TNO Energy Transition
Princetonlaan 6
3584 CB Utrecht
P.O. Box 80015
3508 TA Utrecht
The Netherlands

www.tno.nl

T +31 88 866 42 56
F +31 88 866 44 75

TNO report

TNO2020 R11052 | 2

Probabilistic Seismic Hazard and Risk Analysis in the TNO Model Chain Groningen

Date 17 december 2020

Author(s) _____

Copy no 1
Number of pages 72 (incl. appendices)
Project name Model Chain Groningen
Project number 060.43342/01

All rights reserved.
No part of this publication may be reproduced and/or published by print, photoprint, microfilm or any other means without the previous written consent of TNO.

In case this report was drafted on instructions, the rights and obligations of contracting parties are subject to either the General Terms and Conditions for commissions to TNO, or the relevant agreement concluded between the contracting parties. Submitting the report for inspection to parties who have a direct interest is permitted.

© 2020 TNO

Management summary

Production from the Groningen gas field induces earthquakes and ground motion at the earth's surface. The TNO Model Chain is a Probabilistic Seismic Hazard and Risk Analysis (PSHRA) tool, specifically developed for the Groningen area to predict personal risk from future induced earthquakes. The tool is based on the NAM Hazard and Risk Assessment (HRA), but implemented independently in the public domain using a different numerical methodology. Barring acceptable numerical differences, the tool is able to reproduce the NAM HRA results exactly.

This report summarises the physical-statistical theory, numerical methods, and computational implementation of the Seismic Source Model (to forecast induced earthquakes), the Ground Motion Model (to translate the earthquakes at reservoir depth to ground motions at the surface), and the Damage Model (to translate ground motions to building damage/collapse and the risk to people inside those buildings).

The TNO Model Chain is designed to be modular, such that the chain elements, i.e., the abovementioned model components, can conveniently be updated and be replaced by state-of-art model alternatives. Based on the outcomes of these models, control measures such as building strengthening, can be designed to ensure the public safety in the region.

Contents

Management summary **2**

Preface **4**

1 Introduction **5**

2 Probabilistic Seismic Hazard Analysis **12**
2.1 A probabilistic model for seismic hazard and risk analysis 12
2.2 Elements of TNO Model Chain Groningen 17

3 Implementation of the TNO Model Chain **26**
3.1 Seismological Source Model (V5) 26
3.2 Hazard Model (V5) 41
3.3 Risk Model (V5) 47

4 References **59**

5 Signature **60**

Appendix A: Numerical methods **61**
A.1 Numerical integration: Monte Carlo vs quadrature 61
A.2 Discretization options 65

Appendix B: Follow-up actions external review **69**

Preface

Commissioned by the Ministry of Economic Affairs and Climate (EZK), TNO has developed a Model Chain that calculates the hazard and risk due to induced seismicity in the Groningen gas field in the public domain. The TNO Model Chain is largely based on the models that underlie NAM's hazard and risk assessment (HRA). The present report is part of a series of three TNO reports. The other two reports describe respectively: (1) the infrastructure of the IT platform for the TNO calculations; and (2) a comparison of the NAM and TNO implementation of the numerical methods and calculated risks.

The scope of the present report is to provide insight into the principles and methods behind the development of the public Model Chain. This report consists of three Chapters and an Appendix:

- Chapter 1 is an accessible, introductory description of the functioning of and the relationship between the different components of the Model Chain.
- Chapter 2 is a summary of the theoretical physical-statistical background of the model components.
- Chapter 3 describes the practical implementation of the numerical methods applicable to the different components of the Model Chain in detail.
- The Appendix describes some assumptions and choices behind the numerical methods used.

As a result of the aforementioned scope of this report, Chapter 1 has been written for a wide audience and Chapters 2, 3 and the Appendix contain the high-level details for experts.

After publication of version 1 of this report, dated June 30 2020, a code review has been performed by the external company Tessella[1], as initiated by the State Supervision of the Mines. The current report version 2 is an update of the June 30 version. Updates concern textual corrections of equations in Box 2 and Box 8, and the first equation on page 35, which already were correctly implemented in the Model Chain code and are typo's in the previous report. Appendix B has been added with a list of follow-up points from the external reviewer.

[1] Tessella – PSHRA Software review – Software assessment report. Reference: NPD/10826/CL/OP, Issue V1.R1.M0, September 29 2020.

1 Introduction

Gas production from the Groningen gas field leads to induced seismicity. To ensure public safety in the region, ground motions and building damage related to future induced earthquakes need to be modelled. Based on the outcomes of these models, control measures, such as production strategies and building strengthening, can be designed.

This report summarises the computational models that form the basis of the current version of the TNO Model Chain. The Model Chain has been developed by TNO in the period 2017-2019 to be able to perform state-of-the-art Probabilistic Seismic Hazard and Risk Analyses (PSHRA) for induced seismicity in the Groningen gas field in the public domain. The Model Chain is designed to be modular, such that the chain elements, i.e., the component models, can conveniently be replaced by alternatives. Alternative models are being developed as scientific knowledge on induced seismicity and associated hazard and risk continues to evolve. To ensure traceability and reproducibility of these models as part of Quality Assurance, an IT Platform is designed that serves as a computation infrastructure to perform the TNO PSHRA-calculations. The design of the IT Platform is described in a separate report (TNO, 2020).

The version of the TNO Model Chain described in this report is to a large extent based on the NAM Hazard and Risk Assessment (HRA) and its model components, but implemented independently, using a different numerical methodology. Being able to reproduce the NAM HRA results exactly (within some specified numerical tolerance) has been an important design criterion (TNO, 2019). In the following, we describe the model components as developed and proposed by NAM in the way they are implemented by TNO. It is important to note that in this report we do not discuss the quality and/or appropriateness of the NAM model choices, and by our description and implementation we do not (necessarily) endorse them.

The TNO Model Chain consists of a series of physical-statistical models that forecast the seismic hazard and risk above the Groningen gas field for a given production scenario. A hazard is defined as a cause of potential harm or damage, while risk is the probability of occurrence of harm or damage due to that hazard. In the TNO Model Chain the seismic hazard is posed by the ground motions caused by the induced seismicity due to gas depletion. The seismic risk is the probability of a damage or fatality as a consequence of the ground motion hazard.

The TNO Model Chain is organised hierarchically. At the coarsest level the chain is subdivided into three main components (Figure 1), that each comprise several sub-models:

1. The Seismic Source Model (SSM), forecasts the spatial and temporal distribution of induced earthquakes, as well as their magnitudes, conditional on a production scenario.
2. The Ground Motion Model (GMM), relates the earthquakes at depth to ground motions at the surface.
3. The Damage Model (DM), translates ground motions to building damage/collapse and the risk to people inside those buildings.

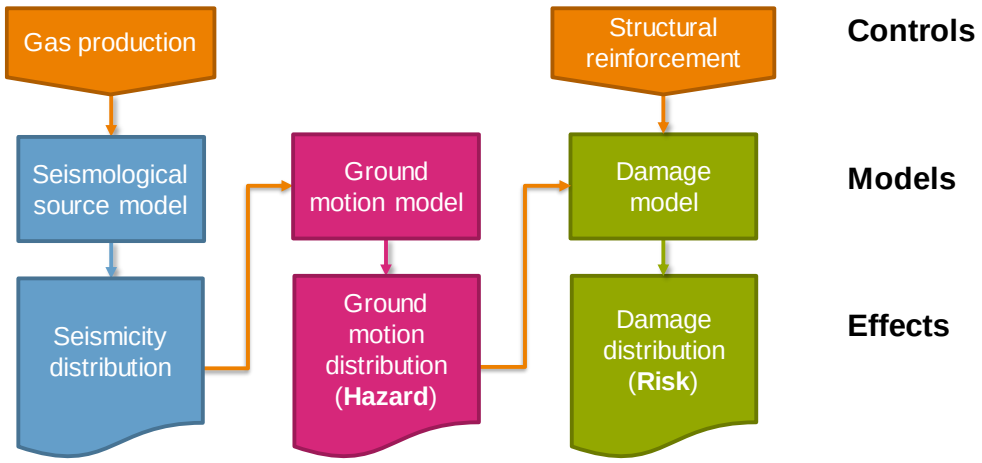


Figure 1 Schematic overview of the TNO Model Chain Groningen.

The model components described in this report all have the “V5” version designation. This combination of models served as the reference in the comparison study between the NAM and TNO implementations (TNO, 2019). A brief summary of the models is provided below. A more rigorous description of the theory behind the probabilistic model chain is provided in Chapter 2. The details of the implementation are provided in Chapter 3. Some additional notes on numerical implementation are provided in Appendix A.

Seismic Source Model (SSM)
The Seismic Source Model (SSM) is based on the work by Stephen Bourne, Steve Oates and co-workers. The model has been developed in a number of stages as reported in several peer reviewed journal papers (Bourne et al., 2014; 2015; 2018; Bourne & Oates, 2017; 2018). None of the paper describes the exact model version “V5”, however. The best reference for implementation has been the technical report (Bourne et al., 2019), supplemented with personal communication.
Induced earthquakes as a result of gas depletion in the Groningen gas field are assumed to be caused by differential compaction along existing faults. To model induced earthquakes we therefore need to know how the gas depletion translates to reservoir compaction (vertical strain) and the locations and properties of existing faults in the subsurface. In addition, to forecast induced earthquakes we also need historic induced earthquakes and corresponding past gas production pressure changes to train the model. Note that the model calibration to forecast induced earthquakes is not addressed in the comparison between the NAM HRA and the TNO Model Chain (TNO, 2019), but is nevertheless part of the TNO Model Chain.

Inputs to the SSM include static data such as reservoir geometry, compressibility and fault data, as well as the dynamic data on pore pressure changes as a result of a gas production scenario (past or future), which are all provided to TNO by NAM. The catalogue of observed induced earthquakes originates from the seismological service of the KNMI.

The compaction model computes vertical strains from the pore pressure changes of a given gas production scenario. These vertical strains in combination with the fault properties and fault locations are translated to a spatial and temporal distribution of Coulomb stress changes. The distribution of Coulomb stress changes is converted to a seismicity distribution in time and space.

The Seismic Source Model includes an Epidemic-Type Aftershock Sequence model (ETAS) to compute an aftershock seismicity distribution dependent on the main shock distribution. The aftershock distribution is added to the main distribution to form a total seismicity distribution. A b-value model is then used to define the magnitude distribution of these seismic events. The magnitude distribution is bounded by a maximum possible magnitude (M_{max}). The result is a distribution of total (main + aftershocks) seismicity in time, space and magnitude.

The Seismic Source Model contains certain parameters, which are determined by training of the model using past gas production years with all possible combinations of model parameters and comparing the outcome of expected seismicity with the monitored seismicity in the past. Model parameter combinations that can retroactively predict the past seismicity well, are assumed to be more likely. The trained model weighs these more likely combinations of parameters more heavily when forecasting seismicity for a future gas production scenario. This results in a probability of seismicity occurring in space-time-magnitude.

Finally, a rupture model translates the distribution of hypocenter locations into a rupture plane distribution with associated magnitudes (Bourne & Oates, 2018). This rupture model is included to reflect that earthquakes (i.e. the sources of seismic waves) do not occur on an infinitesimally small point, but rather on a rupture plane of finite size. The rupture model describes a probabilistic spatial extent for a rupture plane, given a hypocenter location. The length of the rupture depends on the magnitude of the earthquake, while the orientation is based on an average fault strike representative in the Groningen subsurface and the associated variability. The final output of the Seismic Source Model is a statistical distribution of seismicity of a certain magnitude, at a certain distance and within a certain year, for every point at the surface.

Ground Motion Model (GMM)

The Ground Motion Model is based on the collective work of Bommer et al. (2015-2018) and describes how an earthquake at a certain (rupture) distance and of a certain magnitude contributes to a statistical distribution of ground motions. As ground motions are complex, we don't simulate the complete expected ground motion for every earthquake, but only the attributes of the motion that are likely to affect infrastructure and buildings. These attributes are horizontal spectral accelerations (SA) at 23 different periods, peak ground velocity (PGV) and the durations of these movements. The spectral accelerations are simulated at multiple periods, because different types of buildings have different natural vibration periods. To simulate ground motions at the surface, we first model them at a hypothetical surface at the base of the North Sea Group (NS_B), located at a depth of around 800 m. This surface is the top of the sequence of hard rocks in the subsurface. The shallower formations are soil layers that amplify (or attenuate) the ground motions of the solid rock below (Figure 2). The amount of amplification strongly depends on

the type of soil the wave propagates through. To account for spatial variation in soils, and therefore the spatial variation in amplifications factors, a site-response zonation model (Figure 3) is used. The outlines of these site response regions are pre-defined (Bommer et al., 2018) and are used in the TNO Model Chain to determine which point at the surface belongs to which region. For each region, the site response translates the motions at NS_B level to ground motions at the surface, resulting in different ground motion distributions at equal distances from the hypocentre in different regions.

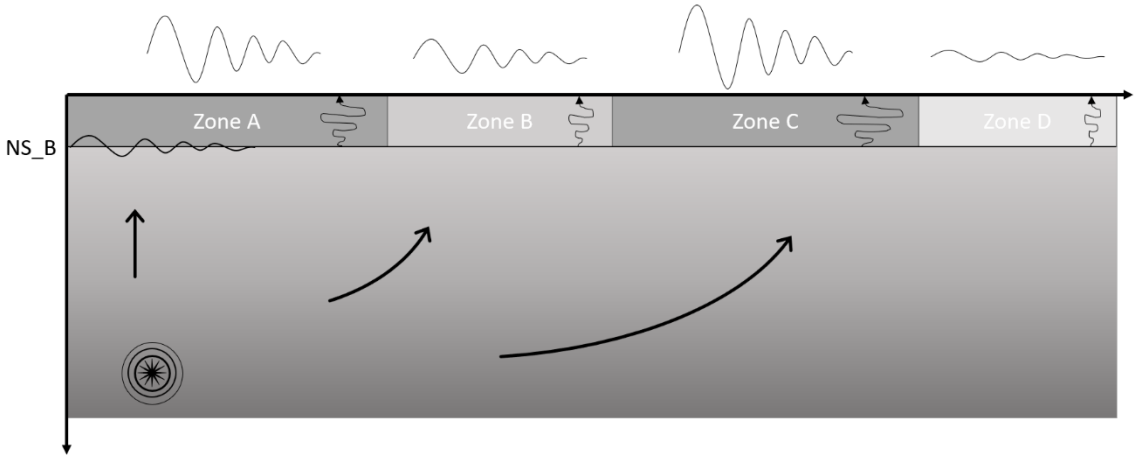


Figure 2 Schematic overview of the ground motion model. Showing the translating from a hypothetical earthquake to ground motion at the base North Sea Group (NS_B) and the translation to the surface ground motion through amplification for different types of soil. Zones A, B, C, and D are hypothetical zones to visualize the amplification for different types of soil.

Model parameters used to model the ground motions are calibrated by ground motion measurements and provided by NAM (Bommer et al., 2018). The calibration of these model parameters is not part of the GMM. The calibrated model parameters are used as input for the GMM.
After combination with the Seismic Source Model, the output of the GMM are annual probability of exceedances of spectral accelerations per grid point at the surface, for all 23 spectral periods. These can be visualised in hazard curves (Figure 4) per grid point or as hazard maps for a given spectral period and annual frequency (return period). For a hazard map, the spectral acceleration is sampled from the hazard curve at a certain return period for all grid points. The default return periods in the GMM in the TNO Model Chain are 475 and 2475 year, but can be specified by the user. A return period of 475 year corresponds to an annual probability of exceedance of 1/475 = 0.002 (a 10% probability of exceedance in 50 years). The return period is the average time in between exceedances of a certain spectral acceleration of a certain frequency.

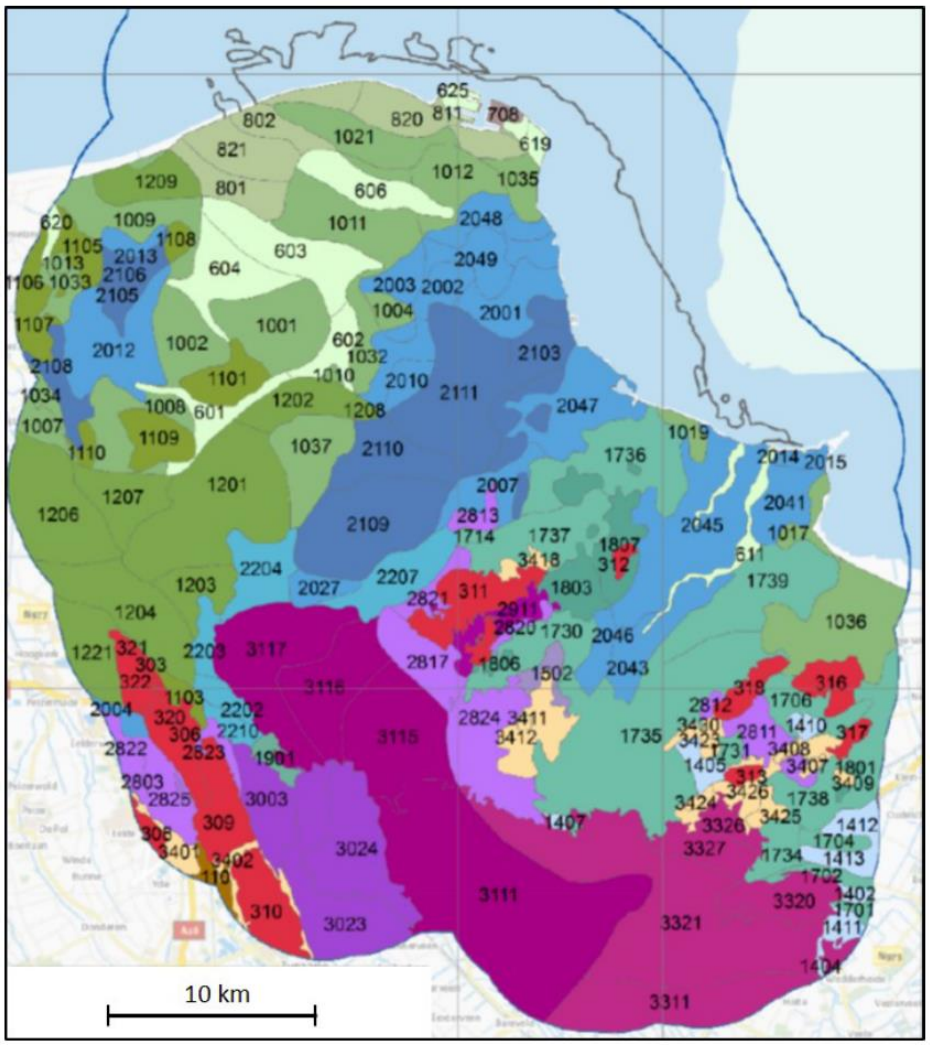


Figure 3 Site-response regions based on shallow (above base North Sea Group), sharp geological boundaries (from Bourne et al., 2018). Different colours indicate different geological profiles and the numbers are the zone ID numbers.

The Damage Model (DM)

The Damage Model consists of two components (Crowley et al., 2017; Crowley & Pinho, 2017):

- Fragility model
- Consequence model

The fragility model describes the behaviour of 54 building types (typologies) when subjected to a certain ground motion. These typologies are the result of a categorisation of all the buildings in the Groningen area, based on structural attributes that are likely to have a big impact on the response of the building to a certain ground motion, e.g. the number of storeys. The model works with three damage states and three collapse states and calculates the probability of exceedance of every damage/collapse state per typology for a given ground motion.

The consequence model describes the probability of dying (fatality) as the result of structural collapse of a building. The output of the consequence model is local personal risk (LPR) per typology. This is the risk of a single hypothetical person dying, who is assumed to be permanently present within/around the building. The person is assumed to be 99% of the time inside the building and 1% of the time outside (within 5 m of the building). The local personal risk is computed based on the probability of exceedance of the collapse states. The damage states do not contribute to the local personal risk, as only the collapse of a building is assumed to cause the death of a person.

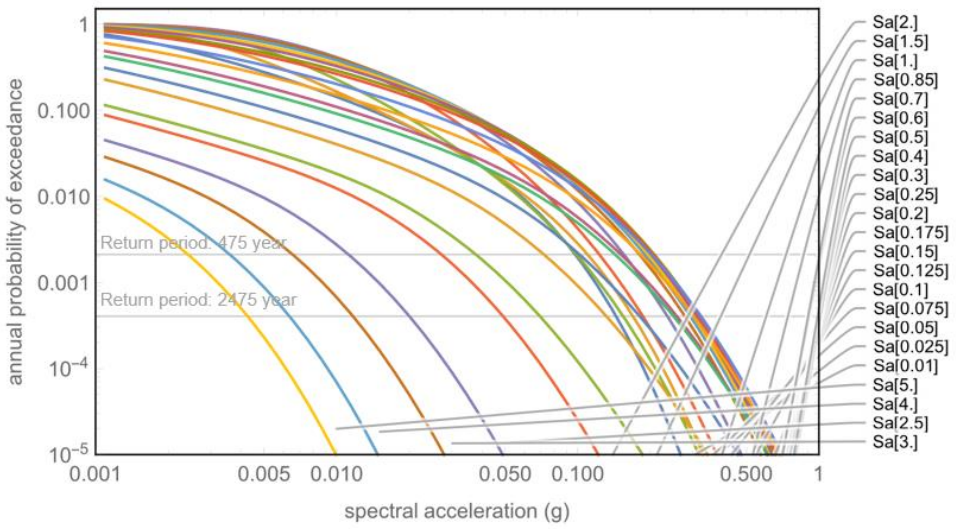


Figure 4 Example of hazard curves of all 23 frequencies (shown as period (s), ranging from 0.01 to 5 s, as listed on the right side of the graph) for one grid point. The two horizontal grey lines indicate the 475 and 2475 year return periods. The intersection between the return period line and the hazard curve is the spectral acceleration for this grid point to be visualized on a hazard map.

After combining the probability of exceedance of the collapse states with an exposure database, which contains information of which building type is located at which coordinates in the Groningen area, the LPR is used to compute the number of buildings that exceed the Meijdam-norm (LPR = 10^{-5}/year). The Meijdam-norm defines the threshold of tolerable risk, where an individual person has a yearly probability of 1 in 100.000 to die due to an earthquake. This risk threshold is similar to the thresholds defined for fatality risk of natural hazards, such as storms or floods.

The input for the Damage Model is:

- Probabilistic ground motion forecast (output) from the GMM.
- Model parameters per typology for the fragility model, translating ground motions to probability of collapse/damage states. These parameters are provided by NAM (Crowley & Pinho, 2017) and have been defined by calibrating numerical models of building damage to experimental results.
- Model parameters per typology for the consequence model, translating the probability of the collapse states to the probability of fatality. These parameters have also been provided by NAM (Crowley & Pinho, 2017).

- Building database of Groningen, containing the locations of 150.000 buildings in the Groningen area and the probability of membership to a certain typology. This database is provided by Arup.

Logic tree

The TNO Model Chain is a probabilistic model that aims to capture all uncertainties. We distinguish two types of uncertainties: epistemic (model) and aleatory. Aleatory uncertainties are statistical uncertainties related to the randomness of the system that is being investigated, for example the time and location of earthquakes. Aleatory uncertainties are captured in the TNO Model Chain by probability distributions. Epistemic or model uncertainties are associated with inadequacies of the model, such as simplifications, theoretical assumptions and limitations in the accuracy of data used for calibration of the model.

To account for epistemic uncertainties a logic tree is used (Figure 5). Every logic tree branching level represents a number of model alternatives. The weights assigned to every logic tree branch should add up to one for each branching level. . Determination of model alternatives of the logic tree, as well as the logic tree weights are not part of the TNO Model Chain. Logic tree parameters and weights are considered as input.

The TNO Model Chain can be run for a single combination of branches, or directly for the full range of 7x4x2x3x3 = 504 branch combinations, computing the mean of the entire logic tree. The Model Chain output of the mean of the logic tree includes all epistemic uncertainties captured in the logic tree.

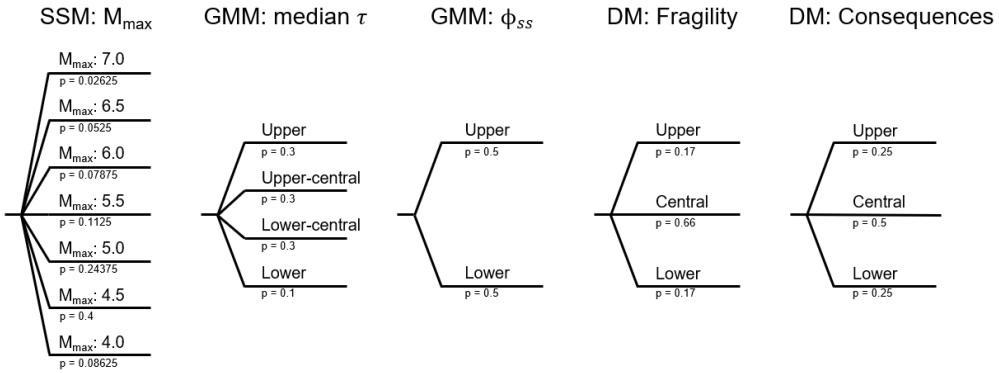


Figure 5 Logic tree used in the TNO Model Chain, consisting of 5 branching levels, each containing 2-7 levels, resulting in a total of 504 (7x4x2x3x3) branch combinations. The p-values are the weights per level, the discrete probabilities.

2 Probabilistic Seismic Hazard Analysis

Probabilistic Seismic Hazard and Risk Analysis (PSHRA) involves calculating the probabilities of possible consequences of earthquake activity, taking into account - as much as possible and/or practically feasible – the various sources of uncertainty. In the current chapter we provide a theoretical framework for PSHRA, with a generic description in Section 2.1, and a more detailed realisation as is used in the TNO Model Chain in Section 2.2. However, many details of the practical implementation are left for Chapters 3 and 4.

2.1 A probabilistic model for seismic hazard and risk analysis

2.1.1 *Earthquake consequences and uncertainty*

The consequences that are being addressed in the PSHRA concern various types of events[2] that may or may not occur as a direct result of a single earthquake. Examples of these events are the exceedance of a particular ground motion intensity level at a specific location, the exceedance of a particular damage state for a specific building, or the decease of a person present inside a specific building as a result of building collapse.

Whether or not an event occurs depends on a wide range of (physical) circumstances, including properties of the earthquake process and properties of the subsurface that propagates the ground motions, but also, depending on the type of event considered, properties of the building or the person exposed, or even the state of infrastructure and the quality of emergency response efforts. These circumstances are to a large extent unknown.

One of the central tasks of PSHRA is to capture all uncertainties in terms of a probabilistic model. It is common to distinguish two categories of uncertainty: epistemic and aleatory.

The first category, epistemic uncertainty, is mainly associated to inadequacies of the model, such as theoretical assumptions and simplifications, computational or numerical compromises, and limitations in the amount or accuracy of the data used for calibration. A frequently occurring limitation is a lack of data in the range of the model that really matters for the future consequences, such that extrapolation is required. A trivial example is the lack of future data: a forecast is always an extrapolation into the future. However, in risk analysis in general, there is always a relative lack of data in the range of the rare events that have the worst consequences. In seismic risk analysis, for example, it is difficult to extrapolate data obtained from lower magnitude earthquakes to forecast the effects of higher magnitude events. Epistemic uncertainty is often interpreted as a lack of knowledge that could, in principle, be minimized by doing more experiments or observations (including also: in the future), developing better models, or making a better effort in computation.

The second category of uncertainty, the aleatory variability, is complementary to epistemic uncertainty, in the sense that it refers to uncertainties that are considered to be irreducible in practice. Examples include the time, place and magnitude of the

[2] To prevent confusion: the term “event” is used in a generic sense as something that may happen with some associated probability. The term “event” does not refer to an earthquake here.

next (or any) earthquake, and the value of a ground motion attribute relative to a (presumably) predictable median.
The distinction between the two categories of uncertainty is inherently subjective. From a practical point of view, it is often more relevant to make a distinction in the quantitative treatment of the uncertainties. Aleatory variability, on the one hand, is usually quantified by means of (parameterized) probability distributions, that are, at least to some extent, calibrated by empirical or experimental data. Inadequacies in the probabilistic descriptions, such as in the choice of distribution and uncertainties in the distribution parameters, belong to the epistemic realm. Epistemic uncertainty, on the other hand, is inherently difficult to quantify. It is usually handled by providing a limited set of model alternatives, with associated weights based on simple heuristics or "expert opinion". Various sources of epistemic uncertainty are usually combined as branching levels in a logic tree framework.

2.1.2 *Random variables, events and probability distributions*

A probabilistic model represents a natural process or experiment in terms of a number of random variables and their mutual relationships, described by their joint probability distribution. In addition, a number of events are defined. These events are specific sets of outcomes of the aforementioned process or experiment to which a probability is assigned. For each outcome of the process/experiment, i.e., for each realization of the random variables, the model predicts whether an event occurs or not. The probability of the event is then defined as the probability-weighted fraction of event occurrences among all possible realizations.
The number and type of random variables included in the probabilistic model may vary per application. Typical examples in PSHRA are ground motion attributes, earthquake magnitudes, hypocentre coordinates, rupture lengths and orientations, building fragility, etc. It may depend on the application which variables are considered to be random and which are chosen fixed. For example, in a scenario-type assessment, where an earthquake of a specific magnitude is assumed to take place at a specific distance from an object of interest, both earthquake location and magnitude will be fixed as a condition, whereas in a personal safety assessment, the probability distributions of both will need to be accounted for.
In the mathematics of probability theory, a probabilistic model is defined by three main parts, collectively called the probability space, being the sample space Ω, a set of events $\mathcal{E}$, and a probability measure P.
Let Ω be the sample space of the random variables associated with the probabilistic model, and ω an arbitrary element of that space, i.e., $\omega \in \Omega$. Note that ω can equivalently be regarded as a vector of interdependent scalar random variables or as single, multivariate random variable. Also note that Ω can contain continuous as well as discrete dimensions.
The probabilistic model associates the sample space Ω with a probability measure $P(\omega)$ such that

$$\int_{\Omega} dP(\omega) = 1. \tag{1}$$

This expression basically states – trivially – that the total probability of all possible realizations of $\omega \in \Omega$ in this probabilistic model equals 1.
Under certain assumptions this integral may also be expressed in terms of a (joint) probability density function (PDF) $f_{\Omega}(\omega)$:

$$\int_{\Omega} f_{\Omega}(\omega) d\omega = 1. \tag{2}$$

Formally, this only holds if all dimensions of Ω correspond to continuous variables and the probability distributions are absolutely continuous. For discrete variables, the integral over a PDF in (2) should actually be replaced with a discrete sum over a probability mass function (PMF). Also, in some cases, the probability density "function" has the properties of generalized function (such as the Dirac delta distribution). This is the case, for example, when a random variable reduces to a deterministic value (e.g., when evaluating a conditional probability). The advantage of the notation in (1) is that it includes all cases and is rather compact, especially when more variables are made explicit; the advantage of the notation in (2) is that it, by being more verbose, is sometimes more convenient to express the interdependence of the random variables.

Any arbitrary function over the sample space, say $g: \Omega \to \mathbb{R}$, defines, in conjunction with the probability measure $P(\omega)$, a random variable. The expectation value $\mathrm{E}(g)$ can be found by integration over the full extent of the probability distribution,

$$\mathrm{E}(g) = \int_{\Omega} g(\omega) dP(\omega) = 1, \tag{3}$$

a process also referred to as marginalization.

To define an event $\varepsilon \in \mathcal{E}$, let g be a quantity of interest, such as a ground motion attribute, a measure of the (excess) seismic demand on a building, or a measure of damage cost. The event ε can now be defined by a predicate on g:

$$\varepsilon: \{\omega \in \Omega \mid g(\omega) > \tilde{g}\}, \tag{4}$$

which defines a subset of the sample space Ω where the predicate holds ($\varepsilon \subset \Omega$). In this example, the predicate defines the exceedance of some reference value $\tilde{g}$ If the parameter $\tilde{g}$ can take on arbitrary values, the formulation (4) basically defines a parameterized range of events. As an example, consider a (continuous) range of (reference) ground motion attributes for which the exceedance probabilities need to be determined. The set $\mathcal{E}$ contains all events that are relevant to the hazard and/or risk assessment.

To determine the probability of the occurrence of event ε, we make use of an indicator function $\mathbf{1}_{\varepsilon}(\omega)$:

$$\mathbf{1}_{\varepsilon}(\omega) = \begin{cases} 1 & \text{if } \omega \in \varepsilon \\ 0 & \text{if } \omega \notin \varepsilon \end{cases}. \tag{5}$$

The probability $\mathcal{P}(\varepsilon)$ can now be expressed as the expectation value of $\mathbf{1}_{\varepsilon}(\omega)$:

$$\mathcal{P}(\varepsilon) = \mathrm{E}(\mathbf{1}_{\varepsilon}) = \int_{\Omega} \mathbf{1}_{\varepsilon}(\omega) dP(\omega), \tag{6}$$

or the probability-weighted fraction of occurrences of event ε for all possible realizations in sample space Ω. Similar to the equivalence of (1) and (2) also this expression may – under the circumstances mentioned above – be expressed in terms of the PDF:

$$\mathcal{P}(\varepsilon) = \int_{\Omega} \mathbf{1}_{\varepsilon}(\omega) f_{\Omega}(\omega) d\omega. \tag{7}$$

It interesting to remark that the two distinct notations in (6) and (7), although mathematically nearly equivalent, point in the direction of two different approaches for numerical implementation. The first approach, following notation (6), is to approximate the integral by a finite sum over discrete samples, where each sample represents a part of Ω with equal probability ($dP_{\Omega} \to \Delta P_{\Omega}$). The second approach, following notation (7), is to approximate the integral by a finite sum over discrete samples, where each sample represents a part of Ω with equal "volume" ($d\omega \to \Delta\omega$). To properly account for the probability structure, each sample must then be weighted by the local probability density ($f_{\Omega}(\omega)\Delta\omega$).

2.1.3 *Earthquake activity and event recurrence rates*

The events discussed in the previous section are causal effects directly associated to a single earthquake occurrence. In practice, however, it is often important to assess the hazard or risk associated to being exposed continuously to all seismicity in a seismically active region. These assessments require taking into account the seismic activity rate, i.e., the expected number of earthquakes above a certain minimum magnitude m_0, per unit time. Also, when considering the spatial distribution of the earthquakes it is useful to specify the activity rate density in space.

Let λ be the activity rate for all earthquakes in the area of interest above a certain minimum magnitude m_0. Then let $\mathcal{P}$ be the probability of occurrence of some defined probabilistic event, taking into account both earthquake origin location and magnitude as random variables. In its simplest form, the event recurrence rate $\mathcal{R}$ due to seismicity anywhere in the region may then simply be expressed as the product:

$$\mathcal{R}(\varepsilon) = \lambda \mathcal{P}(\varepsilon), \tag{8}$$

or in other words, the event recurrence rate $\mathcal{R}(\varepsilon)$ is a fraction of the seismic activity rate λ, with a proportionality factor $\mathcal{P}(\varepsilon)$, being the probability of event occurrence per earthquake.

In many circumstances it is useful to explicitly address the spatial dependence, using the spatial seismic activity rate density $\lambda_X(x)$:

$$\mathcal{R}(\varepsilon) = \int_X \mathcal{P}(\varepsilon|x)\lambda_X(x)dx, \tag{9}$$

with X the spatial domain of the earthquake locations, and $\mathcal{P}(\varepsilon|x)$ the conditional event probability, condition on the occurrence of an earthquake at location $x \in X$. The spatial activity rate density $\lambda_X(x)$ satisfies the following relation:

$$\lambda = \int_X \lambda_X(x)dx, \tag{10}$$

and we can define:

$$f_X(x) = \frac{\lambda_X(x)}{\lambda}, \tag{11}$$

such that $f_X(x)$ may act as the PDF for x (i.e., $\int_x f_X(x)dx = 1$).

To take this one step further, also the magnitude dependence can explicitly be factored out:

$$\mathcal{R}(\varepsilon) = \iint_{X,M} \mathcal{P}(\varepsilon|x,m)\lambda_{XM}(x,m)dxdm, \tag{12}$$

with M the domain of the magnitudes (M:$\{m \in \mathbb{R}|m \geq m_0\}$) and $\mathcal{P}(\varepsilon|x,m)$ the conditional event probability for an earthquake of magnitude m at location x. The earthquake activity rate density in both space and magnitude λ_{XM} is defined as

$$\lambda_{XM}(x,m) = \lambda_X(x)f_M(m|x), \tag{13}$$

with $f_M(m|x)$ the PDF of m, conditional on x, such that:

$$\lambda_X(x) = \int_M \lambda_{XM}(x,m)dm. \tag{14}$$

Similar to above, the earthquake rate density in both space and magnitude can be normalized to a PDF:

$$f_{XM}(x,m) = \frac{\lambda_{XM}(x,m)}{\lambda} = f_X(x)f_M(m|x), \tag{15}$$

such that the event recurrence rate can be written as:

$$\mathcal{R}(\varepsilon) = \lambda \iint_{X,M} \mathcal{P}(\varepsilon|x,m)f_{XM}(x,m)dxdm, \tag{16}$$

or, more succinctly,

$$\mathcal{R}(\varepsilon) = \lambda \iint_{X,M} \mathcal{P}(\varepsilon|x,m)dP(x,m). \tag{17}$$

which both (still) evaluate to (8), which is the marginalized form when magnitude and location are integrated out.

2.1.4 *Reference time frames and event probabilities*

The earthquake activity rates in the previous section have been expressed without a specific reference to time, although it is understood that in general circumstances, but especially in the case of induced seismicity, the seismic activity rate itself varies with time. The activity rate density $\lambda_X(x)$ in (9) can be interpreted in two major ways. First, it can be interpreted as the instantaneous rate density:

$$\lambda_X(x) \equiv \lambda_X(x,t), \tag{18}$$

where we explicitly express the dependence on time. Second, it can be interpreted as the average rate density over a time interval T, with length $|T|$:

$$\lambda_X(x) \equiv \frac{1}{|T|}\int_T \lambda_X(x,t)dt. \tag{19}$$

Note that expression (19) includes expression (18) as a special case, in the limit $|T| \to 0$, an infinitesimal time interval. Also note that, in general, the magnitude distribution may change over time as well. In that case, either a similar interpretation can be made for the rate density $\lambda_{XM}(x,m)$ in (13), or time dependence of conditional probability $\mathcal{P}(\varepsilon|x)$ in (9) – in which the magnitude distribution is incorporated – should be properly accounted for (i.e., $\mathcal{P}(\varepsilon|x) \equiv \mathcal{P}(\varepsilon|x,t)$).

For a hazard or risk assessment it is common practice to refer to probabilities within a specific – hypothetical – reference time frame. This time frame may be, for example, 1 year, or 50 years, but in general it may be different from both the actual time frame of analysis (instantaneous as in (18), or an interval, as in (19)), and the basic unit of time (seconds, years, ..).

In the reference time frame, say Δt, the region-wide activity rate $\lambda = \int_X \lambda_X(x)dx$ is then considered stationary, and the expectation value of the total number of earthquakes is equal to $\lambda\Delta t$.

In (hypothetical) realizations of seismicity in the reference time frame the actual number of earthquakes will vary. The earthquake count n is a random variable that can be described using a discrete probability distribution.

If all earthquakes are mutually independent, in the sense that the spatio-temporal and magnitude probability distribution of each earthquake does not depend on previous earthquake occurrences, then the earthquake count in any finite interval follows a Poisson distribution. For an expected count of ν, the probability mass function (PMF) of the Poisson distribution reads

$$p_N(n|\nu) = \frac{\nu^n}{n!}e^{-\nu}, \tag{20}$$

and indeed, the probability-weighted count – or expectation value – equals $\sum_n n\, p_N(n|\nu) = \nu$. Hence, assuming the Poisson distribution, the probability of encountering n earthquakes (above magnitude m_0) in the reference time frame equals $p_N(n|\lambda\Delta t)$. A similar argument can subsequently be used for the event count, replacing activity rate λ by event recurrence rate $\mathcal{R}(\varepsilon)$, such that the probability of encountering n events in the reference time frame equals $p_N(n|\mathcal{R}(\varepsilon)\Delta t)$.

The most common probability metric is the probability $\mathcal{P}(n_\varepsilon > 0)$ – with n_ε the number of occurrences of event ε – that an event happens at least once in the

reference time frame. This probability is the complement of the probability that the event does not occur at all:

$$\mathcal{P}(n_\varepsilon > 0) = 1 - \mathcal{P}(n_\varepsilon = 0) = 1 - p_N(0|\mathcal{R}(\varepsilon)\Delta t) = 1 - e^{-\mathcal{R}(\varepsilon)\Delta t}. \quad (21)$$

In seismic hazard analysis, this relationship is often used in an inverse manner. When both the probability $\mathcal{P}$ and the reference period Δt are fixed, then the rate $\mathcal{R}(\varepsilon)$ is determined by:

$$\mathcal{R}(\varepsilon) = -\frac{\ln(1-\mathcal{P})}{\Delta t}. \quad (22)$$

In case of the choices $\mathcal{P} = 10\%$ and $\Delta t = 50$ years, the corresponding rate $\mathcal{R}(\varepsilon)$ is 2.11×10^{-3} per year. The inverse of the event rate, the event return period, equals 4.75×10^{2}. years. Likewise, for $\mathcal{P} = 2\%$, the numbers are 2.010×10^{-4} per year and 2.475×10^{3}. years. These numbers explain the remarkably return periods of 475 and 2475 years that are frequently used in seismic hazard assessments. In seismic hazard assessment the event ε stands for the exceedance of some ground motion attribute value. The remaining task is a search to find the attribute value for which the exceedance rate equals the specified rate.

Note that the Poisson assumption of independent events does not hold universally, especially in the presence of earthquake clustering, i.e., in the presence of aftershock/foreshock mechanisms. In those circumstances equations (21) and (22) do not hold and either should be replaced, or should be treated as approximations. For the communication of PSHRA results it is often even more convenient to refer directly to the (mean/expected) rates of Section 2.2, rather than the probabilities of the current section. For rare events, annual probabilities and annual rates are approximately equal.

2.1.5 *Aggregate event rates*

The equations in the previous sections describe events that may or may not occur as a direct result of an earthquake, but in all cases, the events are one-to-one associated with a single earthquake. However, some type of events may occur more than once for a single earthquake. For example, many buildings (and people) are exposed to the same seismicity at once. Let's say, for example, that ε represents the collapse of a building of a specific type. If the distribution of buildings of that type in the area can be represented by a density function $\alpha(y)$, where y represents the spatial coordinates of the building ($y \in Y$), then the aggregate building collapse rate $\mathcal{R}_\alpha(\varepsilon)$ may be expressed as:

$$\mathcal{R}_\alpha(\varepsilon) = \iiint_{Y,X,M} \alpha(y)\mathcal{P}(\varepsilon|y,x,m)\lambda(x,m)dmdxdy. \quad (23)$$

Similarly, when the event ε represents a fatality due to an earthquake and $\alpha(y)$ the population density, then $\mathcal{R}_\alpha(\varepsilon)$ represents the aggregate fatality rate for the region. Note that aggregate rates cannot simply be translated into probabilities for a reference time frame as in Section 2.1.4, because at the exposure side, many events may occur synchronously, so that the Poisson model does not apply.

2.2 Elements of TNO Model Chain Groningen

The Probabilistic Seismic Hazard and Risk (PSHRA) model for induced seismicity in the Groningen gas field consists of a chain of models that represents the causal chain of processes that relates gas production to seismic hazard and seismic risk. In a relatively coarse form the TNO Model Chain Groningen is illustrated in Figure 1, showing as basic chain elements the Seismological Source Model (SSM), the Ground Motion Model (GMM), and the Damage Model (DM). The TNO Model Chain

in reality has a hierarchical structure, where each of the model chain elements by itself can represent a chain of sub-models.

The current section introduces the main model components, as well as the respective sub-models, on a rather global level, with emphasis on the general structure of the models, the interdependence of the model components, and the random variables considered. This exposition is intended to be more or less independent of the precise version of the model components. Details on the functional forms are discussed further in Chapter 4. Sections 2.2.1 to 2.2.3 introduce the SSM, the GMM and the DM respectively. Section 2.2.4 describes the integration of the model components to obtain hazard and risk metrics.

2.2.1 *Seismological source model*

As displayed in Figure 1, the seismic source model (SSM) makes a forecast of the seismicity distribution using a gas production scenario as input. The seismicity distribution, in this context, involves not only hypocentre locations in both space and time, but also the earthquake magnitudes, and the geometry of the rupture planes associated with the earthquakes. The SSM is the most complex part of the model chain. The following bullet list provides an hierarchical overview of the components.

- Dynamic subsurface model (section 2.2.1.1)
    - Reservoir fluid flow model
    - Compaction model
    - Shear strain/stress model
    - Covariate conditioning model
- Seismicity rate model (section 2.2.1.2)
    - Main shock rate model
    - Magnitude model
    - Clustering model
- Rupture model (section 2.2.1.3)
    - Rupture geometry model
    - Rupture distance model

As indicated in the bullet list, the subsections first discuss the model components. To conclude, subsection 2.2.1.4 discusses the calibration of the SSM on observed data.

2.2.1.1 Dynamic subsurface model

The activity rate of seismicity in the Groningen gas field has, according to the observations, been variable both in time and in space. A seismological source model (SSM) that is to be used to forecast seismicity in terms of gas production scenario's for the future, should also be able explain the observations of the past, conditional on the gas production realized in the past. The term "explanation" in this context should be understood in probabilistic sense. The SSM provides a probabilistic forecast of both the number of earthquakes that are to be expected in a given time interval, as well as their distribution in time, space, and magnitude domain.

In brief, the task for the SSM is to provide a quantitative relation between gas production parameters and the seismicity rate density in space and magnitude, that is λ_{XM}, and its variation in time. In other words, λ_{XM} should depend explicitly on subsurface attributes that vary both in space and time as a result of the gas production. The approach taken in the Groningen SSM consists of two steps The first steps to define a set of subsurface attributes that will act as a predictor

variables, or covariates for the second step. The second step subsequently expresses the seismicity rate in terms of those covariates.
For the Groningen SSM, various attributes have been proposed, tested and applied as covariates (Bourne & Oates, 2018). In the calculation of the covariates two mechanical models play an important role. First, a reservoir fluid flow model predicts the spatio-temporal pressure evolution within the reservoir from time series of gas volumes extracted at the wells. Second, a compaction model predicts the deformation of the reservoir as a result of the pressure variations within the reservoir. From the compaction, the Coulomb stress field is computed using information on the existing fault structures in the reservoir.
For the recent SSM models (Bourne et al., 2019), the Coulomb stress and fault density are conditioned with a spatial smoothing scale parameter and a filter on the contributing fault segments based on the fault-throw/reservoir-thickness ratio. The model parameter set used in this conditioning is denoted with γ. After the conditioning, these quantities act as covariates for the seismicity rate model. The exact expression of the covariates in terms of subsurface attributes depends on the version of the model used and is discussed in more detail in Chapter 4. For the purpose of the current section it suffices to specify the dependencies of variables involved. The statement:

$$c \leftarrow \{x, t, \gamma, \mathcal{S}\} \tag{24}$$

expresses the dependency of covariates c on spatial coordinate x and time t. In addition, the dependency statements includes the model parameter set γ, as well as the symbol $\mathcal{S}$, that represents the input data imposed by the gas production scenario.

2.2.1.2 Seismicity rate model

In the seismicity rate model a distinction is made between the main-shock seismicity rate $\hat{\lambda}$ and the total seismicity rate λ. The main-shock seismicity rate explains the mutually independent earthquakes that follow a Poisson process, while total seismicity rate also includes additional earthquakes that are causally related to previous earthquakes in the vicinity, commonly referred to as aftershocks.
The (instantaneous) main-shock rate density in space, $\hat{\lambda}_X$, is fully determined by the subsurface covariates c and a set of seismicity rate model main shock parameters θ as specified in the following dependency statement:

$$\hat{\lambda}_X \leftarrow \{c, \theta\}. \tag{25}$$

The combination with (24) gives a nested dependency that may be expanded into:

$$\hat{\lambda}_X \leftarrow \{x, t, \gamma, \theta, \mathcal{S}\}. \tag{26}$$

For the total main-shock rate (the spatial density integrated over space) this gives:

$$\hat{\lambda} \leftarrow \{t, \gamma, \theta, \mathcal{S}\}. \tag{27}$$

As discussed in section 2.1.4, when it comes to forecasts, it is sometimes useful to suppress the explicit time dependence of the seismicity rate. In this way, the same formulation holds both for instantaneous rates, and for average rates in some time interval. Therefore, for forecasts, we may drop t from the list in (26) and (27), although the dependence on instantaneous time or some definite time interval is still present implicitly.
In section 2.1.3 we discussed the treatment of hypocenter location x and magnitude m as random variables, and in section 2.1.4 we discussed the earthquake count n, a random variable representing the number of earthquakes expected in a reference time frame. Let $\hat{n}$ be the main shock count. To highlight both the random character

of variables and their dependencies, without having to make their functional form explicit, we introduce the following notation, demonstrated here for $\hat{n}$ and $\hat{x}$:

$$\hat{n} \leftsquigarrow \{\gamma, \theta, \Delta t, \mathcal{S}\}, \tag{28}$$

$$\hat{x} \leftsquigarrow \{\gamma, \theta, \mathcal{S}\}, \tag{29}$$

with $\leftsquigarrow$ used to indicate a probabilistic dependency, contrasting with the $\leftarrow$ for a deterministic dependency.Statement (28) says that the distribution of $\hat{n}$ is is conditioned on the covariate conditioning parameter set γ, the main-shock rate parameter set θ, the chosen reference time frame Δt, and the gas production strategy $\mathcal{S}$. In fact, we know in this case that $\hat{n}$ is distributed as a Poisson distribution (20) with rate $\hat{\lambda}\Delta t$, such that $\int \hat{n} dP(\hat{n}) = \hat{\lambda}\Delta t$. Statement (29) says that the main-shock hypocentre location $\hat{x}$ has similar dependency except that it does not depend on the reference time frame.

The magnitude model defines the magnitude distribution ("magnitude-frequency relation") as a function of the subsurface covariates c as well. In addition, the model depends on a parameter set ψ and a discrete index variable $\mathscr{b}_s$, that labels a number of alternative magnitude model choices in the context of epistemic uncertainty. The index variable represents a branching level in the logic tree. By assigning weights to the individual branches, the index variable becomes a categorical random variable.

Magnitude m is a random variable, with the following conditional dependencies:

$$m \leftsquigarrow \{c, \psi, \mathscr{b}_s\}. \tag{30}$$

As for the spatial rate density, also in this case the nested dependency of c could be expanded. However, there is no advantage in doing that at this point.

The total seismicity rate, i.e., including non-Poissonian clustering, is obtained using the Epidemic-Type Aftershock Sequence (ETAS) model. In this model, each earthquake that occurs raises the seismicity rate locally and temporarily. Given a catalogue of n earthquakes, with origin times t_i, magnitudes m_i, and hypocenter coordinates x_i, $(i = 1..n)$, the observation-conditioned total seismicity rate density λ_X^{obs} becomes:

$$\lambda_X^{\mathrm{obs}}(x, t, \gamma, \theta, \zeta, \mathcal{S}) = \hat{\lambda}_X(x, t, \gamma, \theta, \mathcal{S}) + \sum_{i=1}^{n} g_\lambda(t - t_i, |x - x_i|, m_i, \zeta), \tag{31}$$

where $|x - x_i|$ is the epicentral distance, ζ is the ETAS model parameter set, with $\zeta = \{a, K\}$, and g_λ is the spatio-temporal aftershock triggering function defined as:

$$g_\lambda(t, r, m, \zeta) = \begin{cases} 0 & \text{if } t \le 0 \\ K e^{a(m - m_0)} f_T(t) f_R(r) & \text{if } t > 0 \end{cases}. \tag{32}$$

where a and K parameterize the magnitude dependent aftershock productivity.

The triggering function (32) includes PDF's for time (f_T) and distance (f_R), such that the integral over time and space provides the aftershock productivity:

$$\iint_0^\infty g_\lambda(t, r, m, \zeta)(2\pi r) dr dt = K e^{a(m - m_0)}. \tag{33}$$

In a seismicity rate forecast for a time interval at some distance in the future, the actual earthquake occurrences are not known, and therefore should be assumed distributed everywhere and any time, with the frequency imposed by the seismicity rate. The effective increase in activity rate due to aftershocks is therefore distributed as well. Under the relatively mild assumption that the main-shock rate is smooth in time and space compared to the length scales of $f_T(t)$ and $f_R(r)$, the effects of aftershocks can be represented by the aftershock productivity only. Note that this does not hold for short-term forecasting, where the enhanced rates due to the

earthquakes that occurred in the recent past should be included as prior conditions explicitly.
The effective aftershock productivity can be seen as the product of a sequence of aftershock generations, each generation increasing the previous generation by a constant factor. The productivity factor ξ for each generation can be calculated as follows:

$$\xi = \int_M K e^{a(m-m_0)} dP(m). \tag{34}$$

Using (30) and with $\zeta = \{a, K\}$ we find:

$$\xi \leftarrow \{c, \psi, \zeta, \mathscr{b}_s\}. \tag{35}$$

For physically realistic models, the factor ξ is smaller than 1, which puts a prior constraint on $\zeta = \{a, K\}$. The effective productivity factor for all generations Γ then becomes a geometric series:

$$\Gamma = 1 + \xi + \xi^2 + \xi^3 + \cdots = \frac{1}{1-\xi},$$

such that:

$$\Gamma \leftarrow \{c, \psi, \zeta, \mathscr{b}_s\}. \tag{36}$$

The (unconditioned) total seismicity rate λ can now be found by integrating Γ over the spatial domain:

$$\lambda = \hat{\lambda} \int_X \Gamma dP(\hat{x}), \tag{37}$$

with the dependencies composed from (27),suppressing t, (29) and (36):

$$\lambda \leftarrow \{\gamma, \theta, \psi, \zeta, \mathscr{b}_s, \mathcal{S}\}. \tag{38}$$

This expression makes clear that the total seismicity rate depends on many parameters: parameters of the covariate conditioning model, the main-shock seismicity rate model, the magnitude model and the clustering model, as well as on the magnitude model choice, and ultimately the production parameters. It is interesting to note that the coupling between the magnitude model and the total seismicity rate model, as witnessed by the ψ parameter in (38), is induced by the ETAS clustering model, because the aftershock activity depends on the magnitude distribution. This coupling is not yet present in the main-shock rate (27), nor in the observation-conditioned seismicity rate of (31). This means that if the rate model is conditioned on an observed dataset, there is no explicit coupling (correlation) between the activity rate and the magnitude model parameters.
From (38) follow the distributions for the total number of earthquakes n in a time interval Δt, and the spatial location of any earthquake (main or aftershock), analogous to (28) and (29):

$$n \nleftarrow \{\gamma, \theta, \psi, \zeta, \mathscr{b}_s, \Delta t, \mathcal{S}\}, \tag{39}$$

$$x \nleftarrow \{\gamma, \theta, \psi, \zeta, \mathscr{b}_s, \mathcal{S}\}, \tag{40}$$

as well as the notion that

$$dP(x) = \Gamma dP(\hat{x}). \tag{41}$$

Finally, given either deterministic values or probability distributions for $\gamma, \theta, \psi, \zeta$, and $\mathscr{b}_s$, the expectation of the total seismicity rate can be found by marginalization over the full probability space:

$$\lambda^{\text{MEAN}}(\mathcal{S}) = \iiiint \lambda dP(x, \gamma, \theta, \psi, \zeta, \mathscr{b}_s). \tag{42}$$

In section 2.2.1.4 we discuss the ways to obtain the probability distributions for the parameter sets. The probability distribution for the logic tree branching level $\mathscr{b}_s$ is usually set by means of expert judgment/elicitation.

2.2.1.3 Rupture model

The seismicity rate density forecasts discussed above describe the spatial distribution of earthquake hypocentres. A hypocentre is defined as the point in space where an earthquake starts, or in other words, where the earthquake rupture nucleates. After nucleation, the rupture may and will propagate in various directions. In the Groningen model chain it is assumed that earthquakes nucleate at reservoir depth, assumed at 3 km. The rupture takes place at existing fault planes that are sub-vertical, and ruptures may propagate laterally along the strike directions and down dip, but not upwards. Since the ground motions in the Groningen model are conditioned on the nearest distance of the observation point to the rupture plane (see Section 2.2.2), the geometry of the rupture plane is an important element of the forecast, especially for higher magnitudes.

The fault geometry model used in the TNO Model Chain Groningen is relatively simple. The fault planes are assumed to be perfectly vertical such that the point on the rupture plane with the shortest distance to an observation point is always at the top of the fault. This means that only the horizontal trace of the rupture is relevant, not its extent in depth. Also, the fault plane is always assumed to be planar, so the trace is a straight line segment. The rupture trace segment geometry is described by three random variables: (1) its length, which is magnitude dependent, (2) its orientation (azimuth) relative to the median value φ_{rup}, and (3) its position relative to the hypocenter. The random variables have a lognormal, a normal and a bounded constant distribution, respectively. In the following, these parameters are summarised in the rupture model parameter set ρ. More details are provided in Chapter 4. In the seismological models proposed by NAM the median rupture azimuth and its variations have been chosen fixed for the entire field, based on the strike of the dominant fault systems.

Let x be a hypocentre, and y a point at the surface, i.e., an observation point or the location of some exposed structure. Then the hypocentral distance r_{hyp} and azimuth φ_{hyp}, and the relative angle of observation of the rupture plane φ are well defined:

$$\{r_{\mathrm{hyp}}, \varphi_{\mathrm{hyp}}\} \leftarrow \{x, y\}, \tag{43}$$

$$\varphi = \varphi_{\mathrm{hyp}} - \varphi_{\mathrm{rup}}. \tag{44}$$

As a result, given the geometry of the rupture plane, i.e., an instance of the rupture model parameters ρ, the rupture distance, or the distance to the nearest point on the rupture is determined as well:

$$r_{\mathrm{rup}} \leftarrow \{r_{\mathrm{hyp}}, \varphi, \rho\}. \tag{45}$$

The rupture model also defines the probability distributions for the rupture model parameters, which depend only on magnitude, i.e.,

$$\rho \nleftarrow \{m\}. \tag{46}$$

As a result, by marginalizing the rupture model parameters, a rupture distance distribution is obtained:

$$r_{\mathrm{rup}} \nleftarrow \{r_{\mathrm{hyp}}, \varphi, m\}, \tag{47}$$

which shows that the rupture distribution depends only on hypocentral distance, the azimuth relative to the median rupture orientation, and the magnitude.

2.2.1.4 Seismic source model calibration

The success of the seismicity rate model depends to a large degree on the various parameter settings. The parameters can be calibrated on the observed seismic data using a hindcast based on the historic production scenario. The hindcasted seismicity rate, with aftershock rate distributions conditional on the actual events is

shown in equation (31). Extended with the magnitude distribution (30), with PDF f_m this gives:

$$\lambda_{XM}^{\mathrm{obs}}(x,t,m,\gamma,\theta,\psi,\zeta,\mathscr{b}_s,\mathcal{S}) = \lambda_X^{\mathrm{obs}}(x,t,\gamma,\theta,\zeta,\mathcal{S}) f_m(m|c(x,t),\psi,\mathscr{b}_s). \quad (48)$$

The total expected number of events over the observation periods is found by a temporal and spatial integral:

$$\Lambda^{\mathrm{obs}}(\gamma,\theta,\zeta,\mathcal{S}) = \iint_{X,T} \lambda_X^{\mathrm{obs}}(x,t,\gamma,\theta,\zeta,\mathcal{S}) dt dx. \quad (49)$$

The combination of (48) and (49) gives a probability distribution in space, time and magnitude for all events:

$$f_{XTM}(x,t,m|\gamma,\theta,\psi,\zeta,\mathscr{b}_s,\mathcal{S}) = \frac{\lambda_X^{\mathrm{obs}}(x,t,\gamma,\theta,\zeta,\mathcal{S})}{\Lambda^{\mathrm{obs}}(\gamma,\theta,\zeta,\mathcal{S})} f_m(m|c(x,t),\psi,\mathscr{b}_s). \quad (50)$$

In the context of parameter estimation the probability distribution is a likelihood function that can be applied to all observed earthquakes. Because of the normalization, the likelihood function above is not sensitive to the event count. However, a complementary likelihood expression for the number of observed earthquakes is found in (20):

$$p_N\left(n^{\mathrm{obs}} \middle| \Lambda^{\mathrm{obs}}(\gamma,\theta,\zeta,\mathcal{S})\right). \quad (51)$$

The total likelihood is a product of (51) and n^{obs} evaluations of (47), one for every earthquake in the catalogue.

According to the Bayesian approach to parameter estimation, the posterior probability distribution for the parameters is obtained by a multiplication of the prior probability distribution and the likelihood.

2.2.2 *Ground motion model*

Various generations of ground motion models for Groningen seismicity have been developed by Bommer et al. (2015-2018). The ground motion attributes used in the TNO Model Chain Groningen are the spectral accelerations and the significant duration. In the following, the attributes are summarised in the multivariate ground motion attribute g:

$$g \equiv (Sa[0.01s], Sa[0.025s], \ldots, Sa[5.0s], D), \quad (52)$$

which contains peak spectral accelerations Sa for a range of spectral periods, and significant duration D.

The ground motion model is split in two stages that capture two parts of the propagation model. The first stage describes the ground motions at the reference level, the base North Sea Group (~800 m depth). These motions do not depend on the spatial coordinates of either the hypocentre, or the observation point. The ground motions at reference level, g_{ref}, are described as lognormal distributions, conditioned on the rupture distance r_{rup}, the magnitude m and the logic tree index $\mathscr{b}_g$, which represents epistemic uncertainties in both the median ground motion and the ground motion variability:

$$g_{\mathrm{ref}} \rightsquigarrow \{r_{\mathrm{rup}}, m, \mathscr{b}_g\}. \quad (53)$$

The second stage describes the propagation of ground motion from the reference level to the free surface, also referred to as the site response. The site response amplification factors are lognormally distributed as well, conditional on the ground motion at reference level g_{ref}, but also, for a number of spectral periods, on the rupture distance r_{rup} and the magnitude m. The shallow subsurface is subdivided in a number of site response zones with different propagation characteristics. The site response zone is indicated by the discrete index variable $\mathscr{s}$:

$$g \leftsquigarrow \{r_{\text{rup}}, m, g_{\text{ref}}, \mathcal{s}\}. \tag{54}$$

The functional forms of the ground motion models are discussed in Chapter 3.

2.2.3 *Damage Model*

The damage model in the Groningen model chain is based on the work of Crowley & Pinho (2017) and Crowley et al. (2017). It comprises the fragility and consequence models.

2.2.3.1 Fragility model

The first step in the fragility models is the definition of an intensity measure for the seismic demand imposed on the buildings by the seismic ground motions, in terms of the displacement. The measure is different for the each building typology, but in all cases it is defined as a linear combination of a number of ground motion attributes in the log scale. The intensity measure η^D therefore depends on the ground motion attributes g, and a set of coefficients for each typology, represented by the typology index variable $\mathcal{t}$:

$$\eta^D \leftarrow \{g, \mathcal{t}\}, \tag{55}$$

where the superscript D is used to indicate that this is the (seismic) demand on the structure. Since the classification of a specific building as being member of a certain typology is often uncertain, the typology index $\mathcal{t}$ is itself a categorical random variable.

Depending on the seismic demand and the seismic capacity of the structure, various degrees of damage may occur. The fragility framework of Crowley & Pinho (2017) defines up to seven consecutive damage/collapse states for each typology. The first four states represent various degrees of damage, while the last three states represent various stages of building collapse.

The thresholds between the states are defined by limit values (displacement limits) of the intensity measure. The state of the building corresponding to a limit value is referred to as a limit state. The set of six limit values is represented by the multivariate random variable η^C, where the superscript C stands for capacity. The variable η^C is parameterised by a set of six sequential reference values η^C_{ref}, and a single aleatory (building-to-building) variability parameter β that is common to all limit values:

$$\eta^C \leftsquigarrow \{\eta^C_{\text{ref}}, \beta\}. \tag{56}$$

All components of η^C, are perfectly correlated, such that the values never cross. The limit state reference values in turn depend on the typology index $\mathcal{t}$ as well as the logic tree branch index $\mathcal{b}_f$ that represents the epistemic uncertainty:

$$\eta^C_{\text{ref}} \leftarrow \{\mathcal{t}, \mathcal{b}_f\}, \tag{57}$$

while the building-to-building variability depends on the typology only:

$$\beta \leftarrow \{\mathcal{t}\}. \tag{58}$$

In the context of seismic risk analysis a typical example of a probability that is being assessed is the probability of exceeding a certain limit state. In the language of the probabilistic model of Section 2.1, the exceedance of limit state (i) is an “event”, say, $\varepsilon_{\text{EXC},i}$, defined as:

$$\varepsilon_{\text{EXC},i}: \{\omega \in \Omega \,|\eta^D(\omega) > \eta^C_i(\omega)\}, \tag{59}$$

where ω and Ω represent all relevant random variables in the model. The event occurs when the seismic demand exceeds the capacity associated with limit state i.

2.2.3.2 Consequence model

The second part of the damage model is the consequence model, which describes the probability of a hypothetical person present inside or just outside a particular building that either exceeded a particular collapse limit state, or experienced a chimney collapse to die or survive. The consequence model of Crowley & Pinho (2017) conditions the mortality (probability of dying) on the damage state. Therefore we define the ordinal random variable $\mathcal{u}$, representing the current damage/collapse state of the building by the value of the seismic demand relative to the capacity limit states. It therefore depends on both the seismic demand and the building capacity in terms of the intensity measure:

$$\mathcal{u} \leftarrow\!\!\sim \{\eta^D, \eta^C\}. \tag{60}$$

We next define the binary categorical consequence variable κ, which represents the two possible states of the hypothetical person under consideration, being "alive" (κ^*) or "dead" ($\kappa^\dagger$), and has the following dependencies:

$$\kappa \leftarrow \{\mathcal{t}, \mathcal{u}, \ell, \mathcal{b}_c, g\}, \tag{61}$$

among which the typology $\mathcal{t}$, the collapse state $\mathcal{u}$, and the epistemic uncertainty represented by the logic tree branching index $\mathcal{b}_c$. In addition, the categorical variable ℓ represents the probability that the hypothetical person is either inside, or outside of the building. Finally, a direct dependence on surface ground motion g is included, since the fatality due to chimney collapse is conditioned directly on the peak ground acceleration (PGA), which is included in g.

The probabilistic event that represents the fatality of the hypothetical person, ε_{FAT}, can now be defined as:

$$\varepsilon_{\text{FAT}}: \{\omega \in \Omega \,|\kappa(\omega) = \kappa^\dagger\}, \tag{62}$$

where ω and Ω represent all relevant random variables in the model.

2.2.4 *Integration*

In the previous three sections we have discussed the various (random) variables that play a role in the hazard and risk assessment. To quantify the rate or probability of some probabilistic event, according to equations (6), (8) and (21) in Chapter 2, the ultimate task is to integrate over the random variables of interest to that event.

As we have seen, for example in (42), the total seismicity rate λ depends on a subset of the random variables. Therefore, it is important to include λ in the rate integral (8), e.g.:

$$\mathcal{R}(\varepsilon, \mathcal{S}) = \iint_{X,M} \int_{\Omega} \mathbf{1}_{\varepsilon}(\tilde{\omega}, x, m)\lambda(\tilde{\omega}, \mathcal{S}) dP(\tilde{\omega}|x, m)\, dP(x, m), \tag{63}$$

where $\tilde{\omega}$ represents all random variables of interest, except x and m, and $\mathcal{S}$ represents the production parameters as in (42). The probability distributions of x and m may also be relegated to the rate density function in space and magnitude λ_{XM}, using (15), as in:

$$\mathcal{R}(\varepsilon, \mathcal{S}) = \int_{\Omega} \iint_{X,M} \mathbf{1}_{\varepsilon}(\tilde{\omega}, x, m)\lambda_{XM}(x, m, \tilde{\omega}, \mathcal{S}) dx dm\, dP(\tilde{\omega}). \tag{64}$$

From a practical point of view, the order of integration over the various random variables has a large influence on the computational efficiency. This further elaborated in Chapter 3.

3 Implementation of the TNO Model Chain

3.1 Seismological Source Model (V5)

The Seismological Source Model (SSM) V5 is based on theory presented in Bourne & Oates (2017) and Bourne et al. (2018). The essential elements are discussed here together with their implementation in the Groningen model chain.

3.1.1 *Input files*

Since the Seismological Source Model is the first component of the TNO Model Chain, all input files to the SSM are not produced by other chain elements and need to be parsed before they can be used. The parsing is described in this section.

Earthquake catalogue

The earthquakes that are observed in the Groningen area are recorded by the Royal Netherlands Meteorological Institute (*Koninklijk Nederlands Meteorologisch Instituut*, KNMI). The induced earthquake records are available in csv-format at link. This record contains the timing (date and time of day specified to hundreds of seconds), latitude and longitude (specified in decimal format to a thousandth of a degree), magnitude (specified with one decimal place) and depth (all induced earthquakes in the Groningen area are assumed to occur at 3.0 km depth). Other information in the record, such as municipality and PMF mode are not parsed.

The date and time are transformed to a decimal format, taking into account that leap years are 366 days long. Example: *12:00:00 01-05-2016* (leap year): 2016.33196721, *12:00:00 01-05-2017* (non-leap year): 2017.33013699.

Latitude and longitude (WGS-84, EPSG:4326) are transformed to Rijksdriehoek (RD coordinates, EPSG:28992) using the pyproj library.

The catalogue is filtered spatially based on whether it falls within the boundary of Groningen gas field (Groningen_field_outline.csv), temporally on whether it falls within the specified date range (date range is considered to be inclusive, e.g.: 1-jan-2000 to 24-may-2010 means earthquakes occurring between 1-jan-2000 00:00:00 up to 24-may-2010 23:59:59 will be included), and on minimum magnitude (if M_{min} =1.5, earthquakes of M1.5 and above will be included).

After these steps, the result is an array of earthquakes within the specified date range, with a magnitude of M_{min} or higher, falling within the field outline. Each earthquake has a location (in RD coordinates), a timing (decimal year format) and a magnitude.

Reservoir thickness, reservoir compressibility and pore pressure

The relevant reservoir properties are provided as csv files, giving values for RDx, RDy and the respective reservoir thickness (in meters), compaction coefficient (in MPa^{-1}), or pore pressure (in bar, with every column denoting the pore pressure for a single snapshot in time).

Typically these files are given on a regular 2D grid, although this is not obligatory. If the data is supplied on a regular grid (constant and equal dx and dy between points), this grid is used to define a base grid. If the data is not on a regular grid, the base grid will be based on the extent of the data and a user-supplied dxdy (see Figure 6). Finally, the original data is linearly interpolated to the base grid. Grid points that cannot be interpolated (i.e. grid points that are outside the convex shape described by the original points), get assigned a value of NaN (not a number). Note that for data that was originally on a grid, this means that original grid points are maintained.

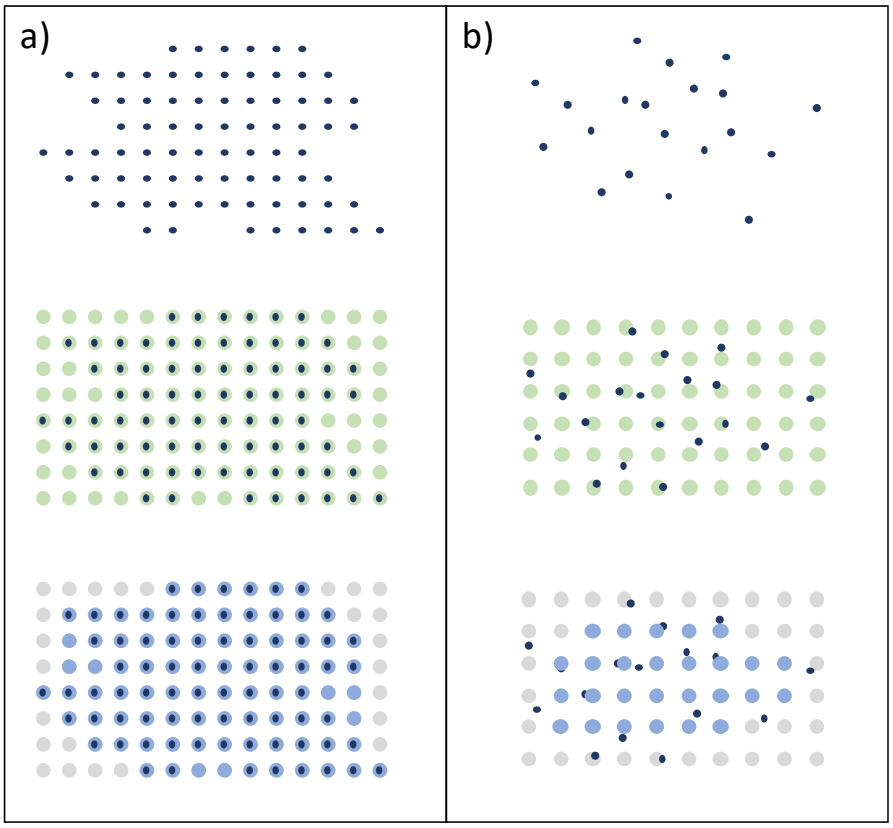


Figure 6: Gridding procedure for data that is already on a grid (a) and data that is originally not on a grid (b). Top: original data locations. Middle: definition of base grid (green). Bottom: grid points that are assigned a numerical value (blue) are inside the convex hull of the original data, grid points that are assigned NaN (grey) are outside the convex hull of the original data.

The pore pressure grid is provided in units of bar. It is translated to units of MPa by dividing each pore pressure value by 10. The pore pressure grid is then transformed into a pore pressure change grid by subtracting the pore pressure for each time step from the initial pore pressure (i.e. depletion is positive).

When performing the modelling, it is important that the spatial extent and positions for the reservoir thickness, reservoir compressibility and pore pressure grids are identical (i.e. that the grid points for all these grids are the same). Since they are provided through different input files, this is not necessarily the case. After reading and gridding each file, a check is performed whether all grids are identical. If they are, no action is needed. If they are not:

- If the user provides a grid definition, all gridded data is remapped to this grid, using linear interpolation. (N.B. other interpolation methods have been tested, with no significant impact on the resulting source distribution. The supplied input files are relatively smooth, resulting in negligible differences due to choice of interpolation method).
- If the user does not provide a grid, the grid definition of the pore pressure grid is taken as the base grid. All other grids are remapped to this grid, using linear interpolation. (N.B. the choice for the pore pressure grid as base grid is arbitrary).

Field outline
The field outline (the projection to the earth's surface of the pre-production position of the gas-water contact) is provided as a sequential set of RDx, RDy coordinates. These points are stored as provided without requiring any parsing.

Fault data
Fault data is provided as an sqlite3/csv-table. Each row describes a point in space where a fault has been interpreted at reservoir level. Beside fault location, the following properties are supplied and stored: a number describing which fault the point belongs to; a number describing it's position within the fault; the offset of the fault; the thickness of two reservoir layers in the footwall; the thickness of two reservoir layers in the hanging wall (four thickness values in total). Other properties (such as dip and dip azimuth) are not used in the model.

The data is read into the model and the following properties are calculated at each point:

- Average thickness (t_{avg}): arithmetic mean of the four thickness values supplied in the input file.
- Throw/thickness ratio: offset/thickness.
- Representative length (l_{repr}): each point represents a certain length (along strike) of fault, for which we assume that the properties of the point are representative. Points that are closer together have a smaller *representative length* per point than points that are further apart. The procedure for determining the representative length per point (see also Figure 7) is:
    - Determine the midpoints between the original points making up the fault.
    - Find the distance from each original point to **both** of its neighboring midpoints.
    - Assign each original point the sum of the distance to both of its neighboring midpoints (the ends of the faults only have one neighboring mid-point, and are assigned the distance to that single neighboring mid-point).
    - Repeat this process for each fault.

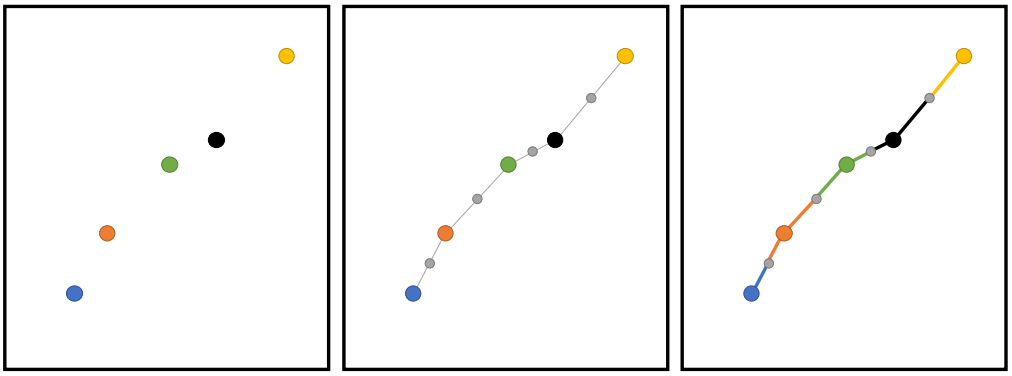


Figure 7: Visual representation of local fault length calculation. In the right-most sub-figure, the representative fault length for each point is colored .

3.1.2 *Input parameters*

The input parameters described here are used to configure the SSM. Here, their keywords and format are described. Their purpose will be described in the ´Implementation' section, whenever the parameter in question becomes relevant. All input parameters are supplied in a single .json file.

```
{
  "base_dir": <string, full path to directory containing all
input files>,
  "eq_file": <string, name of the input file>,
  "outline_file": <string, name of the input file>,
  "daterange_training": <list, [yyyymmdd (start), yyyymmdd
(end)]>,
  "daterange_testing": <list, [yyyymmdd (start), yyyymmdd
(end)]>,
  "forecast_period":<list, [yyyy (start), yyyy (end)]>,
  "dsm":{
      "type": <string, name of the model>,
       "compr_file": <string, name of the input file>,
       "fault_file": <string, name of the input file>,
  "thickness_file": <string, name of the input file>,
       "press_file": <string, name of the input file>,
       },
  "srm":{
   "type": <string, name of the model>,,
   "subtype": <string, name of the model>,
   "bval_model": <string, name of the model>
  }
}
```

3.1.3 *Implementation*

The SSM performs two major tasks:

1) Training (or calibrating) the model on historic seismicity data (i.e. obtaining the posterior distribution of model parameters).
2) Creating a forecast of future seismicity (i.e. integrating the posterior distribution of model parameters and convolving it with a forecast of the pore pressures).

For both tasks, the SSM should be able to create a forecast of seismicity, given a set of input files and model parameters. For computational efficiency and flexibility, the SSM is split into two models:

1) The Dynamic Subsurface Model (DSM). This model calculates a physical subsurface property (e.g. Coulomb Stress) from the input files and input parameters.
2) The Seismicity Rate Model (SRM). This model calculates the expected rate of events from the output of the DSM (e.g. Coulomb Stress) and additional input parameters. Again, two sub-models can be distinguished:
    a. The activity rate model, describing the rate of events as a function of a physical subsurface property and model parameters.
    b. The magnitude model, describing the relative probability of a given earthquake having a certain magnitude as a function of a physical subsurface property and model parameters.

Training

During the training phase, a Bayesian framework is applied to assign a likelihood score to each set of model parameters. Since during training, the activity rate model and the magnitude model are independent of each other (see also Section 2.2.1.2), but both models rely on the DSM, two independent posterior likelihood defined: first, $LL_{AR}(\gamma, \theta, \zeta)$: the log-likelihood function depending on a combination of DSM covariate conditioning parameters (γ), main-shock activity rate parameters (θ) and ETAS clustering model parameters ($\boldsymbol{\zeta}$), and second $LL_M(\gamma, \psi)$**,** the log-likelihood function depending on a combination of DSM covariate conditioning parameters and magnitude parameters (ψ).

For any activity rate model (a model describing the number of events per unit time, independent of magnitude), the log-likelihood is given by:

$$LL_{AR}(\gamma, \theta, \zeta) = -\int_t \int_S \lambda_X(x,t)\, dSdt + \sum_{i=1}^{n} \log\big(\lambda_X(x_i, t_i)\big),$$

where $\lambda_X(\boldsymbol{x}, t)$ is the spatio-temporal event rate density (units: number of events per unit time per unit area, e.g. $m^{-2}year^{-1}$), n is the number of observed events in the time period under consideration and $\lambda_X(\boldsymbol{x}_i, t_i)$ is the event rate density at the time-space location of an observed event.

For training we use the observation-conditioned total seismicity rate density λ_X^{obs} of Equation (31). The ETAS model functions $f_T(t)$ and $f_R(r)$ are the probability density function for temporal and spatial triggering defined as:

$$f_T(t) = \frac{p-1}{c}\left(\frac{t}{c}+1\right)^{-p},$$

$$f_R(r) = \frac{q-1}{\pi d}\left(\frac{r^2}{d}+1\right)^{-q},$$

where c, p respectively are the characteristic time and temporal power-law exponent parameters, defining the speed at which the aftershock rate decays over time. Also, d, q respectively are the characteristic area and spatial power-law exponent parameters of the ETAS model, defining the speed at which the aftershock rate decays spatially.

For any b-value model (a model describing the slope of the Gutenberg-Richter frequency magnitude distribution), the log-likelihood is given by:

$$LL_M(\gamma, \psi) = \sum_{i=1}^{n} \log[b(x_i, t_i)\log(10)] - \sum_{i=1}^{n} b(x_i, t_i)\log(10)\,(m_i - m_0), \qquad (10)$$

where $b(x_i, t_i)$ is the b-value at the space-time location of the i^{th} event and m_i is the magnitude of the i^{th} event.

Box 1 outlines the numerical procedures followed in the implementation of the SSM V5 DSM. Box 2 outlines the training of the activity rate model and the b-value model.

BOX 1 IMPLEMENTATION OF SSM V5 DYNAMIC SUBSURFACE MODEL

DSM: Calculating smoothed incremental Coulomb stress change from input files (pore pressure change, reservoir thickness, reservoir compressibility, fault geometry) and model parameters $\gamma = \{r_{max}, \sigma\}$.

1. Obtain the topographic gradient $\Gamma(\boldsymbol{x})$ and fault density $\rho(\boldsymbol{x})$ on the base grid. To do so, only the points in the fault data which have a throw/thickness ratio $r \leq r_{max}$ are considered.
   a. These points are assigned to the nearest base grid point, weighted by the fault area $A = l_{repr} t_{avg}$ to obtain the fault density grid $\rho(\boldsymbol{x})$.
   b. These same points are assigned to the nearest base grid point, weighted by offset $\times\, l_{repr} t_{avg}$ to obtain the grid $\Gamma\rho(\boldsymbol{x})$.
   c. The topographic gradient is obtained by $\Gamma(\boldsymbol{x}) = \frac{\Gamma\rho(\boldsymbol{x})}{\rho(\boldsymbol{x})}$.
2. Obtain the elastic modulus grid $H(\boldsymbol{x})$ by:
$$H(\boldsymbol{x}) = (H_s^{-1} + C_m(\boldsymbol{x}))^{-1},$$
where $H_s = 10^{+5}$ and $C_m(\boldsymbol{x})$ is the reservoir compressibility grid.
3. Calculate the scalar value $\gamma = \frac{1-2\nu}{2-2\nu}$, where $\nu = 0.2$ is the Poisson ratio. Note that this results in a scalar factor on the incremental Coulomb stress change and therefore does not impact the seismicity forecast after model training. It is included for completeness only.
4. Calculate the vertical strain grid $\epsilon_{zz}\,(\boldsymbol{x}, t) = dP(\boldsymbol{x}, t) C_m(\boldsymbol{x})$, where $dP(\boldsymbol{x}, t)$ is the pore pressure change grid.
5. Calculate the (spatio-temporal) incremental Coulomb stress change $\Delta C(\boldsymbol{x}, t)$:
$$\Delta C(\boldsymbol{x}, t) = \gamma H(\boldsymbol{x}) \epsilon_{zz}(\boldsymbol{x}, t) \Gamma(\boldsymbol{x}).$$
6. Set any negative and NaN values in $\Delta C(\boldsymbol{x}, t)$ to zero.
7. Obtain the smoothed incremental Coulomb stress change and smoothed fault density by applying a Gaussian kernel with characteristic length scale σ to the spatial $(\boldsymbol{x})$ dimensions of the $\Delta C(\boldsymbol{x}, t)$ grid and the $\rho(\boldsymbol{x})$ grid. This is implemented using scipy.ndimage.gaussian_filter with sigma $= \frac{\sigma}{dx}$, where dx is the grid spacing of the $\Delta C(\boldsymbol{x}, t)$ and $\rho(\boldsymbol{x})$ grid and mode = "constant".

BOX 2 IMPLEMENTATION OF SSM V5 MODEL TRAINING

SRM activity rate V5 training: Obtain log-likelihood for parameter vectors $\gamma = \{r_{max}, \sigma\}$*, and* $\theta = \{\theta_0, \theta_1\}$*, and* $\zeta = \{K, a\}$*, and covariate* ΔC

1. Obtain the smoothed incremental Coulomb stress change $\Delta C(\boldsymbol{x}, t)$ as described in Box 1, using parameters $\{r_{max}, \sigma\}$.
2. Obtain $\Delta C(\boldsymbol{x}, t_{start})$, $\Delta C(\boldsymbol{x}, t_{end})$, $\Delta C(\boldsymbol{x}_i, t_i)$, $\dot{\Delta C}(\boldsymbol{x}_i, t_i)$ and $\rho(\boldsymbol{x}_i)$ through spatial nearest neighbor interpolation and cubic spline temporal interpolation (scipy.interpolate.CubicSpline, bc = "natural"). $\dot{\Delta C}(\boldsymbol{x}_i, t_i)$ is the time-derivative of $\Delta C(\boldsymbol{x}_i, t_i)$ and is obtained by using the scipy.interpolate.CubicSpline functionality nu = 1.
3. Obtain $\int_t \int_S \lambda(\boldsymbol{x}, t)\, dSdt$ numerically: $A = \Delta S \sum_{\boldsymbol{x}}\ e^{\theta_0}\rho(\boldsymbol{x})(e^{\theta_1 \Delta C(\boldsymbol{x}, t_{end})} - e^{\theta_1 \Delta C(\boldsymbol{x}, t_{start})})$, with ΔS the surface area of a grid cell.
4. Obtain $K \sum_{i=1}^{n} e^{a(M_i - M_0)}$ numerically: $B = K \sum_i \left[e^{a(M_i - M_0)}\right]$.
5. Obtain $\lambda(\boldsymbol{x}_i, t_i)$ numerically: $C_i = \rho(\boldsymbol{x}_i)\theta_0 \dot{\Delta C}(\boldsymbol{x}_i, t_i) e^{\theta_0 + \theta_1 \Delta C(\boldsymbol{x}_i, t_i)}$.
6. Obtain $\sum_{j=1}^{i-1} K e^{a(M_j - M_{min})} \left(\frac{p-1}{c}\left(\frac{t_i - t_j}{c} + 1\right)^{-p}\right)\left(\frac{q-1}{\pi d}\left(\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{d} + 1\right)^{-q}\right)$ numerically:
   $D_i = \frac{p-1}{c}\frac{q-1}{\pi d} K \sum_j \left[\left(\frac{1 + \delta t_{ij}}{c}\right)^{-p}\left(\frac{1 + \delta r_{ij}}{d}\right)^{-q} e^{a(M_{ij})}\right]$,
   where $p = 1.35$, $q = 3.16$, $d = 4 \times 10^6$ m^2, $c = 0.3$ days. δt_{ij}, δr_{ij}, and M_{ij} are lower triangle matrices of inter-event time, inter-event distance and normalized event magnitude $(M - M_0)$. E.g.:
$$\delta t_{ij} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ t_{12} & \ddots & 0 & 0 \\ \vdots & t_{ij} & \ddots & 0 \\ t_{1n} & t_{2n} & \dots & 0 \end{pmatrix}.$$
7. Obtain the log-likelihood numerically: $LL_{AR}(\gamma, \theta, \zeta) = -A - B + \sum_i \log(C_i + D_i)$.

SRM b-value model V5 training: Obtain log-likelihood for parameter vectors $\gamma = \{r_{max}, \sigma\}$ *and* $\psi = \{\beta_0, C_0, n\}$.

1. Obtain the smoothed incremental Coulomb stress change $\Delta C(\boldsymbol{x}, t)$ as described in Box 1, using parameters $\{r_{max}, \sigma\}$.
2. Obtain $\Delta C(\boldsymbol{x}_i, t_i)$ through spatial nearest neighbor interpolation and cubic spline temporal interpolation (scipy.interpolate.CubicSpline, bc = "natural").
3. Obtain $b(\boldsymbol{x}_i, t_i)$ numerically: $b_i = \beta_0 + \left(\frac{\Delta C(\boldsymbol{x}_i, t_i)}{C_0}\right)^{-n}$.
4. Obtain $b_i^* = \log(10)\, b_i$.
5. Obtain the log-likelihood numerically:
   $LL_M(\gamma, \psi) = \sum_i [b_i^*] - \sum_i [b_i^* \times (m_i - m_0)]$.

N.B.: In order to calculate the log-likelihood of a set of model parameters, the forward model does not have to be evaluated. The only information about the forward model that is required is:

- The number of events in the forward model (i.e. $\int_t \int_S \lambda_X(\boldsymbol{x}, t)\, dSdt$).
- The smoothed incremental Coulomb stress rate at the space-time location of the observed events.
- The time-derivative of the smoothed incremental Coulomb stress rate at the space-time location of the observed events.

Forecasting

The training described in Box 2 is performed for a range of model parameters $\{r_{max}, \sigma, \theta_0, \theta_1, K, a\}$ and$\{r_{max}, \sigma, \beta_0, C_0, n\}$. Each parameter is discretized by supplying a range and step size, a 6D (activity rate) and 5D (b-value model) log-likelihood matrix can be calculated. These matrices can be transformed to probability mass functions (which sum to 1):

$$P(\boldsymbol{a}_i) = \frac{e^{LL(\boldsymbol{a}_i)}}{\sum_i^m e^{LL(\boldsymbol{a}_i)}},$$

where $\boldsymbol{a}_i$ is a single parameter vector (e.g. one scalar value for each parameter), $LL(\boldsymbol{a}_i)$ is the log-likelihood of that single parameter vector, and m is the number of members of the total n-dimensional grid. For very large negative values of $LL(\boldsymbol{a}_i)$ (e.g. less than -1000), typical computers will return $e^{-1000} = 0$. To avoid this, the probability mass function is calculated as:

$$P(\boldsymbol{a}_i) = \frac{e^{LL(\boldsymbol{a}_i) - \max\left(LL(\boldsymbol{a})\right)}}{\sum_i^m e^{LL(\boldsymbol{a}_i) - \max\left(LL(\boldsymbol{a})\right)}}.$$

Once the probability mass function (the posterior distribution) of model parameters is known, these can be used to calculate a forecast. The most naïve way to implement this is by calculating the forecast for each individual combination of model parameters, and weighting each model forecast by the probability mass of that combination of model parameters. The advantage of this method is that it's simple to implement and that it works for any kind of model. The downside is that it's computationally intensive to calculate for a large matrix of model parameters, which typically has $\sim 10^7$ members for the activity rate model and $\sim 10^6$ members for the b-value model. This would result in $\sim 10^{13}$ combinations of model parameters to evaluate.

Since SSM V5 is a film-rate model[3], it can be implemented in a more efficient manner.

For a given *smoothed incremental Coulomb Stress change*[4], the activity rate density is given by:

$$\lambda_X(\boldsymbol{x}, t) = \rho(\boldsymbol{x})\theta_1 \dot{\Delta C}(\boldsymbol{x}, t) e^{\theta_0 + \theta_1 \Delta C(\boldsymbol{x}, t)}.$$

The total number of events since the beginning of production:

$$\Lambda_X(\boldsymbol{x}, t_0) = \int_0^{t_0} \lambda_X(\boldsymbol{x}, t)\, dt = \rho(\boldsymbol{x}) e_0^{\theta} (e^{\theta_1 \Delta C(\boldsymbol{x}, t_0)} - 1).$$

This can be differentiated with respect to compaction to obtain an event density per unit of *smoothed incremental Coulomb Stress change*:

$$\frac{d\Lambda}{d\Delta C}(\Delta C, \boldsymbol{\theta}) = \theta_1 e^{\theta_0 + \theta_1 \Delta C},$$

where $\boldsymbol{\theta} = \theta_0, \theta_1$.

And the mean posterior:

$$\frac{d\Lambda^D}{d\Delta C}(\Delta C) = \int_{\Omega_{\boldsymbol{\theta}}} \frac{d\Lambda}{d\Delta C}(\Delta C, \boldsymbol{\theta})\, f_\theta(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

with $f_\theta(\boldsymbol{\theta})$ the joint posterior PDF of the model parameters and $\Omega_{\boldsymbol{\theta}}$ its domain.

[3] The activity rate only depends on the *current* value of the incremental Coulomb stress, and not on the history.

[4] Note that the *smoothed incremental Coulomb Stress change* depends on model parameters $\{r_{max}, \sigma\}$. The solution to this slight complication will be discussed in the description of the numerical implementation. For now, we discuss the case of a single *smoothed incremental Coulomb Stress change* realization.

To determine the distribution of the events over magnitudes, the Magnitude Frequency Distribution (MFD) needs to be taken into account. This is influenced by the b-value, which gives the slope of the MFD. The b-value, in turn, depends on the *smoothed incremental Coulomb Stress change* and the model parameters $\boldsymbol{\psi} = \{\beta_0, C_0, n\}$.

$$\overline{F_M}(m, \Delta C, \boldsymbol{\psi}) = \begin{cases} 1 & m \le m_{min} \\ \left(1 - \frac{1 - 10^{b(\Delta C, \boldsymbol{\psi})(m - M_{min})}}{1 - 10^{b(\Delta C, \boldsymbol{\psi})(M_{max} - M_{min})}}\right) & m_{min} \le m \le m_{max} \\ 0 & m > m_{max} \end{cases}.$$

And the mean posterior:

$$\overline{F_M}^D(m, \Delta C) = \int_{\Omega_{\boldsymbol{\psi}}} \overline{F_M}(m, \Delta C, \boldsymbol{\psi})\, f_{\psi}(\boldsymbol{\psi}) d\boldsymbol{\psi},$$

with $f_{\psi}(\boldsymbol{\psi})$ the joint posterior PDF of the model parameters and $\Omega_{\boldsymbol{\psi}}$ its domain.
Combining:

$$\frac{d\Lambda_M^D}{d\Delta C}(m, \Delta C) = \frac{d\Lambda^D}{d\Delta C}(\Delta C)\overline{F_M}^D(m, \Delta C).$$

And integrating with respect to ΔC:

$$\Lambda_M^D(m, \Delta C) = \int \frac{d\Lambda_M^D}{d\Delta C}(m, \Delta C)\, d\Delta C.$$

Spatio-temporal clustering according to ETAS will be randomly located in space and time. The effective increase in activity rate is therefore distributed. Assuming the background rate is smooth in time and space (smooth compared to the aftershock influence length scales), the effects of aftershocks can be represented by the enhanced productivity only. The aftershocks can be seen as a sequence of generations, each generation increasing the previous generation by a constant factor. The effective productivity factor Γ can be calculated according to Section 2.2.1.2.
Combining:

$$\frac{d\tilde{\Lambda}_M^D}{dc}(m, \Delta C) = \int_{\Omega_{\boldsymbol{\psi}}} \iint_{\Omega_{\boldsymbol{\theta\zeta}}} \Gamma(\Delta C, \boldsymbol{\zeta}, \boldsymbol{\psi}) \frac{d\Lambda}{d\Delta C}(\Delta C, \boldsymbol{\theta}) \overline{F_M}(m, \Delta C, \boldsymbol{\psi}) f_{\boldsymbol{\theta\zeta}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) f_{\boldsymbol{\psi}}(\boldsymbol{\psi}) d\boldsymbol{\theta} d\boldsymbol{\zeta} d\boldsymbol{\psi}.$$

And finally, for a given *smoothed increment Coulomb stress* interval $\Delta C_0 \to \Delta C_1$:

$$\tilde{\Lambda}_{M\ \Delta C_1}^D(m, \Delta C_1) - \tilde{\Lambda}_M^D(m, \Delta C_0) = \int_{\Delta C_0} \int_{\Omega_{\boldsymbol{\psi}}} \iint_{\Omega_{\boldsymbol{\theta\zeta}}} \Gamma(\Delta C, \boldsymbol{\zeta}, \boldsymbol{\psi}) \frac{d\Lambda}{d\Delta C}(\Delta C, \boldsymbol{\theta}) \overline{F_M}(m, \Delta C, \boldsymbol{\psi}) f_{\boldsymbol{\theta\zeta}}(\boldsymbol{\theta}, \boldsymbol{\zeta}) f_{\boldsymbol{\psi}}(\boldsymbol{\psi}) d\boldsymbol{\theta} d\boldsymbol{\zeta} d\boldsymbol{\psi} d\Delta C.$$

For a given posterior distribution of activity rate parameters and b-value model parameters (i.e. $f_{\boldsymbol{\theta\zeta}}(\boldsymbol{\theta}, \boldsymbol{\zeta})$ and $f_{\boldsymbol{\psi}}(\boldsymbol{\psi})$) this expression can be tabulated for ranges of values of m and ΔC. Such a lookup table can then be convolved with a spatio-temporal *smoothed incremental Coulomb Stress change* realization to obtain a mean posterior forecast of seismicity. This should be done for all logic tree assumptions (prior distributions) of M_{max} and all *smoothed incremental Coulomb Stress change* realizations (which depend on DSM parameters $\{r_{max}, \sigma\}$).

Note that this expression, with effective aftershock productivity, is to be used for forecasting only, not for calibration. Calibration requires the full ETAS formulation since the likelihood expression depends on the observed events.

In Box 3, the implementation of the forecasting is described. This results in a forecast/hindcast of seismicity for each year that pore pressures are provided. If the pore pressures contain a forecast into the future, the resulting seismicity forecast is also a forecast into the future. This seismicity forecast is a 5D matrix:

- For each value of M_{max} (one dimensional; values from logic tree).
- For each time step (one dimensional; typical time step is 1 year).
- For every spatial grid cell (two dimensional; RDx, RDy).
- A Complementary Cumulative Density Function (CCDF, or survival function) over the magnitude dimension (one dimensional).

Example:
Pick a value of $M_{max} = 6.0$. Pick a time step (2010), pick a set of spatial coordinates (248000.0; 586500.0). You now get a one-dimensional array, where each value gives the expected number of events of magnitude equal or higher than m, assuming an M_{max} of 6.0, in 2010 in a 500x500m grid cell centered around 248000.0; 586500.0.

m	1.45	1.50	1.55	1.60	1.65	1.70	1.75	…	6.00
Number of events $M \geq m$	0.00306	0.00274	0.00245	0.00220	0.00197	0.00177	0.00158	…	0.000

BOX 3 IMPLEMENTATION OF SSM V5 FORECASTING

1. Discretize and transform the likelihood function $LL_{AR}(r_{max}, \sigma, \theta_0, \theta_1, K, a)$ into a probability mass function through: $P(\boldsymbol{a}_i) = e^{LL(\boldsymbol{a}_i) - \max(LL(\boldsymbol{a}))} / \sum_i^m e^{LL(\boldsymbol{a}_i) - \max(LL(\boldsymbol{a}))}$.
2. Sum $P(\boldsymbol{a}_i)$ over the axes of SRM parameters $\{\theta_0, \theta_1, K, a\}$ to obtain $P(r_{max}, \sigma)$.
3. For each member of $\{r_{max}, \sigma\}$, obtain the annual event count in spatio-temporal-magnitude bin for each value of M_{max}: $\lambda(\boldsymbol{x}, t, m, M_{max})$:
   a. Obtain $\Delta C(\boldsymbol{x}, t)$ and $\rho(\boldsymbol{x})$ and as described in Box 1.
   b. Obtain $P(\theta_0, \theta_1, K, a \mid r_{max}, \sigma)$ and $P(\beta_0, C_0, n \mid r_{max}, \sigma)$ by repeating Step 1 for the selected vector $\{r_{max}, \sigma\}$.
   c. For each value of M_{max}:
      i. Set up a representative vector of ΔC_{bin} values (bin edges) and associated ΔC_{cen} (bin centers). Set up a representative vector of magnitude value $m_{discrete}$.
      ii. Create a table of b-values $b_{val}(\Delta C_{cen}, \beta_0, C_0, n)$.
      iii. Set up a representative table of b-values $b_{discrete}$ based on the b-values in b_{val}.
      iv. Convolve the table $b_{val}(\Delta C_{cen}, \beta_0, C_0, n)$ with $P(\beta_0, C_0, n)$ to obtain a probability mass function (PMF) of b-value as a function of ΔC_{cen} (i.e. one b-value PMF defined on $b_{discrete}$ for each value of ΔC_{cen}): $b_{pmf}(b_{discrete}, \Delta C_{cen})$.
      v. Obtain the Complementary Cumulative Density Function (CCDF, or survival function) $F_M(b_{discrete}, m_{discrete})$ and the numerical derivative $f_M(b_{discrete}, m_{discrete})$. E.g.
      $$f_M(1.0, 2.3) = F_M(1.0, 2.25) - F_M(1.0, 2.35).$$
      vi. Obtain aftershock productivity factor $\xi(K, a, b_{discrete}) = \sum_m K e^{a(m - M_{min})} \times f_M(m)$ for each combination of $\{K, a, b_{discrete}\}$.
      vii. Obtain effective aftershock productivity $\Gamma(K, a, b_{discrete}) = \min\left(\frac{1}{1 - (K, a, b_{discrete})}; 10\right)$ (Capped at 10 to prevent numerical problems).
      viii. Obtain $\frac{d\Lambda}{d\Delta C}(\Delta C_{cen}, \boldsymbol{\theta}) = \theta_1 e^{\theta_0 + \theta_1 \Delta C_{cen}}$ each combination of $\{\theta_0, \theta_1, \Delta C_{cen}\}$.
      ix. Obtain lookup table $\tilde{\tilde{\Lambda}}_M^D(m_{discrete}, \Delta C_{bin}) =$
      $$\left[\sum_0^{\Delta C}\left[d\Delta C \times \sum_{(b_{discrete}, \theta_0, \theta_1, K, a)} b_{pmf}(b_{discrete}, \Delta C_{cen}) \frac{d\Lambda}{d\Delta C}(\Delta C_{cen}, \theta_0, \theta_1) \times R(K, a, b_{discrete}) F_M(b_{discrete}, m_{discrete}) P(\theta_0, \theta_1, K, a)\right]\right].$$
      Note that this is effectively a mid-point integration scheme (by obtaining the values up to the bin edge by using the bin centers).
      x. Obtain the cumulative event density grid $\Lambda_{\text{scaled}}(\boldsymbol{x}, t, m)$ by performing a linear interpolation of the lookup table obtained in Step 3.a.ix. for each grid point of $\Delta C(\boldsymbol{x}, t)$ (obtained in Step 3.a.). That is: for each value of ΔC, obtain a distribution of events over $m_{discrete}$.
      xi. Obtain = $\Lambda(\boldsymbol{x}, t, m) = dx \times dy \times \rho(\boldsymbol{x}) \times \Lambda_{\text{scaled}}(\boldsymbol{x}, t, m)$.
      xii. Finally, obtain $\lambda(\boldsymbol{x}, t, m)$ by numerical differentiation of $\Lambda(\boldsymbol{x}, t, m)$. E.g:
      $$\Lambda(\boldsymbol{x}, 2010, m) = \lambda(\boldsymbol{x}, 2011.01.01, m) - \lambda(\boldsymbol{x}, 2010.01.01, m).$$
4. Repeat Step 3 for each member of $\{r_{max}, \sigma\}$ and weight the resulting forecast by $P(r_{max}, \sigma)$ to obtain $\lambda^D(\boldsymbol{x}, t, m, M_{max})$.

Rupture model

The 5D seismicity forecast described above gives a *hypocenter* distribution. However, earthquakes occur on planes (or lines in map view). To capture this, the SSM V5 contains a rupture model. Note that this model is not trained on data, but rather informed by geological knowledge of the Groningen gas field and geophysical knowledge of earthquake ruptures in general. The rupture model for Groningen is described in Bourne & Oates (2018). In the TNO Model Chain, it is implemented in two steps:

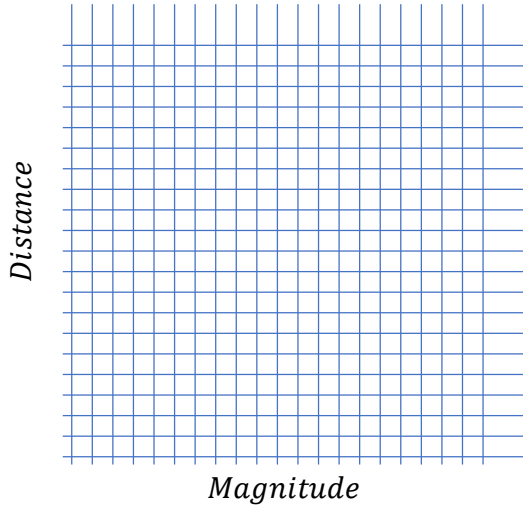
1) Calculation of a lookup table, which gives a probability mass function of distance to a rupture plane, given a distance to a hypocenter, a magnitude and an azimuth (angle with respect to North).
2) Convolution of this lookup table with the hypocenter distribution and a spatial 2D grid of virtual observation points at the earth's surface.

The final result of this twostep approach (implementation discussed in Box 4 and Box 5) is a 5D matrix:

- For each value of M_{max} (one dimensional; values from logic tree).
- For each time step (one dimensional; typical time step is 1 year).
- For every spatial observation point (one dimensional; unique combinations of {RDx, RDy}).
- For every magnitude bin (one dimensional; typically in bins of width 0.1, centered around 1.5, 1.6, 1.7, …).
- The number of events in a distance bin (one dimensional).

Note that this matrix no longer uses a CCDF in the magnitude dimension. A way to visualize this matrix is as follows:

1) Pick a value for M_{max} from the logic tree.
2) On January first of a year (e.g. 2020), go stand on a point in the Groningen earthquake area.
3) For all of 2020, whenever an earthquake occurs, make a mark in the box corresponding to its distance (to the rupture plane) and its magnitude:



If you repeat this exercise for all logic tree values of M_{max}, for every timestep and for all observation points, you obtain the 5D matrix that is produced by the rupture model. The only difference is that the matrix produced by the rupture model contains *expectation values*, rather than observations.

The output of the rupture model is considered to be the final output of the SSM, which serves as input for the GMM.

BOX 4 IMPLEMENTATION OF SSM V5 RUPTURE MODEL LOOKUP TABLE

Note: The rupture model as described in Bourne & Oates (2018) is based on faults having a mean strike, with a certain variance. Here, we describe the rupture model based on 'relative angle', which is geometrically identical, but numerically more advantageous in the numerical integration scheme (see Figure 4).

1. Define a vector of magnitudes $\boldsymbol{m}$ for which to perform the calculation (e.g.: 1.4, 1.5,…,7.1).
2. Define a vector of rupture plane lengths $\boldsymbol{l}$ (e.g. 1m, 5m, 10m, 50m, 100m,….,100km).
3. For each magnitude, calculate the 1D PMF of the rupture length $P(l|m)$, based on the model parameters $\{c, d, L, a_1, b_1, a_2, b_2, \sigma_L\}$.
4. Define a vector of normalized positions of the hypocenter on the rupture plane $\boldsymbol{n}$ [0.0,…,1.0] and its associated PMF $P(n)$ based on a uniform distribution.
5. For each magnitude, calculate the 2D PMF of the rupture length and normalized hypocenter position on the rupture plane (sums to 1.0 for each magnitude) $P(l, n|m)$.
6. Define a vector of relative angles $\boldsymbol{\phi}$, and a vector of hypocentral distances $\boldsymbol{r}^{hyp}$.
7. Calculate equivalent epicentral distances based on a constant rupture depth d_{rup}:

$$\boldsymbol{r}^{epi} = \sqrt{(\boldsymbol{r}^{hyp})^2 - d_{rup}^2}.$$

8. For each relative angle ϕ_i:
   a. Calculate normalized epicentral distances $\tilde{\boldsymbol{r}}_{jk}^{epi} = \frac{\boldsymbol{r}_j^{epi}}{\boldsymbol{l}_k}$.
   b. Calculate the crossline component $\tilde{\boldsymbol{r}}_{jk}^{epi}{}_{\perp} = \sin(\phi_i)\, \tilde{\boldsymbol{r}}_{jk}^{epi}$.
   c. Calculate the clipped inline component $\tilde{\boldsymbol{r}}_{jkl}^{epi}{}_{\parallel} = \max(\cos(\phi_i)\, \tilde{\boldsymbol{r}}_{jk}^{epi} - \boldsymbol{n}_l; 0.0)$.
   d. Calculate the surface projection of the relative rupture plane distance:

$$(\boldsymbol{r}_*^{rup})_{jkl} = \sqrt{\boldsymbol{l}_k^2 \times \left(\tilde{\boldsymbol{r}}_{jkl}^{epi\,2}{}_{\parallel} + \tilde{\boldsymbol{r}}_{jkl}^{epi\,2}{}_{\perp}\right) + d^2}.$$

   e. Define a vector of rupture distances $\boldsymbol{r}^{rup}$. The rupture distances calculated at Step 8.d are then assigned to the members of $\boldsymbol{r}^{rup}$ through linear interpolation in log-space, multiplied by the PMF calculated in Step 5. This results in a 3D array, in which each magnitude m_m and hypocentral distance r_j has an associated PMF of rupture distances $\boldsymbol{r}^{rup}$.
9. Repeat Step 8 for all members of $\boldsymbol{\phi}$ to obtain a 4D lookup table where each combination of azimuth ϕ_i, magnitude m_m and hypocentral distance r_j has an associated PMF of rupture distances $\boldsymbol{r}^{rup}$.
10. Finally, apply a Gaussian kernel over the azimuth dimension of the lookup table to account for the variability in rupture strike.

Note: The lookup table is 2-fold symmetric (e.g. the distribution of rupture distances of an event at 30° relative angle is identical to those at 150°, 210° and 330°). This means that the lookup table only needs to be calculated over a range of 90 degrees.

BOX 5 IMPLEMENTATION OF SSM V5 RUPTURE MODEL INTEGRATION

1. Calculate the source distribution $\lambda^D(\boldsymbol{x}, t, m, M_{max})$ as described in Box 3, and the lookup table $L(r^{hyp}, \phi, m, r^{rup})$ as described in Box 4.
2. Transform the source distribution into $\lambda_c^D(\boldsymbol{x}, t, m, M_{max})$ by numerically differentiating along the magnitude dimension of λ^D. λ_c^D therefore simple contains the expectation number of events in each spatio-temporal-magnitude bin, for each prior value of M_{max}.
3. Define a vector of observation points $\boldsymbol{q}$.
   For each member $q_i = x_i, y_i$:
   a. For each location $\boldsymbol{x}$ in the source distribution λ_c^D:
      i. Determine the distance r^{hyp} and azimuth ϕ.
      ii. Select a sub-table $L^*(r^{hyp}, m, r^{rup})$ from L, based on simple 0th order interpolation of the azimuth (i.e. select the sub-table calculated for the azimuth closest to ϕ.
      iii. Obtain $L^{**}(m, r^{rup})$ based on linear interpolation of L^* in log-space in the r^{hyp} dimension.
      iv. Calculate $f(q_i, \boldsymbol{x}, t, M_{max}, m, r^{rup}) = L^{**}(m, r^{rup})\lambda_c^D(\boldsymbol{x}, t, m, M_{max})$.
   b. Calculate $f(q_i, t, M_{max}, m, r^{rup}) = \sum_{\boldsymbol{x}} f(q_i, \boldsymbol{x}, t, M_{max}, m, r^{rup})$.
4. Repeat Step 3 for all observation points $\boldsymbol{q}$ to obtain $f(q, t, M_{max}, m, r^{rup})$.

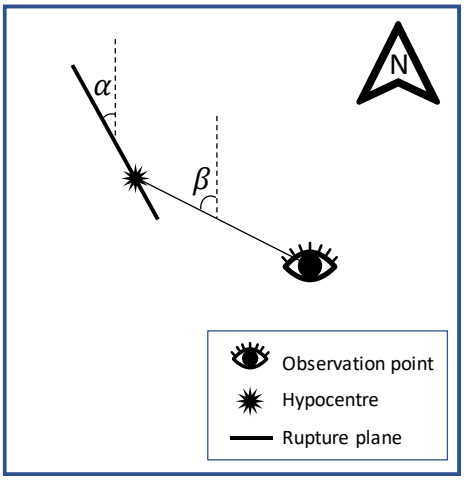


Figure 8: Relation between the mean strike of rupture planes α and relative angle ($\phi = \beta - \alpha$). If the relative angle is 0°, rupture planes have a mean strike parallel to the line connecting the observation point and the hypocenter. Note that variability around the mean strike is geometrically identical to variability in relative angle.

3.2 Hazard Model (V5)

The Hazard Model is based entirely on the GMM V5 (Geometric mean[5]), as described in Bommer et al. (2018). It is implemented in two distinct steps. The first step (preparation) is *independent* of the gas production and can therefore be pre-calculated. It calculates a lookup table of exceedance probabilities of ground motions. The second step (integration) combines this lookup table with the output of the SSM, which is *dependent* on the gas production.

3.2.1 *First step: Preparation*

3.2.1.1 Input files

Model parameters for the hazard preparation calculation are provided in input yaml files[6].

V5_GMM_Medians_NSB.yaml
This file contains the model parameters $m_0, m_1, m_2, m_3, m_4, m_5$ and $r_0, r_1, r_2, r_3, r_4, r_5$ for every logic tree branch GMM_{med} and spectral period T (Figure 5). These parameters are used to compute the source and path term, which together define the ground motion at the reference level, base of the North Sea Group (NS_B).

GmpeSurfaceAmplificationModel_20170826_V5.yaml
This file contains the model parameters:
$a_0, a_1, b_0, b_1, f_2, f_3, \phi_{S2S,1}, \phi_{S2S,2}, Sa_{high}, Sa_{low}, A_{min}, A_{max}$ for every site response region $\mathcal{s}$ and every spectral period T. These parameters are used to compute the probability of exceedance at the surface per site response region and spectral period, given a spectral acceleration at the reference level (NS_B).

V5_GMM_Sigmas_NSB_Tau.yaml
This file contains model parameter τ for every spectral period T and GMM_{med} logic tree branch.

V5_GMM_Sigmas_NSB_PhiSS.yaml
This file contains the model parameter ϕ_{ss} for every spectral period T and Φ_{ss} logic tree branch.

3.2.1.2 Implementation

For each site response region $\mathcal{s}$, the probability of exceeding acceleration A at the earth's surface, for spectral period T, due to an earthquake of magnitude m occurring at distance r^{rup}, for a selection of logic tree branch members GMM_{med}, Φ_{ss} , (i.e. $P(A_0^{sur} > A^{sur} | m, r^{rup}, T, GMM_{med}, \Phi_{ss}, \mathcal{s})$ can be calculated. This is done based on the model parameters in the input files, and based on discretized values for A, m, r^{rup}. The numerical implementation is described in Box 6.

[5] GMM V5 contains the possibility for calculate ground motions according to the Geometric Mean or Arbitrary Component. By convention, the Geometric Mean is used for hazard calculations, while the Arbitrary Component is used for risk calculations.
[6] YAML is a commonly used file format for configuration and input files

BOX 6 IMPLEMENTATION OF HAZARD MODEL PREPARATION

1. Define vectors for spectral acceleration bin edges A_{edge}, and associated bin centers in log-space A. These values will be used for both surface level and reference level.
2. Define vectors for magnitude m, and distance r^{rup}.
3. For every site response region $\mathcal{s}$:
   a. For every spectral period T_j:
      i. Calculate the site response probability of exceedance, given a spectral acceleration at reference level (NS_B):
         1. Calculate $\log\left[(AF_{Sa})^{median}(r^{rup}, m, A)\right]$ (for all unique combinations of r^{rup}, m, A):
            a. $M_{ref} = M_1 - \frac{\log(r^{rup}) - \log(3)}{\log(60) - \log(3)}(M_1 - M_2)$ where M_1 and M_2 are model parameters depending on $\mathcal{s}$ and T_j.
            b. $f_1 = (a_0 + a_1 \log(r^{rup})) + (b_0 + b_1 \log(r^{rup}))(\min(m, M_{ref}) - M_{ref})$ where a_0, a_1, b_0, and b_1 are model parameters depending on $\mathcal{s}$ and T_j.
            c. $\log\left[(AF_{Sa})^{median}\right] = f_1 + f_2 \log\left(\frac{A + f_3}{f_3}\right)$ where f_2 and f_3 are model parameter depending on $\mathcal{s}$ and T_j.
            d. Set all values of $\log\left[(AF_{Sa})^{median}\right]$ lower than $\log(A_{min})$ to $\log(A_{min})$, and all values $\log\left[(AF_{Sa})^{median}\right]$ higher than $\log(A_{max})$ to $\log(A_{max})$, where A_{min} and A_{max} are model parameter depending on $\mathcal{s}$ and T_j.
         2. Calculate $(AF_{Sa})^{\sigma}(A)$:
            a. $(AF_{Sa})^{\sigma}(A) = \phi_{S2S,1} + (\phi_{S2S,2} - \phi_{S2S,1})\left[\frac{\log(A) - \log(Sa_{low})}{\log(Sa_{high}) - \log(Sa_{low})}\right]$, where $\phi_{S2S,1}, \phi_{S2S,2}, Sa_{high}$ and Sa_{low} are model parameter depending on $\mathcal{s}$ and T_j.
            b. Set all values of $(AF_{Sa})^{\sigma}$ lower than Sa_{low} to Sa_{low}, and all values $(AF_{Sa})^{\sigma}$ higher than Sa_{high} to Sa_{high}.
         3. For every unique combination of of A^{ref}, r^{rup}, m, calculate:
            $$P\left(A_0^{sur} > A^{sur} \middle| A_m^{ref}, m_n, r_o^{rup}, T_j, \mathcal{s}\right) = \frac{1}{2} - \frac{1}{2}\operatorname{erf}\left[\frac{\log\left(A_{edge}^{sur}\right) - (\log(AF_{Sa})^{median} + \log(A^{ref}))}{\sqrt{2}(AF_{Sa})^{\sigma}}\right].$$
      ii. For every logic tree branch $(GMM_{med})_k$:
         1. For every member of m, obtain $g_{source} =$:
            $$\begin{cases} m_0 + m_1(m - 4.7) + m_2(m - 4.7)^2 & \text{if } m \le 4.7 \\ m_0 + m_3(m - 4.7) & \text{if } 4.7 < m \le 5.45, \\ m_0 + m_3(5.45 - 4.7) + m_4(m - 5.45) + m_5(m - 5.45)^2 & m > 5.45 \end{cases}$$
            where m_0, m_1, m_2, m_3, m_4 and m_5 are model parameter depending on $(GMM_{med})_k$ and spectral period T_j.
         2. For every unique combination of m, r^{rup}, obtain $g_{path} =$:
            $$\begin{cases} (r_0 + r_1 m)\log\left(\frac{r^{rup}}{3}\right) & \text{if } r^{rup} \le 7 \\ (r_0 + r_1 m)\log\left(\frac{7}{3}\right) + (r_2 + r_3 m)\log\left(\frac{r^{rup}}{7}\right) & \text{if } 7 < r^{rup} \le 12, \\ (r_0 + r_1 m)\log\left(\frac{7}{3}\right) + (r_2 + r_3 m)\log\left(\frac{12}{7}\right) + (r_4 + r_5 m)\log\left(\frac{r^{rup}}{12}\right) & r^{rup} > 12 \end{cases}$$
            where r_0, r_1, r_2, r_3, r_4 and r_5 are model parameter depending on $(GMM_{med})_k$ and spectral period T_j.

BOX 6 IMPLEMENTATION OF HAZARD MODEL PREPARATION (CONT.)

3. $\log(Y(r^{rup}, m)) = g_{path}(r^{rup}, m) + g_{source}(m) + \log\left(\frac{0.01}{9.807}\right)$.
4. For every logic tree branch $(\Phi_{ss})_l$:
   a. Calculate $\sigma_Y = \sqrt{\tau^2 + \phi_{ss}^2}$ where τ depends on $(GMM_{med})_k$ and T_j and ϕ_{ss} depends on $(\Phi_{ss})_l$ and T_j.
   b. Calculate the probability of exceeding A at reference depth (NS_B), for every unique combination of m, r^{rup}:
   $P(A_0^{ref} > A^{ref} | m_n, r_o^{rup}, T_j, (GMM_{med})_k, (\Phi_{ss})_l, \mathcal{s}) =$
   $$\frac{1}{2} - \frac{1}{2}\mathrm{erf}\left[\frac{\log\left(A_{edge}^{ref}\right) - \log\left(Y(r_o^{rup}, m_i)\right)}{\sqrt{2}\sigma_Y}\right].$$
   c. Numerically differentiate $P(A_0^{ref} > A^{ref} | m_n, r_o^{rup}, T_j, (GMM_{med})_k, (\Phi_{ss})_l, \mathcal{s})$ along the A dimension to obtain the PMF $P(A_0^{ref} = A^{ref} | m_n, r_o^{rup}, T_j, (GMM_{med})_k, (\Phi_{ss})_l, \mathcal{s})$.
   d. Obtain $P\left(A_0^{sur} > A^{sur} \middle| m_n, r_o^{rup}, T_j, (GMM_{med})_k, (\Phi_{ss})_l, \mathcal{s}\right) =$ $P\left(A_0^{sur} > A^{sur} \middle| A_m^{ref}, m_n, r_o^{rup}, T_j, \mathcal{s}\right) \times P(A_0^{ref} = A^{ref} | m_n, r_o^{rup}, T_j, (GMM_{med})_k, (\Phi_{ss})_l, \mathcal{s})$ (i.e. multiplication of matrix obtained in Step 3.a.i.3. with the matrix obtained in Step 3.a.ii.4.b.).
   e. Repeat for all logic tree branches $(\Phi_{ss})_l$.
5. Repeat for all logic tree branches $(GMM_{med})_k$.

iii. Repeat for all spectral periods T_j.

b. Repeat for all site response regions $\mathcal{s}$.

4. Save the resulting matrix $P(A_0^{sur} > A^{sur} | m, r^{rup}, T, GMM_{med}, \Phi_{ss}, \mathcal{s})$.

3.2.2 *Second step: Integration*

3.2.2.1 Input files

Lookup table exceedance values of ground motions
This lookup table (produced in the GMM preparation code) contains exceedance probabilities spectral accelerations A, for all site response regions $\mathcal{s}$, spectral periods T, GMM_{med} and Φ_{ss} logic tree branches, due to a hypothetical earthquake at rupture distance r^{rup} and of magnitude m:
$P(A_0 > A|m, r^{rup}, T, GMM_{med}, \Phi_{ss}, \mathcal{s})$.

Output rupture model (SSM)
A probability mass function (PMF) matrix of earthquake expectation values per evaluation point q, year t, maximum magnitude logic tree branch M_{max}, magnitude m and rupture distance r^{rup}:
$p(q, t, M_{max}, m, r^{rup})$.

Zonation shape files
A shape file containing the polygons of all site response regions.

3.2.2.2 Input parameters

```
{
  "logictree": {
    "Mmax": {
            "4.0": 0.08625,
          "4.5": 0.4,
          "5.0": 0.24375,
          "5.5": 0.1125,
          "6.0": 0.07875,
          "6.5": 0.0525,
          "7.0": 0.02625},
    "GMMMedian": {
    "Upper": 0.3,
    "CentralUpper": 0.3,
    "CentralLower": 0.3,
    "Lower": 0.1
    },
    "GMMPhiSS": {
    "phi_ss_high": 0.5,
    "phi_ss_low": 0.5
    } },
  "basedir": <string, full path to directory containing all
input files>,
  "gmmsiteresponseregionsfile": <string, name of the input
file>,
  "ssmmrdistributionsfile": <string, name of the input file>,
  "gmmsurfacepoefile": <string, name of the input file>,
  "hazard_outputfile": <string, name of the output file>,
  "returnperiods": [
    475.0,
    2475.0
  ]
}
```

3.2.3 *Implementation*

The calculations in the hazard integrator are carried out per site response region $\mathcal{s}$. To do this, we first define which evaluation points $q_{|\mathcal{s}_i}$ lay within a site response region $\mathcal{s}_i$. Points within the polygons are considered as well as the ones surrounding the polygon (Figure 9), so that bilinear interpolation can be applied when visualizing hazard maps.

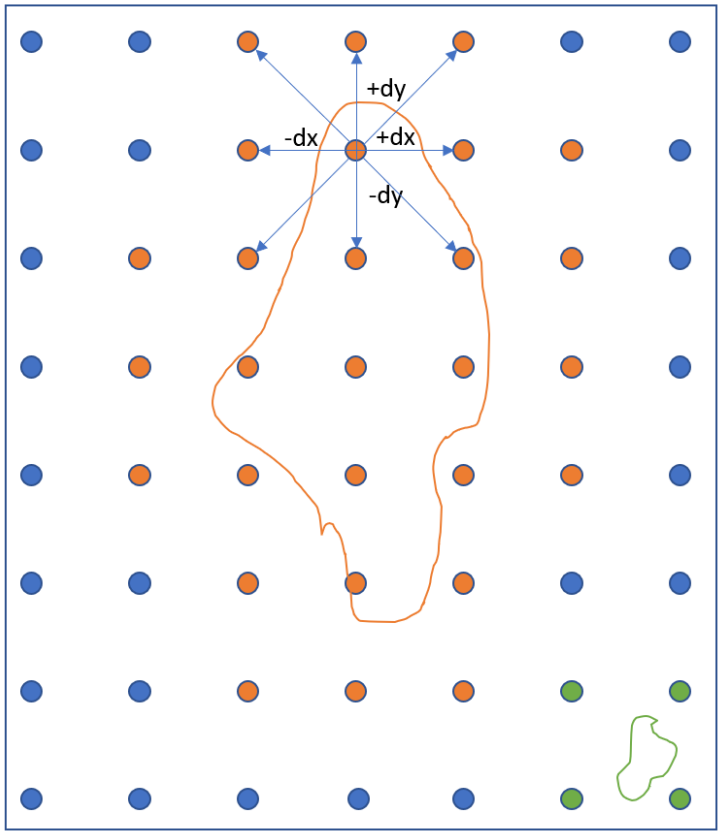


Figure 9: Visual representation of how evaluation points per site response zone ($\mathcal{s}$) are defined. The orange and green polygons define two hypothetical site response regions and all orange and green points are the evaluation points for which hazard is computed according to amplification of that site response region, respectively.

The exceedance probabilities of ground motions $P_{exc}(A_0 > A \mid m, r^{rup}, T)$ over all ground motion logic tree branches (GMM_{med} and Φ_{ss}) and the integrated occurrence probabilities of earthquakes $p_{int}(q, t, m, r^{rup})$ over the M_{max} logic tree branches are computed by matrix multiplication with the respective branch weights $P(GMM_{med})$, $P(\phi_{ss})$ and $P(M_{\max})$.

The annual exceedance probabilities of the ground motions due to the actual forecasted earthquakes are also computed by matrix multiplication of P_{exc} with p_{int}.

To compute the hazard values A_0 (ground motions) corresponding to the pre-defined return periods $1/P(A_0 > A)$, linear interpolation in log space is used to obtain the ground motion associated with that exceedance probability. These can then be plotted as maps. The numerical implementation is described in Box 7.

BOX 7 IMPLEMENTATION OF HAZARD INTEGRATION

1. Import $p(q, t, M_{max}, m, r)$ and associated grids $q, t, M_{max}, m, r^{rup}$.
2. For every site response region $\mathcal{s}$:
   a. Define the evaluation grid points of that site response region $q_{\mathcal{s}}$.
   b. Extract the earthquake PMF only for those evaluation points $p(q_{\mathcal{s}}, t, M_{max}, m, r^{rup})$.
   c. Compute the mean earthquake PMF over the M_{max} logic tree branches: $p_{int}(q_{\mathcal{s}}, t, m, r^{rup}) = \sum_l P(M_{\max_l}) p(q_{\mathcal{s}}, t, M_{max_l}, m, r^{rup})$.
   d. Import $P(A_0 > A | m, r^{rup}, T, GMM_{med}, \Phi_{ss})$.
   e. Compute the probability of exceeding ground motions over GMM_{med} and ϕ_{ss} logic tree branches:
   $$P_{exc}(A_0 > A \mid m, r^{rup}, T) = \sum_j \sum_k P((GMM_{med})_j) P((\phi_{ss})_k) P(A_0 > A \mid m, r^{rup}, T, (GMM_{med})_j, (\phi_{ss})_k).$$
   f. Compute the probability of exceeding ground motions for the forecasted earthquake occurrence:
   $$P(A_0 > A \mid q_{\mathcal{s}}, t, T) = \sum_m \sum_n P_{exc}(A_0 > A \mid m_m, {r^{rup}}_n, T)\, p_{int}(q_{\mathcal{s}}, t, m_m, {r^{rup}}_n).$$
   g. Sample hazard values $A_0(q_{\mathcal{s}}, t, T)$ at return periods $1/P(A_0 > A)$, defined in the input parameter file by linear interpolation in log space.
3. Repeat for every site response region $\mathcal{s}$ and save the resulting matrices: $A_0(q_{\mathcal{s}}, t, T, \mathcal{s})$ of all site response regions $\mathcal{s}$.

3.3 Risk Model (V5)

The Risk Model is a combination of GMM V5 (Arbitrary Component[7]) and the Fragility & Consequence Model V5 (Crowley & Pinho, 2017; Crowley et al., 2017). By combining these models, the output of the SSM can directly by convolved with a risk-lookup table to obtain the Local Personal Risk (LPR) for each location in the field, for each building typology. This result can be used as-is (i.e. what is the risk of type of building at a given location, assuming that type of building is present) or it can be convolved with an exposure database to obtain an LPR value per building.

3.3.1 *First step: Preparation*

3.3.1.1 Input files

Model parameters for the hazard preparation calculation are provided in input yaml files.

GmpeSurfaceZonationVs30_20170826_V5.yaml
This file contains model parameter $V_{S,30}$ for every site response region $\mathcal{s}$. $V_{S,30}$ is the shear velocity at 30 m depth per site response region $\mathcal{s}$, which is used to compute the duration of the ground motion.

Im2im_V5.yaml
This file contains the correlation matrices for spectral periods and duration and for spectral periods to spectral periods.

fragilityV5.yaml
This file contains the model parameters $T_1, T_2, b_0, b_1, b_2, b_3, \beta_s, DL_u, \beta_{ch}, \overline{PGA_{ch}}$ where u are the limit (damage and collapse) states $\{DS1, DS2, DS3,\ CS1, CS2, CS3\}$. These parameters are defined per fragility branch l_{frag} and typology $\mathcal{t}$.

V5_GMM_Medians_NSB.yaml
This file contains the model parameters $m_0, m_1, m_2, m_3, m_4, m_5$ and $r_0, r_1, r_2, r_3, r_4, r_5$ for every logic tree branch GMM_{med} and spectral period T. These parameters are used in the same way as in the hazard preparation.

GmpeSurfaceAmplificationModel_20170826_V5.yaml
This file contains the model parameters:
$a_0, a_1, b_0, b_1, f_2, f_3, \phi_{S2S,1}, \phi_{S2S,2}, Sa_{high}, Sa_{low}, A_{min}, A_{max}$ for every site response region $\mathcal{s}$ and every spectral period T. These parameters are used in the same way as in the hazard preparation. In addition this file also contains parameters M_1, M_2 for every site response region $\mathcal{s}$ and every spectral period T.

V5_GMM_Sigmas_NSB_Tau.yaml
This file contains model parameter τ for every spectral period T and GMM_{med} logic tree branch.

[7] GMM V5 contains the possibility for calculate ground motions according to the Geometric Mean or Arbitrary Component. By convention, the Geometric Mean is used for hazard calculations, while the Arbitrary Component is used for risk calculations.

V5_GMM_Sigmas_NSB_PhiSS.yaml
This file contains the model parameter ϕ_{ss} for every spectral period T and Φ_{ss} logic tree branch.

3.3.1.2 Implementation

Independent of the event density (forecast), and therefore independent of gas production, a lookup table can be created. This lookup table contains a complementary cumulative mass function (CCMF) which contains exceedance probabilities of each limit state u, for every typology t, for every site response region s , every combined ground motion logic tree branch $GMM_{med}\phi_{ss}$, every fragility logic tree branch l_{frag}, due to a hypothetical earthquake at distances r^{rup} and magnitudes m ($P(u_0 > u| t, s, m, r^{rup}, GMM_{med}\phi_{ss}, l_{frag})$).

The lookup table also contains the probability of dying due to chimney collapse $P_{d_{chimney}}$ for all consequence logic tree branches l_{cons}, for every typology t, for every site response region s, every combined ground motion logic tree branch $GMM_{med}\phi_{ss}$, every fragility logic tree branch l_{frag} and fatality logic tree branch l_{fat}, due to a hypothetical earthquake at distances r^{rup} and magnitudes m.

The lookup table is created based on the model parameters in the input files, and based on discretized values for m , r^{rup}. The numerical implementation is described in Box 8.

In order to rapidly convolve the risk output (LPR for each location in the field, for each building typology) with the building exposure database, a preparatory calculation is performed to obtain the contribution of each surrounding grid point to each building in the exposure database (see Figure 11 and Figure 12).

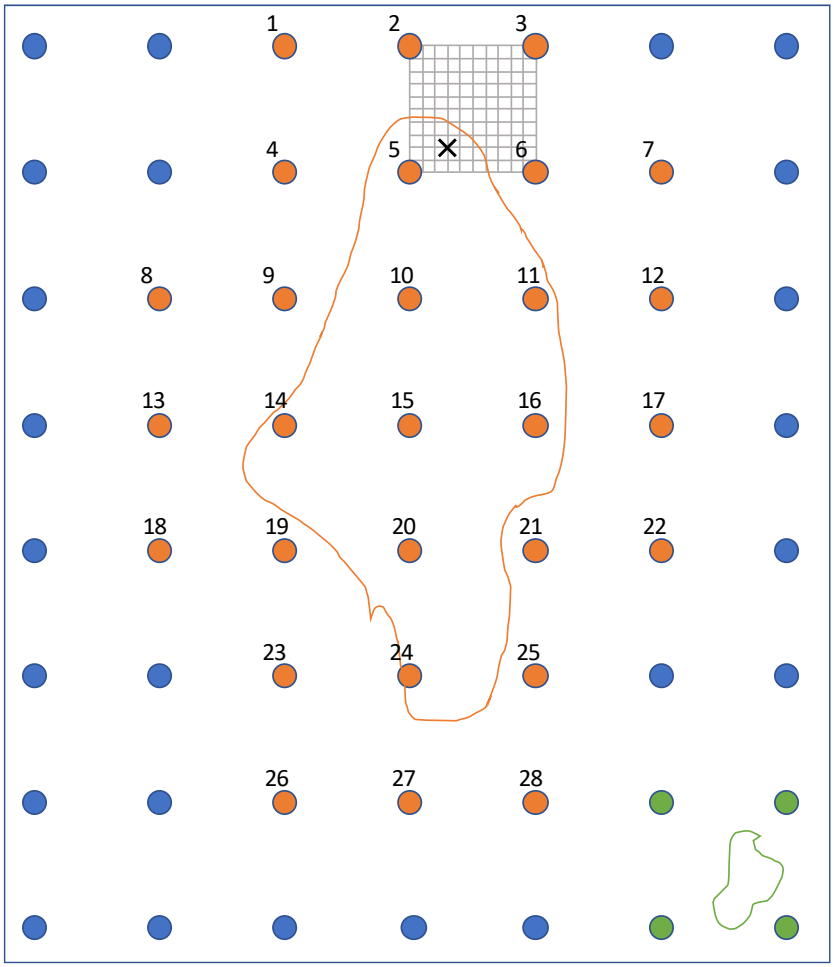


Figure 10: Visualization of calculation of risk at a given building (denoted by the black cross) in the exposure database. 1) Find the zone in which the building is located (the orange zone in this case). 2) Find the four grid points surrounding this point. These grid points have a non-zero contribution to the risk at the location of the building. 3) Find the contributions of points 2, 3, 5, and 6 to the building. In this case, since the building lies at 30% between the vertical lines through 2&5 and 3&6, and at 20% between the horizontal lines through 2&3 and 5&6, the contributions are: Point 2: 0.2x0.7 = 0.14, Point 3: 0.2x0.3 = 0.06, Point 5: 0.7x0.8 = 0.56, Point 6: 0.3x0.8 = 0.24.

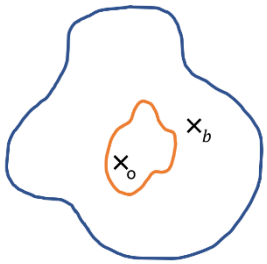


Figure 11: Building $\times_o$ is located in the orange zone, while building $\times_b$ is located in the blue zone, even though both buildings lie within the borders of the blue zone. This is accounted for in the TNO Model Chain code.

BOX 8 IMPLEMENTATION OF RISK MODEL PREPARATION

1. Define vectors for magnitude m, and distance r^{rup}. Define a number of synthetic catalogues N_{cat} and samples per catalogue $N_{samples}$.
2. For every site response region $\mathcal{s}$:
   a. For every synthetic catalogue cat_j:
      i. For every $(GMM_{med})_k$.
         1. For every logic tree branch $(\Phi_{ss})_l$:
            a. For every unique combination of m, r^{rup}:
               i. Repeat N_{sample} times:
                  1. Draw a correlated sample from a multivariate normal distribution with mean $\mu = 0$ and cov is given by the correlation matrix for spectral periods and duration (Im2im_V5.yaml input file): ϵ_{ref}.
                  2. Draw a correlated sample from a multivariate normal distribution with mean $\mu = 0$ and cov is given by the correlation matrix for spectral periods only (Im2im_V5.yaml input file): ϵ_{AF}.
                  3. For every spectral period T_m required for fragility
                     a. Calculate $g_{source} =$:

$$\begin{cases} m_0 + m_1(m-4.7) + m_2(m-4.7)^2 & \text{if } m \le 4.7 \\ m_0 + m_3(m-4.7) & \text{if } 4.7 < m \le 5.45, \\ m_0 + m_3(5.45-4.7) + m_4(m-5.45) + m_5(m-5.45)^2 & m > 5.45 \end{cases}$$

where m_0, m_1, m_2, m_3, m_4 and m_5 are model parameter depending on $(GMM_{med})_k$ and spectral period T_m.

                     b. Calculate $g_{path} =$

$$\begin{cases} (r_0 + r_1 m)\log\left(\frac{r^{rup}}{3}\right) & \text{if } r^{rup} \le 7 \\ (r_0 + r_1 m)\log\left(\frac{7}{3}\right) + (r_2 + r_3 m)\log\left(\frac{r^{rup}}{7}\right) & \text{if } 7 < r^{rup} \le 12, \\ (r_0 + r_1 m)\log\left(\frac{7}{3}\right) + (r_2 + r_3 m)\log\left(\frac{12}{7}\right) + (r_4 + r_5 m)\log\left(\frac{r^{rup}}{12}\right) & r^{rup} > 12 \end{cases}$$

where r_0, r_1, r_2, r_3, r_4 and r_5 are model parameter depending on $(GMM_{med})_k$ and spectral period T_m.

                     c. $\log(Y(r^{rup}, m)) = g_{path}(r^{rup}, m) + g_{source}(m) + \log\left(\frac{0.01}{9.807}\right)$.

                     d. Calculate $\sigma_Y = \sqrt{\tau^2 + \phi_{ss}^2 + \sigma_{c2c}^2}$ where τ depends on $(GMM_{med})_k$, ϕ_{ss} depends on $(\Phi_{ss})_l$, and $\sigma_{c2c}^2(T_m, m, r^{rup}) =$

$$\begin{cases} 0.026 + 1.03[5.6 - \min(5.6; \max(m; 3.6)](r^{rup})^{-2.22} & \text{if } T \le 0.1 \\ \left[\sigma_{c2c}^2(0.1, m, r^{rup}) + \left[\frac{\log(T_j) - \log(0.1)}{\log(0.85) - \log(0.1)}\right] \times \right. & \text{if } 0.1 < T < 0.85. \\ \left. [\sigma_{c2c}^2(0.85, m, r^{rup}) - \sigma_{c2c}^2(0.1, m, r^{rup})]\right] & \\ 0.045 + 5.315[5.6 - \min(5.6; \max(m; 3.6)](r^{rup})^{-2.92} & T \ge 0.85 \end{cases}$$

                     e. $A^{ref} = \exp(\log(Y) + \sigma_{c2c}^2 \times (\epsilon_{ref})_m)$.

                     f. $M_{ref} = M_1 - \frac{\log(r^{rup}) - \log(3)}{\log(60) - \log(3)}(M_1 - M_2)$, where M_1 and M_2 are model parameters depending on $\mathcal{s}$ and T_m.

                     g. $f_1 = (a_0 + a_1 \log(r^{rup})) + (b_0 + b_1 \log(r^{rup}))(\min(m, M_{ref}) - M_{ref})$, where a_0, a_1, b_0, and b_1 are model parameter depending on $\mathcal{s}$ and T_m.

BOX 8 IMPLEMENTATION OF RISK MODEL PREPARATION (CONT.)

h. $\log\left[(AF_{Sa})^{median}\right] = f_1 + f_2 \log\left(\frac{A^{ref}+f_3}{f_3}\right)$ where f_2 and f_3 are model parameter depending on $\mathcal{s}$ and T_m.

i. Set all values of $\log\left[(AF_{Sa})^{median}\right]$ lower than log (A_{min}) to log (A_{min}), and all values $\log\left[(AF_{Sa})^{median}\right]$ higher than log (A_{max}) to log (A_{max}), where A_{min} and A_{max} are model parameter depending on $\mathcal{s}$ and T_m.

j. Calculate $(AF_{Sa})^{\sigma}$:

i. $(AF_{Sa})^{\sigma} = \phi_{S2S,1} + (\phi_{S2S,2} - \phi_{S2S,1})\left[\frac{\log(A^{ref}) - \log(Sa_{low})}{\log(Sa_{high}) - \log(Sa_{low})}\right]$, where $\phi_{S2S,1}, \phi_{S2S,2}, Sa_{high}$ and Sa_{low} are model parameter depending on $\mathcal{s}$ and T_m.

ii. Set all values of $(AF_{Sa})^{\sigma}$ lower than Sa_{low} to Sa_{low}, and all values $(AF_{Sa})^{\sigma}$ higher than Sa_{high} to Sa_{high}.

k. $A^{sur} = \exp\left[\log\left(A^{ref}\right) + \log\left((AF_{Sa})^{median}\right) + (AF_{Sa})^{\sigma} \times (\epsilon_{AF})_m\right]$.

4. Calculate $g_{source} =$:

$$\begin{cases} m_6 + m_7(m - 5.25) & \text{if } m \leq 5.25 \\ m_6 + m_8(m - 5.25) + m_9(m - 5.25)^2 & \text{if } m > 5.25, \end{cases}$$

where m_6, m_7, m_8 and m_9 are model parameter depending on $(GMM_{med})_k$ and spectral period T_m.

5. Calculate $g_{path} =$

$$\begin{cases} (r_6 + r_7 m)\left[\log\left(\frac{r^{rup}}{3}\right)\right]^{r_8} & \text{if } r^{rup} \leq 12 \text{ km} \\ (r_6 + r_7 m)\left[\log\left(\frac{12}{3}\right)\right]^{r_8} + (r_9 + r_{10} m)\log\left(\frac{r^{rup}}{12}\right) & \text{if } r^{rup} > 12 \text{ km}, \end{cases}$$

where r_6, r_7, r_8, r_9 and r_{10} are model parameter depending on $(GMM_{med})_k$ and spectral period T_m.

6. $\log(D_{5-75}(r^{rup}, m)) = g_{path}(r^{rup}, m) + g_{source}(m)$.

7. $\sigma^2_{Dc2c} = 0.0299 + 2.434[5.6 - \min(5.6; \max(3.6; m))](r^{rup})^{-1.95}$.

8. $(D_{5-75})^{\sigma} = \sqrt{\tau^2 + \phi_{ss}^2 + \sigma^2_{Dc2c}}$, where τ depends on $(GMM_{med})_k$, ϕ depends on $(\Phi_{ss})_l$.

9. Calculate $g_{site} = -0.2246 \log\left(\frac{\min(V_{S,30}; 600)}{600}\right)$, where $V_{S,30}$ is a property of site response region $\mathcal{s}$.

10. $D_{5-75} = \exp\left(\log(D_{5-75}) + (D_{5-75})^{\sigma} \times (\epsilon_{ref})_m + g_{site}\right)$.

b. The matrix obtained in 2.a.i.1.a.i. (up until the line above) contains $N_{samples}$ samples of ground motion for every spectral period T (and D_{5-75}) for every unique combination of m, r^{rup}.

c. Use this matrix to calculate, for every fragility logic tree branch $(l_{frag})_m$:

i. For every typology t_n :

1. Calculate $\log(IM) = b_0 + b_1 \log(A^{sur}_{T1}) + b_2 \log(D_{5-75}) + b_3 \log(A^{sur}_{T2})$, where b_0, b_1, b_2, and b_3 are dependent on $(l_{frag})_m$ and t_n, and A^{sur}_{T1}, A^{sur}_{T2} are the sampled ground motions for the spectral periods for which the fragility curves of typology t_n are defined for.

2. For each limit state u_p:

a. $X_u = \frac{\log(DL_u) - \log(\mathrm{IM})}{\beta_s}$ where DL_u is dependent on limit state u, $(l_{frag})_m$ and t_n and β_s is dependent on $\left(l_{frag}\right)_m$ and t_n.

BOX 8 IMPLEMENTATION OF RISK MODEL PREPARATION (CONT.)

b. $P_f = 0.5(1 - \text{erf}\left(\frac{X_u}{\sqrt{2}}\right)$.

c. $P\left(u_0 > u_p \middle| cat_j, t_n, s, m, r, (GMM_{med})_k, (\phi_{ss})_l, \left(l_{frag}\right)_m\right) =$

$$\frac{1}{N_{samples}} \sum_{i}^{N_{samples}} (P_f)_i.$$

3. For each consequence branch $(l_{cons})_q$:

a. Calculate $P_{ch} =$

$$\begin{cases} 0 & \text{if } \beta_{ch} = 0 \\ 0.5(1 + \text{erf}\left(\frac{X_{ch}}{\sqrt{2}}\right) & \text{if } \beta_{ch} > 0, \end{cases}$$

where $X_{ch} = \min(\log(A^{sur}_{Sa\,0.01}) ; \log(0.75)) - \frac{\log\left(\overline{PGA_{ch}}\right)}{\beta_{ch}}$,

β_{ch} and $\overline{PGA_{ch}}$ depend on $(l_{cons})_q$ and t_n, and $A^{sur}_{Sa\,0.01}$ sampled ground motions for the spectral period 0.01 sec.

b. $P\left(d_{chimney} \middle| cat_j, t_n, s, m, r, (GMM_{med})_k, (\phi_{ss})_l, \left(l_{frag}\right)_m, (l_{cons})_q\right) =$

$$\frac{1}{N_{samples}} \sum_{i}^{N_{samples}} ((1 - P_f^{CS1})_i \times (P_{ch})_i.$$

2. Repeat for every logic tree branch $(\Phi_{ss})_l$.

ii. Repeat for every logic tree branch $(GMM_{med})_k$.

b. Repeat for every synthetic catalogue cat_j.

3. Repeat for every site response region s.

4. $P(u_0 > u | t, s, m, r, GMM_{med}, \phi_{ss}, l_{frag}) =$

$\frac{1}{N_{cat}} \sum_j^{N_{cat}} P\left(u_0 > u_p \middle| cat_j, t, s, m, r, GMM_{med}, \phi_{ss}, l_{frag}\right)$.

5. $P\left(d_{chimney} \middle| t, s, m, r, GMM_{med}, \phi_{ss}, l_{frag}, l_{cons}\right) =$

$\frac{1}{N_{cat}} \sum_j^{N_{cat}} P\left(d_{chimney} \middle| cat_j, t, s, m, r, GMM_{med}, \phi_{ss}, l_{frag}, l_{cons}\right)$.

BOX 9 IMPLEMENTATION OF BUILDING EXPOSURE PREPARATION

1. Load the building exposure database from the original csv-file. This contains the location of each building ($\boldsymbol{x}_b$), as well as the PMF of typologies $P(t_b = t_i)$.
2. For each building b:
   a. Find the zone s_b in which building b is located, based on $\boldsymbol{x}_b$ and the shapefiles of the site-response zonation. Care is taken to correctly assign buildings that lie in a 'zone-within-a-zone' (see Figure 11).
   b. Obtain the contribution to the risk of building b of each grid point defined for s_b. This is done by:
      i. Finding the four grid points surrounding building b. The contribution of all other grid points defined for zone s_b is set to zero.
      ii. The four grid points ($q_{leftupper}, q_{rightupper}, q_{leftlower}, q_{rightlower}$) define four bounds: $RDx_{left}, RDx_{right}, RDy_{lower}, RDy_{upper}$. The contributions of the grid points are defined by:

$$c_{q_{leftupper},b} = \frac{|x_b - RDx_{right}|}{|RDx_{right} - RDx_{left}|} \times \frac{|y - RDy_{lower}|}{|RDy_{upper} - RDy_{lower}|}$$
$$c_{q_{rightupper},b} = \frac{|x_b - RDx_{left}|}{|RDx_{right} - RDx_{left}|} \times \frac{|y - RDy_{lower}|}{|RDy_{upper} - RDy_{lower}|}$$
$$c_{q_{leftlower},b} = \frac{|x_b - RDx_{right}|}{|RDx_{right} - RDx_{left}|} \times \frac{|y - RDy_{upper}|}{|RDy_{upper} - RDy_{lower}|}$$
$$c_{q_{rightlower},b} = \frac{|x_b - RDx_{left}|}{|RDx_{right} - RDx_{left}|} \times \frac{|y - RDy_{upper}|}{|RDy_{upper} - RDy_{lower}|}.$$

   c. Obtain the PMF of typologies $P(t_b = t_i)$ (i.e. the probability of building b belonging to typology t_i).
3. The above step results in two matrices for each zone:
   a. A 2D matrix of size $nr_{buildings_in_zone} \times nr_{gridpoints_in_zone}$.
   b. A 2D matrix of size $nr_{buildings_{in_{zone}}} \times nr_{typologies_{defined_{fieldwide}}}$.

   These matrices are saved and used in the integration step.

3.3.2 *Second step: Integration*

3.3.2.1 Input files

Lookup table exceedance values of damage and collapse states
This lookup table is a complementary cumulative mass function (CCMF), which contains exceedance probabilities of six limit states u (three damage states and three collapse states, $\{DS1, DS2, DS3, CS1, CS2, CS3\}$) of typologies t, for all site response regions s, logic tree ground motion branches $GMM_{med}\phi_{ss}$, logic tree fragility branches l_{frag}, due to a hypothetical earthquake at rupture distances r^{rup} and of magnitudes m ($P(u_0 > u|\, t, s, m, r^{rup}, GMM_{med}\phi_{ss}, l_{frag}\,)$).

The lookup table also contains the probability of dying due to chimney collapse $P_{d_{chimney}}$ for the consequence logic tree branches l_{cons}, for typologies t, all site response regions s, logic tree ground motion branches $GMM_{med}\phi_{ss}$, logic tree fragility branches l_{frag}, due to a hypothetical earthquake at distances r^{rup} and magnitudes m.

Exposure lookup table
Lookup table containing per site response region s, the contribution of evaluation points q to a building b from the database $c_{q,b}(\boldsymbol{x}_b)$ and of every building the probability of belonging to a certain typology $P(t_b = t_i).c_{q,b}(\boldsymbol{x}_b)$ is computed in the exposure prep by bilinear interpolation of the location of the building $\boldsymbol{x}_b$ to the surrounding evaluation points q.

Consequence input file
Input file containing per consequence branch l_{cons} for all typologies t, the probabilities of dying inside and outside, given the occurrence of one of the three collapse states, $P_{d_{inside|CSi}}$ and $P_{d_{outside|CSi}}$, where CSi is one of the three collapse states $\{CS1, CS2, CS3\}$.

Output rupture model (SSM)
A 5D probability mass function (PMF) matrix of earthquake expectation values ($f(q, t, m_{max}, m, r^{rup})$) per evaluation point q, year t, maximum magnitude logic tree branch m_{max}, magnitude m, and distance r^{rup}.

3.3.2.2 Input parameters


```
{
  "logictree": {
    "Mmax": {
            "4.0": 0.08625,
          "4.5": 0.4,
          "5.0": 0.24375,
          "5.5": 0.1125,
          "6.0": 0.07875,
          "6.5": 0.0525,
          "7.0": 0.02625},
    "GMMMedian": {
    "Upper": 0.3,
    "CentralUpper": 0.3,
    "CentralLower": 0.3,
    "Lower": 0.1
```

```
    },
    "GMMPhiSS": {
    "phi_ss_high": 0.5,
    "phi_ss_low": 0.5
    },
    "Fragility": {
          "Middle": 0.66,
          "Lower": 0.17,
          "Upper": 0.17},
    "Consequence":{
          "Middle": 0.5,
        "Lower": 0.25,
          "Upper": 0.25},
  "basedir": <string, full path to directory containing all
input files>,
  "ssmmrdistributionsfile": <string, name of the input file>,
  "consequencefile": <string, name of the input file>,
  "building_prep_file": <string, name of the input file>,
  "gmmdmpoefile": <string, name of the input file>,
  "risk_outputfile": <string, name of the output file>,
}
```

3.3.2.3 Implementation

Similar as in the hazard integrator, all calculations are carried out per site response region $\mathcal{s}$. However, the assignment of evaluation points $q_{|\mathcal{s}_i}$ to site response regions $\mathcal{s}_i$ is already done in the exposure preparation and can therefore be imported from the exposure lookup table.

Computation of the mean occurrence probabilities of earthquakes $f_{mean}(q,t,m,r^{rup})$ of the m_{max} logic tree branches is computed by matrix multiplication. Similarly, the mean exceedance probabilities of limit states $P_{mean}(u_0 > u \mid t,m,r^{rup})$, where $u = \{DS1, DS2, DS3, CS1, CS2, CS3\}$ and the mean probability of dying due to chimney collapse $P_{d\,chimney}(t,m,r,l_{cons})$ are also computed by matrix multiplication, computing the mean over the GMMMedian GMM_{med}, GMMPhiSS ϕ_{ss} and fragility l_{frag} logic tree branches, by using the logic tree weights $P(GMM_{med})$ and $P(\phi_{ss})$ and $P(l_{frag})$, defined in the input parameter file.

$P_{mean}(u_0 > u \mid t,m,r^{rup})$ is the mean probability of exceeding limit state u, given the occurrence of a hypothetical earthquake at rupture distance r^{rup} of magnitude m. To determine the probability of exceeding limit states due to the forecasted earthquake occurrence $P(u_0 > u \mid t,q,t)$ as a result of gas production, we combine the P_{mean} with the mean PMF of earthquake occurrence $f_{mean}(q,t,m,r^{rup})$ and integrate over the magnitudes and distances.

The consequence of building collapse $\{CS1, CS2, CS3\}$ is computed as Local Personal Risk (LPR), the probability of a hypothetical person dying P_d due to building collapse. To compute this, we make use of the consequence input file, which contains the probabilities of dying P_d inside or outside, given one of the collapse states $\{CS1, CS2, CS3\}$, per typology t and per consequence logic tree branch l_{cons}. The probability of dying due to chimney collapse $P_{d\,chimney}$ is already computed in the preparation script and imported through the lookup table.

The probability of dying inside and outside due to gas production is then computed, by combining the P_d's from the consequence input file with the probability of exceedance of the limit states $P(u_0 > u \,|t, q, t)$. Then, the mean probabilities of dying, inside, outside, and due to chimney collapse over the consequence logic tree branches l_{cons} are computed by matrix multiplication with the logic tree weights $P(l_{cons})$. We now have an LPR_{inside}, $LPR_{outside}$ and $LPR_{chimney}$ per typology t, evaluation point q and year t. To compute the total LPR, we assume the hypothetical person to be 99% of the time inside the house and 1% of the time within 5 m outside the house.

The last step is to determine the $LPR(b, t)$, the local personal risk per actual building b in the Groningen area per year t. For this we use the exposure lookup table, which contains the contribution of evaluation points q to a building b from the database $c_{q,b}(\boldsymbol{x}_b)$ and of every building b the probability of belonging to a certain typology $P(t_b = t_i)$.

The numerical implementation is described in Box 10.

BOX 10 IMPLEMENTATION OF RISK INTEGRATION

1. Import $f(q, t, M_{max}, m, r)$ and associated grids $q, t, M_{max}, m, r^{rup}$.
2. Import evaluation points q_s per site response region s.
3. For every site response region s:
   a. Extract the earthquake PMF only for those evaluation points $f(q_s, t, M_{max}, m, r^{rup})$.
   b. Compute the mean earthquake PMF over the M_{max} logic tree branches: $f_{mean}(q_{|}, t, m, r^{rup}) = \sum_l P(M_{\max_l}) f(q_s, t, M_{max_l}, m, r^{rup})$.
   c. Import $P(u_0 > u| t, m, r, GMM_{med}, \phi_{ss}, l_{frag})$ and $P_{d_{chimney}}(t, m, r, GMM_{med}, \phi_{ss}, l_{frag}, l_{cons})$ from the risk lookup table. Both are dependent on site response region s.
   d. Compute the mean over the logic tree branches of the distributions : $P_{mean}(u_0 > u \mid t, m, r^{rup}) = \sum_j \sum_l \sum_k P\left(GMM_{med_j}\right) P(\phi_{ss_l}) P\left(l_{frag_k}\right) P(u_0 > u|, m, r^{rup}, GMM_{med_j}, \phi_{ss_l}, l_{frag_k})$,

$$P_{d_{chimney}}(t, m, r^{rup}, l_{cons}) = \sum_j \sum_l \sum_k P\left(GMM_{med_j}\right) P(\phi_{ss_l}) P\left(l_{frag_k}\right) P_{d_{chimney}}\left(t, m, r^{rup}, GMM_{med_j}, \phi_{ss_l}, l_{frag_k}, l_{cons}\right).$$

   e. Combine with PMF of earthquake occurrence: $P(u_0 > u |t, q_s, t) = \sum_n \sum_m P_{mean}(u_0 > u |t, m_n, r^{rup}{}_m) f_{mean}(q_s, t, m_n, r^{rup}{}_m)$ and $P_{d_{chimney}}(t, q_s, t, l_{cons}) = \sum_n \sum_m P_{d_{chimney}}(t, m_n, r_m, l_{cons}) f_{mean}(q_s, t, m_n, r_m)$.
   f. For every consequence branch l_{cons_o}:
      i. For every typology t_p:
         1. $P_{d_{inside}}(q_s, t) = [P(u_0 > CS1 \mid q_s, t) - P(u_0 > CS2 \mid q_s, t)] * P_{d_{inside|CS1}} + [P(u_0 > CS2 \mid q_s, t) - P(u_0 > CS3 \mid q_s, t)] * P_{d_{inside|CS2}} + P(u_0 > CS3 \mid q_s, t) * P_{d_{inside|CS3}}$.
         2. $P_{d_{outside}}(q_s, t) = [P(u_0 > CS1 \mid q_s, t) - P(u_0 > CS2 \mid q_s, t)] * P_{d_{outside|CS1}} + [P(u_0 > CS2 \mid q_s, t) - P(u_0 > CS3 \mid q_s, t)] * P_{d_{outside|CS2}} + P(u_0 > CS3 \mid q_s, t) * P_{d_{outside|CS3}}$.

            All above probabilities are dependent on l_{cons_o}, t_p and s.

BOX 10 IMPLEMENTATION OF RISK INTEGRATION (CONT.)

g. Compute inside, outside and chimney local personal risk:

$$LPR_{inside}(t, q_s, t) = \sum_o P(l_{cons_o})\ P_{d_{inside}}(t, q_s, t, l_{cons_o}),$$

$$LPR_{outside}(t, q_s, t) = \sum_o P(l_{cons_o})\ P_{d_{outside}}(t, q_s, t, l_{cons_o}),$$

$$LPR_{chimney}(t, q_s, t) = \sum_o P(l_{cons_o})\ P_{d_{chimney}}(t, q_s, t, l_{cons_o}).$$

h. Compute total local personal risk:

$$LPR(t, q_s, t) = 0.99\, LPR_{inside}(t, q_s, t) + 0.01\left[LPR_{outside}(t, q_s, t) + LPR_{chimney}(t, q_s, t)\right].$$

5. Repeat for every site response region s and save the resulting separate matrices per site response region s: $P(u_0 > u\,|t, q_s, t, s)$ and $LPR(t, q_s, t, s)$.
6. For every site response region s:
   a. $LPR(t, b_s, t) = \sum_j LPR\left(t, q_{s\,j}, t\right)\ c_{q_{s\,j}, b_s}(\boldsymbol{x}_{b_s})$,
   where b_s are the buildings positioned within the site response region s.
7. Sum over all the site response regions: $LPR(t, b, t) = \sum_i LPR(t, b_s, t, s)$.
4. Compute the local personal risk per building from the database:

$$LPR(b, t) = \sum_p P(t_b = t_0)_p\ LPR(t_p, b, t).$$

4 References

Bommer, J. J., Stafford, P. J., Edwards, B., Dost, B., & Ntinalexis, M. (2015). *Development of GMPEs for Response Spectral Accelerations and for Strong-Motion Durations (V1).* Retrieved February 5, 2017, from http://feitenencijfers.namplatform.nl/download/rapport/1179fefc-bd80-4489-aa06-b71ca2e5da3f?open=false

Bommer, J., Edwards, B., Kruiver, P., Rodriquez-Marek, A., Stafford, P., Dost, B., . . . Spetzler, J. (2017). *"V5 Ground-Motion Model (GMM) for the Groningen Field.*

Bommer, J., Stafford, P., Edwards, B., Dost, B., van Dedem, E., Rodriguez-Marek, A., . . . Ntinalexis, M. (2017). Framework for a Ground-Motion Model for Induced Seismic Hazard and Risk Analysis in the Groningen Gas Field, The Netherlands. *Earthquake Spectra*. doi:10.1193/082916EQS138M

Bourne, S. J., & Oates, S. J. (2017). Extreme Threshold Failures Within a Heterogeneous Elastic Thin Sheet and the Spatial-Temporal Development of Induced Seismicity Within the Groningen Gas Field. *Journal of Geophysical Research, 122*(12). Retrieved September 18, 2018, from https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017jb014356

Bourne, S. J., & Oates, S. J. (2018). *Note for File: Finite rupture simulation for ground motion modelling and probabilistic seismic hazard and risk analysis for the Groningen gas field.* DRAFT Internal NAM report.

Bourne, S. J., Oates, S. J., Scheefhals, R., Nash, P. C., Mar-Or, A., Storck, T., . . . Van Elk, J. (2019). *A Monte Carlo method for probabilistic seismic hazard and risk analysis of induced seismicity in the Groningen gas field.* Shell, Alten Technical Software and NAM.

Bourne, S., Oates, S., & Van Elk, J. (2018). The exponential rise of induced seismicity with increasing stress levels in the Groningen gas fiedl and its implications for controlling seismic risk. *Geophysical Journal International*, 1693-1700.

TNO. (2019). *Comparative analysis of the NAM and TNO implementations in the Groningen Seismic Hazard and Risk Assessment.* TNO 2019 R11997.

TNO. (2020). *IT-platform for the TNO Groningen Model Chain PSHRA calculations.* TNO 2020 R10474.

5 Signature

Utrecht, December 17 2020 TNO

Drs. J.A.J. Zegwaard
Head Advisory Group for Economic Affairs

Appendix A: Numerical methods

A.1 Numerical integration: Monte Carlo vs quadrature

At many points in the TNO Model Chain, a function needs to be evaluated with a (probability) distribution as input rather than a scalar value. When performing such an operation on a distribution, the aim is obtain the mean value (i.e. the expectation value), which is found by integrating over the (probability) distribution.

Such an integration can be performed through different methods. In the TNO Model Chain, this is achieved through direct numerical integration. Here, we compare direct numerical integration to another method that is often applied: Monte Carlo integration.

A.1.1 Monte Carlo integration

Evaluating an integral through Monte Carlo integration is done by drawing a sample from the input distribution, evaluating the function, and storing the result. This process is repeated many times. The result in then simply the mean of all individual function evaluations:

$$S = \frac{1}{N}\sum_{i=1}^{N} g(x_i),$$

where x_i is a random sample from the input distribution, g is the function to be evaluated, and N is the number of samples used.
Monte Carlo integration is a non-deterministic approach to integration, as each realization of the integration provides a different result (i.e. if a function is evaluated 1 million times to obtain a mean result, and this whole process is repeated, the mean result will not be exactly identical). It is usually applied when direct numerical integration becomes unfeasible due to the number of integration points involved. This is often the case when the number of dimensions becomes large, resulting in an impractical /impossible number of integration points. When enough samples are used, the resulting mean can be brought arbitrarily close to the ground truth.

A.1.2 Quadrature

The classical approach to numerically evaluating an integral is through numerical quadrature, effectively using a Riemann sum (or a simple extension to higher dimensions). The probability distribution is discretized and transformed into a probability mass function. For each member of the probability mass function, the function is evaluated and multiplied with its mass μ_i. Finally, all individual members are summed to obtain the integration result:

$$S = \sum_{i} g(x_i)\mu(x_i),$$

where x_i is a member of the discretized input distribution,g is the function to be integrated, and μ is the probability mass function.

A.1.3 Example

As an example, consider the function $g(x) = x^2$. Let u be normally distributed with a mean value of 15 and a standard distribution of 2. The function to be evaluated is then:

$$E\big(g(u)\big) = \int_{\Omega_u} g(u) f_U(u) du,$$

where f_U is de probability density function of u and Ω_u its domain.

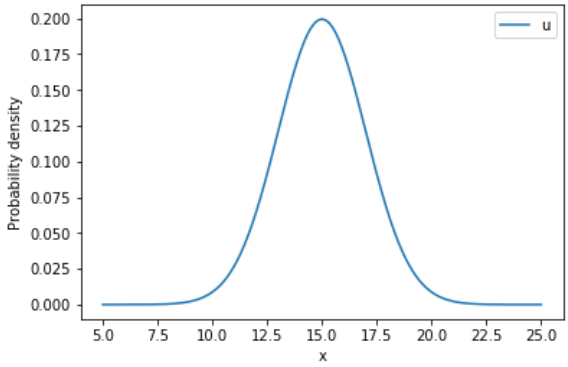


Figure 12: The probability density function of the input distribution.

Analytically, the function integrates to:

$$E\big(g(u)\big) = \int_{-\infty}^{\infty} x^2 \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-15}{2}\right)^2\right) dx = 229.0.$$

Note $mean\big(g(u)\big) \neq g\big(mean(u)\big)$ since $g(x)$ is not a linear function.

A.1.3.1 Monte Carlo integration demonstration

For the sake of this example, the integration will initially be performed with 10 samples:

Sample nr (i)	Sample (x_i)	Function value (x_i^2)
1	15.0496214	226.49110416
2	15.42085428	237.80274676
3	12.07933037	145.91022219
4	16.08003403	258.56749443
5	19.17949013	367.85284153
6	10.62454195	112.88089172
7	18.41537726	339.12611972
8	16.95214992	287.37538685
9	13.14547428	172.80349399
10	16.08384112	258.68994521
	Total	2407.50024656
Integration result =	Total/nr_samples	**240.750024656**

This is a relatively poor approximation of the analytical result. However, with an increasing number of samples, the approximation becomes increasingly better:

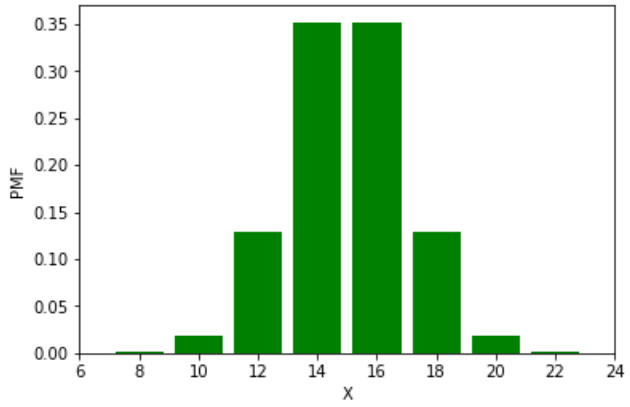
Number of samples	Monte Carlo integration result
100	227.26850012748906
1000	228.00546769993346
10000	229.31856812990537
100000	229.24401152063456
1000000	229.06487488734854
10000000	229.0212108127112

100000000	229.0012567734311
1000000000	228.99640057412972

A.1.3.2 Direct numerical integration demonstration:

In order to perform the numerical integration, the input distribution first needs to be transformed into a probability mass function, $p_X(x)$. This means that the domain of integration needs to be finite (note that this step was not required for Monte Carlo integration). Based on Figure 12, the integration domain is chosen as $[6,24]$. An initial choice for the discretization (with $dx = 2$) looks like this:

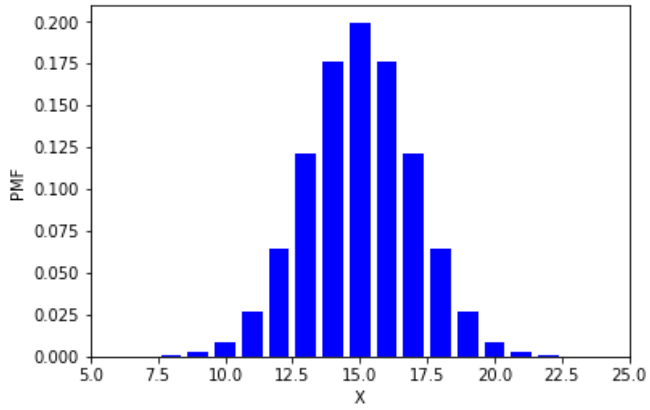
Integration point (x_i)	Function value (x_i^2)	Probability mass function $p_X(x_i)$	$(x_i)^2 p_X(x_i)$
6	36	1.59837E-05	0.000575
8	64	0.000872683	0.055852
10	100	0.0175283	1.75283
12	144	0.129517596	18.65053
14	196	0.352065327	69.0048
16	256	0.352065327	90.12872
18	324	0.129517596	41.9637
20	400	0.0175283	7.01132
22	484	0.000872683	0.422378
24	576	1.59837E-05	0.009207
			+
Total			**228.994429**



An alternative discretisation with domain $[5,25]$ and $dx = 1$ gives:

Integration point (x_i)	Function value (x_i^2)	Probability mass function $p_X(x_i)$	$(x_i)^2 p_X(x_i)$
5	25	7.4336E-07	0.000018584
6	36	7.9919E-06	0.00028771
7	49	6.6915E-05	0.00327884
8	64	0.00043634	0.02792585
9	81	0.00221592	0.17948986
10	100	0.00876415	0.87641502
11	121	0.02699548	3.26645347
12	144	0.0647588	9.32526689
13	169	0.12098536	20.4465262

14	196	0.17603266	34.502402
15	225	0.19947114	44.8810065
16	256	0.17603266	45.0643618
17	289	0.12098536	34.9647697
18	324	0.0647588	20.9818505
19	361	0.02699548	9.74536946
20	400	0.00876415	3.5056601
21	441	0.00221592	0.97722257
22	484	0.00043634	0.21118921
23	529	6.6915E-05	0.03539809
24	576	7.9919E-06	0.00460332
25	625	7.4336E-07	0.0004646
			+
Total			**228.999987**



The example considered here demonstrates that if the number of samples is equal to the number of integration points, the direct numerical integration method yields a better approximation of the analytical result than the Monte Carlo integration method.

A.2 Discretization options

In order to apply direct numerical integration effectively and efficiently to a given problem, the discretization of the input distribution needs to be chosen carefully. The fundamental question is: How well does the numerical (discretized) representation approximate the analytical (continuous) solution, while using a reasonable/computable number of operations. To achieve this, it is useful to again consider the general Riemann sum that is being computed:

$$S = \sum_i g(\boldsymbol{x}_i) p_X(\boldsymbol{x}_i),$$

where $\boldsymbol{x}_i$ are the discrete sample points of the input space, $g(\boldsymbol{x})$ is the function to be evaluated, and $p_X(\boldsymbol{x}_i)$ is the probability mass function associated with the input space. Note that $\boldsymbol{x}_i$ may be a scalar value or a vector of any length, depending on the functional form of $g(\boldsymbol{x})$.

The contribution of each individual grid point $\boldsymbol{x}_i$ to the total integral S is the product of the function value $g(\boldsymbol{x}_i)$ and the probability mass $p_X(\boldsymbol{x}_i)$. The probability mass is given by:

$$p_X(\boldsymbol{x}_i) = f_X(\boldsymbol{x}_i) d\boldsymbol{x}_i,$$

where $f_X(\boldsymbol{x}_i)$ is the value of the probability density function and $d\boldsymbol{x}_i$ is the *measure* of the underlying set. This *measure* can be thought of as the length of the interval in 1D, the area in 2D, the volume in 3D, etc.

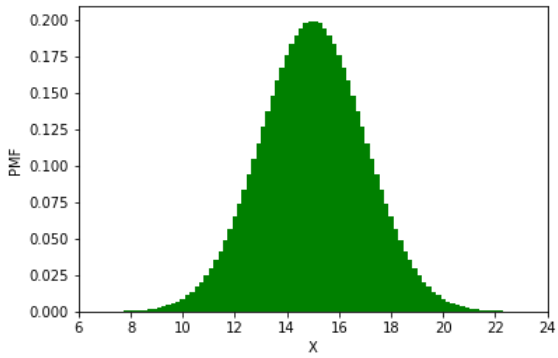
A.2.1 Choice of domain

The domain of each dimension of $\boldsymbol{x}$ should be chosen in such a way that it sufficiently encompasses the underlying distribution. Ideally, $\sum_i p_X(\boldsymbol{x}_i) = 1$. However, in practice many distributions (such as the normal distribution) have an infinite domain, which means that a choice of domain is inevitable, and that $\sum_i p_X(\boldsymbol{x}_i) < 1$. An often applied rule of thumb is that the domain should be chosen in such a way that further extension of the domain does not appreciably affect the final result. To account for the fact that practicality often dictates a finite domain, the integral is commonly normalized:

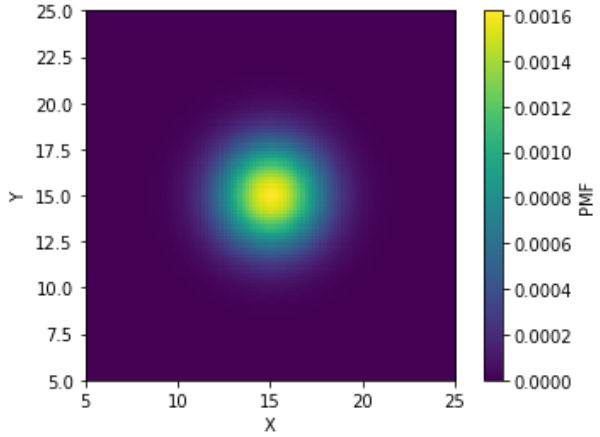
$$S = \frac{\sum_i f(\boldsymbol{x}_i) p_X(\boldsymbol{x}_i)}{\sum_i p_X(\boldsymbol{x}_i)}.$$

A.2.2. Number of grid points

The number of grid points is another important choice. Since the number of dimensions of a probability distribution in the TNO Model Chain often exceeds 5, it becomes important not to use more grid points per dimension than are strictly needed. For example, it may be tempting to use 100 grid points per dimension. A one-dimensional normal distribution then looks like:



The 2D-extension of the same normal distribution looks like this:



However, the so-called *curse of dimensionality* prevents this gridding from remaining feasible in higher dimensions. Storing a single precision number requires 8 bytes of memory, making the 1D probability mass function 800 bytes (~1 kB). The 2D probability mass function requires 100 times as much (~0.08 MB). An extension to 3D would require ~8 MB, and a 4D extension would require 800 MB. This is still reasonable for a modern computer. However, further extension would become more difficult, as a 5D extension would require 80 GB which is not feasible on most systems.

Therefore, in order to computational effort reasonable, and memory load physically possible, it is advisable to keep the number of points along each dimension as small as possible. Again, the often applied rule of thumb is that the discretization should be chosen in such a way that further refinement of the grid does not appreciably affect the final result.

A.2.3 Spacing of points

By using the Riemann approximation, the underlying assumption is that the integral $\int_{x_i-0.5dx}^{x_i+0.5dx} g(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x}$ *is reasonably well-approximated by* $g(\boldsymbol{x}_i) p_X(\boldsymbol{x}_i)$*. This is the case when the integrand is approximately linear over this domain. In general, this assumption becomes increasingly valid for smaller values of* dx*, and therefore for a larger number of grid points within the same domain. For some functions, it is possible to achieve a better approximation of the analytical solution without increasing the number of grid points, but rather by changing their spacing within the domain from a linear spacing (the distance from one grid point to the next is defined by a constant* increment*) to a logarithmic spacing (the distance from one grid point to the next is defined by a constant* factor*). For example, a function that benefits from integration using log-spacing is:*

$$\int_1^{10} \sqrt{\log_{10}(x)}\, dx.$$

Analytically:

$$\int_1^{10} \sqrt{\log_{10}(x)}\, dx = 10 - \frac{1}{2}\,\mathrm{erfi}(\ln(10)) \sqrt{\frac{\pi}{\ln(10)}} \approx 7.2105\ldots$$

Integration point (x_i)	Function value $\sqrt{\log_{10}(x)}$	dx (linear)	$\sqrt{\log_{10}(x)}\, dx$
1.25	0.311303731	0.5	0.155652
1.75	0.492988893	0.5	0.246494
2.25	0.593449676	0.5	0.296725
…			
…			
9.75	0.994487	0.5	0.497244
			+
		Integral	7.223931

Integration point (x_i)	Function value $\sqrt{\log_{10}(x)}$	dx (logarithmic)	$\sqrt{\log_{10}(x)}\, dx$
1.06605	0.166667	0.136464	0.022744
1.211528	0.288675	0.155086	0.044769
1.376857	0.372678	0.17625	0.065684
…			
…			
9.380419	0.986013	1.200775	1.18398
			+
		Integral	7.209009

Here, the linearly spaced numerical integral gives an error of $\sim 0.18\%$ while the log-spaced integral gives an error of $\sim 0.02\%$, which is an order of magnitude less.

A.2.4 Summarizing

In order to numerically integrate a given function, three fundamental choices need to be made:

1. Which domain should be considered (i.e. where should the integration grid start and stop)?
2. How many grid points are required to adequately approximate the true value?
3. How should the grid points be spaced within the domain to adequately approximate the true value?

In many cases, the answers to these questions are not immediately obvious, and require careful testing and the analysis to be well-defined.

Appendix B: Follow-up actions external review

Action No.	Priority	Find and Action	Comment	Completed / to-do / no follow up
[RA-1]	High	Align the specified version numbers for external packages across the packages and environments of the system and implement mechanisms to exert control over these versions in the future.	-	Completed
[RA-2]	Medium	Refactor the framework classes in the ssm package to clearly separate code for different versions of the models. Similar refactorings are recommended for the Reader class in the ha zard_risk_prep package.	Classes are separated but follow the following philosophy: inherit as much as possible. Classes inherit from parent classes whenever possible for two reasons: 1) avoid duplicate code, 2) make explicit how models overlap	No immediate follow up
[RA-3]	Low	Refactor the class design for assessments in the hazardintegrator and riskintegrator packages with subclasses.	Matter of style, but no inherently quality change.	No immediate follow up
[RA-4]	Low	Deduplicate the SiteResponseRegionsData classes present in both the hazardintegrator and riskintegrator packages by moving one copy to the chainutils package.	-	Planned for development cycle 2021
[RA-5]	Low	Remove the obsolete code module json_generator_logictree.py in the riskintegrator package, or document its use.	Will be documented	Planned for development cycle 2021
[RA-6]	Low	Move each class into a separate code file to improve transparency of the codebase.	Not recommended practice (not part of PEP). Matter of style.	No immediate follow up
[RA-7]	Low	Consider refactoring more non-OO code into the OO design.	The framework is specifically modular to allow multi-paradigm programming. Within	No immediate follow up

			modules, the coding style is consistent.	
[RA-8]	Medium	Use a static analysis tool (such as pylint) in the local development environment and as a quality gate on the continuous integration server to be warned about violations of standards in the code.	We opt to use analysis provided by our IDE (PyCharm).	No immediate follow up
[RA-9]	Low	Review the appropriate use of exception types, in particular instances of NotImplementedError.	-	Planned for development cycle 2021
[RA-10]	Medium	Provide docstrings for classes detailing the class attributes and methods.	-	Planned for development cycle 2021
[RA-11]	Low	Use the docstring of methods to detail the dimensions of multidimensional numpy arrays where applicable.	This can be useful in some cases, but in practicality has never proven useful for this project and is error-prone.	No immediate follow up
[RA-12]	Low	Review the naming of variable, class, and function names and align them with Python standards.	-	Planned for development cycle 2021
[RA-13]	Low	Review cases of code duplication and, where appropriate, factor out common code to classes or methods.	Continued point of attention, but can always improve.	No immediate follow up
[RA-14]	Medium	Review reported cases of high complexity and, where appropriate, refactor the code to a more modular structure.	Continued point of attention, but can always improve.	No immediate follow up
[RA-15]	Low	Review class structures for use of public and private methods, as well as further opportunities to modularise the code through use of inheritance.	In Python, nothing is truly private. No benefit expected in changing the approach here	No immediate follow up
[RA-16]	Low	Review hardcoded parameters and move those that are subject to potential future change to configuration files.	Parameters are only hardcoded when they are part of a model specification	No immediate follow up
[RA-17]	Low	Document the expected format of the input and output files, including the expected physical units,	-	Planned for development cycle 2021

		see Table 6-1 and Table 6-2.		
[RA-18]	Low	Add units to the headers of the input files defined in Table 6-1.	-	Planned for development cycle 2021
[RA-19]	Medium	Implement validation of expected type and unit of data in methods that load the input files, raising exceptions early in case of unexpected data. This would also address all cases where data is loaded by an assumed column order.	-	Planned for development cycle 2021
[RA-20]	Low	Review the source code for opportunities to implement classes for custom data types.	Not planned to offer support for custom data types. We are trying to be as 'main stream' as possible to avoid support issues.	No immediate follow up
[RA-21]	Low	Use pytest fixture methods more widely to reduce the size of some long (>100 lines) test methods.	Continued point of attention, but can always improve.	No immediate follow up
[RA-22]	Low	Refactor test methods that contain multiple test cases. These methods should be parameterized, with test cases as arguments.	-	Planned for development cycle 2021
[RA-23]	Low	Focus test cases more precisely on edge cases, rather than iterating over large grids of parameters.	-	Planned for development cycle 2021
[RA-24]	Medium	Move methods that are only called by unit tests from production code to the test modules.	-	Planned for development cycle 2021
[RA-25]	Medium	Improve test coverage where there are notable gaps, such as: the rupturemodel and chainutils repositories, and the CalibrationFramework BruteForce, HazardAssessment and RiskAssessment classes.	-	Planned for development cycle 2021
[RA-26]	Low	Improve test coverage for intermediate steps of calculations. Where this is not practical due to the structure of the targeted	Not always practically achievable.	No immediate follow up

		code, refactoring should be considered.		
[RA-27]	Medium	Review code comments of complex methods and document their origin in the commentary where possible.	-	Planned for development cycle 2021
[RA-28]	Low	Review the names of methods and ensure they accurately reflect the variable that is calculated or returned, in particular where the implementation of an equation in code differs from the documentation.	-	Planned for development cycle 2021
[RA-29]	Medium	Review and update the documentation where complex steps of the calculations are not described in full detail in the implementation boxes.	The code is written to be self-explanatory. Docstrings are there as a guide, not as main explanation. Complex calculations are implemented, which are inherently complex to understand.	No immediate follow up
[RA-30]	Low	Review exception handlers to ensure exceptions are consistently logged.	All exceptions are logged in the platform framework.	No immediate follow up