tebbens, S.F. (2005). Power law scaling and probabilistic forecasting of tsunami run-up heights. *Pure and Applied Geophysics*, 162, pp. 331–342. DOI: 10.1007/s00024-004-2603-5.

Caccavale, M., Matano, F. & Sacchi, M. (2017). An integrated approach to earthquake-induced landslide hazard zoning based on probabilistic seismic scenario for Phlegraean Islands (Ischia, Procida and Vivara), Italy. *Geomorphology*, 295, pp. 235-259. DOI: 10.1016 /j.geomorph.2017.07.010.

Caellum (1991) & Subsequent updates. *A History of notable Weather Events in South Africa: 1500-1990*. The Weather Bureau, Department of Environment Affairs.

Cao, A. & Gao, S.S. (2002). Temporal variation of seismic b-values beneath northeastern Japan island arc. *Geophysical Research Letters*, 29(9), 48-1–48-3. DOI: 10.1029/2001GL013775.

Campbell, K.W. (1982). Bayesian analysis of extreme earthquake occurrences. Part I. Probabilistic hazard model. *Bulletin of the Seismological Society of America*, 72, pp. 1689–1705.

Campbell, K.W. (1983). Bayesian analysis of extreme earthquake occurrences. Part II. Application to the San Jacinto Fault zone of southern California. *Bulletin of the Seismological Society of America*, 73, pp. 1099–1115.

Changnon, S.A. (1967) Areal-temporal variations in hail intensity in Illinois. *Journal Applied Meteorology*, 6, pp. 536–541. DOI: 10.1175/1520-0450(1967)006<0536:ATVOHI>2.0.CO;2.

Changnon, S.A. (2008). Temporal and spatial distributions of damaging hail in the continental United States. *Physical Geography*, 29(4), pp. 341–350. DOI: 10.2747/0272-3646.29.4.341.

Changnon, S.A., Changnon, D. & Hilberg, S.D. (2009). *Hailstorms across the nation: An atlas about hail and its damages*. ISWS Contract Report 2009-12.

Changnon, S.A. & Stout G. (1967). Crop-hail intensities in Central and Northwest United States. *Journal of Applied Meteorology*, 6, pp. 542–548. DOI: 0.1175/1520-0450 (1967)006<0542:CHIICA>2.0.CO;2.

Chave, A.D. (2017). *Computational Statistics in the Earth Sciences: With Applications in MATLAB*. Cambridge University Press.

Cheng, R.H., & Traylor, L. (1995). Non-regular maximum likelihood problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, pp. 3–44. jstor.org/stable/2346086.

Chinnery, M.A. (1979). *Investigations of the seismological input to the safety design of nuclear power reactors in New England*. U.S. Nuclear Regulatory Commission Report NUREG/CR-0563, 72 pp.

Clauset, A., Shalizi, C.R. & Newman, M.E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), pp. 661–703. DOI: 10.1137/070710111.

Cloete, G., Benito, G., Grodek, T., Porat, N. & Enzel, Y. (2018). Analyses of the magnitude and frequency of a 400-year flood record in the Fish River Basin, Namibia. *Geomorphology*, 320, pp.1-17. DOI: 10.1016/j.geomorph.2018.07.025.

Cohn, T.A., & Stedinger, J.R. (1987). Use of historical information in a maximum-likelihood framework. *Journal of Hydrology*, 96(1-4), pp.215-223. DOI: 10.1016/0022-1694(87)90154-5.

Coles, S. (2001). *An introduction to statistical modelling of extreme values*. Springer Series in Statistics, London.

Cooke, P. (1979). Statistical inference for bounds of random variables. *Biometrika*, 66(2), pp. 367–374. DOI: 10.1093/biomet/66.2.367.

Cooke, P. (1980). Optimal linear estimation of bounds of random variables. *Biometrika*, 67(1), pp. 257–258. DOI: 10.1093/biomet/67.1.257.

Cooley, D., Nychka, D. & Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479), pp. 824–840. DOI: 10.1198/016214506000000780.

Coppersmith, K.J. (1994). *Conclusions regarding maximum earthquake assessment. In:* The Earthquakes of Stable Continental Regions, Vol. 1. Assessment of Large Earthquake Potential, Electric Power Research Institute, Palo Alto, pp. 6.1–6.24.

Cornell, C.A. (1968). Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*, 58, pp. 1583–1606.

Cornell, C.A. (1994). Statistical analysis of maximum magnitudes in the earthquakes of stable continental regions, in The Earthquakes of Stable Continental Regions, Vol. 1. Schneider J, editor. Seismic hazard methodology for the Central and Eastern United States. "The Earthquakes of Stable Continental Regions.Vol.1. Assessment of large Earthquake Potential.". EPRI, Palo Alto, Calif. 1994; NP-4726: 5-1-5-27.

Cornell, C.A. & Winterstein, S.R. (1988). Temporal and magnitude dependence in earthquake recurrence models. *Bulletin of the Seismological Society of America*, 78, pp. 1522–1537.

Corominas, J. & Moya, J. (2008). A review of assessing landslide frequency for hazard zoning purposes. *Engineering Geology*, 102(3-4), pp.193-213. DOI: 10.1016/j.enggeo.2008.03.018.

Cosentino, P., Ficara, V. & Luzio, D. (1977). Truncated exponential frequency – magnitude relationship in the earthquake statistics. *Bulletin of the Seismological Society of America*, 67, pp. 1615-1623.

Cramér, H. (1961). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, New Jersey.

Cumming, S.G. (2001). A parametric model of the fire-size distribution. *Canadian Journal of Forest Research*, 31(8), pp.1297-1303. DOI: 10.1139/x01-032.

Cunningham, R.J., Herzog, T.N. & London, R.L. (2012). *Models for quantifying risk*. ACTEX publications.

Dargahi-Noubary, G.R. (1983). A procedure for estimation of the upper bound for earthquake magnitudes. *Physics of the Earth and Planetary Interiors*, 33(2), 91–93. DOI: 10.1016/0031-9201(83)90142-5.

Davies, N. & A Kijko (2003). Seismic risk assessment: with an application to the South African insurance industry. *South African Actuarial Journal*, 3, pp. 1–28. hdl.handle.net/10520/EJC17093.

Davison, A.C. (2003). *Statistical models, Cambridge. Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Daykin, C.D., Pentikainen, T. & Pesonen, M. (1993). *Practical risk theory for actuaries*. Chapman and Hall/CRC.

De Haan, L. & Ferreira, A. (2007). *Extreme value theory: an introduction*. Springer Science & Business Media.

Deluca, A. and Corral, Á. (2013). Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophysica*, 61(6), pp.1351-1394. DOI: 10.2478/s11600-013-0154-9.

De Witt, C.J. (2013). *Investigating the threshold of event detection with application to earthquake and operational risk theory*. MSc Actuarial Science Dissertation, University of Pretoria.

Der Kiureghian, A. & Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2), pp. 105–112. DOI: 10.1016/j.strusafe.2008.06.020.

Dong, W., Shah, H.C., Bao, A. & Mortgat, C.P. (1984). Utilization of geophysical information in Bayesian seismic hazard model. *International Journal of Soil Dynamics and Earthquake Engineering*, 3(2), pp.103-111. DOI: 10.1016/0261-7277(84)90005-6.

Downes, G. L. & Stirling, M.W. (2001). Groundwork for development of a probabilistic tsunami hazard model for New Zealand. *International Tsunami Symposium 2001, Seattle, Washington*, pp. 293–301.

Dyson, L.L., Pienaar N., Smit, A., Kijko, A. & Poolman E. (unpublished). An ERA-interim hail climatology for South Africa. Submitted to the International Journal of Climatology (2019).

Dzierma, Y. & Wehrmann, H. (2010). Eruption time series statistically examined: Probabilities of future eruptions at Villarrica and Llaima Volcanoes, Southern Volcanic Zone, Chile. *Journal of Volcanology and Geothermal Research*, 193(1), pp. 82–92. DOI: 10.1016/j.jvolgeores. 2010.03.009.

Eadie, W.T. Drijard, D. James, F.E., Roos, M. & Sadoulet, B. (1971). *Statistical methods in experimental physics*. Second reprint 1982. North-Holland, Amsterdam.

Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press. New York, p. 235.

Egüen, M., Aguilar, C., Solari, S. & Losada, M.A. (2016). Non-stationary rainfall and natural flows modeling at the watershed scale. *Journal of Hydrology*, 538, pp. 767–782. DOI: 10.1016/j.jhydrol.2016.04.061.

Eichner, J. (2015). Severe convective storms and hail – Icy cricket balls from above. For Munich Re. Available at www.munichre.com/site/australia-hazards/get/documents_E-166132984/mr/Australia-Hazards/documents/Downloads/MR-Australia-Hailstorms.pdf. Last accessed 2018/07/24.

El-Isa, Z. H. & Eaton, D. W. (2014). Spatiotemporal variations in the b-value of earthquake magnitude-frequency distributions: Classification and causes. *Tectonophysics*, 615-616, pp. 1–11. DOI: 10.1016/j.tecto.2013.12.001.

EM-DAT: The Emergency Events Database - Universite catholique de Louvain (UCL) - CRED, D. Guha-Sapir - www.emdat.be, Brussels, Belgium; Accessed 30 April, 2018

Epstein, B. & Lomnitz C. (1966) A model for occurrence of large earthquakes. *Nature*, 211, pp. 954–956. DOI: 10.1038/211954b0.

Fedorov, V.V. (2013). *Theory of optimal experiments*. Elsevier.

Fernandes, W., Naghettini, M. & Loschi, R. (2010). A Bayesian approach for estimating extreme flood probabilities with upper-bounded distribution functions. *Stochastic Environmental Research and Risk Assessment*, 24(8), pp. 1127–1143. DOI: 10.1007/s00477-010-0365-4.

Fernandez, L.M. & Guzman, J.A. (1979a). Earthquake hazard in southern Africa. *Geological Survey of South Africa, Seismologic Series*, 10, 22 pp.

Fernandez, L.M. & J.A. Guzman, J.A. (1979b). Seismic history of southern Africa. *Geological Survey of South Africa, Seismologic Series*, 9, pp. 1–38.

Field, E.H., Jackson, D.D. & Dolan, J.F. (1999). A mutually consistent seismic-hazard source model for Southern California. *Bulletin of the Seismological Society of America*, 89(3), pp. 559–578.

Fraile, R., Berthet, C. & Sanchez, J.L. (2003). Return periods of severe hailfalls computed from hailpad data. *Atmospheric Research*, 67–68, pp. 189–202. DOI: 10.1016/S0169-8095(03)00051-6.

Fröhlich, C. (1998) Does maximum earthquake size depend on focal depth? *Bulletin Seismological Society of America*, 88, pp. 329-336.

Furumura, T., Takemura, S., Noguchi, S., Takemoto, T., Maeda, T., Iwai, K. & Padhy, S. (2011). Strong ground motions from the 2011 off-the Pacific-Coast-of-Tohoku, Japan (Mw=9.0) earthquake obtained from a dense nationwide seismic network. *Landslides*, 8, pp. 333–338. DOI: 10.1007/s10346-011-0279-3.

Galli, P., Galadini, F. & Pantosti, D. (2008). Twenty years of paleoseismology in Italy. *Earth-Science Reviews*, 88(1), pp.89-117. DOI: 10.1016/j.earscirev.2008.01.001.

Gardner, J.K. & Knopoff, L. (1974). Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bulletin of the Seismological Society of America*, 64, pp. 1363-1368.

Geist, E.L. (2005). *Local tsunami hazards in the Pacific Northwest from Cascadia subduction zone earthquakes*. U.S. Geological Survey Professional Paper 1661-B, 17 pp.

Geist, E.L. & Parsons, T. (2006). Probabilistic analysis of tsunami hazards. *Natural Hazards*, 37, pp. 277-314. DOI: 10.1007/s11069-005-4646-z.

Geist, E.L. & Parsons, T. (2014). Undersampling power-law size distributions: Effect on the assessment of extreme natural hazards. *Natural Hazards*, 72(2), pp. 565-595. DOI: 10.1007/s11069-013-1024-0.

Geist, E.L. & Uri, S. (2012). NRC/USGS Workshop Report: Landslide Tsunami Probability. USGS administrative report to the US Nuclear Regulatory Commission. nrc.gov/docs/ ML1227/ML12272A130.pdf.

Gibowicz, S.J. & Kijko, A. (1994). *An introduction to mining seismology*. Academic Press, San Diego. pp. 396.

González, F.I., Geist, E.L., Jaffe, B., Kânoğlu, U., Mofjeld, H., Synolakis, C.E., Titov, V.V., Arcas, D., Bellomo, D., Carlton, D. & Horning, T. (2009). Probabilistic tsunami hazard assessment at

seaside, Oregon, for near-and far-field seismic sources. *Journal of Geophysical Research: Oceans*, 114(C11). pp. 1978–2012. DOI: 10.1029/2008JC005132.

Green, R.W.E., & McGarr, A. (1972). A comparison of the focal mechanism and aftershock distribution of the Ceres, South Africa, earthquake of September 29, 1969. *Bulletin of the Seismological Society of America*, 62(3), pp. 869–871.

Green, R.W.E. & Bloch, S. (1971). The Ceres, South Africa, earthquake of September 29, 1969. I. Report on some aftershocks. *Bulletin of the Seismological Society of America*, 61(4), pp. 851–859.

Grezio, A., Marzocchi, W., Sandri, L. & Gasparini, P. (2010). A Bayesian procedure for probabilistic tsunami hazard assessment. *Natural Hazards*, 53, pp. 159–174. DOI: 10.1007/s11069-009-9418-8.

Grezio, A., Sandri, L., Marzocchi, W., Argnani, A. & Gasparini, P. (2012). Probabilistic tsunami hazard assessment for Messina Strait area (Sicily-Italy). *Natural Hazards*, 64, pp. 329–358. DOI: 10.1007/s11069-012-0246-x

Grezio, A., Babeyko, A., Baptista, M.A., Behrens, J., Costa, A., Davies, G., Geist, E.L., Glimsdal, S., González, F.I., Griffin, J. & Harbitz, C.B. (2017). Probabilistic tsunami hazard analysis: multiple sources and global applications. *Reviews of Geophysics*, 55(4), pp. 1158-1198. DOI: 10.1002/2017RG000579.

Grilli, S. T., Harris, J. & Tajalli Bakhsh, T. (2011). *Literature Review of Tsunami Sources Affecting Tsunami Hazard Along the US East Coast*, NTHMP Progress report, Res. Rept. CACR-11-08, p.60 Center for Applied Coastal Research, University of Delaware, Newark, 2011.

Grobler R.R. (2001). A Framework for modelling losses arising from natural catastrophes in South Africa. Masters Thesis/Dissertation University of Pretoria. repository.up.ac.za/bitstream/handle/ 2263/23749/dissertation.pdf?sequence=1.

Grünthal, G. & Wahlström, R. (2012). The European-Mediterranean earthquake catalogue (EMEC) for the last millennium. *Journal of Seismology*, 16(3), pp. 535-570. DOI 10.1007/s10950-012-9302-y.

Guan, X. & Chen, C. (2014). Using social media data to understand and assess disasters. *Natural Hazards*, 74(2), pp. 837-850. DOI: 10.1007/s11069-014-1217-1.

Gumbel, E.J. (1962). Statistical theory of extreme values (main results). *Contributions to Order Statistics*, pp.56-93.

Gunturi, P. & Tippett, M.K. (2017). Managing severe thunderstorm risk: Impact of ENSO on US tornado and hail frequencies. Technical Report Willis Re Inc, willisre.com/Media_Room/Press_Releases_(Browse_All)/2017/WillisRe_Impact_of_ENSO_on_US_Tornado_and_Hail_frequencies_Final.pdf. Last Accessed 2018/07/24.

Gusiakov, V.K. (2001). Basic Pacific tsunami catalogs and database, 47 BC–2000 AD: results of the first stage of the project. *Proceedings of the International Tsunami Symposium, August 7–9, 2001, Seattle, USA, PMEL/NOAA*, pp. 263– 272.

Gusiakov, V.K. (2005). Tsunami generation potential of different tsunamigenic regions in the Pacific. *Marine Geology*, 215(1), pp. 3-9. DOI:10.1016/j.margeo.2004.05.033.

Gusiakov, V.K. (2009). *Tsunami history: recorded. In the sea.* Bernard, E.N., Robinson, A.R., Eds.; Harvard University Press: Boston, MA, USA; pp. 1-15.

Gutenberg, B. & Richter, C.F. (1942). Earthquake magnitude, intensity, energy, and acceleration. *Bulletin of the Seismological Society of America*, 32, pp. 163–191.

Gutenberg, B. & Richter, C.F. (1956). Earthquake magnitude, intensity, energy, and acceleration: (Second paper). *Bulletin of the Seismological Society of America*, 46, pp. 105-145.

Guttorop, P. & Hopkins, D. (1986). On estimating varying b values. *Bulletin of the Seismological Society of America*, 76(3), pp. 889-895.

Guzzetti, F., Ardizzone, F., Cardinali, M., Galli, M., Reichenbach, P. & Rossi, M. (2008). Distribution of landslides in the Upper Tiber River basin, central Italy. *Geomorphology*, 96(1-2), pp.105-122. DOI: 10.1016/j.geomorph.2007.07.015.

Hall, P. (1982). On estimating the endpoint of a distribution. *The Annals of Statistics*, pp.556-568. jstor.org/stable/2240689.

Hamada, M. S., Wilson, A. G., Reese, C.S., & Martz, H. F. (2008). *Bayesian reliability*. Springer, New York, pp. 430.

Hamilton, R.M. (1967). Mean magnitude of an earthquake sequence. *Bulletin of the Seismological Society of America*, 57(5), pp. 1115-1116.

Harbitz, C. B., Glimsdal, S., Bazin, S., Zamora, N., Løvholt, F., Bungum, H., Smebye, H., Gauer, P. & Kjekstad, O. (2012). Tsunami hazard in the Caribbean: regional exposure derived from credible worst case scenarios. *Continental Shelf Research*, 38, pp. 1-23. DOI: 10.1016/j.csr.2012.02.006.

Hébert, H., Schindelé, F., Altinok, Y., Alpar, B. & Gazioglu, C. (2005). Tsunami hazard in the Marmara Sea (Turkey): A numerical approach to discuss active faulting and impact on the Istanbul coastal areas. *Marine Geology*, 215, pp. 23–43. DOI: 10.1016/j.margeo.2004.11.006.

Heidarzadeh, M. & Kijko, A. (2011). A probabilistic tsunami hazard assessment for the Makran subduction zone at the northwestern Indian Ocean. *Natural Hazards*, 56(3), pp. 577-593. DOI 10.1007/s11069-010-9574-x.

Hewitt, K. (1970). Probabilistic approaches to discrete natural events: a review and theoretical discussion. *Economic Geography*, 46(sup1), pp. 332-349. jstor.org/stable/143148.

Holmes, T.P., Huggett, R.J. & Westerling, A.L. (2008). Statistical analysis of large wildfires. *In The Economics of Forest Disturbances*, Springer, Dordrecht. pp. 59-77. DOI: 10.1007/978-1-4020-4370-3_4.

Horikawa, K. & Shuto, N. (1983). Tsunami disasters and protection measures in Japan, In: K. Iida and T. Iwasaki (eds), *Tsunamis — their science and engineering*, Tokyo, Terra Scientific Publishing Company, pp. 9–22.

Houston, J.R., Carver, R.D. & Markle, D.G. (1977). *Tsunami-wave elevation frequency of occurrence for the Hawaiian Islands*. Technical Report H-77-16, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS, 66 pp.

Howard, L.S. (2018). Natural Catastrophe Claims in 2017 Reached a Record $135B: Munich Re, https://www.insurancejournal.com/news/national/2018/01/04/476093.htm. Last accessed 2019/01/25.

Hu, F. & Zidek, J.V. (2002). The weighted likelihood. *Canadian Journal of Statistics*, 30, pp. 347–372. DOI: 10.2307/3316141.

IAEA TECDOC-1767 (2015). Library Cataloguing in Publication Data. The contribution of palaeoseimology to seismic hazard assessment in site evaluation for nuclear installations - Vienna: International Atomic Energy Agency. IAEA-TECDOC series, ISSN 1011-4289; No 1767. ISBN: 978-92-0-105415-9; 208 pp.

IMechE (2013). Natural Disasters Saving Lives Today, Building Resilience for Tomorrow. Institution of Mechanical Engineers. www.imeche.org/policy-and-press/reports/detail/natural-disasters-saving-lives-today-building-resilience-for-tomorrow. Last accessed 2019-05-25.

International Seismological Centre, On-line Bulletin, http://www.isc.ac.uk, Internatl. Seis. Cent., Thatcham, United Kingdom, 2010. Last accessed 2015/05/01.

Jaynes, E.T. (1957). Information theory and statistical mechanics. II. *Physical Review*, 108(2), p.171-190. DOI: 10.1103/PhysRev.108.171.

Johnson, N.L., Kemp, A.W. & Kotz, S. (2005). *Univariate discrete distributions*. John Wiley & Sons, 672 pp.

Kagan, Y.Y. (1991). Seismic moment distribution. *Geophysical Journal International*, 106(1), pp. 123-134. DOI: 10.1111/j.1365-246X.1991.tb04606.x.

Kagan, Y.Y. (1997). Seismic moment-frequency relation for shallow earthquakes: Regional comparison. *Journal of Geophysical Research: Solid Earth*, 102(B2), pp. 2835-2852. DOI: 10.1029/96JB03386.

Kagan, Y.Y. (2002a). Seismic moment distribution revisited: I. Statistical results. *Geophysical Journal International*, 148(3), pp. 520-541. DOI: 10.1046/j.1365-246x.2002.01594.x.

Kagan, Y.Y. (2002b). Seismic moment distribution revisited: II. Moment conservation principle. *Geophysical Journal International*, 149(3), pp. 731-754. DOI: 10.1046/j.1365-246X.2002.01671.x.

Kagan, Y.Y. (2010). Statistical distributions of earthquake numbers: consequence of branching process. *Geophysical Journal International*, 180(3), pp. 1313-1328. DOI: 10.1111/j.1365-246X.2009.04487.x.

Kagan, Y.Y. & Schoenberg, F. (2001). Estimation of the upper cutoff parameter for the tapered Pareto distribution. *Journal of Applied Probability*, 38(A), pp. 158-175. DOI: 10.1239/jap/1085496599.

Karakaisis, G.F., Papazachos, C.B., Scordilis, E.M., Papaioannou, C.A. & Papazachos, B.C. (2002). Time variation of seismicity and seismic hazard in Hellenic Arctrench system. *XXXVIII General Assembly of ESC, pp.1-6. 1-6 September 2002, Genoa, Italy*.

Karlis, D. & Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review*, 73(1), pp. 35-58. DOI: 10.1111/j.1751-5823.2005.tb00250.x

Kendall, M.G. & Stuart, A. (1967). *The Advanced Theory of Statistics. Vol. 2, Inference and Relationship*, Griffin, London.

Khaliq, M.N., Ouarda, T.B.M.J., Ondo, J.C., Gachon, P. & Bobée, B. (2006). Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review. *Journal of Hydrology*, 329(3-4), pp. 534-552. DOI: 10.1016/j.jhydrol.2006.03.004.

Kijko, A. (2004). Estimation of the maximum earthquake magnitude m_{max}. *Pure and Applied Geophysics*, 161, pp. 1-27. DOI: 10.1007/s00024-004-2531-4.

Kijko, A. (2011). Seismic hazard. In *Encyclopedia of Solid Earth Geophysics* (pp. 1107-1121). Gupta, H.K (Ed.), Springer, Dordrecht.

Kijko, A. (2012). On Bayesian procedure for maximum earthquake magnitude estimation. *Research in Geophysics*, 2(1), p.7. DOI: 10.4081/rg.2012.e7.

Kijko, A. & Dessokey, M.M. (1987). Application of the extreme magnitude distribution to incomplete earthquake files. *Bulletin of the Seismological Society of America*, 77, pp. 1429-1436.

Kijko, A. & Graham, G. (1998). "Parametric-historic" procedure for probabilistic seismic hazard analysis. Part I: Assessment of maximum regional magnitude m_{max}. *Pure and Applied Geophysics*, 152, pp. 413-442. DOI: 10.1007/s000240050161.

Kijko, A. & Graham, G. (1999). "Parametric-historic" Procedure for Probabilistic Seismic Hazard Analysis Part II: Assessment of Seismic Hazard at Specified Site. *Pure and Applied Geophysics*, 154(1), pp. 1-22. DOI: 10.1007/s000240050218.

Kijko, A., Lasocki, S. & Graham, G. (2001). Non-parametric seismic hazard in mines. *Pure and Applied Geophysics*, 158(9-10), pp. 1655-1675. DOI: 10.1007/PL00001238.

Kijko, A., Retief, S.J.P. & Graham, G. (2003). Seismic hazard and risk assessment for Tulbagh, South Africa: Part II – Assessment of Seismic Risk. *Natural Hazards*, 30, pp. 25-41. DOI: 10.1023/A:1025036409750.

Kijko, A. & Sellevoll, M.A. (1989). Estimation of earthquake hazard parameters from incomplete data files, Part I, Utilization of extreme and complete catalogues with different threshold magnitudes. *Bulletin of the Seismological Society of America*, 79, pp. 645-654.

Kijko, A. & Sellevoll, M.A. (1992). Estimation of earthquake hazard parameters from incomplete data files. Part II. Incorporation of magnitude heterogeneity. *Bulletin of the Seismological Society of America*, 82, pp. 120-134.

Kijko, A. & Singh, M. (2011). Statistical tools for maximum possible earthquake magnitude estimation. *Acta Geophysica*, 59, pp. 674-700. DOI: 10.2478/s11600-011-0012-6.

Kijko. A. & Smit, A. (2012). Extension of the Aki-Utsu b-value estimator for incomplete catalogs. *Bulletin of the Seismological Society of America*, 102, pp. 1283-1287. DOI: 10.1785/0120110226.

Kijko, A. & Smit A. (2017) Estimation of the frequency–magnitude Gutenberg-Richter b-value without making assumptions on the level of completeness. *Seismological Research Letters*. 102(3), pp.1283-1287. DOI: 10.1785/0220160177.

Kijko, A., Smit, A. & Sellevoll, M.A. (2016). Estimation of earthquake hazard parameters from incomplete data files. Part III. Incorporation of uncertainty of earthquake-occurrence

model. *Bulletin of the Seismological Society of America*, 106(3), pp. 1210-1222. DOI: 10.1785/0120150252.

Kijko, A., Smit, A. & Sellevoll, M.A. (2017). Reply to "Comment on 'Estimation of Earthquake Hazard Parameters from Incomplete Data Files. Part III. Incorporation of Uncertainty of Earthquake-Occurrence Model' by Andrzej Kijko, Ansie Smit, and Markvard A. Sellevoll" by Gert Zöller. *Bulletin of the Seismological Society of America*, 107(4), pp. 1979-1982. DOI: 10.1785/0120160303.

Kimball, B.F. (1946). Sufficient statistical estimation functions for the parameters of the distribution of maximum values. *The Annals of Mathematical Statistics*, 17, pp. 299-306.

Klugman, S.A., Panjer, H.H. & Willmot, G.E. (2008). *Loss models. From data to decisions*. John Willey & Sons, Inc., Hoboken, New Jersey, pp.726.

Kohli, D., Sliuzas, R., Kerle, N. & Stein, A. (2012). An ontology of slums for image-based classification. *Computers, Environment and Urban Systems*, 36(2), pp. 154-163. DOI: 10.1016/j.compenvurbsys.2011.11.001.

Koravos, G.C., Yadav, R.B.S. & Tsapanos, T.M. (2015). Evaluation of tsunami potential based on conditional probability for specific zones of the Pacific tsunamigenic rim. *Tectonophysics*, 658, pp. 159-168. DOI: 10.1016/j.tecto.2015.07.018.

Kossobokov, V.G., Peresan, A. & Panza, G. (2015). Reality check: seismic hazard models you can trust. *Earth & Space Science News*, pp. 9-11. DOI: 10.1029/2015EO031919.

Kousky, C. (2014). Informing climate adaptation: A review of the economic costs of natural disasters. *Energy Economics*, 46, pp. 576-592. DOI: 10.1016/j.eneco.2013.09.029.

Kron, W., Steuer, M., Löw, P. & Wirtz, A. (2012). How to deal properly with a natural catastrophe database–analysis of flood losses. *Natural Hazards and Earth System Sciences*, 12(3), pp. 535-550. DOI: 10.5194/nhess-12-535-2012.

Krüger, F., & Scherbaum, F. (2014). The 29 September 1969, Ceres, South Africa, earthquake: Full waveform moment tensor inversion for point source and kinematic source parameters. *Bulletin of the Seismological Society of America*, 104, pp. 576-581. DOI: 10.1785/0120130209.

Kuffer, M., Pfeffer, K. & Sliuzas, R. (2016). Slums from space—15 years of slum mapping using remote sensing. *Remote Sensing*, 8(6), pp. 1-29. DOI: 10.3390/rs8060455.

Kunz, M. & Puskeiler, M. (2010). High-resolution assessment of the hail hazard over complex terrain from radar and insurance data. *Meteorologische Zeitschrift*, 19(5), pp. 427-439. DOI: 10.1127/0941-2948/2010/0452.

Lasocki, S., & Urban, P. (2011). Bias, variance and computational properties of Kijko's estimators of the upper limit of magnitude distribution, M_{max}. *Acta Geophysica*. 59, pp. 659–673. DOI: 10.2478/s11600-010-0049-y.

Lam, D., Thompson, C., Croke, J., Sharma, A. & Macklin, M. (2017). Reducing uncertainty with flood frequency analysis: The contribution of paleoflood and historical flood information. *Water Resources Research*, 53(3), pp. 2312-2327. DOI: 10.1002/2016WR019959.

LeCam L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Statistics,* 41, pp. 802-28. jstor.org/stable/2239236.

Leigh, R. (1998). *Hail damage to motor vehicles: an examination of economic costs*. Natural Hazards Research Centre, Macquarie University Sydney, Australia.

Leigh, R. & Kuhnel, I. (2001). Hailstorm loss modelling and risk assessment in the Sydney region, Australia. *Natural Hazards*, 24(2), pp. 171-185. DOI: 10.1023/A:1011855801345.

Leonard, M. (2010). Earthquake fault scaling: Self-consistent relating of rupture length, width, average displacement, and moment release. *Bulletin of the Seismological Society of America*, 100, pp. 1971-1988. DOI: 10.1785/0120090189.

Lin, I.-C. & Tung, C.C. (1982). A preliminary investigation of tsunami hazard. *Bulletin of the Seismological Society of America*, 72, pp. 2323–2337.

Liu, H., Chen, W. & Sudjianto, A. (2004), January. Relative entropy based method for global and regional sensitivity analysis in probabilistic design. In ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (pp. 983-992). American Society of Mechanical Engineers.

Liu, Y., Santos, A., Wang, S. M., Shi, Y., Liu, H. & Yuen, D. A. (2007). Tsunami hazards along Chinese coast from potential earthquakes in South China Sea. *Physics of the Earth and Planetary Interiors*, 163, pp. 233–244. DOI: 10.1016/j.pepi.2007.02.012.

Liu, Z., Xu, J. & Shi, K. (2014). Self-organized criticality of climate change. *Theoretical and Applied Climatology*, 115(3-4), pp. 685-691. DOI: 10.1007/s00704-013-0929-6.

Lockridge, P.A. & Smith, R.H. (1984). Tsunamis in the Pacific Basin, 1900–1983 (map). Boulder, CO, NOAA/NGDC.

Lomnitz, C. (1973). Poisson processes in earthquake studies. *Bulletin of the Seismological Society of America*, 63, pp. 735-758.

Lord, D. & Park, P.Y.J. (2008). Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. *Accident Analysis and Prevention*, 40(4), pp. 1441-1457. DOI: 10.1016/j.aap.2008.03.014.

Lorito, S., Tiberti, M.M., Basili, R., Piatanesi, A. & Valensise, G. (2008). Earthquake-generated tsunamis in the Mediterranean Sea: Scenarios of potential threats to Southern Italy. *Journal of Geophysical Research*, 113, B01301, pp. 1-14. DOI: 10.1029/2007JB004943.

Lyons, L. (1989) *Statistics for Nuclear and Particle Physicists*. Cambridge University Press.

Lyubushin, A.A. & Parvez, I.A. (2010). Map of seismic hazard of India using Bayesian approach. *Natural Hazards*, 55(2), pp. 543-556. DOI: 10.1007/s11069-010-9546-1.

Lyubushin, A.A., Tsapanos, T.M., Pisarenko, V.F. & Koravos, G.C. (2002). Seismic hazard for selected sites in Greece: a Bayesian estimate of seismic peak ground acceleration. *Natural Hazards*, 25(1), pp. 83-98. DOI: 10.1023/A:1013342918801.

Machado, J.T., Pinto, C.M. & Lopes, A.M. (2015). A review on the characterization of signals and systems by power law distributions. *Signal Processing*, 107, pp. 246-253. DOI: 10.1016/j.sigpro.2014.03.003.

Main, I. (1996). Statistical physics, seismogenesis, and seismic hazard. *Reviews of Geophysics*, 34(4), pp. 433-462. DOI: 10.1029/96RG02808.

Main, I.G. & Burton, P.W. (1984). Information theory and the earthquake frequency–magnitude distribution. *Bulletin of the Seismological Society of America*, 74(4), pp. 1409-1426.

Malamud, B.D., Turcotte, D.L., Guzzetti, F. & Reichenbach, P. (2004). Landslide inventories and their statistical properties. *Earth Surface Processes and Landforms*, 29(6), pp. 687-711. DOI: 10.1002/esp.1064.

Manfré, L.A., Hirata, E., Silva, J.B., Shinohara, E.J., Giannotti, M.A., Larocca, A.P.C. & Quintanilha, J.A. (2012). An analysis of geospatial technologies for risk and natural disaster management. *ISPRS International Journal of Geo-Information*, 1(2), pp. 166-185. DOI: 10.3390/ijgi1020166.

Markatou, M., Basu, A. & Lindsay, G.G. (1998). Weighted likelihood estimating equations with a bootstrap root search. *Journal of the American Statistical Association*, 93, pp. 740-750. DOI: 10.1080/01621459.1998.10473726.

Marin, J.M., Mengersen, K. & Robert, C.P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics*, 25, pp. 459-507. DOI: 10.1016/S0169-7161(05)25016-2.

McCalpin, J.P. (2009). Application of paleoseismic data to seismic hazard assessment and neotectonic research. *International Geophysics*, 95, pp. 1-106. DOI: 10.1016/S0074-6142(09)95009-4.

McDonald, J.B. & Butler, R.J. (1987). Some generalized mixture distributions with an application to unemployment duration. *The Review of Economics and Statistics*, pp. 232-240. jstor.org/stable/1927230.

McGarr, A. (1984). Some applications of seismic source mechanism studies to assessing underground hazard. In *Rockbursts and Seismicity in Mines*. pp. 199-208.

McGuire, R.K. (1976). *Fortran computer program for seismic risk analysis*. USGS. Open File Rep. 76-67.

McGuire, R.K. (2004). *Seismic hazard and risk analysis*. Earthquake Engineering Research Institute, MNO-10, pp. 221.

McGuire, R.K. & Arabasz, W.J. (1990). An introduction to probabilistic seismic hazard analysis. *Geotechnical and Environmental Geophysics*, 1, pp. 333-353.

McGuire, R.K. & Toro, G. (1986). *Methods of earthquake ground motion estimation for the eastern United States.* Electric Power Research Institute Research Project No. RP2556-16, prepared by Risk Engineering, Inc.

Mignan, A. & Woessner, J. (2012). Estimating the magnitude of completeness for earthquake catalogs. *Community Online Resource for Statistical Seismicity Analysis*. DOI: 10.5078/corssa-00180805.

Molchan, G.M., Keilis-Borok, V.L. & Vilkovich, V. (1970). Seismicity and principal seismic effects. *Geophysical Journal International,* 21, pp. 323–335. DOI: 10.1111/j.1365-246X.1970.tb01795.x.

Mora, P., Matsu'ura, M., Madariaga, R. & Minster, J.B. (2000). Microscopic and macroscopic simulation: Towards predictive modelling of the earthquake process. *Pure and Applied Geophysics*, *157*(11/12).

Mortgat, C.P. & Shah, H.C. (1979). A Bayesian model for seismic hazard mapping. *Bulletin of the Seismological Society of America*, 69(4), pp. 1237-1251.

Mulargia, F. & Tinti, S. (1985). Seismic sample areas defined from incomplete catalogs: An application to the Italian territory. *Physics of the Earth and Planetary Interiors*, 40 (4), pp. 273–300.

Nelder, J.A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4), pp. 308-313. DOI: 10.1093/comjnl/7.4.308.

Newman, M.E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5), pp. 323-351. DOI: 10.1080/00107510500052444.

Nomura, S., Ogata, Y., Komaki, F. & Toda, S. (2011). Bayesian forecasting of recurrent earthquakes and predictive performance for a small sample size. *Journal of Geophysical Research: Solid Earth,* 116(B4). DOI: 10.1029/2010JB007917.

Norio, O., Ye, T., Kajitani, Y., Shi, P. & Tatano, H. (2011). The 2011 Eastern Japan Great Earthquake Disaster: Overview and Comments. *International Journal of Disaster Risk Science*, 2, 1. p. 34-42. DOI: 10.1007/s13753-011-0004-9.

Nott, J.F. (2003). Intensity of prehistoric tropical cyclones. *Journal of Geophysical Research: Atmospheres*, 108(D7). DOI: 10.1029/2002JD002726.

Nuttli, O.W. (1981). On the problem of maximum magnitude of earthquakes. In Evaluation of regional seismic hazards and risk. *Proceedings of Conference XIII: US. Geological Survey Open-File Report*, Vol. 437, pp. 111-123.

Ogata, Y. (1999a). Estimating the hazard of rupture using uncertain occurrence times of paleoearthquakes. *Journal of Geophysical Research: Solid Earth*, 104(B8), pp. 17995-18014. DOI: 10.1029/1999JB900115.

Ogata, Y. (1999b). Seismicity analysis point-process modeling: A review. *Pure and Applied Geophysics*. 155, pp. 471-507. DOI: 10.1007/978-3-0348-8677-2_14.

Ogata, Y. & Abe, K. (1991). Some statistical features of the long-term variation of the global and regional seismic activity. *International Statistical Review/Revue Internationale de Statistique*, 59, pp. 139-161. jstor.org/stable/1403440.

Olivier, J. & Van Rensburg, P.A. (1992). The impact of hail on Transvaal maize production. *Agrekon*, 31(2), pp.62-73. ageconsearch.umn.edu/record/267518/files/agrekon-31-02-003.pdf.

Orfanogiannaki, K. & Papadopoulos, G.A. (2007). Conditional probability approach of the assessment of tsunami potential: application in three tsunamigenic regions of the Pacific Ocean. *Pure and Applied Geophysics*, 164(2-3), pp. 593-603. DOI: 10.1007/s00024-006-0170-7.

Page, R. (1968). Aftershocks and microaftershocks. *Bulletin of the Seismological Society of America*, 58, pp. 1131-1168.

Papadopoulos, G.A. & Imamura, F. (2001). A proposal for a new tsunami intensity scale. *Proceedings of the International Tsunami Symposium* 2001. pp. 569–577.

Papastamatiou, D. (1980). Incorporation of crustal deformation to seismic hazard analysis. *Bulletin of the Seismological Society of America*, 70(4), pp. 1321-1335.

Papoulia, J., Stavrakakis, G. & Papanikolaou, D. (2001). Bayesian estimation of strong earthquakes in the Inner Messiniakos fault zone, southern Greece, based on seismological and geological data. *Journal of Seismology*, 5(2), pp. 233-242. DOI: 10.1023/A:1011491103922.

Parsons T., Geist E.L. (2008) Tsunami Probability in the Caribbean Region. In: Cummins P.R., Satake K., Kong L.S.L. (eds) *Tsunami Science Four Years after the 2004 Indian Ocean Tsunami*. Pageoph Topical Volumes. Birkhäuser Basel. DOI: 10.1007/978-3-0346-0057-6_7.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33, pp. 1065-1076. jstor.org/stable/2237880.

Parvez, I.A. (2007). On the Bayesian analysis of the earthquake hazard in the North-East Indian peninsula. *Natural Hazards*, 40(2), pp. 397-412. DOI: 10.1007/s11069-006-9002-4.

Patskoski, J. & Sankarasubramanian, A. (2018). Reducing uncertainty in stochastic streamflow generation and reservoir sizing by combining observed, reconstructed and projected streamflow. *Stochastic Environmental Research and Risk Assessment*, 32(4), pp. 1065-1083. DOI: 10.1007/s00477-017-1456-2.

Paulatto, M., Pinat, T. & Romanelli, F. (2007). Tsunami hazard scenarios in the Adriatic Sea domain. *Natural Hazards and Earth System Sciences*, 7, pp. 309–325, DOI: 10.5194/nhess-7-309-2007.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*. A, 185, pp. 71-110. jstor.org/stable/90667.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical transactions of the Royal Society of London*, 186(Part I), pp. 343-424.

Petersen, M.D., Cramer, C.H. & Frankel, A.D. (2002). Simulations of seismic hazard for the Pacific Northwest of the United States from earthquakes associated with the Cascadia subduction zone. In *Earthquake Processes: Physical Modelling, Numerical Simulation and Data Analysis Part I*, Basel: Birkhäuser, pp. 2147–2168. DOI: 10.1007/978-3-0348-8203-3_15.

Pienaar, N., Dyson, L.L. & Klopper, E. (2015) September. Atmospheric anomalies during the 2012/13 extreme hail seasons over Gauteng. In *Proceedings of the 31st Annual conference of South African Society for Atmospheric Sciences*, pp. 21-22.

Pisarenko, V.F. (1991). Statistical evaluation of maximum possible magnitude, Izvestiya. *Earth Physics*. 27, pp. 757–763.

Pisarenko, V.F. & Lyubushin, A.A. (1997). Statistical estimation of maximum peak ground acceleration at a given point of a seismic region. *Journal of Seismology*, 1, pp. 395-405. DOI: 10.1023/A:1009795503733.

Pisarenko, V.F., Lyubushin, A.A., Lysenko, V.B. & Golubeva, T.V. (1996). Statistical estimation of seismic hazard parameters: Maximum possible magnitude and related parameters. *Bulletin of the Seismological Society of America*, 86(3), pp. 691-700.

Pisarenko, V.F & Rodkin, M.V. (2010). *Heavy-tailed distributions in disaster analysis* (Vol. 30). Springer Science & Business Media.

Pisarenko, V.F. & Rodkin, M.V. (2014). *Statistical analysis of natural disasters and related losses*. Cham: Springer.

Pisarenko, V.F. & Sornette, D. (2003). Characterization of the frequency of extreme earthquake events by the generalized Pareto distribution. *Pure and Applied Geophysics*, 160(12), pp. 2343-2364. DOI: 10.1007/s00024-003-2397-x.

Pisarenko, V.F., Sornette, A., Sornette, D. & Rodkin, M.V. (2008). New approach to the characterization of M max and of the tail of the distribution of earthquake magnitudes. *Pure and Applied Geophysics*, 165(5), pp. 847-888. DOI: 10.1007/s00024-008-0341-9.

Pisarenko, V.F., Sornette, A., Sornette, D. & Rodkin, M.V. (2014). Characterization of the tail of the distribution of earthquake magnitudes by combining the GEV and GPD descriptions of extreme value theory. *Pure and Applied Geophysics*, 171(8), pp. 1599-1624. DOI: 10.1007/s00024-014-0882-z.

Poolman, E.R. (1992). *Die voorspelling van haelkorrelgroei in Suid-Afrika (The forecasting of hail growth in South Africa).* M.S. Thesis, Faculty of Engineering, University of Pretoria, 113 pp.

Power, W., Downes, G. & Stirling, M. (2007). Estimation of tsunami hazard in New Zealand due to South American earthquakes. *Pure and Applied Geophysics*, 164, pp. 547–564, DOI: 10.1007/s00024-006-0166-3. Also In: Satake K., Okal E.A., Borrero J.C. (eds) Tsunami and Its Hazards in the Indian and Pacific Oceans. Pageoph Topical Volumes. Birkhäuser Basel. DOI: 10.1007/978-3-7643-8364-0_15.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1986). *Numerical recipes. The art of scientific computing (FORTRAN version)*. Cambridge University Press, 818 pp.

Punge, H.J., Bedka, K.M., Kunz, M. & Werner, A. (2014). A new physically based stochastic event catalog for hail in Europe. *Natural Hazards*, 73(3), pp. 1625-1645. DOI: 10.1007/s11069-014-1161-0.

Punge, H.J. & Kunz, M. (2016). Hail observations and hailstorm characteristics in Europe: A review. *Atmospheric Research*, 176, pp. 159-184. DOI: /10.1016/j.atmosres.2016.02.012.

PWC (2014). Opportunity knocks Insurance industry analysis Analysis of major South African insurers' results for the year ended 31 December 2013 https://www.pwc.co.za/en/assets/pdf/insurance-industry-analysis-march-2014.pdf; last accessed 2018/09/13.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 3-4, pp 353-360, DOI: 10.1093/biomet/43.3-4.353.

Rao, C. R. (1973). *Linear statistical inference and its application*, Edition 2. John Willey and Sons, New York, pp.625.

Rhoades, D.A. (1996). Estimation of the Gutenberg-Richter relation allowing for individual earthquake magnitude uncertainties. *Tectonophysics*, 25, pp. 871-83. DOI: 10.1016/0040-1951(95)00182-4.

Rhoades, D.A. & Dowrick, D.J. (2000). *Effects of magnitude uncertainties on seismic hazard estimates.* In Proceedings of the 12th World Conference on Earthquake Engineering. Auckland. Publisher: Upper Hutt, N.Z.: New Zealand National Society for Earthquake Engineering.

Rhoades, D.A., Van Dissen, R.J. & Dowrick, D.J. (1994). On the handling of uncertainties in estimating the hazard of rupture on a fault segment. *Journal of Geophysical Research: Solid Earth*, 99(B7), pp. 13701-13712. DOI: 10.1029/94JB00803.

Rikitake, T. & Aida, I. (1988). Tsunami hazard probability in Japan. *Bulletin of the Seismological Society of America*, 78, pp. 1268–1278.

Roberts, W.J.R. & Alexander, W.J.R. (1982). Lessons learnt from the 1981 Laingsburg flood. *Civil Engineering (Siviele Ingenieurswese)*, 24(1), pp. 17-27. hdl.handle.net/10520/AJA10212019_20986.

Robson, D.S. & Whitlock, J.H. (1964). Estimation of a truncation point. *Biometrika*, 51(1/2), pp. 33-39. jstor.org/stable/2334193.

Rosenblueth, E. (1986). Use of statistical data in assessing local seismicity. *Earthquake Engineering & Structural Dynamics*, 14, pp. 325–337. DOI: 10.1002/eqe.4290140302.

Rosenblueth, E. & Ordaz, M. (1987). Use of seismic data from similar regions. *Earthquake Engineering & Structural Dynamics*, 15, pp. 619–634. DOI: 10.1002/eqe.4290150507.

Rotondi, R. & Varini, E. (2007). Bayesian inference of stress release models applied to some Italian seismogenic zones. *Geophysical Journal International*, 169, pp. 301-314. DOI: 10.1111/j.1365-246X.2006.03314.x.

Rougier, J., Sparks, R.S.J., Cashman, K.V. & Brown, S.K. (2018). The global magnitude–frequency relationship for large explosive volcanic eruptions. *Earth and Planetary Science Letters*, 482, pp. 621-629. DOI: 10.1016/j.epsl.2017.11.015.

Sachs, M.K., Yoder, M.R., Turcotte, D.L., Rundle, J.B. & Malamud, B.D. (2012). Black swans, power laws, and dragon-kings: Earthquakes, volcanic eruptions, landslides, wildfires, floods, and SOC models. *The European Physical Journal Special Topics*, 205(1), pp. 167-182. DOI: 10.1140/epjst/e2012-01569-3.

Sahakian, V., Bormann, J., Driscoll, N., Harding, A., Kent, G. & Wesnousky, S. (2017). Seismic constraints on the architecture of the Newport-Inglewood/Rose Canyon fault: Implications for the length and magnitude of future earthquake ruptures. *Journal of Geophysical Research: Solid Earth*, 122(3), pp. 2085-2105. DOI: 10.1002/2016JB013467.

Sander, J., Eichner, J.F., Faust, E. & Steuer, M. (2013). Rising variability in thunderstorm-related US losses as a reflection of changes in large-scale thunderstorm forcing. *Weather, Climate, and Society*, 5(4), pp. 317-331. DOI: 10.1175/WCAS-D-12-00023.1.

Saunders, I., Brandt, M., Steyn, J., Roblin, D. & Kijko, A. (2008). The South African National Seismograph Network. *Seismological Research Letters*, 792, pp. 203-210. DOI: 10.1785/gssrl.79.2.203.

Schorlemmer, D., Gerstenberger, M.C., Wiemer, S., Jackson, D.D. & Rhoades, D.A. (2007). Earthquake likelihood model testing. *Seismological Research Letters*, 78(1), pp.17-29.

Schnebele, E. & Cervone, G. (2013). Improving remote sensing flood assessment using volunteered geographical data. *Natural Hazards and Earth System Sciences*, 13, pp. 669–677. DOI: 10.5194/nhess-13-669-2013.

Scholz, C.H. (2015). On the stress dependence of the earthquake b value. *Geophysical Research Letters*, 42(5), pp. 1399-1402. DOI: 10.1002/2014GL062863.

Sharma, S., Baruah, S., Sahu, O.P., Bora, P.K. & Duarah, R. (2013). Low b-value prior to the Indo-Myanmar subduction zone earthquakes and precursory swarm before the May 1995 M 6.3 earthquake. *Journal of Asian Earth Sciences*, 73, pp. 176-183. DOI: 10.1016/j.jseaes.2013.04.019.

Shaw, B., Ambraseys, N.N., England, P.C., Floyd, M.A., Gorman, G.J., Higham, T.F.G., Jackson, J.A., Nocquet, J.-M., Pain, C.C. & Piggott, M.D. (2008). Eastern Mediterranean tectonics and tsunami hazard inferred from the AD 365 earthquake. *Nature Geoscience*, 1, pp. 268–276, DOI: 10.1038/ngeo151.

Shi, K. & Liu, C.Q. (2009). Self-organized criticality of air pollution. *Atmospheric Environment*, 43(21), pp. 3301-3304. DOI: 10.1016/j.atmosenv.2009.04.013.

Shi, Y. & Bolt, B.A. (1982). The standard error of the magnitude-frequency b value. *Bulletin of the Seismological Society of America*, 72, pp. 1677-1687.

Sieberg, A. (1927). *Geologische, physikalische und angewandte Erdbebenkund*. Verslag von Gustav Fisher, Jena.

Silva, A.T., Portela, M.M., Naghettini, M. & Fernandes, W. (2017). A Bayesian peaks-over-threshold analysis of floods in the Itajaí-açu River under stationarity and nonstationarity. *Stochastic*

Environmental Research and Risk Assessment, 31(1), pp. 185-204. DOI: 10.1016/10.1007/s00477-015-1184-4.

Simpson, D.W. & Richards, P.G. (1981). *Earthquake prediction, an international review*. Maurice Ewing series IV, Eds: D. W. Simpson, P.G. Richards. American Geophysical Union, Washington, D.C., 680 pp.

Sioutas, M., Meaden, T. & Webb, J.D.C. (2005). Hail Frequency, distribution and intensity in Northern Greece. Atmospheric Research, 93(1-3), pp.526-533. DOI: 10.1016/j.atmosres.2008.09.023 .

Smit, A., Kijko, A. & Stein A. (2017). Probabilistic Tsunami Hazard Assessment from Incomplete und Uncertain Historical Catalogues with an Application to tsunamigenic regions in the Pacific Ocean. *Pure and Applied Geophysics*, Global Tsunami Science: Past and Future, Volume II, 174(8), pp. 3065-3081. DOI: 10.1007/s00024-017-1564-4.

Smit A., Stein A. & Kijko A. (2019). Bayesian Inference in Natural Hazard Analysis for Incomplete and Uncertain Data. Environmetrics. p.e2566. DOI: 10.1002/env.2566.

Soloviev, S.L. (1970). Recurrence of tsunamis in the Pacific. In *Tsunamis in the Pacific Ocean*, edited by W.M. Adams, East-West Centre Press, Honolulu, pp. 149-163.

Sørensen, M.B., Spada, M., Babeyko, A., Wiemer, S. & Grünthal G. (2012). Probabilistic tsunami hazard in the Mediterranean Sea. *Journal of Geophysical Research*, 117, B01305, pp. 1-15. DOI: 10.1029/2010JB008169.

Sornette, D. & Sornette, A. (1999). General theory of the modified Gutenberg-Richter law for large seismic moments. *Bulletin of the Seismological Society of America*, 89(4), pp. 1121-1130.

Stavrakakis, G.N. & Tselentis, G.A. (1987). Bayesian probabilistic prediction of strong earthquakes in the main seismogenic zones of Greece. *Bollettino Di Geofisica Teorica Ed Applicata*, 113, pp. 51-63.

Stedinger, J.R. & Cohn, T.A. (1986). Flood frequency analysis with historical and paleoflood information. *Water Resources Research*, 22(5), pp. 785-793. DOI: 10.1029/ WR022i005p00785.

Stein, R.S. & Hanks, T.C. (1998). M ≥ 6 earthquakes in southern California during the twentieth century: No evidence for a seismicity or moment deficit. *Bulletin of the Seismological Society of America*, 88(3), pp. 635-652.

Stirling, M., Goded, T., Berryman, K. & Litchfield, N. (2013). Selection of earthquake scaling relationships for seismic-hazard analysis. *Bulletin of the Seismological Society of America*. 103, pp. 2993-3011. DOI: 10.1785/0120130052.

Sun, X., Thyer, M., Renard, B. & Lang, M. (2014). A general regional frequency analysis framework for quantifying local-scale climate effects: A case study of ENSO effects on southeast Queensland rainfall. *Journal of Hydrology*, 512, pp. 53-68. DOI: 10.1016/j.jhydrol.2014.02.025.

Talbi, A. & Yamazaki, F. (2009). A mixed model for earthquake inter-event times. *Journal of Seismology*, 10, pp. 289-307. DOI: 10.1007/s10950-009-9166-y.

Tatard, L., Grasso, J.R., Helmstetter, A. & Garambois, S. (2010). Characterization and comparison of landslide triggering in different tectonic and climatic settings. *Journal of Geophysical Research: Earth Surface*, H 115(F4), pp. 1-18. DOI: 10.1029/2009JF001624.

Tate, R.F. (1959). Unbiased estimation: functions of location and scale parameters. *The Annals of Mathematical Statistics*, pp. 341-366. jstor.org/stable/2237086.

Tatsumi, D., Calder, C.A. & Tomita, T. (2014). Bayesian near-field tsunami forecasting with uncertainty estimates. *Journal of Geophysical Research: Oceans*, 119, pp. 2201–2211. DOI: 10.1002/2013JC009334.

Terblanche D.E. (1985) *Die hailstorm van 1 November 1985*. South African Weather Bureau monthly newsletter. November 1985. 1–6.

Thio, H. K., Somerville, P. & Ichinose, G. (2007). Probabilistic analysis of strong ground motion and tsunami hazard in Southeast Asia. *Journal of Earthquake and Tsunami*, 1(2), pp. 119–137. DOI: 10.1142/S1793431107000080.

Tinti, S. (1993). *Evaluation of tsunami hazard in Calabria and eastern Sicily, Italy, Tsunamis in the world*. In (ed. Tinti, S.) Advances in Natural and Technological Hazards Research (Kluwer Academic Publishers, Dordrecht), pp. 141–157.

Tinti, S. & Armigliato, A. (2003). The use of scenarios to evaluate tsunami impact in Southern Italy. *Marine Geology*, 199, pp. 221–243. DOI: 10.1016/ S0025-3227(03)00192-0.

Tinti, S., Armigliato, A., Tonini, R., Maramai, A. & Graziani, L. (2005). Assessing the hazard related to tsunamis of tectonic origin: a hybrid statistical-deterministic method applied to southern Italy coasts. *ISET Journal of Earthquake Technology*, 42, pp. 189-201.

Tinti, S. & Maramai, A. (1996). Catalogue of tsunamis generated in Italy and in Cote d' Azur, France: a step towards a unified catalogue of tsunamis in Europe. *Annali di Geofisica*, 39, pp. 1253–1299.

Tinti, S., Maramai, A. & Graziani, L. (2001). A new version of the European tsunami catalogue: updating and revision. *Natural Hazards and Earth System Science*, Copernicus Publications on behalf of the European Geosciences Union 1(4): 255-262.

Tinti, S. & Mulargia, F. (1985). Effects of magnitude uncertainties in the Gutenberg-Richter frequency–magnitude law. *Bulletin of the Seismological Society of America*, 75, pp. 1681-1697.

Tonini, R., Armigliato, A., Pagnoni, G., Zaniboni, F. & Tinti, S. (2011). Tsunami hazard for the city of Catania, eastern Sicily, Italy, assessed by means of Worst-case Credible Tsunami Scenario Analysis (WCTSA). *Natural Hazards and Earth System Sciences*, 11(5), p. 1217-1232.

United Nations Office for Disaster Risk Reduction (UNISDR) (2018). Economic Losses, Poverty and Disasters 1998-2017. Report by United Nations Office for Disaster Risk Reduction (UNISDR) and the Centre for Research on the Epidemiology of Disasters (CRED), part of the Institute of Health and Society (Université catholique de Louvain). unisdr.org/2016/iddr/IDDR2018_Economic%20Losses.pdf. Last accessed 2019/01/12.

Urlaub, M., Talling, P.J. & Masson, D.G. (2013). Timing and frequency of large submarine landslides: Implications for understanding triggers and future geohazard. *Quaternary Science Reviews*, 72, pp. 63–82. DOI: 10.1016/j.quascirev.2013.04.020.

Utsu, T. (1965). A method for determining the value of b in the formula log(n) = a–bM showing the magnitude-frequency relation for earthquakes (with English summary). *Geophysics Bulletin Hokkaido University*. 13, pp. 99-103.

Vere-Jones, D., Robinson, R. & Yang, W. (2001). Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation. *Geophysical Journal International*, 144(3), pp. 517-531. DOI: 10.1046/j.1365-246x.2001.01348.x.

Vermeulen, P.J. & Kijko, A. (2017). More statistical tools for maximum possible earthquake magnitude estimation. *Acta Geophysica*, 65(4), pp. 579-587. DOI: 10.1007/s11600-017-0048-3.

Vicini, L., Hotta, L.K. & Achar, J.A. (2013). Non-homogeneous Poisson process in the presence of one or more change points: an application to air pollution data. *Journal of Environmental Statistics*, 5(3), pp. 1-22. jenvstat.org/v05/i03.

Viglione, A., Merz, R., Salinas, J.L. & Blöschl, G. (2013). Flood frequency hydrology: 3. A Bayesian analysis. *Water Resources Research*, 49(2), pp. 675-692. DOI: 10.1029/2011wr010782.

Vinet, F. (2001). Climatology of hail in France. *Atmospheric Research*. 56, pp. 309–323.

Visser, P.J. & Van Heerden, J. (2000). Comparisons of hail kinetic energy derived from radar reflectivity with crop damage reports over the eastern Free State. *Water SA-Pretoria*, 26(1), pp. 91-96.

Von der Linden, W., Dose, V. & Von Toussaint, U. (2014). *Bayesian Probability Theory. Applications in the Physical Sciences*. Cambridge University Press.

Wald, D.J., Quitoriano, V., Worden, C.B., Hopper, M. & Dewey, J.W. (2011). USGS "Did You Feel It?" internet-based macroseismic intensity maps. *Annals of Geophysics*, 54(6), pp. 688-707. DOI: 10.4401/ag-5354

Wang, X. (1992). *Combining Information: Statistical Issues and Opportunities for Research. U.S. National Research Council*. National Academy Press, Washington, DC.

Wang, X. (2001). *Weighted Likelihood Estimation*. Doctoral Dissertation, Department of Statistics. University of British Columbia, Vancouver, Canada.

Wang, X. (2006). Approximating Bayesian inference by weighted likelihood. *Canadian Journal of Statistics*, 34(2), pp. 279-298. DOI: 10.1002/cjs.5550340206.

Wang, X, Van Eeden, C. & Zidek, J.V. (2004). Asymptotic properties of the maximum weighted likelihood estimators. *Journal of Statistical Planning and Inference*, 119, pp. 37–54. DOI: 10.1016/S0378-3758(02)00410-X.

Ward, S.N. (1997). More on M_{max}. *Bulletin of the Seismological Society of America*, 87(5), pp. 1199-1208.

Ward, S.N. (2001). Landslide tsunami. *Journal of Geophysical Research*, 106, pp. 11201–11215. DOI: 10.1029/2000JB900450.

Ward, S.N. (2003). Tsunamis. In: *The Encyclopedia of physical science and technology*, R. A. Meyers (ed.), Academic Press, pp. 175–191.

Ward, S. N. & Asphaug, E. (2000). Asteroid impact tsunami: a probabilistic hazard assessment. *Icarus* 145, pp. 64–78. DOI: 10.1006/icar.1999.6336.

Wartman, J., Dunham, L., Tiwari, B. & Pradel, D. (2013). Landslides in eastern Honshu induced by the 2011 Tohoku earthquake. *Bulletin of the Seismological Society of America*, 103(2B), pp. 1503-1521. DOI: 10.1785/0120120128.

Webb, J.D.C., Elsom, D.M. & Meaden, G.T. (1986). The TORRO hailstorm intensity scale. *Journal of Meteorology*, 11, pp. 337-339.

Webb, J.D.C., Elsom, D.M. & Meaden, G.T. (2009). Severe hailstorms in Britain and Ireland, a climatological survey and hazard assessment. *Atmospheric Research*, 93(1-3), pp. 587-606. DOI: 10.1016/j.atmosres.2008.10.034.

Weichert, D. (1980). Estimation of the earthquake recurrence parameters for unequal observation periods for different magnitudes. *Bulletin of the Seismological Society of America*, 70, pp. 1337-1346.

Weiss, L. & Wolfowitz, J. (1973). Maximum likelihood estimation of a translation parameter of a truncated distribution. *The Annals of Statistics*, 1(5), pp. 944-947.

Wells, D.L. & Coppersmith, K.J. (1994). New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the Seismological Society of America*, 84, pp. 974-1002.

Wheeler, R.L. (2009). *Methods of Mmax estimation east of the Rocky Mountains*. U.S. Geological Survey Open-File Report 2009-1018, 44 pp.

Wijetunge, J.J. (2014). A deterministic analysis of tsunami hazard and risk for the southwest coast of Sri Lanka. *Continental Shelf Research*, 79, pp. 23-35. DOI: 10.1016/j.csr.2013.09.009.

Witt, A., Malamud, B.D., Rossi, M., Guzzetti, F. & Peruccacci, S. (2010). Temporal correlations and clustering of landslides. *Earth Surface Processes and Landforms*, 35(10), pp. 1138-1156. DOI: 10.1002/esp.1998.

Woodroofe, M. (1972). Maximum likelihood estimation of a translation parameter of a truncated distribution. *The Annals of Mathematical Statistics*, pp. 113-122. jstor.org/stable/2239904.

Woodroofe, M. (1974). Maximum likelihood estimation of translation parameter of truncated distribution II. *The Annals of Statistics*, pp. 474-488. jstor.org/stable/2958134.

Working Group on Central California Earthquake Probabilities (WGCEP) (1995). Seismic Hazard in Southern California: Probable Earthquakes, 1994 to 2024. *Bulletin of the Seismological Society of America*, 85, pp. 379–439.

Working Group MPS (2004). Drafting of the seismic hazard map provided by the Ordinance PCM 3274 of 20 March 2003. Final Report for the Department of Civil Protection, INGV, Milan-Rome, April 2004, 65 pp. + 5 appendices.

Wyss, M., Nekrasova, A. & Kossobokov, V. (2012). Errors in expected human losses due to incorrect seismic hazard estimates. *Natural Hazards*, 62(3), pp. 927–935. DOI: 10.1007/s11069-012-0125-5.

Yadav, R.B.S., Tsapanos, T.M., Tripathi, J.N. & Chopra, S. (2013). An evaluation of tsunami hazard using Bayesian approach in the Indian Ocean. *Tectonophysics*, 593, pp. 172-182. DOI: 10.1016/j.tecto.2013.03.004.

Yazdani, A. & Kowsari, M. (2013). Bayesian estimation of seismic hazards in Iran. *Scientia Iranica*, 20(3), pp. 422-430. DOI: 10.1016/j.scient.2012.12.032.

Youngs, R.R. & Coppersmith, K.J. (1985). Implications of fault slip rates and earthquake recurrence models to probabilistic seismic hazard estimates. *Bulletin of the Seismological Society of America*, 75(4), pp.939-964.

Zhang, J.Z. & Song, L.Y. (1981). On the method of estimating b-value and its standard error. *Acta Seismologica Sinica*, 3(3), pp. 292-301.

Zöller, G. (2018). A statistical model for earthquake recurrence based on the assimilation of paleoseismicity, historic seismicity, and instrumental seismicity. *Journal of Geophysical Research: Solid Earth*, 123(6), pp. 4906-4921. DOI: 10.1029/2017JB015099.

Zawada, P.K. (1994). Palaeoflood hydrology of the Buffels River, Laingsburg, South Africa: was the 1981 flood the largest? *South African Journal of Geology*, 97(1), pp.21-32. hdl.handle.net/10520/AJA10120750_807.

Appendix: Proofs of Equations

Equation 3.5a and b.

Following the definition of the cumulative distribution function,

$$F_X(X > x) = 1 - F_X(X \leq x).$$

Assuming e^{α} in Eq. 2.3 is a constant, $n_{X \geq x} = e^{-\beta(x-0)}$ and $n_{TOT} = e^{-\beta(x-x)}$ such that

$$F_X(X \leq x) = \ 1 - F_X(X > x)$$

$$= 1 - \frac{e^{-\beta(x-0)}}{e^{-\beta(x-x)}}.$$

$$F_X(x) = 1 - e^{-\beta x}. \tag{3.5a}$$

$$f_X(x) = \frac{d}{dx} F_X(x).$$

$$f_X(x) = \beta e^{-\beta x}. \tag{3.5b}$$

■

Equation 3.6a and b.

$$F_X(X > x | x \geq x_{min}) = 1 - F_X(X \leq x | x \geq x_{min}).$$

Assuming e^{α} in Eq. 2.3 is a constant, $n_{X>x|x \geq x_{min}} = e^{-\beta(x-x_{min})}$ is defined on the interval $[x, \infty)$ and $n_{TOT|x \geq x_{min}} = e^{-\beta(x_{min}-x_{min})} = 1$ defined on the interval $[x_{min}, \infty)$ such that

$$F_X(x|x_{min}) = \ 1 - F_X(X > x | x \geq x_{min}) = \begin{cases} 0 & x < x_{min} \\ 1 - e^{-\beta(x-x_{min})} & x \geq x_{min} \end{cases}. \tag{3.6a}$$

$$f_X(x) = \frac{d}{dx} F_X(x|x_{\min})$$

$$f_X(x) = \begin{cases} 0 & x < x_{\min} \\ \beta e^{-\beta(x-x_{\min})} & x \geq x_{\min} \end{cases}. \tag{3.6b}$$

■

Equation 3.7a and b.

The shifted-truncated exponential distribution is derived by normalising the conditional PDF of the shifted exponential distribution derived in Eq. 3.6a.

$$f_X(x|x_{min},x_{max}) = \begin{cases} 0 & x < x_{min} \\ \dfrac{f_X(x|x_{min})}{F_X(x_{max}|x_{min})} & x_{min} \le x \le x_{max} \\ 0 & x > x_{max} \end{cases}.$$

$$f_X(x|x_{min},x_{max}) = \begin{cases} \dfrac{\beta e^{-\beta(x-x_{min})}}{1-e^{-\beta(x_{max}-x_{min})}} & x_{min} \le x \le x_{max} \\ 0 & x < x_{min}\ and\ x > x_{max} \end{cases}. \quad (3.7b)$$

The conditional CDF of the normalised, shifted-truncated exponential distribution equals

$$F_X(x|x_{min},x_{max}) = \int_{x_{min}}^{x} f_X(u)\,du$$

$$F_X(x|x_{min},x_{max}) = \int_{x_{min}}^{x} \begin{cases} \dfrac{\beta e^{-\beta(u-x_{min})}}{1-e^{-\beta(x_{max}-x_{min})}} \\ 0 \end{cases} du$$

$$F_X(x|x_{min},x_{max}) = \frac{\beta e^{\beta x_{min}}}{1-e^{-\beta(x_{max}-x_{min})}}\left(\frac{-1}{\beta}\right)\left(e^{-\beta u}\Big|_{x_{min}}^{x}\right)$$

$$F_X(x|x_{min},x_{max}) = \begin{cases} 0 & x < x_{min} \\ \dfrac{-e^{-\beta(x-x_{min})}+e^{-\beta(x_{min}-x_{min})}}{1-e^{-\beta(x_{max}-x_{min})}} & x_{min} \le x \le x_{max} \\ 1 & x > x_{max} \end{cases}$$

$$F_X(x|x_{min},x_{max}) = \begin{cases} 0 & x < x_{min} \\ \dfrac{1-e^{-\beta(x-x_{min})}}{1-e^{-\beta(x_{max}-x_{min})}} & x_{min} \le x \le x_{max} \\ 1 & x > x_{max} \end{cases}$$

or alternatively as

$$F_X(x|x_{min},x_{max}) = \frac{1-e^{-\beta(x-x_{min})}}{1-e^{-\beta(x_{max}-x_{min})}} \times \frac{e^{-\beta x_{min}}}{e^{-\beta x_{min}}}$$

$$F_X(x|x_{min},x_{max}) = \frac{e^{-\beta x_{min}}-e^{-\beta x}}{e^{-\beta x_{min}}-e^{-\beta x_{max}}}. \quad (3.7a)$$

■

Equations 3.9a and b.

Following Assumption 8, the extreme distribution for prehistoric data is defined.

$$F_X^{MAX}(x_0) = P(X \le x_0).$$

$$F_X^{MAX}(x_0) = P[X_1 \le x_0]P[X_2 \le x_0] \cdots [X_{n_P} \le x_0].$$

$$F_X^{MAX}(x_0) = F_{X_1}(x_0)F_{X_2}(x_0) \cdots F_{X_{n_P}}(x_0).$$

$$F_X^{MAX}(x_0) = [F_X(x_0)]^{n_P}. \tag{3.9a}$$

$$f_X^{MAX}(x_0) = \frac{d}{dx_0}F_X^{MAX}(x_0),$$

$$f_X^{MAX}(x_0) = n_P[F_X(x_0)]^{n_P-1}f_X(x_0). \tag{3.9b}$$

■

Equation 3.13.

Following Assumption 9,

$$P(X_j \le x_0|\Delta t) = P_{N_P}(n_P)[F_X(x_0)]^{n_P},$$

and the theorem of total probability (Cramer, 1961), the conditional CFD for extreme events x_0 in a specific time interval Δt are defined as

$$F_X^{MAX}(x_0) = P(N_P = 0)[F_X(x_0)]^0 + P(N_P = 1)[F_X(x_0)]^1 + \cdots,$$

$$F_X^{MAX}(x_0|\Delta t) = \sum_{n_P=0}^{\infty} P_{N_P}(n_P)[F_X(x_0)]^{n_P}. \tag{3.13}$$

■

Equation 3.14 a and b.

Following Eqs 3.8 and 3.13,

$$F_X^{MAX}(x_0|\Delta t_P) = \sum_{n_P=0}^{\infty} \left(\frac{e^{-\lambda_P \Delta t_P}(\lambda_P \Delta t_P)^{n_P}}{n_P!} \right) [F_X(x_0)]^{n_p}.$$

Let $\Lambda = \lambda_P \Delta t_P F_X(x_0)$ such that

$$F_X^{MAX}(x_0|\Delta t) = \sum_{n_P=0}^{\infty} \left(\frac{e^{-\lambda_P \Delta t_P} \Lambda^{n_P}}{n_P!} \right),$$

where

$$\sum_{n_P=0}^{\infty} \frac{\Lambda^{n_P}}{n_P!} = e^{\Lambda}.$$

Then

$$F_X^{MAX}(x_0|\Delta t_P) = e^{-\lambda_P \Delta t_P} e^{\Lambda},$$

$$F_X^{MAX}(x_0|\Delta t_P) = e^{-\lambda_P \Delta t_P} e^{\lambda_P \Delta t_P F_X(x_0)},$$

$$F_X^{MAX}(x_0|\Delta t_P) = e^{-\lambda_P \Delta t_P [1 - F_X(x_0)]}. \tag{3.14a}$$

$$f_X^{MAX}(x_0|\Delta t_P) = \frac{d}{dx_0} F_X^{MAX}(x_0|\Delta t_P)$$

$$= e^{-\lambda_P \Delta t_P (1 - F_X(x_0))} \frac{d}{dx_0} \left(-\lambda_P \Delta t_P (1 - F_X(x_0)) \right).$$

$$f_X^{MAX}(x_0|\Delta t) = \lambda_P \Delta t_P f_X(x_0) e^{-\lambda_P \Delta t_P (1 - F_X(x_0))}. \tag{3.14b}$$

■

Equation 3.24.

$$P_N\left(n|\bar{\lambda}, \Delta t, v_\lambda\right) = \int_0^{\infty} P_N(n|\lambda, \Delta t) f_\Lambda(\lambda) d\lambda$$

$$\int_0^{\infty} \frac{e^{-\lambda \Delta t} (\lambda \Delta t)^n}{n!} \left[\frac{\lambda^{q-1} p^q e^{-p\lambda}}{\Gamma(q)} \right] d\lambda$$

$$= \frac{1}{n! \, \Gamma(q)} \int_0^{\infty} e^{-\lambda \Delta t} (\lambda \Delta t)^n \left(\lambda^{q-1} p^q e^{-p\lambda} \right) d\lambda$$

$$= \frac{1}{n! \, \Gamma(q)} \int_0^{\infty} e^{-\lambda \Delta t} \lambda^n (\Delta t)^n \left(\lambda^{q-1} p^q e^{-p\lambda} \right) d\lambda$$

$$= \frac{(\Delta t)^n p^q}{n! \, \Gamma(q)} \int_0^{\infty} e^{-\lambda \Delta t} \lambda^{n+q-1} e^{-p\lambda} d\lambda.$$

Let

$$u = \lambda^{q+n}, \quad \frac{du}{d\lambda} = (q+n)\lambda^{q+n-1}, \quad \frac{du}{(q+n)\lambda^{q+n-1}} = d\lambda,$$

and set the lower boundary equal to $a(=0)$ and upper boundary equal to $b(=\infty)$. Then

$$P_N\left(n|\bar{\lambda}, \Delta t, v_\lambda\right) = \frac{(\Delta t)^n p^q}{n!\,\Gamma(q)} \int_{u(a)=0}^{u(b)=\infty} e^{-\lambda\Delta t}\lambda^{n+q-1}e^{-p\lambda}\frac{du}{(q+n)\lambda^{q+n-1}}$$

$$= \frac{(\Delta t)^n p^q}{n!\,\Gamma(q)(q+n)} \int_0^\infty e^{-\lambda(\Delta t+p)}du$$

$$= \frac{(\Delta t)^n p^q}{n!\,\Gamma(q)(q+n)} \int_0^\infty e^{-(\Delta t+p)u^{\frac{1}{q+n}}}du.$$

Let

$$v = (\Delta t+p)^{q+n}u, \qquad \frac{dv}{du} = (\Delta t+p)^{q+n}, \qquad \frac{dv}{(\Delta t+p)^{q+n}} = du.$$

Then

$$P_N\left(n|\bar{\lambda}, \Delta t, v_\lambda\right) = \frac{(\Delta t)^n p^q}{n!\,\Gamma(q)(q+n)} \int_{v(u(a))=0}^{v(u(b))=\infty} \frac{e^{-v^{\frac{1}{q+n}}}}{(\Delta t+p)^{q+n}}dv$$

$$= \frac{(\Delta t)^n p^q}{n!\,\Gamma(q)(q+n)(\Delta t+p)^{q+n}} \int_0^\infty e^{-v^{\frac{1}{q+n}}}dv$$

Let

$$y = v^{\frac{1}{q+n}}, \qquad \frac{dy}{dv} = \frac{v^{\frac{1}{q+n}-1}}{q+n}, \qquad \frac{(q+n)dy}{v^{\frac{1}{q+n}-1}} = dv,$$

$$P_N\left(n|\bar{\lambda}, \Delta t, v_\lambda\right) = \frac{(\Delta t)^n p^q}{n!\,\Gamma(q)(q+n)(\Delta t+p)^{q+n}} \int_{y\left(v(u(a))\right)=0}^{y\left(v(u(b))\right)=\infty} \frac{(q+n)e^{-y}dy}{v^{\frac{1}{q+n}-1}}$$

$$= \frac{(\Delta t)^n p^q}{n!\,\Gamma(q)(q+n)(\Delta t+p)^{q+n}} \int_0^\infty \frac{(q+n)e^{-y}dy}{(y)^{1-q-n}}$$

$$= \frac{(\Delta t)^n p^q (q+n)}{n!\,\Gamma(q)(q+n)(\Delta t+p)^{q+n}} \int_0^\infty \frac{e^{-y}dy}{y^{1-q-n}}$$

$$= \frac{(\Delta t)^n p^q (q+n)}{n!\,\Gamma(q)(q+n)(\Delta t+p)^{q+n}} \int_0^\infty \frac{e^{-y}dy}{y^{1-q-n}}$$

$$= \frac{(\Delta t)^n p^q (q+n)}{n!\,\Gamma(q)(q+n)(\Delta t+p)^{q+n}} \int_0^\infty y^{q+n-1}e^{-y}dy$$

$$= \frac{(\Delta t)^n p^q (q+n)}{n!\,\Gamma(q)(q+n)(\Delta t+p)^{q+n}} \int_0^\infty y^{q+n-1}e^{-y}dy$$

$$= \frac{(\Delta t)^n p^q (q+n)}{n!\,\Gamma(q)(q+n)(\Delta t+p)^{q+n}} \Gamma(n+q).$$

Plug in the solved integrals $\int_0^\infty \frac{e^{-v^{\frac{1}{q+n}}}}{(\Delta t+p)^{q+n}} dv$ to obtain the

The Poisson-gamma distribution with gamma hyper-parameters p_λ and q_λ, equal to

$$P_N\left(n|\bar{\lambda},\Delta t,v_\lambda\right) = \frac{\Gamma(n+q_\lambda)}{n!\,\Gamma(q_\lambda)}\left(\frac{p_\lambda}{\Delta t+p_\lambda}\right)^{q_\lambda}\left(\frac{\Delta t}{\Delta t+p_\lambda}\right)^n. \tag{3.24}$$

■

Equation 3.26.

Following from Eqs 3.24 and 3.25

$$P_N\left(n|\bar{\lambda},\Delta t,v_\lambda\right) \equiv P_N\left(n|\bar{\lambda},\Delta t,v_\lambda\right) = \frac{\Gamma(n+q_\lambda)}{n!\,\Gamma(q_\lambda)}\left(\frac{p_\lambda}{\Delta t+p_\lambda}\right)^{q_\lambda}\left(\frac{\Delta t}{\Delta t+p_\lambda}\right)^n$$

$$= \frac{\Gamma(n+q_\lambda)}{n!\,\Gamma(q_\lambda)}\left(\frac{\frac{q_\lambda}{\bar{\lambda}}}{\Delta t+\frac{q_\lambda}{\bar{\lambda}}}\right)^{q_\lambda}\left(\frac{\Delta t}{\Delta t+\frac{q_\lambda}{\bar{\lambda}}}\right)^n$$

$$= \frac{\Gamma(n+q_\lambda)}{n!\,\Gamma(q_\lambda)}\left(\frac{q_\lambda}{\bar{\lambda}}\times\frac{\bar{\lambda}}{\bar{\lambda}\Delta t+q_\lambda}\right)^{q_\lambda}\left(\frac{\Delta t\bar{\lambda}}{\Delta t\bar{\lambda}+q_\lambda}\right)^n$$

$$P_N\left(n|\bar{\lambda},\Delta t,v_\lambda\right)=\frac{\Gamma(n+q_\lambda)}{n!\,\Gamma(q_\lambda)}\left(\frac{q_\lambda}{\bar{\lambda}\Delta t+q_\lambda}\right)^{q_\lambda}\left(\frac{\Delta t\bar{\lambda}}{\Delta t\bar{\lambda}+q_\lambda}\right)^{n}. \tag{3.26}$$

■

Equation 3.27a, b, and c.

First, the mixture distribution for the shifted CDF is derived, as defined in Eq. 3.6a. It is normalised to provide the CDF of the shifted-truncated CDF is according to Eq. 3.7a. Using Eq. 3.6a, the gamma-exponential distribution equals

$$F_X\left(x|x_{min},\bar{\beta},v_\beta\right)=\int_0^\infty\left[1-e^{-\beta(x-x_{min})}\right]\left[\frac{\beta^{q_\beta-1}p^{q_\beta}e^{-p_\beta\beta}}{\Gamma(q_\beta)}\right]d\beta$$

$$=\frac{p_\beta^{q_\beta}}{\Gamma(q_\beta)}\int_0^\infty\left[1-e^{-\beta(x-x_{min})}\right]\beta^{q_\beta-1}e^{-p_\beta\beta}d\beta$$

$$=\frac{p_\beta^{q_\beta}}{\Gamma(q_\beta)}\int_0^\infty\left[\beta^{q_\beta-1}e^{-p_\beta\beta}-\beta^{q_\beta-1}e^{-p_\beta\beta}e^{-\beta(x-x_{min})}\right]d\beta.$$

Apply linearity of integrals

$$=\frac{p_\beta^{q_\beta}}{\Gamma(q_\beta)}\left[\int_0^\infty\beta^{q_\beta-1}e^{-p_\beta\beta}d\beta-\int_0^\infty\left[\beta^{q_\beta-1}e^{-p_\beta\beta}e^{-\beta(x-x_{min})}\right]d\beta\right].$$

Solve $A=\int_0^\infty\beta^{q_\beta-1}e^{-p_\beta\beta}d\beta$.

Let

$$u=\beta^{q_\beta},\qquad\frac{du}{d\beta}=q_\beta\beta^{q_\beta-1},\qquad\frac{du}{q_\beta\beta^{q_\beta-1}}=d\beta,$$

and set the lower boundary equal to $a(=0)$ and upper boundary equal to $b(=\infty)$. Then

$$A=\int_{u(a)=0}^{u(b)=\infty}\beta^{q_\beta-1}e^{-p_\beta\beta}\frac{du}{q_\beta\beta^{q_\beta-1}}$$

$$=\frac{1}{q_\beta}\int_0^\infty e^{-p_\beta u^{\frac{1}{q_\beta}}}du.$$

Let

$$v = p^{q_\beta}u, \qquad \frac{v}{p^{q_\beta}} = u, \qquad \frac{dv}{du} = p^{q_\beta}, \qquad \frac{dv}{p^{q_\beta}} = du,$$

then

$$A = \frac{1}{q_\beta} \int_{v(u(b))=0}^{v(u(b))=\infty} e^{-p_\beta\left(\frac{v}{p^{q_\beta}}\right)^{\frac{1}{q_\beta}}} \frac{dv}{p^{q_\beta}}$$

$$= \frac{1}{q_\beta p^{q_\beta}} \int_0^\infty e^{-v^{\frac{1}{q_\beta}}} dv.$$

Let

$$y = v^{\frac{1}{q_\beta}}, \qquad \frac{dy}{dv} = \frac{1}{q_\beta} v^{\frac{1}{q_\beta}-1}, \qquad \left(\frac{q_\beta}{v^{\frac{1}{q_\beta}-1}}\right) dy = dv,$$

then

$$A = \frac{1}{q_\beta p_\beta^{q_\beta}} \int_{y\left(v(u(a))\right)=0}^{y\left(v(u(b))\right)=\infty} e^{-y} \left(\frac{q_\beta}{v^{\frac{1}{q_\beta}} v^{-1}}\right) dy$$

$$= \frac{1}{q_\beta p_\beta^{q_\beta}} \int_0^\infty e^{-y} \left(\frac{q_\beta}{y y^{-q_\beta}}\right) dy$$

$$= \frac{q_\beta}{q_\beta p_\beta^{q_\beta}} \int_0^\infty y^{-(-q_\beta+1)} e^{-y} dy$$

$$= \frac{1}{p_\beta^{q_\beta}} \int_0^\infty y^{q_\beta-1} e^{-y} dy$$

$$\mathrm{A} = \frac{1}{p_\beta^{q_\beta}} \Gamma(q_\beta).$$

Solve: $B = \int_0^\infty \left[\beta^{q_\beta-1} e^{-p_\beta \beta} e^{-\beta(x-x_{min})}\right] d\beta$.

Let

$$u = \beta^{q_\beta}, \qquad \frac{du}{d\beta} = q_\beta \beta^{q_\beta-1}, \qquad \frac{du}{q_\beta \beta^{q_\beta-1}} = d\beta.$$

Substitute in u

$$B = \int_{u(a)=0}^{u(b)=\infty} \left[\beta^{q_\beta - 1} e^{-p_\beta \beta} e^{-\beta(x - x_{min})}\right] \frac{du}{q_\beta \beta^{q_\beta - 1}}$$

$$= \frac{1}{q_\beta} \int_0^\infty \left[e^{-p_\beta \beta} e^{-\beta(x - x_{min})}\right] du$$

$$= \frac{1}{q_\beta} \int_0^\infty \left[e^{-p_\beta u^{\frac{1}{q_\beta}}} e^{-u^{\frac{1}{q_\beta}}(x - x_{min})}\right] du$$

$$= \frac{1}{q_\beta} \int_0^\infty \left[e^{u^{\frac{1}{q_\beta}}(x_{min} - x - p_\beta)}\right] du.$$

Let

$$v = \left(x_{min} - x - p_\beta\right)^{q_\beta} u, \quad \frac{dv}{du} = \left(x_{min} - x - p_\beta\right)^{q_\beta}, \quad \frac{dv}{\left(x_{min} - x - p_\beta\right)^{q_\beta}} = du,$$

then

$$B = \frac{1}{q_\beta} \int_{v(u(a))=0}^{v(u(b))=\infty} \left[e^{\frac{v}{\left(x_{min} - x - p_\beta\right)^{q_\beta}}^{\frac{1}{q_\beta}} (x_{min} - x - p_\beta)}\right] \frac{dv}{\left(x_{min} - x - p_\beta\right)^{q_\beta}}$$

$$B = \frac{1}{q_\beta \left(x_{min} - x - p_\beta\right)^{q_\beta}} \int_0^\infty \left[e^{v^{\frac{1}{q_\beta}}}\right] dv.$$

Let

$$y = -v^{\frac{1}{q_\beta}}, \qquad \frac{dy}{dv} = -\frac{1}{q_\beta} v^{\frac{1}{q_\beta} - 1}, \qquad \frac{-q dy}{v^{\frac{1}{q_\beta}} v^{-1}} = dv,$$

then

$$B = \frac{1}{q_\beta \left(x_{min} - x - p_\beta\right)^{q_\beta}} \int_{y\left(v(u(a))\right)=0}^{y\left(v(u(b))\right)=\infty} e^{-y} \frac{-q_\beta dy}{v^{\frac{1}{q_\beta}} v^{-1}}$$

$$= \frac{q_\beta}{q_\beta \left(x_{min} - x - p_\beta\right)^{q_\beta}} \int_0^\infty e^{-y} \frac{dy}{(-(y)^{q_\beta})^{\frac{1}{q_\beta}} (-1)(y)^{-q_\beta}}$$

$$= \frac{1}{\left(x_{min} - x - p_\beta\right)^{q_\beta}} \int_0^\infty e^{-y} \frac{dy}{(y)^{1 - q_\beta}}$$

$$= \frac{1}{\left(x_{min} - x - p_\beta\right)^{q_\beta}} \int_0^\infty y^{q_\beta - 1} e^{-y} dy$$

$$= \frac{1}{\left(x_{min} - x - p_\beta\right)^{q_\beta}} \int_0^\infty y^{q_\beta - 1} e^{-y} dy$$

$$B = \frac{1}{\left(x_{min} - x - p_\beta\right)^{q_\beta}} \Gamma(q_\beta).$$

Combining equations A and B into

$$F_X(x|x_{min}, \bar{\beta}, v_\beta) = \frac{p_\beta^{q_\beta}}{\Gamma(q_\beta)} \left[\frac{1}{p_\beta^{q_\beta}} \Gamma(q_\beta) - \frac{1}{\left(x_{min} - x - p_\beta\right)^{q_\beta}} \Gamma(q_\beta) \right]$$

$$= \frac{p_\beta^{q_\beta} \Gamma(q_\beta)}{\Gamma(q_\beta)} \left[\frac{1}{p_\beta^{q_\beta}} - \frac{1}{\left(x_{min} - x - p_\beta\right)^{q_\beta}} \right]$$

$$= \left[\frac{p_\beta^{q_\beta}}{p_\beta^{q_\beta}} - \left(\frac{p_\beta}{x_{min} - x - p_\beta} \right)^{q_\beta} \right].$$

The shifted exponential-gamma distribution is, therefore, defined as

$$F_X(x|x_{min}, x_{max}, \bar{\beta}, v_\beta) = \left[1 - \left(\frac{p_\beta}{x_{min} - x - p_\beta} \right)^{q_\beta} \right]$$

$$= \left[1 - \left(\frac{\frac{q_\beta}{\bar{\beta}}}{x_{min} - x - \frac{q_\beta}{\bar{\beta}}} \right)^{q_\beta} \right]$$

$$= \left[1 - \left(\frac{\frac{q_\beta}{\bar{\beta}}}{\frac{\bar{\beta}(x_{min} - x) - q_\beta}{\bar{\beta}}} \right)^{q_\beta} \right]$$

$$= \left[1 - \left(\frac{q_\beta}{\bar{\beta}(x_{min} - x) - q_\beta} \right)^{q_\beta} \right].$$

The function is normalised with C_β, such that

$$C_\beta [F_X(x_{max}|\beta) - F_X(m_{min}|\beta)] = 1$$

$$C_\beta \left[\left[1 - \left(\frac{q_\beta}{\bar{\beta}(x_{max} - x_{min}) + q_\beta} \right)^{q_\beta} \right] - \left[1 - \left(\frac{q_\beta}{\bar{\beta}(x_{min} - x_{min}) + q_\beta} \right)^{q_\beta} \right] \right] = 1$$

$$C_\beta\left[\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_{max}-x_{min})+q_\beta}\right)^{q_\beta}\right]-\left[1-\left(\frac{q_\beta}{q_\beta}\right)^{q_\beta}\right]\right]=1$$

$$C_\beta=\frac{1}{1-\left(\frac{q_\beta}{\bar{\beta}(x_{max}-x_{min})+q_\beta}\right)^{q_\beta}}$$

$$C_\beta=\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_{max}-x_{min})+q_\beta}\right)^{q_\beta}\right]^{-1}. \qquad (3.27c)$$

Therefore, the exponential-gamma distribution for the shifted-truncated frequency–event-size function equals

$$F_X(x|x_{min},x_{max},\bar{\beta},v_\beta)=C_\beta\left[1-\left(\frac{q_\beta}{\bar{\beta}(x-x_{min})+q_\beta}\right)^{q_\beta}\right]\quad x_{min}\le x\le x_{max}. \qquad (3.27a)$$

Equation 3.27a can be defined also in terms of the p parameter as

$$F_X(x|x_{min},x_{max},\bar{\beta},v_\beta)=\left[1-\left(\frac{p_\beta}{x_{max}-x_{min}+p_\beta}\right)^{q_\beta}\right]^{-1}\left[1-\left(\frac{p_\beta}{x-x_{min}+p_\beta}\right)^{q_\beta}\right],$$

as seen in Eqs 10 and 11 in Kijko *et al*. (2016).

The PDF of the shifted-truncated exponential-gamma distribution equals

$$f_X(x|x_{min},x_{max},\bar{\beta},v_\beta)=\frac{dF_X(x|x_{min},x_{max},\bar{\beta},v_\beta)}{dx}$$

$$=\frac{d}{dx}\left\{\frac{\left[1-\left(\frac{q_\beta}{\bar{\beta}(x-x_{min})+q_\beta}\right)^{q_\beta}\right]}{\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_{max}-x_{min})+q_\beta}\right)^{q_\beta}\right]}\right\}$$

$$=\frac{\left[\frac{d}{dx}\left[-\left(\frac{q_\beta}{\bar{\beta}(x-x_{min})+q_\beta}\right)^{q_\beta}\right]\right]}{\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_{max}-x_{min})+q_\beta}\right)^{q_\beta}\right]}$$

$$=\frac{-q_\beta\left[\frac{d}{dx}\left[\left(\frac{1}{\bar{\beta}(x-x_{min})+q_\beta}\right)\right]q_\beta\left(\frac{q_\beta}{\bar{\beta}(x-x_{min})+q_\beta}\right)^{q_\beta-1}\right]}{\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_{max}-x_{min})+q_\beta}\right)^{q_\beta}\right]}$$

$$= \frac{-q_\beta^2 \left(\frac{q}{\bar{\beta}(x-x_{min})+q_\beta}\right)^{q_\beta-1} \left(\frac{\frac{d}{dx}(\bar{\beta}(x-x_{min})+q_\beta)}{(\bar{\beta}(x-x_{min})+q_\beta)^2}\right)}{\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_{max}-x_{min})+q_\beta}\right)^{q_\beta}\right]}$$

$$= \frac{q_\beta \bar{\beta} \left(\frac{q}{\bar{\beta}(x-x_{min})+q_\beta}\right)^{q_\beta}}{\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_{max}-x_{min})+q_\beta}\right)^{q_\beta}\right](\bar{\beta}(x-x_{min})+q_\beta)},$$

that simplifies to

$$f_X(x|x_{min}, x_{max}, \bar{\beta}, v_\beta) = C_\beta \bar{\beta}\left[1+\frac{\bar{\beta}(x-x_{min})}{q_\beta}\right]^{-(q+1)} \quad x_{min} \le x \le x_{max}. \qquad (3.27b)$$

■

Equation 3.34.

Derivation provided by S. Verryn (2011)

$$F_X(x|x_{min}, x_{max}, \bar{\beta}, v_\beta) = \int_{x_{min}}^{x_{max}} f_Y(\zeta)\left[\int_{-\infty}^{x-\zeta} \frac{1}{\sigma_x\sqrt{2\pi}}\exp(-\frac{\epsilon^2}{2\sigma_x^2})d\varepsilon\right]d\zeta,$$

can be reduced to $G_X(x) = \int_{x_{min}}^{x_{max}} f_X(\zeta)\Phi\left(\frac{x-\zeta}{\sigma_x}\right)d\zeta$

Assume that

$$\Phi\left(\frac{x-\zeta}{\sigma_x}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-\zeta}{\sigma_x}} \exp(-\frac{u^2}{2})\,du,$$

where $\alpha = \frac{x-\zeta}{\sigma_x}$, such that

$$\Phi(\alpha) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\alpha} \exp(-\frac{u^2}{2})du$$

$$= \frac{1}{2}\left\{1+erf\left(\frac{\alpha}{\sqrt{2}}\right)\right\}$$

$$= \frac{1}{2}\left\{1+\left(\frac{2}{\pi}\right)^{1/2}\left[\alpha-\frac{\alpha^3}{2\cdot 1!\,3}+\frac{\alpha^5}{2^2\cdot 2!\,5}-\frac{\alpha^7}{2^3\cdot 3!\,7}+\cdots\right]\right\}.$$

Let $\Delta = \alpha - \frac{\alpha^3}{2\cdot 1!3} + \frac{\alpha^5}{2^2\cdot 2!5} - \frac{\alpha^7}{2^3\cdot 3!7} + \cdots$.

Then $F_X\left(x|x_{min}, x_{max}, \bar{\beta}, v_\beta\right)$ can be re-written as

$$F_X\left(x|x_{min}, x_{max}, \bar{\beta}, p_\beta, q_\beta\right) = -\frac{1}{\sigma_x}\int_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}} f(x-\alpha\sigma_x)\Phi(\alpha)d\alpha$$

$$= -\frac{1}{\sigma_x}(C_\beta \bar{\beta} q_\beta^{q_\beta+1})\int_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}} (q_\beta + \bar{\beta}(x-\alpha\sigma_x - x_{min}))^{-q_\beta-1}\Phi(\alpha)d\alpha$$

$$= -\frac{1}{\sigma_x}\left(C_\beta \bar{\beta} q_\beta^{q_\beta+1}\right)\int_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}} \left[(q_\beta + \bar{\beta}(x-\alpha\sigma_x - x_{min}))^{-q_\beta-1}\right]\left[\frac{1}{2}\left\{1+\left(\frac{2}{\pi}\right)^{1/2}[\Delta]\right\}\right]d\alpha$$

$$= -\frac{1}{2\sigma_x}\left(C_\beta \bar{\beta} q_\beta^{q_\beta+1}\right)\left\{\int_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}} \left[(q_\beta + \bar{\beta}(x-\alpha\sigma_x - x_{min}))^{-q_\beta-1}\right]d\alpha\right.$$

$$\left. + \left(\frac{2}{\pi}\right)^{1/2}\int_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}} \left[(q_\beta + \bar{\beta}(x-\alpha\sigma_x - x_{min}))^{-q_\beta-1}\right]\Delta\, d\alpha\right\},$$

$$= -\frac{1}{2\sigma_x}\left(C_\beta \bar{\beta} q_\beta^{q_\beta+1}\right)\left\{\int_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}} (r_1 + r_2\alpha)^{-q_\beta-1}d\alpha + \left(\frac{2}{\pi}\right)^{1/2}\int_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}} \left[(r_1 + r_2\alpha)^{-q_\beta-1}\right]\Delta\, d\alpha\right\},$$

where $r_1 = q_\beta + \bar{\beta}(x - x_{\min})$, $r_2 = -\bar{\beta}\sigma_x$ and $C_\beta = \left(1 - \left(\frac{q_\beta}{q_\beta + \bar{\beta}(x_{max}-x_{min}}\right)^{q_\beta}\right)^{-1}$.

Now $\int \frac{\alpha^w}{(r_1+r_2\alpha)^u}d\alpha = \frac{1}{r_2^{w+1}}\sum_{j=0}^{w}\frac{w!(-r_1)^j(r_1+r_2\alpha)^{w-u-j+1}}{(w-j)!j!(w-u-j+1)}$,

therefore,

$$\int_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{\max}}{\sigma_x}} (r_1 + r_2\alpha)^{-q_\beta-1}\left[\alpha - \frac{\alpha^3}{2\cdot 1!\,3} + \frac{\alpha^5}{2^2\cdot 2!\,5} - \frac{\alpha^7}{2^3\cdot 3!\,7} + \cdots\right]d\alpha$$

and

$$\int_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}} (r_1 + r_2\alpha)^{-q_\beta-1}d\alpha = -\frac{1}{r_2 q_\beta}(r_1 + r_2\alpha)^{-q_\beta}\Big|_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}}$$

yields

$$= \sum_{h=0}^{\infty} \frac{(-1)^h}{2^h \cdot h!\,(2hh!)} \frac{1}{{r_2}^{2h+2}} \sum_{j=0}^{2h+1} \frac{(2h+1)!\,(-r_1)^j (r_1 + r_2\alpha)^{2h+1-q_\beta-j}}{(2h+1-j)!\,j!\,(2h+1-q_\beta-j)} \Bigg|_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}}.$$

The final result is

$$F_X\left(x|x_{min}, x_{max}, \bar{\beta}, v_\beta\right) = \frac{C_\beta \bar{\beta} q_\beta^{q+1}}{2\sigma_x}\{A + B\} \tag{3.34}$$

$$A = \frac{(r_1 + r_2\alpha)^{-q_\beta}}{r_2 q_\beta} \Bigg|_{\frac{x-x_{\min}}{\sigma_x}}^{\frac{x-x_{\max}}{\sigma_x}}$$

$$B = \left(\frac{2}{\pi}\right)^{1/2} \sum_{h=0}^{\infty} \frac{(-1)^h}{2^h h!\,(2h+1)} \frac{1}{b^{2h+2}} \sum_{j=0}^{2h+1} \frac{(2h+1)!\,(-r_1)^j (r_1 + r_2\alpha)^{2h+1-q_\beta-j}}{(2h+1-j)!\,j!\,(2h+1-q_\beta-j)} \Bigg|_{\frac{x-x_{min}}{\sigma_x}}^{\frac{x-x_{max}}{\sigma_x}}$$

in which $C_\beta = \left[1 - \left(\frac{q_\beta}{q_\beta + \bar{\beta}(x_{max} - x_{min})}\right)^{q_\beta}\right]^{-1}$ (Eq. 3.27c), $r_1 = q_\beta + \bar{\beta}(x - x_{min})$, $r_2 = \bar{\beta}\sigma_x$, $\alpha = q_\beta + \bar{\beta}(x - x_{min})$ and $b = -\bar{\beta}\sigma_X$.

■

Equation 3.39a.

Following Eq. 3.26

$$P_N\left(n|\bar{\lambda}, \Delta t, p_\lambda, q_\lambda\right) = \frac{\Gamma(n + q_\lambda)}{n!\,\Gamma(q_\lambda)} \left(\frac{q_\lambda}{\bar{\lambda}\Delta t + \mathrm{q_\lambda}}\right)^{q_\lambda} \left(\frac{\Delta t \bar{\lambda}}{\Delta t \bar{\lambda} + q_\lambda}\right)^{n},$$

the likelihood function for λ_i is defined as

$$L_{\lambda i}\left(\bar{\lambda}\middle|\boldsymbol{\mathcal{I}}_{\lambda i}\right) = \frac{\Gamma(n_i + q_\lambda)}{n_i!\,\Gamma(q_\lambda)} \left(\frac{q_\lambda}{\bar{\lambda}_{\mathrm{i}}\Delta t_i + \mathrm{q_\lambda}}\right)^{q_\lambda} \left(\frac{\bar{\lambda}_{\mathrm{i}}\Delta t_i}{\bar{\lambda}_{\mathrm{i}}\Delta t_i + q_\lambda}\right)^{n_i},$$

for $\boldsymbol{\mathcal{I}}_{\lambda i} = (n_i, \Delta t_i, v_\lambda)$. Since different levels of completeness are defined for each of the sub-datasets ($i = 1, \ldots s$), the likelihood function $L_{\lambda i}$ must be normalised to a single λ. This is done by using

Identity 1, $\bar{\lambda}^{(i)} = \bar{\lambda}(x_{min})[1 - F_X(x)]$, where x_{min} is the smallest, known level of completeness. The likelihood function for $L_{\lambda i}$ for $i = 1, \dots, s$ is, therefore, defined as

$$L_{\lambda i}(\bar{\lambda}|\boldsymbol{\mathcal{I}}_{\lambda i}) = \bar{\lambda}(x_{min})[1 - F_X(x)]\frac{\Gamma(n_i + q_\lambda)}{n_i!\,\Gamma(q_\lambda)}\left(\frac{q_\lambda}{\bar{\lambda}_{\mathrm{i}}\Delta t_i + \mathrm{q}_\lambda}\right)^{q_\lambda}\left(\frac{\bar{\lambda}_{\mathrm{i}}\Delta t_i}{\bar{\lambda}_{\mathrm{i}}\Delta t_i + q_\lambda}\right)^{n_i},$$

or alternatively as

$$L_{\lambda i}(\bar{\lambda}|\boldsymbol{\mathcal{I}}_{\lambda i}) = const_i\left(\frac{1}{\bar{\lambda}_{\mathrm{i}}\Delta t_i + \mathrm{q}_\lambda}\right)^{q_\lambda}\left(\frac{\bar{\lambda}_{\mathrm{i}}\Delta t_i}{\bar{\lambda}_{\mathrm{i}}\Delta t_i + q_\lambda}\right)^{n_i},$$

with

$$const_i = \bar{\lambda}(x_{min})[1 - F_X(x)]\frac{q_\lambda^{q_\lambda}\Gamma(n_i + q_\lambda)}{n_i!\,\Gamma(q_\lambda)}.$$

Assuming that the shifted-truncated exponential-gamma distribution defined in Eq. 3.27a is applied to describe the event size distribution, $L_{\lambda i}(\bar{\lambda}|\boldsymbol{\mathcal{I}}_{\lambda i})$ equals

$$L_{\lambda i}(\bar{\lambda}|\boldsymbol{\mathcal{I}}_{\lambda i}) = const_i\left(\frac{1}{\bar{\lambda}_{\mathrm{i}}\Delta t_i + \mathrm{q}_\lambda}\right)^{q_\lambda}\left(\frac{\bar{\lambda}_{\mathrm{i}}\Delta t_i}{\bar{\lambda}_{\mathrm{i}}\Delta t_i + q_\lambda}\right)^{n_i}$$

with (3.39a)

$$const_i = \bar{\lambda}(x_{min})\left[1 - F_X\left(x|x_{min}^{(i)}, x_{max}, \bar{\beta}, v_\beta\right)\right]\frac{q_\lambda^{q_\lambda}\Gamma(n_i + q_\lambda)}{n_i!\,\Gamma(q_\lambda)}.$$

■

Equation 3.39b.

Following the PDF of Eq. 3.27b

$$f_X(x|x_{min}, x_{max}, \bar{\beta}, v_\beta) = C_\beta\bar{\beta}\left[1 + \frac{\bar{\beta}(x - x_{min})}{q_\beta}\right]^{-(q_\beta+1)} \quad x_{min} \le x \le x_{max},$$

with $C_\beta = \left[1 - \left(q_\beta/\left(\bar{\beta}(x_{max} - x_{min}) + q_\beta\right)\right)^{q_\beta}\right]^{-1}$, the likelihood function for the shifted-truncated gamma-exponential event size distribution equals

$$L_{\beta i}\left(\bar{\beta}\middle|\boldsymbol{\mathcal{I}}_{\beta i}\right)=\prod_{j=1}^{n_i} C_\beta \bar{\beta}\left[1+\frac{\bar{\beta}\left(x_{ij}-x_{min}^{(i)}\right)}{q_\beta}\right]^{-(q_\beta+1)},$$

$$L_{\beta i}\left(\bar{\beta}\middle|\boldsymbol{\mathcal{I}}_{\beta i}\right)=\left[C_\beta \bar{\beta}\right]^{n_i}\prod_{j=1}^{n_i}\left[1+\frac{\bar{\beta}\left(x_{ij}-x_{min}^{(i)}\right)}{q_\beta}\right]^{-(q_\beta+1)}, \tag{3.39b}$$

for $\boldsymbol{\mathcal{I}}_{\beta i}=(\boldsymbol{x}_i, x_{min}^{(i)}, x_{max}, v_\beta,)$ and $v_\beta=\sigma_\beta/\mu_\beta$.

■