Horizon 2020 Programme

METIS

Research and Innovation Action (RIA)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 945121

Start date : 2020-09-01 Duration : 48 Months

Report and code developments for PSHA testing

Authors : Dr. Graeme WEATHERILL (HGF/GFZ), Fabrice Cotton (GFZ), Guillaume Daniel (EDF), Irmela Zentner (EDF)

METIS - Contract Number: 945121

Project officer: Katerina PTACKOVA

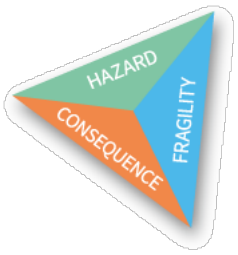
Document title	Report and code developments for PSHA testing
Author(s)	Dr. Graeme WEATHERILL , Fabrice Cotton (GFZ), Guillaume Daniel (EDF), Irmela Zentner (EDF)
Number of pages	88
Document type	Deliverable
Work Package	WP4
Document number	D4.5
Issued by	HGF/GFZ
Date of completion	2024-02-07 10:57:27
Dissemination level	Public

Summary

The use of probabilistic seismic hazard analysis (PSHA) as a means of quantifying the likelihood of exceeding levels of strong shaking at a site is widespread, and applications span a range of sectors in industry and society. Given the breadth of applications, seismic hazard modellers are often confronted with a need to demonstrate consistency between their models and observations of ground motion across a region. Though testing and validation of a probabilistic seismic hazard curve at a single site is not feasible with the limited time window of direct and indirect ground motion measures, clear and practical precedents can be found in the literature for quantitative comparisons of seismic hazard models and data when applied at larger regional scales. Here, observed exceedances of ground motion across many sites can provide a means to calculate the fit of models to data for return periods of engineering relevance. Implementation of existing methods in the literature can be challenging, however, owing to the complexities of modern seismic hazard models, limitations of the observed data, and a lack of available and transparent tools. The deliverable looks in detail at how we can quantitatively compare seismic hazard models for a region and how we can assess whether they are consistent with observed ground motions. This effort has lead to the development of PyPSHATest, an open-source Python toolkit for quantitative model-to-model and model-to-observation comparison probabilistic seismic hazard results (https://gitlab.pam-retd.fr/openmetis/pypshatest). The software features various modules for comparisons, both at the level of the components of a seismic hazard model and the outputs. We address the challenges in compiling data sets for comparison of hazard models against observed ground motions to data, focusing on the issue of observational completeness and how we can make inferences from the available data to fill in the gaps and thus ensure more robust comparisons. PyPSHATest is explained in detail and its usage illustrated using a real-world case study application to compare two existing PSHA models covering the territory of metropolitan France: Drouet et al. (2020) and the 2020 European Seismic Hazard Model (Danciu et al., 2021). These are initially contrasted against each other, using novel metrics to quantify the model-to-model divergence in distributions of seismogenic sources and resulting exceedance of ground motion. Then using a specially compile...

Approval

Date	By
2024-02-27 15:20:32	Dr. Marco PAGANI (GEM)
2024-02-27 15:28:00	Dr. Irmela ZENTNER (EDF)



METIS

Seismic Risk Assessment
for Nuclear Safety

Research & Innovation Action

NFRP-2019-2020

Developments and Tools for PSHA to Data Comparison

Deliverable 4.5

Version No 1

Authors: Graeme Weatherill (GFZ)

Fabrice Cotton (GFZ)

Guillaume Daniel (EDF)

Irmela Zentner (EDF)

This project has received funding from the Horizon 2020 programme under grant agreement

Disclaimer

The content of this deliverable reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

Document Information

Grant agreement	945121
Project title	Methods And Tools Innovations For Seismic Risk Assessment
Project acronym	METIS
Project coordinator	Dr. Irmela Zentner, EDF
Project duration	1st September 2020 – 31st August 2024 (48 months)
Related work package	WP 4
Related task(s)	Task 4.6
Lead organisation	GFZ
Contributing partner(s)	EDF
Due date	January 2024
Submission date	December 2023
Dissemination level	Public

History

Version	Submitted by	Reviewed by	Date	Comments
N°1	1	Luiz Alvarez Marco Pagani	29/11/2023	

Table of Contents

List of figures

List of tables

Summary

The use of probabilistic seismic hazard analysis (PSHA) as a means of quantifying the likelihood of exceeding levels of strong shaking at a site is widespread, and applications span a range of sectors in industry and society. Given the breadth of applications, seismic hazard modellers are often confronted with a need to demonstrate consistency between their models and observations of ground motion across a region. Though testing and validation of a probabilistic seismic hazard curve at a single site is not feasible with the limited time window of direct and indirect ground motion measures, clear and practical precedents can be found in the literature for quantitative comparisons of seismic hazard models and data when applied at larger regional scales. Here, observed exceedances of ground motion across many sites can provide a means to calculate the fit of models to data for return periods of engineering relevance. Implementation of existing methods in the literature can be challenging, however, owing to the complexities of modern seismic hazard models, limitations of the observed data, and a lack of available and transparent tools.

The deliverable looks in detail at how we can quantitatively compare seismic hazard models for a region and how we can assess whether they are consistent with observed ground motions. This effort has lead to the development of PyPSHATest, an open-source Python toolkit for quantitative model-to-model and model-to-observation comparison probabilistic seismic hazard results (https://gitlab.pam-retd.fr/openmetis/pypshatest). The software features various modules for comparisons, both at the level of the components of a seismic hazard model and the outputs. We address the challenges in compiling data sets for comparison of hazard models against observed ground motions to data, focusing on the issue of observational completeness and how we can make inferences from the available data to fill in the gaps and thus ensure more robust comparisons.

PyPSHATest is explained in detail and its usage illustrated using a real-world case study application to compare two existing PSHA models covering the territory of metropolitan France: Drouet et al. (2020) and the 2020 European Seismic Hazard Model (Danciu et al., 2021). These are initially contrasted against each other, using novel metrics to quantify the model-to-model divergence in distributions of seismogenic sources and resulting exceedance of ground motion. Then using a specially compiled database of ground motion observations using both accelerometer and broadband velocity meter records, a quantitative comparison of the models against observed data is undertaken. The potential for novel investigation into PSHA model testing is illustrated in the final chapter by comparing the model-to-data tests for the cases that ergodic regional scale models are used against those when partially-nonergodic models are available. This application demonstrates the complete testing workflow that can be implemented using PyPSHATest, both at a component and product level, and highlights how we can gain deeper insights into both the hazard models and data rapidly and transparently.

Keywords

Probabilistic seismic hazard analysis; testing; comparison; ground motion; PyPSHATest

1. Introduction

Effective mitigation of seismic risk begins with an understanding and the earthquake process and quantification of the likelihood of future events, and their potential impacts, into a framework for decision making. This process contains inherent uncertainty, however, both in the form of the intrinsic natural variability of the system, or *aleatory uncertainty*, and in our knowledge or understanding of it, the *epistemic uncertainty*. The framework for decision making in risk mitigation must therefore operate from a probabilistic perspective, one in which outcomes are specified in terms of probabilities of occurrence conditional upon the uncertainties that we aim to quantify. In the field of earthquake science and engineering probabilistic seismic hazard and risk analysis is the established framework to guide decision making among stakeholders ranging from engineers, insurers, government and many others. Yet while our modelling of seismic hazard and risk aims to integrate the uncertainties inherent in the earthquake system and the response of the built environment, there is also an expectation among stakeholders that the proponents or developers of probabilistic seismic hazard and risk models can demonstrate *at best* their validity, and *at least* show their consistency with current observations of earthquakes and their consequences.

1.1. Probabilistic Seismic Hazard Analysis

Probabilistic seismic hazard analysis is the means by which our knowledge of the earthquake process is translated into a usable framework for decision making by stakeholders. For a given measure of strength of shaking of ground motion (the ground motion *intensity measure*), such a peak ground acceleration (PGA) or the peak acceleration response of a single-degree-of-freedom oscillator with damping ξ and natural period T, PSHA returns the probability of exceeding a defined level of shaking (*intensity measure level*) at a location within a specified period of t years. Detail of how this is determined within a PSHA calculation will be shown in section 2. The core output of a PSHA calculation is the *seismic hazard curve*, which defines the probability of exceedance in t years of a vector of successively increasing levels of the selected intensity measure. From this a suite of products can be derived that can relate this information in a variety of different contexts to meet the needs of stakeholders. Furthermore, when we combine the probabilistic characterisation of ground shaking at a site with models of seismic fragility and vulnerability of a structure or structures situated there, the result is a similar curve indicating the probability of exceedance of a given level or damage or loss to said structure. PSHA is therefore a crucial input for analysis of seismic risk to the built environment.

An important aspect of PSHA in practice is the question of the spatial scale of the model. This is obviously highly dependent on the context of the application, but it is important to set out what we mean by different scales before embarking on the definition of testing PSHA. The smallest spatial scale is a site-specific analysis, which is usually conducted for engineered structures of high consequence of failure. State-of-the-art practice in seismic hazard modelling is using found at this scale, where intensive geotechnical and seismological investigation should yield a high level of detail about the site in question, while the high consequence of structural performance failure justifies intensive focus on the characterisation of the model inputs and uncertainties. As we will demonstrate shortly, however, validation of hazard at a single site using direct observations of ground motion is not achievable owing to the short window of observations at the site with respect to the recurrence of the low probability hazard considered for critical engineering applications. Successful efforts have been made in these cases to exploit the presence of fragile geological structures in order to provide constraints on the levels of shaking observed at these sites over geological timescales on the order of tens or hundreds of thousands of years (Brune, 1999; Baker et al. 2013).

The next scale to consider might be considered urban-scale or localised multi-site. Here one may wish to model seismic hazard for multiple sites located in close proximity to one another, such as for a transportation network or a multi-unit nuclear power plant facility. In this case one would expect a high level of geotechnical and geological information for the sites, though perhaps not so detailed as for a single site. These cases are similar to those of the single-site case with the notable exception that it may be necessary to account for spatial dependencies and correlations in the probabilities of exceedance of multiple sites. As with the site-specific case, validation of localised multi-site models is seldom possible via direct observations of ground motion.

The third spatial scale is that of regional scale PSHA, which might be undertaken for a specific province of a country, the entirety of a country, or even across multiple countries. Here seismic hazard is calculated usually for tens of thousands of sites, but for which measurements of the geophysical and geotechnical characteristics of the sites are available in very exceptional cases among those considered. Regional scale PSHA has a wide variety of applications, including the development of maps describing the expected levels of shaking for a seismic design code, along with estimates of the seismic risk for a geographically diverse portfolio of insured properties in a country (or countries). In contrast to site-specific cases the paucity of geotechnical information or seismological observations at the target sites limits the application of the most state-of-the-art techniques, and the need to characterise hazard at so many locations introduces a computational demand that often necessitates a simpler characterisation of models and model uncertainty compared to site-specific studies. However, regional scale PSHA allows use to consider exceedances of levels of ground motion over a large geographical area, increasing the possibility of using direct observations of ground motion to attempt quantify the differences between model and data. As regional scale PSHA can inform decision making in a legal context, it is here that the question of validation has often been applied.

1.2. An Overview of "Testing" PSHA

While the foundations of PSHA were laid in the 1960s with the pioneering work of Cornell (1968) and Esteva (1968), the means of confronting models against observed data took several decades to emerge, arguably owing to the limited number of sites worldwide with histories of recorded motions sufficient to capture multiple instances of significant acceleration against which the hazard curves could be tested. One of the first direct comparisons was from Ordaz & Reyes (1999), who plotted the observed rates of exceedance of ground motion at a single recording station in Mexico City against a corresponding probabilistic seismic hazard curve. We use the term *direct* comparison here to refer to comparisons of observed ground motions (as recorded by seismic instruments) against hazard curves, as opposed to *indirect* comparison, which refers to an inference of the ground motion to have occurred at a site from other observation such as macroseismic intensity (e.g., Stirling & Petersen, 2006; Rota & Rosti, 2017, Rey et al., 2018) or exclusively from a prediction of the distribution of ground motion using an empirical ground motion model (Ward, 1995).

While Ordaz & Reyes (1999) demonstrate a level of consistency between the observations at a site and the seismic hazard curve, this cannot necessarily for a general approach for "validation" of hazard at a specific location. From the basic implications of the Poisson process, Beauval et al. (2008) demonstrate that for a process with a rate of occurrence of 1 / year, a minimum time window of observation of 25 years would be needed ensure an estimated rate within 20 % of the "true" rate. For return periods of engineering relevance (on the on the order of 100 – 2500 years), one would need continuous observation of ground motion at a site on the order of thousands to tens of thousands of years to be able to constrain the true rate of occurrence to within the same 20 % uncertainty. Mak et al. (2014) reach a similar estimated same time-scale needed for "validation" of seismic hazard at a single location using the concept of statistical power. These results support the largely intuitive recognition that at estimates of recurrence of an earthquake process that operates on multi-century (or even multi-millennium) timescales cannot necessarily be strictly "validated" by direct observations of ground motion over a few decades, nor even that effectively by indirect observation over a few centuries.

But if direct (or indirect) observation cannot necessarily "validate" PSHA at a given site, to what extent can we make inferences about the PSHA models in general? Can it be shown in any quantitative way whether a given seismic hazard model for a region can be said to fit better or worse the observations within a region ? Ward (1995) posits that as seismic hazard maps make predictions of the levels of ground motions that would be exceeded with a fixed probability P across thousands of points within an area, then if P percent of the total sites see an exceedance of their target level of acceleration in t years, this can act as a validation on the model. The relation between the size of the area needed for validation of a given return period was later quantified by Iervolino et al. (2021).

While there is broad consensus that exceedance of ground motion at a single location cannot be validated by observations in the time frame available, aggregating observations across many sites to accumulate more data and increase the statistical power of comparisons opens the possibility of making more quantitative inferences on the "fit" of a model. This assumption forms the basis for many

subsequent studies that have been proposed methods to "test" seismic hazard models using direct or indirect observed data. Beauval et al. (2008), Stirling and Petersen (2006), Fujiwara et al. (2009) and Stirling and Gerstenberger (2010) all demonstrate the use comparisons against direct or indirect observations across multiple sites to make inferences on probabilistic seismic hazard models.

Albarello & D'Amico (2008) present a statistical framework for testing hazard models by comparing observations of exceedance of ground motion levels at multiple sites against predictions from a seismic hazard model for Italy. They later formalise this into a "model scoring" strategy for assessing relative performance of different PSHA models in a region (Albarello & D'Amico, 2015; Albarello et al., 2015). An important difference between their framework and those of previous studies is that observation is presented in terms of number of sites with one or more exceedance of a given level of ground motion, rather than the total number of exceedances over multiple sites in a region (e.g., Tasan et al., 2014). Both are valid forms of "aggregated hazard curve" but require different interpretations and statistical distributions for quantification. Mak & Schorlemmer (2016) explain the different forms of aggregated hazard curves clearly. Given their importance in the applications in this report a comprehensive overview of their formulation and its relation to that of Albarello & D'Amico (2008) is provided in Section 4 of this report.

Between 2012 and 2016 the seismological community saw a flurry of articles published addressing the need for testing PSHA (Stein et al., 2011), its relation to the context and developments of PSHA (Stirling, 2012), alongside example applications (Mezcua et al., 2013; Tasan et al., 2014; Mak & Schorlemmer, 2016). Since this period, however, although we did not see a wide expansion of systematic PSHA testing, some notable developments have emerged. Recognising that modern PSHA describes hazard not only in terms of a single curve relating ground motion intensities to their corresponding probabilities of exceedance but rather a suite of curves characterising the epistemic uncertainty, Marzocchi & Jordan (2014) present the theoretical concept of testing hazard models for ontological errors, i.e., errors in the model's quantification of aleatory variability and epistemic uncertainty, which influences how observed exceedance should be compared against a predicted probability of exceedance in the form of probability distribution. This was later developed into a broader framework for application to multi-site PSHA in Marzocchi & Jordan (2018).

Several studies have explored the implications of one of the critical assumptions in the use of aggregating hazard curves for testing, which is that of independence in the probabilities of exceedance at each location. In reality, the probability of exceedance of ground motion at a given site is not necessarily conditionally independent of that at a neighbouring site given that both will be affected by strong shaking during the same event. There also exists spatial correlation in the ground motion variability between two or more sites, which decays with increasing separation distance. But while the statistical framework for comparing observed and predicted exceedances has required independence, Iervolino et al. (2017) explored the effect of spatial dependence when testing hazard in a multi-hazard context, and Albarello & Peruzza (2017) proposed an adaptation to the method of Albarello & D'Amico (2008) to account for this correlation.

More recently, the question of testing PSHA has re-emerged with the publication of multiple national scale models (e.g., Wiemer et al., 2016; Grünthal et al., 2018; Drouet et al., 2021; Mosca et al., 2021, Meletti et al., 2021) and alongside the latest generation European Seismic Hazard Model (ESHM20) (Danciu et al., 2021). As these new models update existing ones for application in their respective countries, and alongside the existing of the pan-European model that may give a different estimate of hazard and its uncertainty in these same regions, hazard modellers are once again being confronted with the need to demonstrate "validity" and make quantitative comparisons of model performance. Iervolino et al. (2023) utilise the approach of counting exceedances of ground motion at strong motions stations in Italy to test the existing national seismic hazard model for Italy MPS04, its proposed update MPS19 (Meletti et al., 2021) and the ESHM20. A similar development has taken place in New Zealand as part of the update to their national seismic hazard map (Stirling et al., 2022). These studies show that there exists a continuing need to apply quantitative procedures to compare probabilistic seismic hazard against observations, but with this comes the need for careful scrutiny of the process, especially if these comparisons are intended to inform legislative decisions arising from the updated models, such as new seismic design codes.

1.3. PyPSHATest: Tools for Quantitative Comparison of PSHA Models

1.3.1. Limitations of Testing PSHA in Practice

The overview of PSHA testing provided in the previous sub-section has not yet addressed some of the limitation and challenges of applying testing methods in practice. Many of these studies limit themselves to comparison against a single hazard model or mean seismic hazard map (e.g., Stirling & Gerstenberger, 2010; Tasan et al., 2014; Mak & Schorlemmer, 2014), or to a simplified form of the model's original logic tree (Iervolino et al., 2023). In modern PSHA a substantial amount of time is dedicated to the characterisation of epistemic uncertainty, and recent developments in seismogenic source and ground motion modelling are carefully adapted into practical logic tree frameworks. Neglect or adaptation of the epistemic uncertainty can lead to erroneous conclusions about the suitability of a PSHA model for application. Although Marzocchi & Jordan (2014; 2018) describe how the testing approach can be applied to the complete representation of epistemic uncertainty, we are unaware of an illustration of this in practice.

This brings us to the second limitation, which is an apparent lack of reproducibility in the testing process. In none of the aforementioned literature has scientific code for implementation of the tests been made public or demonstrated to have undergone a level of quality assurance. Though in some cases one might argue that the computation is trivial (simply counting exceedances), in the process of assimilating ground motion data, seismic hazard data, assessing predicted and observed exceedance and applying statistical distributions to define tests, there is ample scope for error and or misinterpretation of the methodology described in the respective publication. This highlights a significant shortcoming of hazard model "validation", which is a lack of verification in the methodology for testing. As considerable efforts have been undertaken in recent years toward verification of PSHA models themselves (e.g., Thomas et al., 2010; Hale et al., 2018), a reference implementation of the testing methodology that can be openly scrutinised itself would be welcome.

The last major limitation that we encounter in applying these methods in practice is in the compilation of the observed data for testing itself. Indeed, compilation and harmonisation of observational data and PSHA models accounts for the greatest proportion of time and effort in the testing process, to the extent that it can be too costly for hazard modellers to expend effort and resources on testing during the model development. When we are using observational data as a means of quantitative validation of PSHA models we assume that the data itself is sufficiently complete for the context of the test and fully harmonised. Completeness here means that the database of observations is a faithful archive of what has happened in the period and region of interest, without significant gaps. Harmonisation requires that the data are uniformly processed, without apparent spatial or temporal bias, in a manner that the measures for testing derived from it are concordant with those being described by the seismic hazard model. When gathering an archive of processed ground motion recordings across a region we are confronted with significant incompleteness as recording stations suffer outages or ground motion recordings can be too poor quality to derive accurate measures of shaking intensity. This is not a trivial problem, and as resolution of incompleteness problems is a major element of this work we explore this in detail in Section 3 of this report. Within the aforementioned studies only Tasan et al. (2014) address the completeness problem in detail and present a methodology to account for it, although it is also acknowledged but not explicitly addressed by Iervolino et al. (2023).

1.3.2. Objectives of a Toolkit for PSHA Model Comparison and Testing

The limitations of testing PSHA in practice highlighted above form the main motivation for developing a new toolkit for quantitative comparison of PSHA models, both from model to model and between model and data. We do not approach this from the perspective an "independent tester", but rather as a "model developer". The tester may see only the hazard outputs as forecasts to be validated or invalidated irrespective of the advances contained within a model itself. The model developer is instead focused on the advances of the model. For them, confrontation of a model against data should form just one step of the model development process, checking for sanity and checking for consistency as part of a sequence of iterations toward improving the hazard model. To meet their needs, we have broadened

our scope beyond comparisons of statistics of exceedance between hazard model and direct observations of ground motion, but also to look at quantitative assessment of the model components and to make comparisons of different PSHA models themselves irrespective of observational data. Moreover, as we are mindful of the efforts required in the testing process, we aim to make these steps easy to implement, flexible to configure, and dependent upon products and information that is generally compiled in the model development process.

PyPSHATest is developed to fulfil the objectives we have set out above. First and foremost, it is open-source, written in Python (with some R) and distributed under a GNU General Public License v3.0. It is free to access and download from https://gitlab.pam-retd.fr/openmetis/pypshatest. The toolkit leverages upon OpenQuake, an open-source probabilistic seismic hazard and risk calculation software (Pagani et al., 2024), not only as a basis for running PSHA calculations but also to capitalise on many of the functionalities within its own Python modular library of earthquake hazard and risk related functions to implement the features contained in PyPSHATest. The tools themselves within PyPSHATest are designed with modularity in mind, meaning that contain classes and methods for individual steps of the process, which the user can configure, adapt and connect to create customised workflows. We will demonstrate in this report how these modules fit together for the main PSHA testing workflow, but most can be used independently to address different problems. Although there are additional software dependencies needed for PyPSHATest, we would aim for installation and deployment to be straightforward for any PSHA modeller who is themselves already working in with OpenQuake. We do not require that all hazard models are run with OpenQuake and will continue to look to expand to support outputs of PSHA models implemented in different software. However, OpenQuake's structuring of the hazard components does provide a clear and effective template around which to develop a standard or generalisable representation of seismic hazard inputs and outputs. This will be explained in more detail in section 2.

A general overview of the structure and features of PyPSHATest can be seen in Figure 1. We divide the tools into three different categories: 1) those for handling data inputs, 2) those for methods needed to prepare the observed data and models for quantitative comparisons, and 3) those that implement the comparisons themselves. A comprehensive PSHA model to data comparison study requires three inputs: an earthquake catalogue, a database of ground motions and associated metadata, and a set of one or more PSHA models. While earthquake catalogue processing is not a primary goal of this toolkit, we encounter many parts of the comparison process where manipulation of the catalogues is necessary. Useful features added into merging of harmonised earthquake catalogues into the metadata of the ground motion database and simple declustering (although we can support integration of outputs from other declustering software, such as those found within OpenQuake's Hazard Modeller's Toolkit). Next, we have tools for loading the database of ground motions in the form a tabular flatfile, which can be quantitatively compared against ground motion models in OpenQuake to retrieve values of random effects of residuals of the observed ground motions against the models. This is useful as a standalone tool for exploring the fit of ground motion models to data, but more importantly it is a critical input into one of the most innovative features of the toolkit, which is *data imputation by mixed effects regression*. This feature allows us to compensate for incompleteness in the ground motion data, leveraging on inferences about the characteristics of ground motion from each event and each station represented in the ground motion data. Finally, we have tools for managing comprehensive seismic hazard models featuring large logic tree. PyPSHATest can import the critical seismic hazard information from an OpenQuake hazard model, and has features that allow the user to easily compare the distributions of hazard from different models calculated at the same sites. Keeping with the objectives of component-level testing of the models, the hazard tools allow the user to explore and quantify differences in distributions of seismogenic source models, and has interoperability with PyCSEP (Savran et al., 2022), an open-source software for seismicity forecast testing implementing the state-of-the-art methods proposed by the Collaboratory for the Study of Earthquake Predictability (CSEP, Schorlemmer et al., 2018).

The central component PyPSHATest is the "Station and Hazard Database", a single datastore in which gradually all the necessary information needed for model to data comparison is assimilated. This datastore is built in several stages and connects all observations of ground motions across the sites found in the ground motion database, their corresponding mixed effects residuals with respect to the user's choice of ground motion models, and the expected seismic hazard at the sites from different models. The datastore itself is an organised high-density binary (hdf5) file, which is structured according

to the schema shown in Figure 2. With seismic hazard data and observation combined in this file, PyPSHATest boasts a series of tools that can be used to compare models, featuring methods such as those found in Mak & Schorlemmer (2016) and Albarello & D'Amico (2008). Many comparison modules here also contain functions to assist in visualising the comparisons, to help interpret and draw inferences from the analysis.

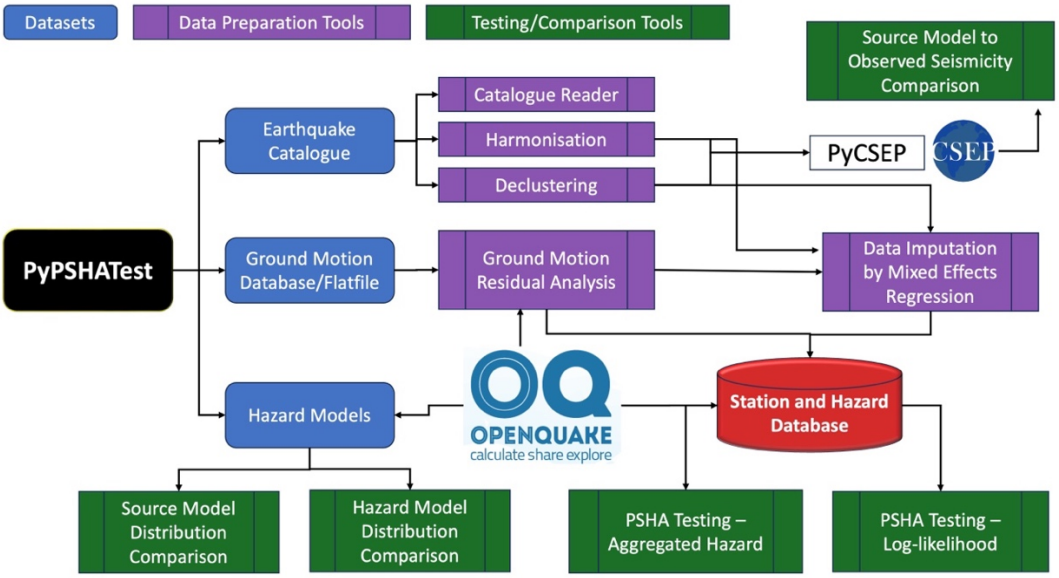


Figure 1: Overview of PyPSHATest Tools and Workflow. Blue rounded boxes indicate inputs, purple boxes the respective tools and green boxes the outputs/applications

1.3.3. Structure of The Report

This report is intended to give an overview of PyPSHATest, addressing section by section some of the potential needs and challenges in model and data comparison and illustrating how these can be approached using PyPSHATest's tools. In Section 2 we begin with an overview of PSHA, including the structure and definition of the inputs that introduces terminology we use throughout the report. We then focus on how hazard information is represented and stored in OpenQuake, which forms the basis for the hazard model assessment tools. We then show how we can assimilate seismic hazard information and perform a set of quantitative comparisons between one model and another in the context of epistemic uncertainty, both at the level of the seismogenic source distribution and the PSHA output distribution. These tools are illustrated with examples from seismic hazard models in France (Drouet et al., 2020) and Germany (Grünthal et al., 2018), alongside the ESHM20.

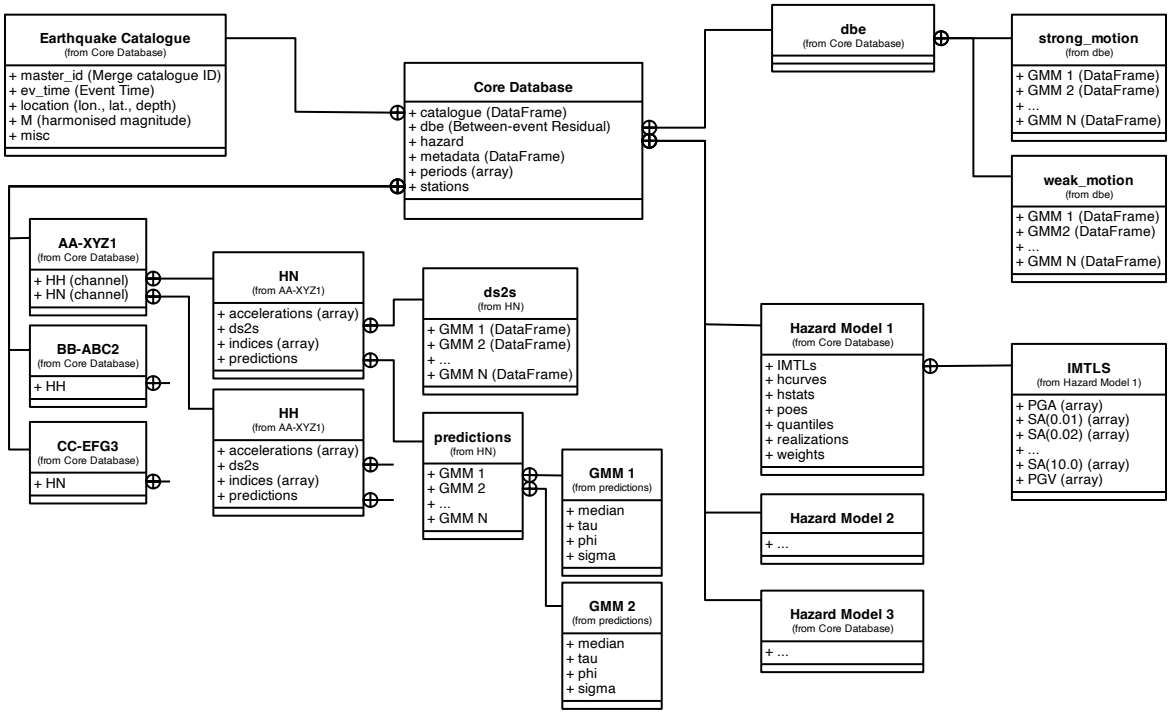


Figure 2: Schema of the code Station and Hazard Database that forms the central component of the PyPSHATest toolkit

In Section 3 we move onto the process of managing direct observations in the form of databases of ground motion data, and we focus on how we can address the completeness issue using a method we term data imputation by mixed effects regression. In the PyPSHATest examples we will show how to load ground motion data from a flatfile, explore the comparisons with GMMs to create databases of residuals and explore them prior to the imputation. Here we illustrate the usage of the toolkit using a combined weak and strong motion ground motion flatfile for northwest Europe. This case study is chosen specifically to illustrate how PyPSHATest can be used to explore ground motion residual trends both separately and together for weak and strong motion data, which can allow for expansion of the data available for quantitative comparison of PSHA models against data, especially in low to moderate seismicity areas.

Section 4 presents the tools for making quantitative comparisons between PSHA models and observed ground motions, showing the general workflow beginning with the data imputation and proceeding with analysis of statistics of exceedance from aggregated hazard curves and log-likelihood analysis. Each step is illustrated in turn, but we also go into detail to show a case study application of the tools to address the question of whether the use of partially and/or fully non-ergodic ground motions have an influence on model to data comparisons. We conclude the report in Section 5 with a brief recap of what we aim for the tools to achieve before then addressing questions that we believe can provide insights to theoretical and practical issues for undertaking quantitative comparisons.

Throughout the report we will be illustrating how to call the Python functions from PyPSHATest with snippets of code. Boxes aligned to the left margin indicate code snippets and are not labelled as figures. Example output images and explanatory diagrams are labelled as figures and centred in the margins in the conventional manner. With the toolkit itself, an additional set of Jupyter Notebooks is included that illustrate the functionalities and show how to combine them into the workflows explained here. Online documentation of the models and their API is under construction and will be made available in due course. As an open-source code PyPSHATest is in active development and new features, modifications, fixes etc. will be added from time to time. We encourage users to test these features out and provide feedback or report problems via the available "Issues" in the online code repository

2. Preparing and Analysing Data for Seismic Hazard Comparisons 1: Seismic Hazard Models

2.1. Structure of a PSHA Model

The central product of a probabilistic seismic hazard analysis is the ***seismic hazard curve***, which defines the probability P of exceedance (or annual frequency of exceedance) of successively increasing ground motion ***intensity measure levels (IMLs)*** in a given time period of t years. In the conventional formulation of Cornell (1968) and McGuire (1976) this is determined via evaluation of the integral:

$$\lambda(a \geq A) = \sum_{i=1}^{N_{SRCS}} \lambda_{m>M_{min},i} \int_M \int_R P[a \geq A|m, r, \boldsymbol{\theta}] \cdot f_{M_i}(m) \cdot f_{R_{ij}}(r|m)\, dm\, dr \qquad 2.1$$

Where $\lambda(a \geq A)$ is the rate of exceedance of ground motion level A at a given site, $\lambda_{m>M_{min},i}$ the rate of occurrence of earthquakes above some minimum magnitude M_{min} from the i^{th} seismic source of N_{SRCS} seismic sources, $f_{M_i}(m)$ the probability density function of magnitude for the i^{th} seismic source, $f_{R_{ij}}(r|m)$ the probability density of earthquake rupture to site distance r, and $P[a \geq A|m, r, \boldsymbol{\theta}]$ the probability of ground motion at a site exceeding A given the occurrence of an earthquake of magnitude m at distance r and any specific source, path or site properties $\boldsymbol{\theta}$. For time-independent PSHA, the probability of exceedance, in a given time period $P(a \geq A|t)$ is usually derived from $\lambda(a \geq A)$ using the Poisson assumption:

$$P(a \geq A|t) = 1 - e^{-\lambda(a \geq A) \cdot t} \qquad 2.2$$

This formulation of PSHA has been present in many existing seismic hazard software for several decades; however, more recent software have adopted an alternative formulation around the concept of the ***earthquake rupture forecast (ERF)*** (Field et al., 2003; Pagani et al., 2014).

$$P(a \geq A|t) = 1 - \prod_i^{N_{SRCS}} \prod_j^{N_{RUP,i}} \left(1 - P_{rup_{ij}}(n \geq 1|t)\right)^{P\left(a \geq A|rup_{ij}\right)} \qquad 2.3$$

Where $P_{rup_{ij}}(n \geq 1|t)$ is the probability of $n = 1$ or more occurrences of rup_{ij}, which is the j^{th} of $N_{RUP,i}$ ruptures generated by the i^{th} seismic source of N_{SRCS} seismogenic sources, and $P\left(a \geq A\middle|rup_{ij}\right)$ the probability of exceedance of ground motion at a site given the occurrence of the earthquake corresponding to rup_{ij}. In the case that $P_{rup_{ij}}(n \geq 1|t)$ is time-independent and follows a Poisson distribution, the two formulations of the PSHA calculation in equations 2.1 and 2.3 are equivalent. The *earthquake rupture forecast* is the complete inventory of ruptures generated by all sources rup_{ij} and their corresponding probabilities of occurrence in a given time period. $P[a \geq A|m, r, \boldsymbol{\theta}]$ and $P\left(a \geq A\middle|rup_{ij}, \boldsymbol{\theta}\right)$ are the same and come from the ***ground motion model*** **(GMM)**. For the purposes of comparing and testing PSHA against observational data there is no difference in terms of the output seismic hazard curves; hence, no special consideration is needed depending on which method is used to generate $P(a \geq A|t)$.

From the representations of probabilistic seismic hazard described above, it is clear that all PSHA models comprise several key components:

1) The ***seismogenic source model (SSM)*** describes the complete set of potential seismogenic sources in a region, their geographic location, rupture geometry (where necessary) and magnitude frequency distribution, comprises both a rate of occurrence of earthquakes above M_{min} and the magnitude probability density $f_{M_i}(m)$. In addition to M_{min}, the earthquake source

must also contain an upper bound magnitude M_{max}. The *earthquake rupture forecast* is constructed given the parameterisation of the seismogenic source model.

2) The ***ground motion model*** **(GMM)** describes the probability of exceeding ground motion at a site given the occurrence of an earthquake of magnitude, m, at distance r from the site. This assumes the form of a lognormal distribution with expected value $\mu(m,r,\boldsymbol{\theta})$ and standard deviation σ_T such that $\ln Y \sim \mathcal{N}(\mu,\sigma)$, where Y is the ground motion intensity measure level of interest. Traditionally the *ground motion model* has taken the form of a parametric ground motion prediction equation (GMPE) that describes μ and σ using mathematical expression dependent on magnitude, distance and collection of parameters $\boldsymbol{\theta}$ describing additional properties of the earthquake source (e.g., depth, style of faulting, dip, width etc.) and those of the target site (e.g., shearwave velocity averaged over the upper 30 m of the crust [V_{S30}], depth to a given shearwave velocity later [Z_{V_S}], etc.). The functional form of the GMPE is based on the fundamental physics of seismic wave propagation from a source, while its coefficients are usually calibrated by regression against observed (or simulated) ground motion data. GMPEs are the most ubiquitous form of GMM, some recent PSHA studies have adopted non-parametric forms of GMM such as tables describing μ and σ for a comprehensive range of scenarios, while the possibility of generating the density function for $P[a \geq A|m,r,\boldsymbol{\theta}]$ directly from suites of physics-based ground motion simulations as a form of GMM is currently being explored (e.g. Milner et al., 2021).

3) The ***site model*** describes the local characteristics of the site in question, which may in the simplest case consist of one or more site properties (such as V_{S30} or Z_{V_S}) that are input into the GMM. For detailed site-specific PSHA studies, the site model may instead take the form of a comprehensive description of the amplification of ground motion with respect to a reference seismic bedrock, and this amplification is usually dependent on the frequency of the shaking, amplitude of the bedrock ground motion, and potentially the properties of the seismic source and path.

4) The ***configuration***, which describes the comprehensive set of parameters and user inputs that control the execution of the PSHA in order to generate the hazard curve inside a PSHA calculation software. This generally contains a miscellanea of parameters controlling, for example, target levels of ground motion for calculation of the hazard curve, discretisation of seismogenic source properties and probability distributions, truncations to the distributions etc.

The PSHA formulations in Equations 2.1 and 2.3 and the components described above yield a single seismic hazard curve, one that integrates over probability distributions of magnitude, source-to-site distance and ground motion that describe the inherent ***aleatory variability*** of the earthquake process. It is now well established in PSHA practice that uncertainties on the probability distributions themselves should be incorporated into the analysis. These are referred to as ***epistemic uncertainties*** and relate to both the choice of model (e.g. the choice of GMM, seismogenic source model, magnitude frequency distribution, local soil amplification etc.) and/or the uncertainty on its parameters (e.g. uncertainty on M_{MAX}, standard deviations σ_a and σ_b of the Gutenberg-Richter distribution parameters a and b respectively). To incorporate such uncertainties into PSHA it is common practice to adopt a ***logic tree***, a framework that describes for each component of the PSHA model a set of alternative models (or alternative parameterisations of a model or models) and their associated weights. A specific model and weight is referred to as a logic tree *branch,* and the complete suite of branches (and weights) representing the epistemic uncertainty on a single component of the model is termed a *branch set*. Evaluation of the *logic tree* yields a suite of $\mathcal{H}$ seismic hazard curves, each curve h corresponding to a specific *logic tree path* (i.e. specific combination of branches from all of the model components represented in the logic tree). $P(a \geq A|t)$ therefore takes the form of a distribution, which we typically represent via an expectation (i.e. "mean" hazard) and a set of quantiles (e.g. 16th, 50th (median), 84th etc.). As the branch sets within a logic tree may represent alternative models or parameter uncertainties (or both) the $\mathcal{H}$is not necessarily well modelled by any specific probability distribution, hence the comprehensive set of curves $P_h(a \geq A|t)$ and their weights w_h for $h = 1, 2, \ldots, N_h$ end branches of the logic tree are considered to represent $\mathcal{H}$ in a non-parameteric form subject to the condition $\sum w_h \equiv 1$. Throughout the rest of the report when we refer to a *PSHA model* we will be referring to the complete logic tree $\mathcal{H}$.

Although the seismic hazard curve is the central product of PSHA, a comprehensive seismic hazard study should yield a suite of outcomes derived from the hazard curves. When considering ground motion in terms of spectral acceleration (Sa) across a set of periods T, individual hazard curves are produced for ground motion at each spectral period T. The vector of spectral accelerations for the set of periods T, i.e., Sa(T), corresponding to a specified probability of exceedance P in time t is referred to as the *uniform hazard spectrum*, a representation of the ground motion relevant for application in seismic engineering design. Similarly, for a given model we may wish to determine seismic hazard across a region, which will comprise a potentially large number of sites S. For a given ***intensity measure type*** (i.e., peak ground acceleration [PGA], peak ground velocity [PGV], response spectral acceleration at period T [Sa(T)]), one may wish to compile the spatial vector of *intensity measure levels* corresponding to a specific probability of exceedance P in time t across all sites S in a region. This is a ***seismic hazard map***, which provides a representation of the spatial variation in seismic hazard across a region. Seismic hazard maps serve a variety of purposes, and while they are insufficient as a basis for site-specific characterisation of hazard for design of structures of high importance, they may form the basis for describing the level of seismic acceleration as input to design of ordinary buildings within a national or regional seismic design code.

In addition to the seismic hazard curve(s) and its (their) derivative products, other important information relevant to describe the earthquake hazard at a site can be extracted from the evaluation of the PSHA integral. For the purposes of comparing probabilistic hazard against observation few of these additional products are used directly in this analysis, and we instead refer the reader to the excellent overview of various seismic hazard products provided by Baker et al. (2021) and also presented in other technical reports within the METIS project. Two products are particularly noteworthy, however:

Disaggregation returns the probability that ground motion exceeding A in time t come from rupture rup_{ij}:

$$P\left(rup_{ij}\middle|a \geq A\right) = \frac{P\left(a \geq A\middle|rup_{ij}\right) \cdot \lambda\left(rup_{ij}\right)}{\lambda(a \geq A)} \qquad 2.4$$

This result is usually determined across all ruptures in a PSHA calculation and aggregated into bins of distance r, magnitude m, and ε number of standard deviations of ground motion above/below the median value μ. In some software, bins may also extend to other parameters of the hazard or even to geographic grid cells. Disaggregation can quantify of the relative contribution of different earthquake scenarios to the seismic hazard, which may have many uses in application for selecting ground motion records representative of the seismic hazard at a site. This is also a key requirement for calculating ***conditional spectra*** (Baker, 2011; Lin et al., 2013).

Stochastic Event Sets (SES) and ***Ground Motion Fields (GMFs)*** are not direct products from a PSHA when executed using either of the formulations in Equations 2.1 and 2.3. However, they can be extracted from the exact same SSM and GMM inputs and can be used to provide an evaluation of the seismic hazard in an alternative manner. This can be useful for testing purposes as we shall see in later sections of the report. Stochastic Event Sets are synthetic realisations of seismicity for a given time period, t, produced by random sampling of either the seismic source model or the earthquake rupture forecast. Assuming the ERF formulation of hazard (Equation 2.3), a synthetic realisation of seismicity from a given SSM or ERF is generated by sampling $P_{rup_{ij}}(n \geq 1|t)$ for each rupture. For statistical analysis we would commonly want to create large numbers of event sets to generate N_{SAMP} realisations of seismicity in time t from a given source model. Similarly, for a given rupture we may wish to generate a one or more realisations of ground motion at S sites. This can be done by sampling the GMM, which involves drawing random samples of the distribution $\ln Y = \mathcal{N}\left(\mu(m,r,\boldsymbol{\theta}),\sigma(m,r,\boldsymbol{\theta})\right)$ for each site to produce a stochastically sampled *ground motion field* from the rupture. An important feature of the SES calculations is that if an appropriate spatial (cross-) correlation model is selected then site-to-site correlation in the residuals of the ground motions can be incorporated into the probabilistic hazard and risk analysis. This feature will be exploited when we discuss the issue of spatial dependence in seismic hazard and observations later in this report.

From the definition of both the SES and GMFs, one can intuit how PSHA can be computed in a manner based on Monte Carlo sampling, which is referred to as ***event-based PSHA***. Using the same SSM and GMF as for the ***classical PSHA*** calculation (Equation 2.1 or 2.3), one can generate multiple (N_{SAMP})

synthetic catalogues of duration *t*, i.e. the SES, and for each rupture in each SES the ground motion at S sites is sampled to create a corresponding GMF. If we define the *effective* duration of a SES, t_{eff}, as the product of the number of synthetic catalogues in an SES and their duration t, and the total number of ruptures from all event sets as N_k, then the rate of exceedance of ground motion at a given site s of S sites is determined via:

$$\lambda_s(a \geq A) = \frac{\sum_{k=1}^{N_k} I\left(a \geq A \middle| rup_{ij}, s\right)}{t_{eff}} \tag{2.5}$$

Where $I\left(a \geq A \middle| rup_{ij}, s\right)$ take the value of 1 if the logical term evaluated by $I(\cdot)$ is true, or zero otherwise. In the case of time-independent PSHA, $P(a \geq A|t)$ can be evaluated from the same Poisson formula in equation 2.2, and if t_{eff} is sufficiently large then the resulting probabilities of exceedance in the seismic hazard curves will converge with those determined from classical PSHA.

2.2. OpenQuake Engine

The general structure of a PSHA calculation given in Section 2.1 is intended to introduce terminology and concepts that will be adopted throughout the rest of the report. These are not necessarily intended to be specific to any given PSHA software, albeit that some of the formulations are presented are consistent with those adopted for OpenQuake (Pagani et al., 2014) and OpenSHA (Field et al., 2003). For many of the PSHA comparison tools contained within PyPSHATest no specific PSHA calculation software need be assumed, and the outputs of any such calculations can be used *provided that they are implemented with the required formats*. However, as PyPSHATest is a Python module that exploits several features offered by the OpenQuake libraries, it is first and foremost intended for compatibility with the OpenQuake-engine software and can capitalise on many features offered by the software. These include standardised file formats for input and outputs for SSMs, SES, GMFs, logic trees etc. using OpenQuakes Natural Risk Markup Language (nrml), all the way through to components of its internal calculation storage (in the form of a structured hdf5 binary file) to retrieve other products from the calculation. In addition, OpenQuake's internal libraries provide an extensive set of modular functions that allow an advanced user to develop highly customised calculation workflows that extract certain functionalities from the PSHA calculations. As OpenQuake is therefore embedded in the tools deeply, we will for much of this report assume that seismic hazard is computed using OpenQuake, and this is the case for all case study applications herein.

With OpenQuake forming our reference hazard tool, we shall elaborate in further detail the structure of OpenQuake's inputs and outputs. The overall schematic of the complete suite of inputs to an OpenQuake seismic hazard calculation is given in Figure 3 (adapted from Weatherill et al., 2023a). Working from left to right, the top-level input is the ***Configuration*** file, which usually takes a variant of the name *job.ini* or similar. This is a simple text file in which the user sets all the parameters that control the general operation of the earthquake hazard and/or risk calculator. This includes the type of calculation to be run, the configuration of the target sites for the calculation (including their location and site properties), the location of the files containing the logic tree for the source model and for the ground motion model, the intensity measure types (PGA, Sa, etc.) and their corresponding vectors of intensity measure levels (for which the probabilities of exceedance will be calculated), the selected outputs and a miscellanea of other parameters that control how the calculations execute. Note that depending on the seismic hazard model in question, the target sites might be a single site, a regular grid of sites with common properties, or a collection of sites each with its own site properties. In the case of the latter, site information is supplied in a separate file (the *site model file*), not shown in the schematic in Figure 3.

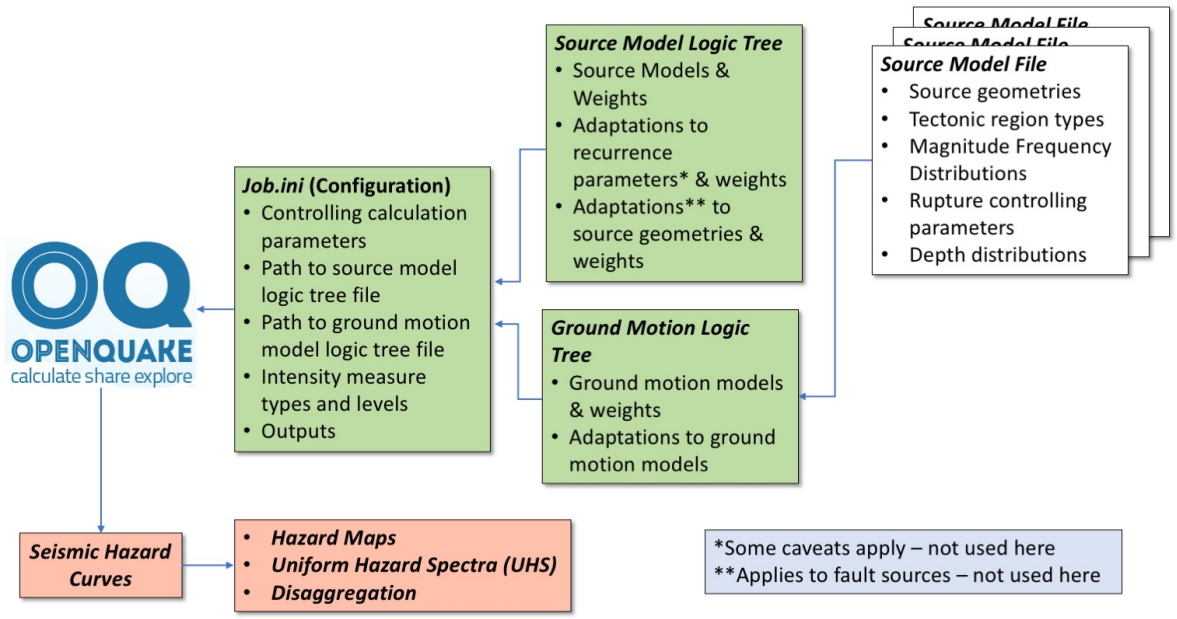


Figure 3: Overview of OpenQuake Inputs

Next, we move onto the seismic hazard model inputs themselves, which are the source model logic tree (a single xml file that describes the source model branches of the calculation and their respective weights) and the ground motion model logic tree (the file describing the choice of ground motion model and respective weights). Many PSHA models may simply define their logic trees as a set of alternative model choices; however, recent years have seen OpenQuake expand its logic tree functionality to support adaptable branches, i.e., branches for which epistemic uncertainty a particular component or parameter can be described.

At the furthest right-hand side of the input configuration are the individual source model files, each describing part or all of a seismogenic source model. As described previously, an individual seismogenic source model contains one or (usually) more seismogenic sources. The seismogenic source contains: i) a unique identifier, name and assigned tectonic region type ii) the geographical location of potential earthquake seismogenesis, which may take the form of a 3D fault surface or a *distributed seismicity* region (such as a uniform area source or grid of point sources), iii) the magnitude frequency distribution (or direct probability of occurrence of k ruptures), and iv) various parameters that may control how ruptures from the source are distributed with depth, scale with magnitude, are oriented, are constrained by crustal thickness etc. A comprehensive description can be found in OpenQuake's documentation (https://docs.openquake.org/oq-engine/manual/latest/).

As the core objective of PSHA is to calculate the probability of exceedance of an increasing vector of ground motion intensity levels to create a hazard curve, this is one of the primary outputs of OpenQuake itself. Depending on the configuration one can return just the mean curve from the logic tree, as well as the set of curves corresponding to a list of user-specified quantiles. OpenQuake calculates mean and quantiles of the probabilities of exceedance for each of the intensity measure levels, and it operates on the absolute values for the IMLs and not the logarithm such that the mean curve is the weighted arithmetic mean. From the seismic hazard curves OpenQuake can then return outputs including the uniform hazard spectrum and seismic hazard maps, which we have defined previously. If the user wishes to undertake a disaggregation calculation for a selected site or sites, then these outputs are also returned in terms of the absolute contribution of each bin to the selected hazard level. In this case the user can define multiple disaggregation outputs to return, for example, the marginal distributions for magnitude, distance, magnitude + distance, magnitude + distance + GMM ε, etc. Once again, an interested reader is referred to the OpenQuake documentation for further information.

While many of the features described are available in many PSHA software, there are several specific characteristics of OpenQuake that are particularly relevant for the comparisons between models and

data, which are exploited in the tools developed here. The first feature is the ground motion model library, which boasts several hundred ground motion models that can all be executed via a common interface. The vast majority of GMMs in the software are rigorously tested against reference ground motion values and/or implementations from other PSHA software, making this an especially robust library. The common interface makes it possible to develop functionalities that can readily calculate mean and standard deviation of ground motion with any number of selected GMMs, which have been exploited in other software such as the OpenQuake Model Building Toolkit (https://github.com/GEMScienceTools/oq-mbtk/) and the eGSIM platform (https://egsim.gfz-potsdam.de/home). However, as most GMMs will have different needs in terms of descriptive source, distance and site parameters, full interoperability and comparison requires a comprehensive set of metadata. As in many cases the metadata are incomplete, this requires filling in gaps or making reasonable judgements as to appropriate source, path and site properties. This will be discussed in further detail in the next section.

The second useful feature is OpenQuake's datastore, which is a high-density binary (hdf5) file created for each calculation in which critical information of the calculation is stored, allowing it to be retrieved and/or re-used in subsequent calculations. When running PSHA calculations it is in the datastore that we find the complete suite of seismic hazard curves, $\mathcal{H}$, stored in an organised and accessible array format. Development of code that interacts directly with the datastore through common hdf5 libraries is usually inadvisable owing to changes to the datastore from version-to-version of OpenQuake, an application programming interface (API) is available inside OpenQuake's code that allows for easy retrieval of some (or all) of its contents. The datastore retains a lot of information about the seismic hazard calculation, but a minimal set of information to describe a complete seismic hazard output model can be readily retrieved and in itself can form a compact general format for storing hazard results from other software if needed. The general structure of the "lightweight" datastore, which also forms the part of the general database for storing PSHA models shown in Figure 2. A more detailed image of the "lightweight" hazard data store is shown in Figure 4, and this comprises eight attributes:

1. ***Intensity Measure Types and Levels (IMTLs)*** which is defined by a directory of N_{IMTS} ***intensity measure types (IMTs)*** (e.g., PGA, PGV, Sa (T) …), each a 1D vector containing N_{IMLS} ***internsity measure levels (IMLs)***.

2. ***Hazard curves (hcurves)*** contains the complete suite of seismic hazard curves as a single 4D array of probabilities of exceedance for each site of N_{SITES} sites, each branch of $N_{BRANCHES}$ branches, each intensity measure type, and each intensity measure level. This array has dimension $[N_{SITES}, N_{BRANCHES}, N_{IMTS}, N_{IMLS}]$. The probability of exceedance corresponds to the investigation time, which is defined in the original config file *and* stored as a top-level attribute of the hazard group in the HDF5 file.

3. ***Hazard curve statistics ("hstats")*** are the set of curves corresponding to the mean curve and each quantile of N_{QNTL} required quantiles. These too are stored in a single 4D array with dimension: $[N_{SITES}, N_{QNTL} + 1, N_{IMTS}, N_{IMLs}]$.

4. ***Probabilities of exceedance (poes)*** refer to a single vector of N_{POES} user-specified probabilities of exceedance (with respect to investigation time) that are intended for production of hazard maps and uniform hazard spectra.

5. ***Quantiles*** refer to the single vector of N_{QNTLS} user-specified quantiles, which must take values between 0 and 1

6. ***realizations*** is a table with each row containing the "ordinal" (i.e. position of the branch), the source model logic tree ID ("smlt_id"), the ground motion model ID ("gsim_id") and the weight of each branch from $N_{BRANCHES}$ branches.

7. ***sites*** is the table of target site properties for each of the N_{SITES} sites. These include the longitude, latitude, soil properties (e.g. V_{S30}, $Z_{1.0}$, etc.) and other attributes needed for the GMM.

8. ***Weights*** is the $N_{BRANCHES}$ vector containing the weights for each branch (also a column in ***realizations***).

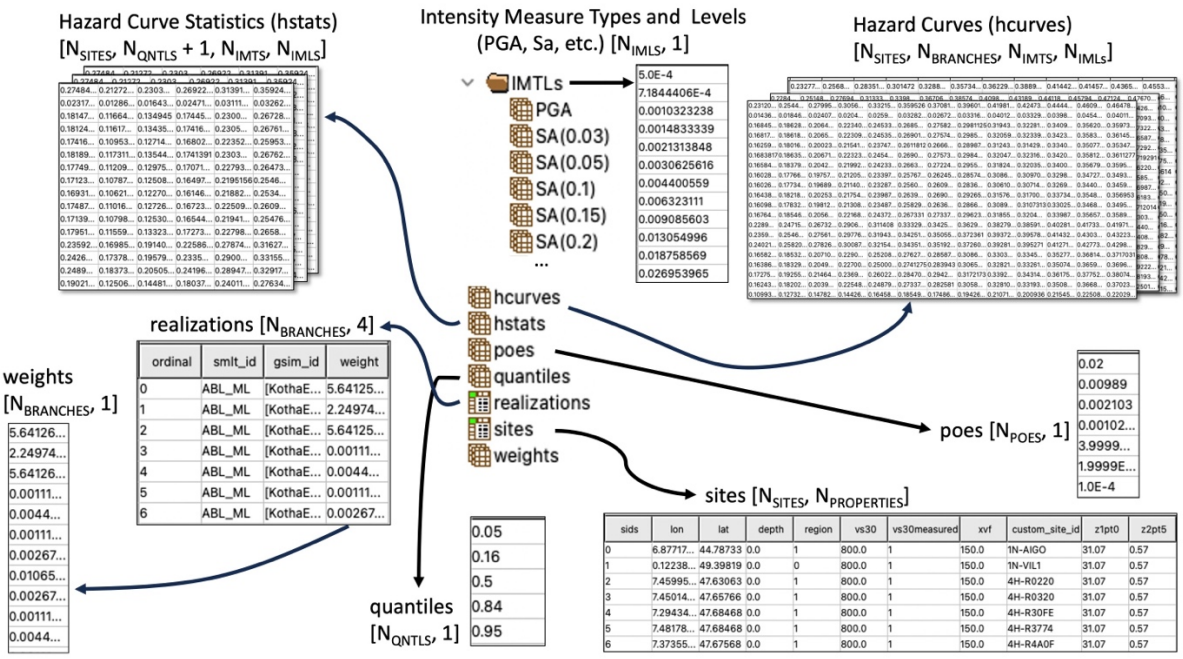


Figure 4: Structure of the Hazard Model Datastore used by PyPSHATest (simplified from OpenQuake's internal datastore)

The last specific feature of OpenQuake that the PSHA comparison tools can capitalise on is the capability to run PSHA using either of the classical *or* stochastic event-based approaches from the same input source models and ground motion models. The only differences between the two are in the configuration file, in which the user must specify the number of stochastic event sets per logic tree branch ("investigation time" in this case now becomes the event set duration), and the choice of spatial ground motion spatial correlation model if the user wishes to apply spatial correlation in the calculations. The same output hazard products (hazard curves, UHS, maps etc.) can be retrieved as for the classical approach, and the results should be equivalent *provided that a sufficient effective catalogue duration is specified*. In practice it is difficult to achieve convergence between the classical and event-based seismic hazard curves for very low annual probabilities of exceedance (e.g., $APoE < 10^{-4}$) owing to the size of the event set needed to adequately capture the representative rates of very low probability events. As for the classical PSHA case, a lightweight reduced datastore can also be constructed from the event-based calculations as shown in Figure 4. Additionally, however, users can store the ground motion fields themselves alongside their identifying rupture information. Even in the lighter weight reduced datastore format and with a small number of sites ($N_{SITES} < 1000$) this can still become a large file, so caution is recommended in using this. We strongly advise against storing event-based calculation results for PSHA from thousands of locations across a region.

2.2.1. Using PyPSHATest to Manage Hazard Calculations

The tools contain functionalities for managing seismic hazard models in an organised manner in order to facilitate comparisons with one another and with respect to data. It does not itself call or execute OpenQuake calculations, but it does retrieve and organise the key information needed for the functionalities we will see later in the book. To retrieve the information from an OpenQuake PSHA calculation and create the lightweight datastore one can use the *OpenQuakeHazardOutputManager*. This tool allows a user to specify an OpenQuake calculation (in the form of the calculation ID or the file of the OpenQuake datastore) from which it constructs the lightweight datastore and adds it to a new database file. The database file can contain an indefinite number of hazard models and is interoperable with the general site and hazard database explained in Section 1. An example of how to create this lightweight database and add hazard model results is shown in below. The process is two-step: in the first step the OpenQuakeHazardOutputManager is instantiated with the name of the datafile to which the results will be stored, then in the second step a model is added via the "add_hazard_results" function, in which the model is specified by the OpenQuake calculation ID or datastore file path and is given a name (by which it will be retrieved in the file).

```python
from pshatest.openquake_hazard_manager import OpenQuakeHazardOutputManager

# Instantiate the hazard output manager with the data store file
manager = OpenQuakeHazardOutputManager(
    dbfile="./path/to/new_hazard_datastore.hdf5",
    num_proc=8,  # Number of processors to set for parallelized calculations
)

# Add a seismic hazard model from an OpenQuake calculation ID ...
manager.add_hazard_results(
    calc_id=1,
    calc_type="classical",
    model_name="Example Seismic Hazard Model",
)

# or, alternatively, add it from an OpenQuake datastore file
manager.add_hazard_results(
    calc_id="../oqdata/calc_1.hdf5"
    calc_type="classical",
    model_name="Example Seismic Hazard Model",
)
```

While the functionality shown above is suitable for most PSHA models, in some cases where a large number of logic tree branches are specified it is not possible to evaluate the logic tree in a single calculation run. This happens for two reasons. Firstly, recent versions of OpenQuake have implemented a cap on the number of permissible branches for a source model/tectonic region or GMM branch set that is smaller than the number of branches required in recent PSHA models (Drouet et al., 2020; Grünthal et al., 2018). Secondly, where large numbers of alternative source models or source model parameters are required the computational demand (both memory usage and CPU) for generating the earthquake rupture forecast can become too large for the computational resources available. In the OpenQuake implementations of both the France PSHA model from Drouet et al. (2020) and the Germany national seismic hazard model from Grünthal et al. (2018) this was the case. To be able to obtain PSHA results one may benefit from splitting the calculation into sub-sets of logic tree branches, executing OpenQuake to obtain the individual curves and then re-combining and re-calculating the mean and quantiles *a posteriori.* This operation is also supported in the tools and split PSHA calculations can be added to the database. For example, if I have hypothetical PSHA model with $N_{BRANCHES} = 500$ and I find that it is preferable to split it into 5 separate calculations, each of 100 branches. In this case I am splitting the total weight evenly among the subsets of branches. To re-combine the calculations, I need to specify the list of calculation IDs, the corresponding positions of the branches in the full set of $N_{BRANCHES}$ realizations and the proportion of the total weight assigned to each of the subsets. This can be stored as a json or input manually in Python, as illustrated below. The split calculation can then be added to the database via the function "add_split_classical_logic_tree_results(…)". Note that depending on $N_{BRANCHES}$ this function may take longer because it will need to re-calculate the mean and quantiles once the full set of branches have been re-combined. This last step calls OpenQuake's functions for mean and quantile calculation and is parallelized in the code. The number of processers to set allocate for this can be set in the call to the OpenQuakeHazardOutputManager (as illustrated above).

```
# JSON contents
{
    {1: {"range": [0, 100], "weight": 0.2}},
    {2: {"range": [100, 200], "weight": 0.2}},
    {3: {"range": [200, 300], "weight": 0.2}},
    {4: {"range": [300, 400], "weight": 0.2}},
    {5: {"range": [400, 500], "weight": 0.2}},
}
```

```python
from pshatest.openquake_hazard_manager import\
    load_split_calculation_info

# Load in the information from JSON
split_hazard_indices_weights = load_split_calculation_info(
    "./path/to/split_calculation_info.json"
)
```

```python
# Define the split information as a list of tuples of
# (calculation ID, indices for positions in logic tree,
# total weight of branches)
split_hazard_indexes_weights = [
    (1, np.array([0, 1, 2, ... 99]), 0.2),
    (2, np.array([100, 101, 102, ... 199]), 0.2),
    (3, np.array([200, 201, 202, ... 299]), 0.2),
    (4, np.array([300, 301, 302, ... 399]), 0.2),
    (5, np.array([400, 401, 402, ... 499]), 0.2),
]

manager.add_split_classical_logic_tree_results(
    calc_properties=split_hazard_indices_weights,
    model_name="Example_Split_Hazard_Model"
)
```

2.3. Comparing Source Model Activity Rate and Seismic Hazard Distributions Across Spatial Domains

The standardised formats for OpenQuake source model inputs and outputs, combined with the tools for handling PSHA results already shown in this section, allow us to interrogate the models further. In particular we may wish to compare two or more PSHA models for a given region and undertake quantitative comparisons not only of the specific result (e.g., mean, median, etc.) but of the whole probability distribution of hazard from the model suite $\mathcal{H}$. The first set of scientific tools we will demonstrate is directed at allowing modellers to explore and quantify differences between PSHA models for a given region of interest, accounting for epistemic uncertainty. There are several different use cases for such tools in the context of understanding changes in seismic hazard for a site and/or a specific region.

The first case comes when a new PSHA model is introduced that is intended to supersede or update an existing model, such as for an update to a national PSHA model. In this case modellers are typically expected to demonstrate the changes in seismic hazard between the old and new models. While previously these comparisons have focused on the mean seismic hazard or a specific quantile, most modern hazard models contain a logic tree, meaning that differences should not necessarily be seen in terms of a percentage or absolute increase or decrease in hazard but in differences in the total probability distribution implied by $\mathcal{H}$.

The second case occurs when a seismic hazard end-user is confronted with overlapping PSHA models for a region. This may be at the border of neighbouring countries where national/regional models overlap, but it may also occur when a local or national scale model needs to be compared against a larger regional or multi-national scale model. Both situations can be seen in Europe, where national seismic hazard models from different countries overlap and where there exists a pan-European seismic hazard model (the European Seismic Hazard Model, ESHM) against which national models are compared. As most national and regional models now contain logic trees of increasing complexity, a user may once again be interested to quantify differences between the models, particularly over a spatial domain, in terms of the distributions rather than mean or median seismic hazard.

The third case may be something that occurs more for site specific PSHA modelling in a SSHAC level 3 or 4 project, in which multiple models may be developed by their respective proponents and an integrator is required to evaluate and assess the models. For example, several proponents may propose different seismic source zonations for a region. As each zonation will yield its own set of logic tree branches to account for other dependent epistemic uncertainties (e.g., MFD, M_{max} depth distribution etc.) then comparison of the resulting distributions (both of activity and of hazard) can help the integrators and/or evaluators understand the difference between the models and get a picture of their distribution in the entire model space.

2.3.1. Comparing source model activity rate distributions across a region

The first set of comparison tools from PyPSHATest that we will consider addresses the seismogenic source models and aims to visualise and compare distributions of seismogenic source models from a logic tree in a quantitative manner. To do this we need to define a common framework through which to represent an individual seismogenic source model, which we do in the form of an activity rate grid. An area of interest is described by the bounding box $(\phi_{min}, \theta_{min}, \phi_{max}, \theta_{max})$ where ϕ_{min}(°E) and ϕ_{max}(°E) are the western and eastern bounds in terms of degrees longitude, and θ_{min} (°N) and θ_{max} (°N) the southern and northern bounds respectively. This is discretised into a grid of N_ϕ evenly spaced cells of longitude and N_θ evenly spaced cells of latitude. We then similarly define a range of magnitudes $M_{min} \leq M \leq M_{max}$, which are then discretised into N_m equally spaced bins. Note here that neither M_{min} nor M_{max} need to correspond to those values in any particular seismic hazard model, and indeed for the current purposes it may be preferable to choose values that envelope those in the models being compared. The two discretisations are combined into a single 3D grid $(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{M})$.

With the grid properties defined, an individual seismogenic source branch from a full seismogenic source model logic tree can be rendered into a grid of activity rate, $\boldsymbol{\lambda}(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{M})$, where the rate of occurrence of seismicity in grid cell of longitude $i = 1, 2, \ldots, N_\phi$, latitude $j = 1, 2, \ldots, N_\theta$ and magnitude $k = 1, 2, \ldots, N_M$

is given by λ_{ijk}. Specific details of how each type of seismogenic source is partitioned into the grid spaced can be found in the documentation strings of the code but briefly: 1) for (multi-)point sources the activity rate of the source is assigned to the cell in which the point itself falls, 2) for uniform area sources the activity rate is distributed across the overlapping grid cells in proportion to the area of the source intersecting with each individual grid cell, and 3) for fault sources the activity rate is distributed in according to the proportion of the area of the fault's surface projection intersecting with the cell. With this approach we can encode each seismogenic source model, $\mathcal{h}$, into its corresponding activity rate array $\boldsymbol{\lambda}(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{M}|\mathcal{h})$, hereafter just $\boldsymbol{\lambda}_{\mathcal{h}}$, and this array is associated with the corresponding source model branch weight $w_{\mathcal{h}}$.

Owing to OpenQuake's standardised format for representation of seismogenic sources we can apply PyPSHATest to easily generate such grids both for a single seismogenic source model (described by a source model in .nrml (xml) format) and for a logic tree of source models from an OpenQuake source model logic tree file. For this we need two tools that can be imported from PyPSHATest via:

```
from pshatest.seismicity_rate import RateGrid2D, RateGrid2DSet
```

In the following example we consider the region of France (and surroundings), which we define according to the longitudes $-5^{\circ}E \leq \phi \leq 16^{\circ}E$ and latitudes $41^{\circ}N \leq \theta \leq 52^{\circ}N$, with each cell having a width of $0.2^{\circ} \times 0.2^{\circ}$. Magnitudes are specified in the range $4.5 \leq M \leq 7.5$, spaced every 0.2 magnitude units. In the first step we define the basic grid:

```
bbox = (-6.0, 41.0, 10.0, 52.0)
dlon = 0.2
dlat = 0.2
mmin = 4.5
mmax = 7.5
dm = 0.2
grid = RateGrid2D(bbox, dlon, dlat, mmin, mmax, dm)
```

An individual source model can be rendered onto the grid via:

```
source_model_file = "./path/to/source_model_1.xml"
grid.parse_source_model(source_model_file,
                        source_model_id = "SOURCE_MODEL1")
```

The class RateGrid2D boasts several functionalities for performing arithmetic operations (+, -, ×, ÷, ^) with one or more grid objects, which can make simple comparisons easy. In addition, a plotting functionality is also added that can allow a user to visualise the spatial distribution of rates within a given magnitude range, e.g.:

```
grid.plot(
    mmin=4.5,  # Minimum magnitude
    mmax = np.inf,  # Maximum magnitude
    **kwargs,  # Plotting configuration
               # and/or matplotlib Axes
)
```

Examples of such grid plots for a two different source models is shown in Figure 5.

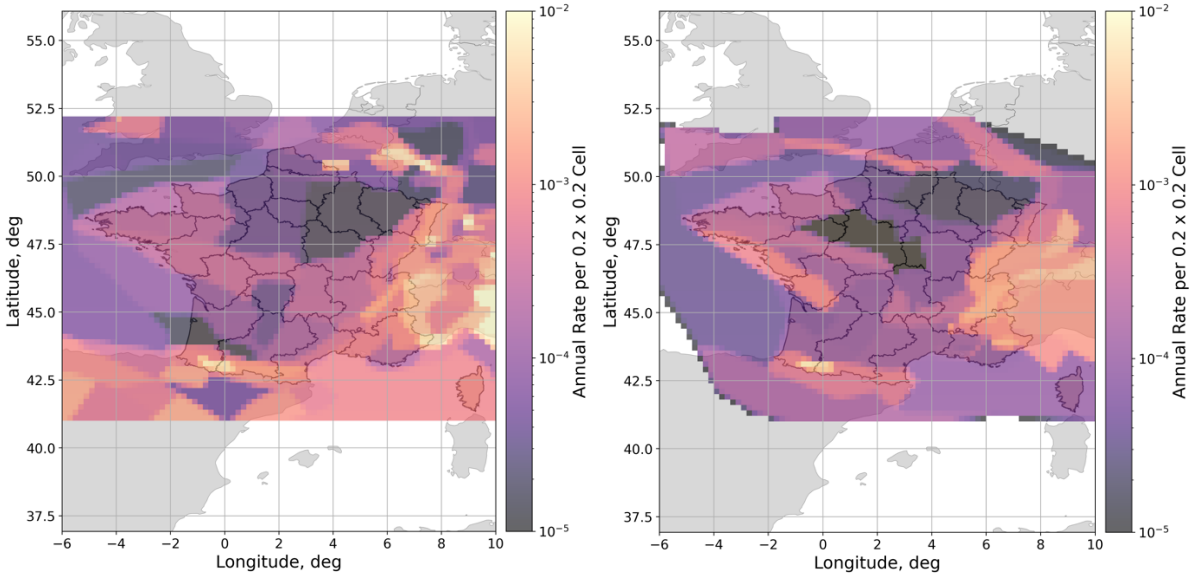


Figure 5: Comparison of the rates of seismicity for M ≥ 4.5 from two example source models using the grid configuration shown in the text.

For two different source models rendered onto their respective grids (Figure 5), a difference map showing the change in activity rate from the first to the second source model be produced via:

```
grid.plot_diff(
    grid2,  # Second grid to be compared
    mmin=4.5,  # Minimum magnitude
    mmax = np.inf,  # Maixmum magnitude
    absolute=False,  # Compare the grid in terms of
                     # absolute values (True)
                     # or relative % change (False)
    **kwargs,  # Plotting configuration
               # and/or matplotlib Axes
)
```

This difference map is shown in Figure 6.

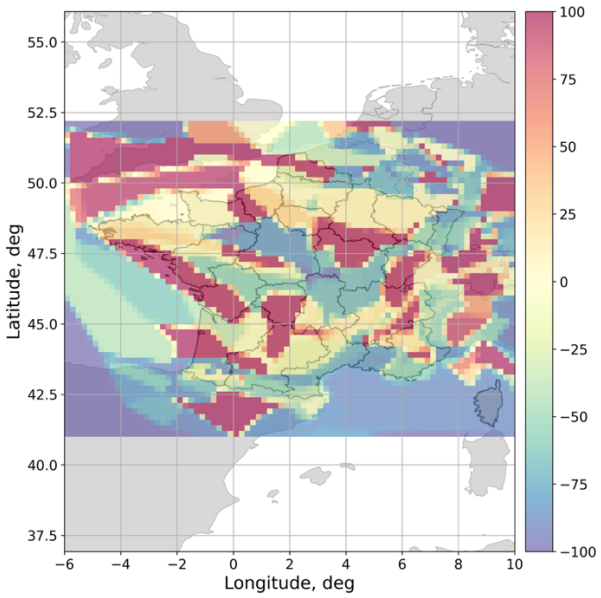


Figure 6: Percent change in activity rate from model 1 (left side of Figure 5) to model 2 (right side of Figure 5)

Comparison of the activity rate grids for individual sources may be interesting but our main aim was to be able to compare distributions of source models. For this we need the second function "RateGrid2DSet", which allows us to build $\lambda_{\hbar}(\phi, \theta, M|\hbar)$ for all $\hbar$ in $\boldsymbol{\mathcal{H}}$. As the size of the logic trees may become large, this may result in excessive RAM consumption depending on the grid size. Instead "RateGrid2DSet" automatically persists the grids to a datastore, which is again another HDF5 binary file, with each grid labelled according to the source model ID found in the logic tree file. This avoids loading the grids into memory, and as loading and rendering many models may take time, it means that results can be retrieved and re-used at a future point.

In the following example we compare for the same geographic region to source model distributions: the 2020 European Seismic Hazard Model [ESHM20] (Danciu et al., 2021) and the PSHA model for the French metropolitan territory prepared by (Drouet et al., 2020). For the region of interest, ESHM20 provides a total of 15 source model branches, while the model of Drouet et al. (2020) uses 400 branches corresponding to 100 samples of MFD from each of four seismogenic source models.

The grid sets are constructed first for ESHM20:

```
# Define a datastore for the model
eshm20_dstore = "./eshm20_source_model_dstore.hdf5"

# Define the path to the source model logic tree file
eshm20_logic_tree_file = "./path/to/eshm20_source_model_logic_tree.xml"

# Get the set of activity rate grids
eshm20 = RateGrid2DSet.from_OQ_logic_tree_file(
    grid=RateGrid2D(bbox, dlon, dlat, mmin, mmax, dm),  # Reference grid
    datastore=eshm20_dstore,  # Datastore
    source_model_logic_tree_file=eshm20_logic_tree_file)  # Logic Tree
```

And then for Drouet et al. (2020):

```
# Define a datastore for the model
france_2020_dstore = "./france_source_model_dstore.hdf5"

# Define the path to the source model logic tree file
france_logic_tree_file = "./path/to/france_2020_source_model_logic_tree.xml"

# Get the set of activity rate grids
france_2020 = sra.RateGrid2DSet.from_OQ_logic_tree_file(
    grid=RateGrid2D(bbox, dlon, dlat, mmin, mmax, dm),  # Reference grid
    datastore=france_2020_dstore,  # Datastore
    source_model_logic_tree_file=france_logic_tree_file)  # Logic Tree
```

With the two sets of grids defined we may be interested to compare models in terms of basic descriptive statistics. The class RateGrid2DSet contains a simple function "get_statistical_rate_set" that allows a user to define a magnitude range and will then return spatial grids of the following statistical descriptors:

- "mean": The weighted mean rate grid:

$$\mu(\boldsymbol{\lambda}_{\hbar|\delta M}) = \frac{\sum_{\hbar \in \mathcal{H}} \boldsymbol{\lambda}_{\boldsymbol{h}|\boldsymbol{\delta M}} \cdot w_{\hbar}}{\sum_{\hbar \in \mathcal{H}} w_{\hbar}} \quad 2.5$$

- "quantiles": The weighted quantiles for q_{pp} where $pp = 0.05, 0.16, 0.25, 0.5, 0.75, 0.84$ and 0.95.
- "minimum" $\min(\boldsymbol{\lambda}_{\hbar|\boldsymbol{\delta M}})$ and "maximum" $\max(\boldsymbol{\lambda}_{\hbar|\boldsymbol{\delta M}})$ (equivalent to q_{00} and $q_{1.00}$ respectively)
- "inter-quartile range": $Q_{75} - Q_{25}$

In these definitions $\lambda_{\hbar|\delta M}$refers to the total activity rate in each geographical cell for the model given the magnitude range $\delta M = M_u - M_l$:

$$\lambda_{\hbar|\delta M}(\phi_i, \theta_j) = \sum_{k=1}^{N_m} I(M_l \leq M_k \leq M_u) \cdot \lambda_{ijk}(\phi_i, \theta_j, M_k|\hbar) \quad 2.6$$

Where $I(\cdot)$ takes the value 1 if the operation $\cdot$ is true, or 0 otherwise.

The suite of descriptors can be retrieved for each model via:

```python
# For ESHM20
eshm20_statistics = eshm20.get_statistical_rate_set(
    mmin=4.5,
    mmax=7.5
)

# For Drouet et al. (2020)
drouet2020_statistics = france_2020.get_statistical_rate_set(
    mmin=4.5,
    mmax=7.5
)
```

Example maps of some of these quantities for the two models are shown in Figure 7

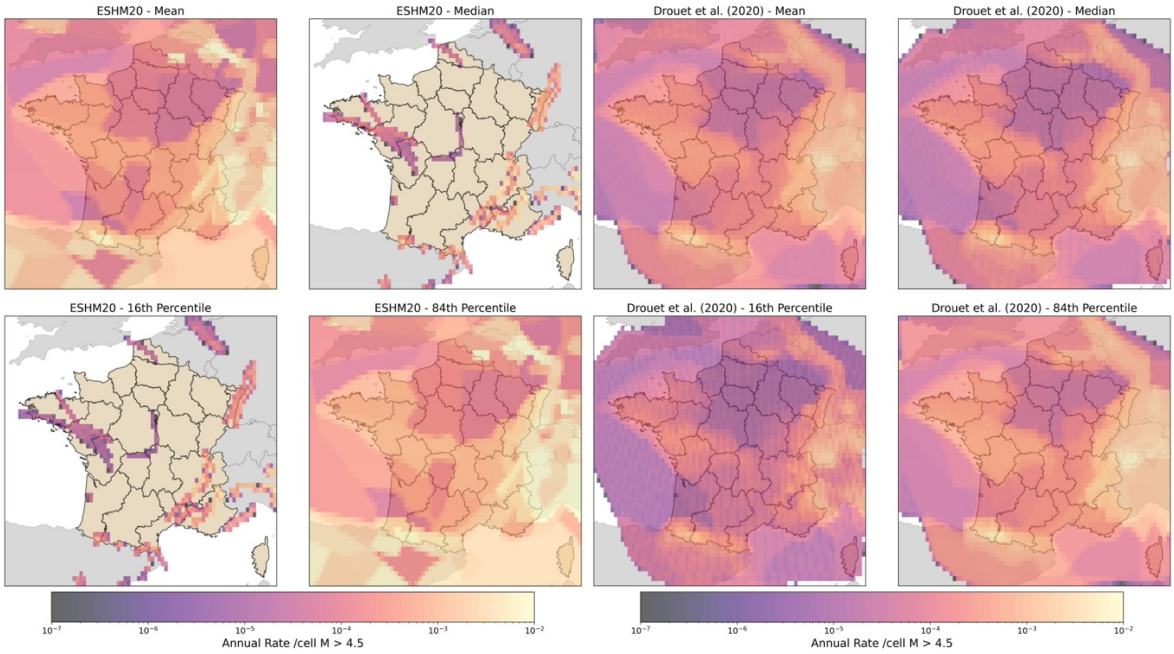


Figure 7: Example descriptive statistics maps for the seismogenic source model distribution of the ESHM20 (left) and Drouet et al. (2020) for France

With these descriptors rendered onto a simple grid, it is then easy to apply simple arithmetical operations to produce difference maps, such as the change in mean or a specific quantile. To compare the range of values one can also consider the inter-quartile range ratio between two models (A and B):

$$IQR_{ratio}(A,B) = \frac{Q_{75}^{B}(\lambda_{\hbar|\delta M}) - Q_{25}^{B}(\lambda_{\hbar|\delta M})}{Q_{75}^{A}(\lambda_{\hbar|\delta M}) - Q_{25}^{A}(\lambda_{\hbar|\delta M})} \quad 2.7$$

These quantities are illustrated for ESHM20 (model A) and Drouet et al. (2020) (model B) in Figure 8.

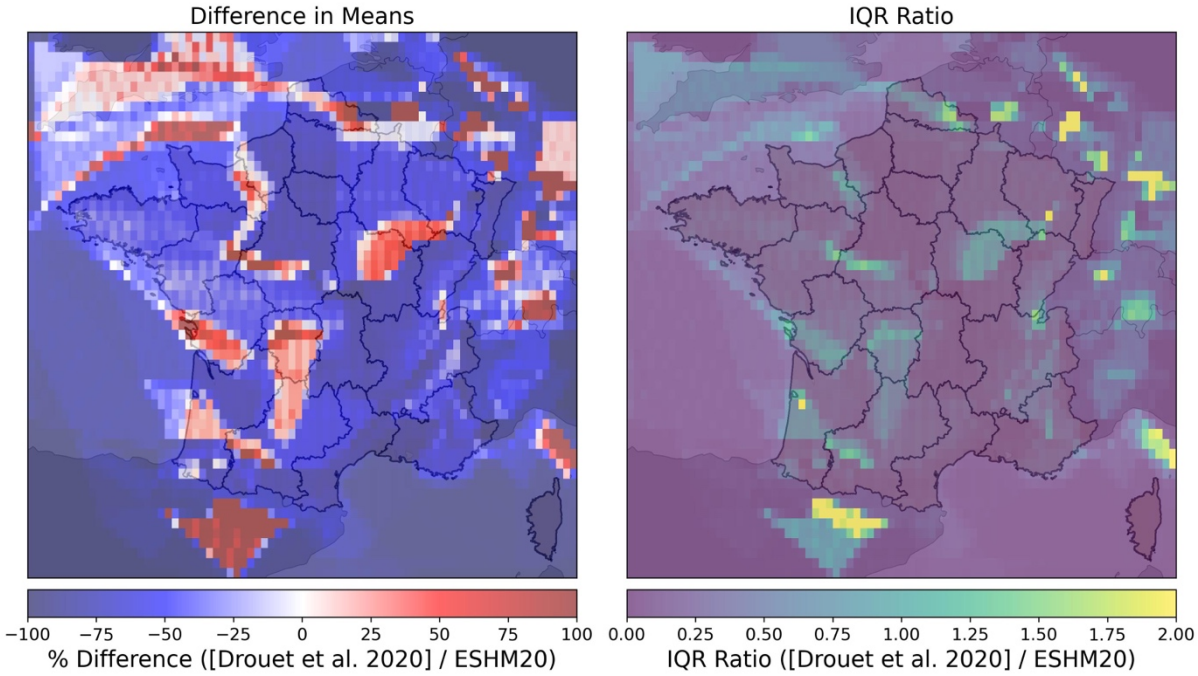


Figure 8: Maps of percent difference between in mean seismicity rate implied by the logic tree of Drouet et al. (2020) over that of the ESHM20 (left), and inter-quartile range ratio of the two models (right)

Descriptive statistics can be useful as a means of describing differences in the general tendency between two models, such as whether one is likely to result in higher or lower activity rate than another, or whether the variance has increased or decreased. If we are looking to understand the extent to which the full distributions differ, we can adopt other non-parametric statistical metrics. For each cell (ϕ_i, θ_j) and given magnitude range δM, the distribution of activity rate $\lambda_{\hbar|\delta M}$ over all $\mathcal{H}$ for a given PSHA model can be described via an empirical probability density $f(\lambda_{\hbar|\delta M})$, with corresponding cumulative density $F(\lambda_{\hbar|\delta M})$. If the cdfs for models A and B are given as $F_A(\lambda_{\hbar|\delta M})$ and $F_B(\lambda_{\hbar|\delta M})$ respectively then for each geographical cell we can define the distance between the two model distributions using either the weighted Kolmogorov-Smirnov distances (D_{KS}) or the Wasserstein Distance (D_{WS}). These are defined as:

$$D_{KS} = \sup_{\lambda_{\hbar|\delta M}} \left| F_B(\lambda_{\hbar|\delta M}) - F_A(\lambda_{\hbar|\delta M}) \right| \quad 2.8$$

$$D_{WS} = \int_{-\infty}^{\infty} \left| F_B(\lambda_{\hbar|\delta M}) - F_A(\lambda_{\hbar|\delta M}) \right| d\lambda_{\hbar|\delta M} \quad 2.9$$

The conceptual meanings of D_{KS} and D_{WS} are illustrated for two empirical CDFs in Figure 9. The former can be recognised as the maximum distance between $F_A(x)$ and $F_B(x)$ over all x, while the latter is the total area enclosed by them. From this definition $D_{KS} = [0, 1]$ (perfect agreement to no overlap in the CDFs over any x) and $D_{WS} = [0, \infty]$ (perfect agreement to an infinitely large distance between $F_A(x)$ and $F_B(x)$).

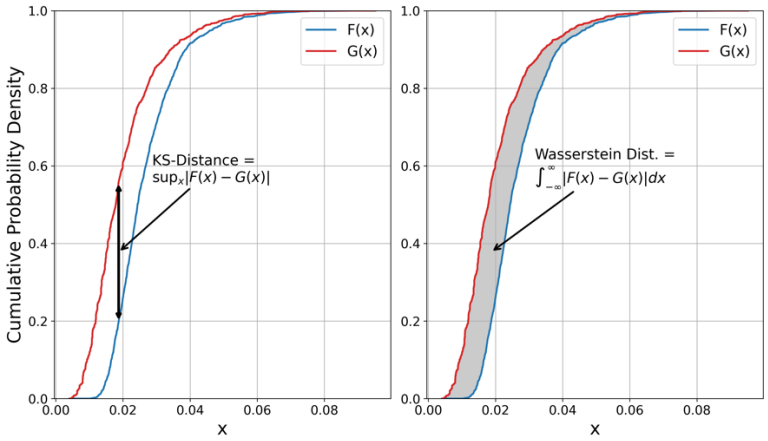


Figure 9: Interpretation of Kolmogorov-Smirnov Distance (left) and Wasserstein Distance (right) in terms of two empirical cumulative distribution fuctions

With these metrics we can analyse model differences over any spatial domain covered by the model, from a single cell (ϕ_i, θ_j), to a grouping of cells or for the whole region. Specifically, by calculating D_{WS} and D_{KS} per cell we can visualise the spatial trends in model similarity in seismogenic source model distributions between two PSHA models. This is illustrated for the case of ESHM20 and Drouet et al. (2020), which can be called via the example below and the resulting maps of D_{KS} and D_{WS} shown in Figure 10.

```
# Get the Kolmogorov-Smirnov Distance and
# p-value from a 2-sample KS-Test
ks_dist, ks_prob = eshm20.get_ks_distance(
    france_2020,
    mmin=4.5,
    mmax=7.5)

# Get the Wasserstein Distance
wasserstein_dist = eshm20.get_wasserstein_distance(
    france_2020,
    mmin=4.5,
    mmax=7.5)
```

Generally, D_{WS} is a smoother metric that can more clearly identify coherent spatial differences between models. D_{KS} can change quite abruptly from one cell to another when using empirical CDFs when the maximum distance may jump from one point in the distribution to another. D_{KS} does, however, have an advantage that it can be used in a statistical 2-Sample Kolmogorov-Smirnov Test (KS-Test) to quantify the p-value for rejection of the null hypothesis that the empirical values in $F_A(\lambda_{\hbar|\delta M})$ and $F_B(\lambda_{\hbar|\delta M})$ are drawn from the same distribution.

2.3.2. Comparing Seismic Source Models to Observed Seismicity

The comparisons of seismogenic source models in terms of activity rate distributions can be insightful in explaining why and where differences may emerge. They do not address the question of how well the source models compare against data. Bearing in mind the potential caveats outlined in section 1, the characterization of the source models in the form of activity rate grids effectively transposes them into time-independent seismicity forecasts that can be statistically tested with respect to observed seismicity. Seismicity forecast testing is a well-established field and efforts to build a standard, harmonized testing framework have been underway for several decades via the Collaboratory for the Study of Earthquake Predictability (CSEP) (Schorlemmer et al., 2018). More recently the CSEP testing framework has itself been implemented in an open-source Python software module named PyCSEP (https://github.com/SCECcode/pycsep) (Savran et al. 2022).

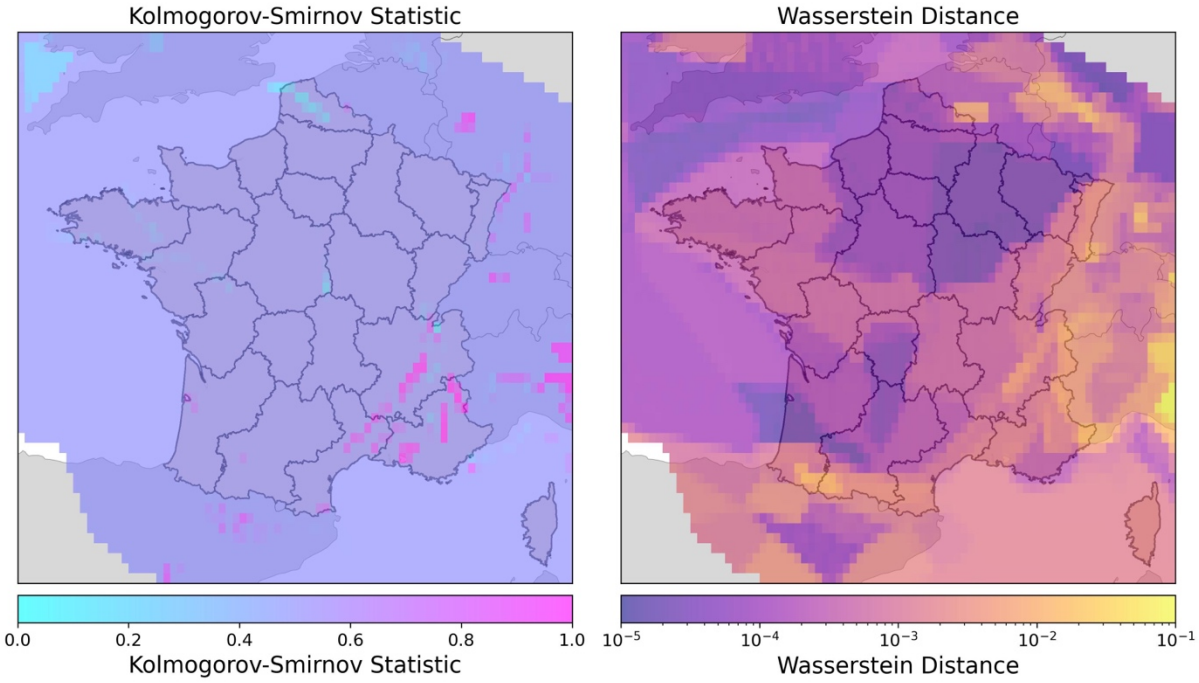


Figure 10: Maps of Kolmorogov-Smirnov Distance (left) and Wassestein Distance (right) between the full distribution of activity rate implied by the logic trees of Drouet et al. (2020) and ESHM20

With such a software available we are afforded the opportunity to leverage on existing tools for testing seismicity models against data, and we see no basis for re-inventing or re-implementing the pyCSEP functionality in the present software. Instead, to allow for interoperability between PyPSHATest and pyCSEP we simply add on to the RateGrid2D and RateGrid2DSet classes a method to transform it into pyCSEP's own GriddedForecast object. Provided that the user has pyCSEP installed, this is done via:

```
# For a single grid
csep_forecast = grid.to_csep_forecast(
    mmin = 4.5,
    mmax=7.5
    start_time="2015-01-01 00:00:00.0",
    end_time="2021-01-01 00:00:00.0"
)

# For a whole seismic source model logic tree
eshm20_forecast_set = eshm20.to_csep_forecast_set(
    mmin = 4.5,
    mmax=7.5
    start_time="2015-01-01 00:00:00.0",
    end_time="2021-01-01 00:00:00.0"
)
```

With the seismicity models exported to GriddedForecasts, a user can then implement forecast testing processes following the examples given in the pyCSEP online documentation and Jupyter notebooks (https://docs.cseptesting.org/tutorials/gridded_forecast_evaluation.html#grid-forecast-evaluation). Further worked examples of how to test seismogenic source models against data using PyPSHATest and pyCSEP will be added into the online documentation of PyPSHATest.

2.3.3. Comparing Seismic Hazard Distributions across a Region

The approaches for comparing distributions of seismogenic source models can also be extended to apply to comparisons of the full distribution of seismic hazard across a region. Further explanation of this is given in Weatherill et al. (2023b); however, from both the descriptive statistics and from the non-parametric measures of model distribution difference for given levels of seismic hazard we can find new ways to quantify differences in the full probabilistic seismic hazard results. To illustrate this, we consider in Figure 11 three models that share an overlapping region: Drouet et al. (2020) for metropolitan France

[FR2020], Grünthal et al. (2018) 2016 National Seismic Hazard Model for Germany [DE2016], and the ESHM20 (Danciu et al. 2021) for Europe. All three models can be seen to overlap along the France and Germany border region, and we define a target area such that all three models contain sources up to 200 km from the full extent of the bounding box. All three models were either implemented directly in OpenQuake (for ESHM20) or were translated from their original implementation into OpenQuake format (FR2020 and DE2016). Again, full details of the translation of models from their original software into OpenQuake can be found in (Weatherill et al., 2023a, b). The hazard maps for PGA with a 10 % probability of exceedance in 50 years are given in Figure 11.

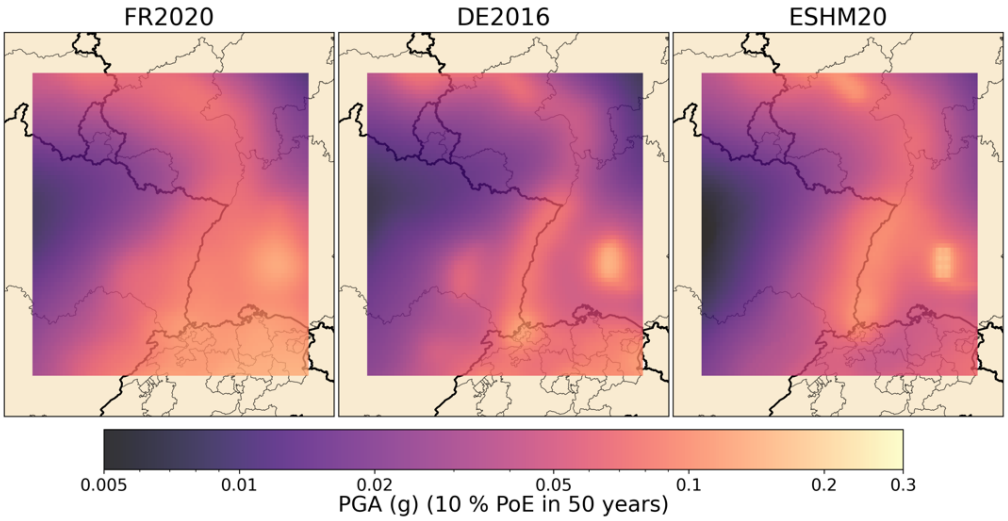


Figure 11: Mean PGA with a 10 % Probability of Exceedance in 50 years for the France-Germany border region according to Drouet et al. (2020) [FR2020] (left), Grünthal et al., (2018) [DE2016] (centre) and ESHM20 (right)

To use PyPSHATest to undertake quantitative comparisons of PSHA models for a given region, we must first run each of the three PSHA models for the same target sites. In this case we used a regular grid of sites within the bounds 5°E – 9.5°E and 47.0°N – 50.5°N, spaced every 5 km. The three PSHA calculations were run, and the outputs stored to a common database using the OpenQuakeHazardOutputManager class shown in Section 2.2.1. From this same class we can implement the comparisons of any two models within the database using the function "compare_hazard_distributions". For this we need to specify the name of the two models as they are stored in the database, the probabilities of exceedance to be used for the comparisons (here we use annual probabilities of exceedance [APoE] of 0.002105 and 0.000404, corresponding to the 10 % and 2 % PoE In 50 years respectively), the choice of intensity measures, and the list of comparison metrics to be calculated. The output of the function is a Python dictionary, containing for each metric a 3D array with the values of the metrics for each site, probability of exceedance and intensity measure type.

```python
# Define the names of the two models in the database
# to be compared
model1 = "NAME_MODEL_1"
model2 = "NAME_MODEL_2"

# Define the list of metrics for comparison
metrics = [
    "mean difference",
    "quantile difference|0.16-0.5-0.84",
    "iqr_ratio",
    "Komogorov-Smirnov",
    "Wasserstein"
]

# Define probabilities of exceedance for the comparisons
poes = [0.002105, 0.000404]  # 10 % and 2 % PoE in 50 years

# Define IMTs
imts = ["PGA", "SA(0.2)", "SA(0.5)", "SA(1.0)"]

# Compare the distributions
model_comparisons = manager.compare_hazard_distributions(
    model1,
    model2,
    poes,
    imts, metrics
)
```

The available metrics are:

- ***"mean difference"*** – The percent change in mean hazard from model 1 to model 2
- ***"quantile difference\|q_1-q_2-q_3"*** - The percent change in hazard for the listed quantiles (q_1, q_2 and q_3 in this example) from model 1 to model 2.
- ***"iqr_ratio"*** – The ratio of the inter-quartile range of model 2 over model 1
- ***"Kolmogorov-Smirnov"*** – The Kolmogorov-Smirnov statistic D_{KS} between the two models
- ***"Wasserstein"*** – The Wasserstein distance D_{WS} between the two models

Comparisons between the combinations of the three models (FR2020 → ESHM20, FR2020 → DE2016, DE2016 → ESHM20) in terms of the difference in the mean hazard (Figure 12), the Kolmogorov-Smirnov distance (Figure 13) and the Wasserstein Distance (Figure 14). All three metrics can produce visually interpretable results showing the difference between PSHA models in the full distribution of a given seismic hazard quantity arising from the models' respective epistemic uncertainty. Maps such that these can highlight where discrepancies are most apparent, indicating where interpretations of the seismicity and tectonics may be most divergent among different modellers.

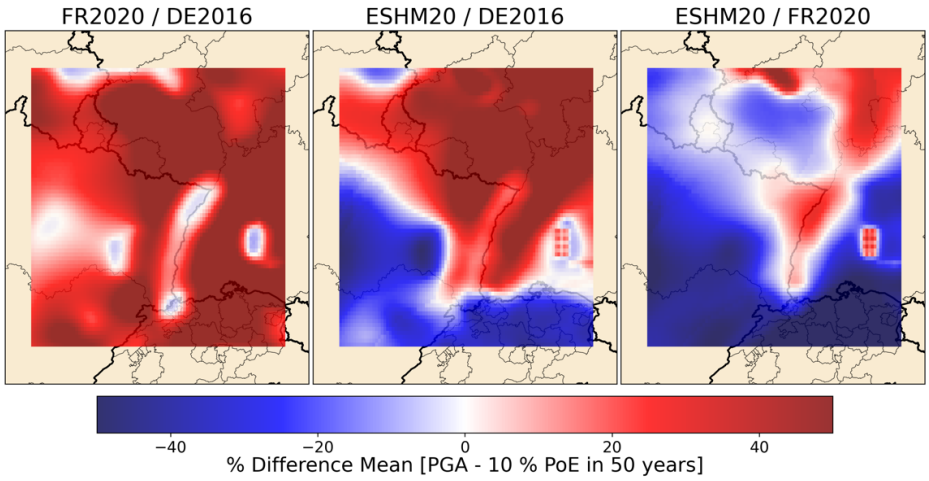


Figure 12: Comparisons of the hazard distribution in terms of differences in means

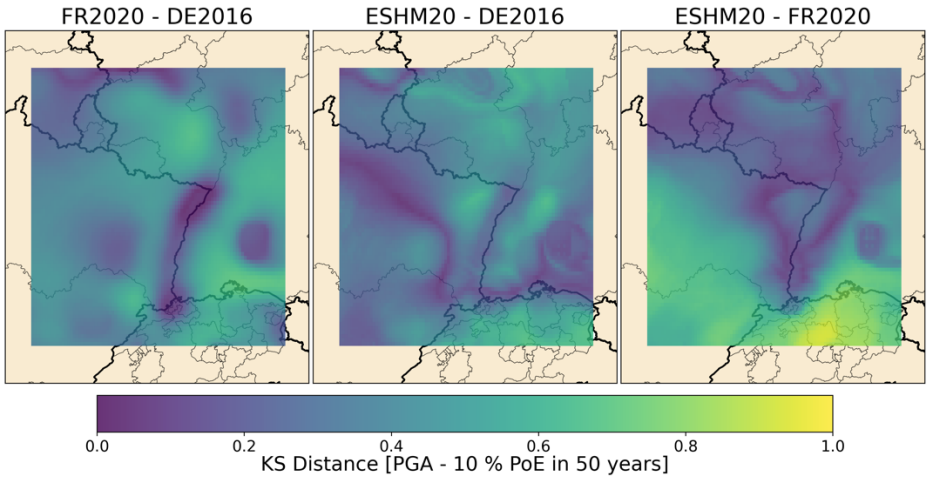


Figure 13: Comparison of the hazard distribution in terms of Kolmogorov-Smirnov distance (D_{KS})

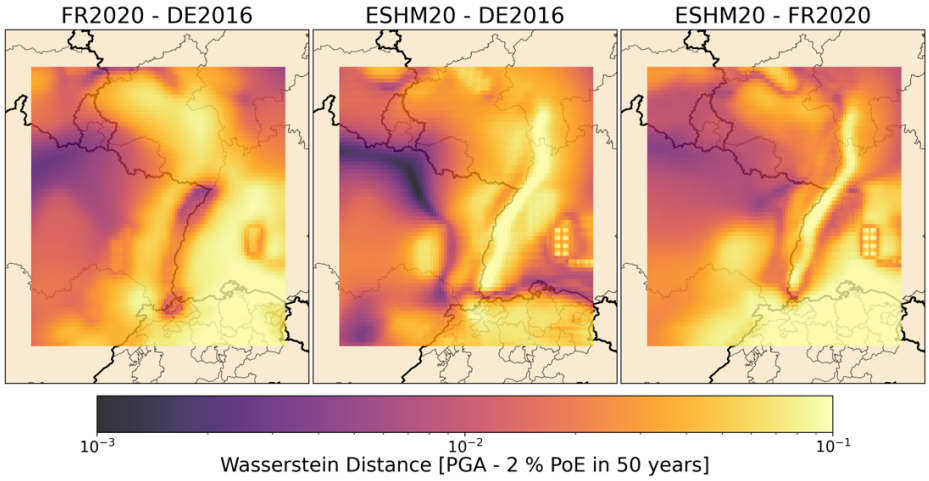


Figure 14: Comparison of the hazard distribution in terms of Wasserstein Distance (D_{WS})

3. Preparing and Analysing Data for Seismic Hazard Comparisons 2: Observed Ground Motions

3.1. Ground Motion Data and Metadata

In the introductory section to this report, we have made the distinction between *direct* and *indirect* observations of ground motion. The former refers to records of ground shaking as detected and quantified by seismic instruments, from which a full representation of the wave can be retrieved and common intensity measure types (e.g., PGA, PGV, Sa (T), Arias Intensity etc.) calculated. The latter refers to data from which we can make inferences of the ground motion at a site but cannot directly retrieve the necessary IMTs. The most common example of this is macroseismic intensity, though one could also apply this to other forms of evidence of shaking, such as recent or paleoliquefaction features, co-seismic geomorphological events and even archeoseismological features. In the current study we focus specifically on direct observations.

For comparison of PSHA against observational ground motion data we are usually concerned with strong ground motions recorded via accelerometer networks. The specific networks will obviously depend on the region or country of interest. In Europe, the developments made during the last decade by the Observatories and Research Facilities for European Seismology (ORFEUS) (a pillar of the European Plate Observation System (EPOS) Seismology) have helped to facilitate access to seismic waveform data from networks across Europe. Among the services that have developed, and have been exploited within this current work, are the European Integrated Data Archive (EIDA) (http://www.orfeus-eu.org/data/eida/), the Engineering Strong Motion Database (ESM) (https://esm-db.eu/) and the Italian Accelerometric Archive v4.0 (ITACA) (https://itaca.mi.ingv.it/). Though not used here, the Rapid Raw Strong Motion database (http://orfeus-eu.org/rrsm/) is also worthy of mention as an additional facility to access automatically processed strong motion data soon after an event. The tools and methods that are presented in this report do not necessarily *require* these services and can indeed be adopted in other regions of the world; however, the data we will be using here are direct products from these services and their use in the context of hazard model comparisons highlights the diverse applications they are serving.

Before entering into detail on the types of data and the formats, we should state what our needs are when assembling ground motion data for the purpose of quantitative comparison against PSHA. First and foremost, we are looking for our data to produce the most comprehensive archive of earthquake induced ground shaking across a set of recording stations in a region. Ultimately hazard curves can only be compared against observation data in terms of exceedances of given levels of ground motion; be it in terms of the number of exceedances of a given level of motion at one or more stations or in terms of the number of stations exceeding a given level of motion (more on this in section 4). For this to be fair the duration of operation of each station should be known and the set of *relevant* ground motions present in the archive for each station should be an accurate depiction of the history of the shaking, ideally without operational gaps or missing observations. As we will see later in this section, this is not simple to achieve in practice.

3.1.1. Flatfiles

For quantitative comparison of hazard models to observed ground motion data our primary information takes the form of a parametric table of ground motion intensity measure values and associated metadata, otherwise referred to as a *flatfile*. The flatfile is a one-row-per-entry listing of ground motion records, with each entry containing event metadata (e.g. event ID, date and time, longitude, latitude, depth, magnitude, focal mechanism, rupture dimensions etc.), site and sensor metadata (e.g. station ID, longitude, latitude, soil/basin properties [e.g. V_{S30}, Z_h], sensor type, housing information, filter frequencies etc.), and source-to-site distances using a variety of distance measures. The ground motion itself is described by its peak properties (e.g., peak ground acceleration [PGA], peak ground velocity

[PGV], peak ground displacements [PGD]), scalar metrics of the record (e.g. duration, Arias Intensity, Cumulative Absolute Velocity) and, finally, values of the spectra across a fixed set of frequencies. The most encountered form of the spectra are pseudo-acceleration response spectra (denoted by Sa), which is the definition of spectral acceleration Sa adopted within seismic hazard analysis. Recent efforts to compile flatfiles for Fourier Amplitude Spectra have been made, which can also serve a variety of applications. Again, the tools and methods presented could be agnostic to the type of spectra used to represent the ground motion, albeit that response spectra are the main type used in PSHA.

Ground motion intensity measures and spectra usually refer to the horizontal ground motion, which results from a combination of the properties from the two horizontal channels of the three component seismometers and accelerometers. Until the latest generation of models, peak intensity measures and spectra found in flatfiles were usually the geometric mean of the values as recorded on the two horizontal components (i.e. $Sa_h(T) = \sqrt{Sa_x(T) \cdot Sa_y(T)}$) or the envelope of the two values (i.e. $Sa_h(T) = \max\left(Sa_x(T), Sa_y(T)\right)$. The most recent generation of flatfiles adopt orientation-independent measures of horizontal motion that are based on quantiles of the values extracted from the combined records rotated over all non-redundant angles (Boore, 2010). This later quantity is termed *RotDnn* where *nn* refers to the quantile in question (00 for the minimum over all angles, 100 for the maximum and 50 for the median). *RotD50* has emerged as the most commonly used for recent GMMs; however, those developed for as-recorded geometric mean ground motion, or the older orientation-independent measures are still in widespread use. To accommodate this flatfiles such as those developed from the ESM and ITACA data archives also include in the same row ground motion intensity values and response spectra for different definitions of horizontal motion. Other flatfiles have attempted to address this issue using different flatfiles for different horizontal definitions of motion. For application in PyPSHATest it is required that a single definition of horizontal spectra is input in the flatfile.

As a single-row-per-entry parametric table, flatfiles such as those used for the tools here are usually stored in spreadsheet (e.g., Microsoft Excel xlsx, or open document format .ods) or as a comma-separated value (.csv) file.

3.1.2. Metadata

While compilation of the set of ground motion parameters and response spectra for each record is challenging, so too is harmonisation of the metadata to the extent that comparison with GMMs is possible. Across the suite of GMMs available to use in the toolkit (and in the scientific literature in general) there are a myriad of parameters to describe the source, path and site, most of which have been demonstrated to be relevant in the prediction of ground motion. To ensure flexibility in the choice of GMM for analysis, PyPSHATest requires that the metadata of the ground motion flatfile contains at a minimum the parameters listed in Table 1, taking note of their unit and validity range.

Table 1: Required metadata and data attributes of a ground motion flatfile for use with PyPSHATest

Column Name	Description	Units	Range
ev_id	Unique identifier of an earthquake (string)		
gm_id	Unique identifier of a ground motion record (string)		
ev_time	Earthquake time (as datetime string YYYY-MM-DD hh:mm:ss.s)		
ev_lon	Longitude of epicentre	°E	$[-180, 180]$
ev_lat	Latitude of epicentre	°N	$[-90, 90]$

ev_depth	Hypocentral depth	km	$[0, \infty]$
mag	Harmonised earthquake magnitude (M_W or proxy M_W^*)		
z_tor	Depth to the top of the earthquake rupture	km	$[0, \infty]$
z_bor	Depth to the bottom of the earthquake rupture	Km	$[z_{tor}, \infty]$
f_length	Length of earthquake rupture	km	$[0, \infty]$
f_width	Down-dip width of earthquake rupture	km	$[0, \infty]$
strike	Strike of earthquake rupture	°	$[0, 360]$
dip	Dip of earthquake rupture	°	$[> 0, 90]$
rake	Rake of earthquake rupture, following Aki & Richards (1992) convention	°	$[-180, 180]$
r_epi	Distance between the earthquake epicentre and the recording site	km	$[0, \infty]$
r_hyp	Distance between the earthquake hypocentre and the recording site	km	$[0, \infty]$
r_jb	Distance along the Earth's surface between the surface projection of the earthquake fault rupture and recording site	km	$[0, \infty]$
r_rup	Shortest direct distance between the earthquake fault rupture and recording site	km	$[0, \infty]$
r_x	Shortest distance between the recording site and a line representing up-dip surface projection of the rupture, measured perpendicular to the fault strike	km	$[-\infty, \infty]$
r_y0	Shortest distance between the recording site and the nearest end of the rupture, measured perpendicular to strike	km	$[0, \infty]$
network	Seismological operating network of the recording station (IRIS two-letter code)		
station	Code of the recording station within the network		
sta	Unique identifier of the station as a concatenation of network-station		

channel	Indicates the instrument channel on which the recording is taken.		[HH, EH HN, HG]
sta_lon	Longitude of the recording station	°E	$[-180,180]$
sta_lat	Latitude of the recording station	°N	$[-90,90]$
vs30	Average shearwave velocity V_S of the upper 30 m of the crust at the station (or inference from proxy)	m/s	$[0,\infty]$
vs30 measured	Indicates if the vs30 is a measured quantity for the station (True) or inferred from proxy (False)		
sta_start_ date	Date of start of operation of the station		
sta_end_ date	Date of end of operation of the station (if not still active)		
Z1.0	Depth to the $V_S = 1.0$ km/s shearwave velocity layer	m	$[0,\infty]$
Z2.5	Depth to the V_S = 2.5 km/s shearwave velocity layer	km	$[0,\infty]$
region	For models that require definition of a region to which the station belongs (e.g. ESHM20), the assigned region identifier is required		
PGA	Peak absolute ground acceleration	g	$[0,\infty]$
PGV	Peak absolute ground velocity	cm/s	$[0,\infty]$
pSA_#.##	Peak spectral acceleration at period #.##	g	$[0,\infty]$

Many of the attributes listed in Table 1 should be retrieved directly in the process of the flatfile compilation. These include the unique identifiers for earthquake, station and record, earthquake hypocentre location and time, the station network, code and location, and the actual ground motion parameters. Distances from the earthquake epicentre/hypocentre to the recording station can usually be obtained easily using appropriate algorithms in geospatial analysis software. Compilation of the other parameters for the metadata can be more challenging.

While all network operators and seismological bulletin will provide a magnitude estimate for the earthquake, it is usually the case that this is reported in different scales depending on the source, and different recording networks may apply different corrections that calibrate the same magnitude scale differently. This problem of heterogeneity in the magnitude scales and its impact in seismic hazard analysis is well known (e.g., Grünthal & Wahlström, 2012; Weatherill et al., 2016) and compilation of a harmonised earthquake catalogue is a necessary step in hazard model development. It is not uncommon to encounter raw databases of ground motion in which the magnitude scale differs within the flatfile

(hopefully placed into different columns). For use in the toolkit "mag" must refer to the earthquake magnitude harmonised into a common reference scale compatible with that used by the GMMs. This is nearly always moment magnitude M_W. If this is not provided in the flatfile but is available in an independently compiled harmonised earthquake catalogue PyPSHATest has functionality to identify the events common to the flatfile and catalogue, thus assigning in this column the harmonised magnitude. An example of this will be shown in the case study later in this section.

Another major limitation of most preliminary compilations of ground motion data is the absence of information describing the finite fault. As the majority of modern GMMs use metrics in relation to the finite fault, such as distances to the fault or depths to the top/bottom of the ruptures, this presents an enormous challenge in flatfile metadata compilation. Models of the three-dimensional fault rupture are usually only available (or compiled routinely) for large earthquakes with M ≥ 6.5. Rupture models for moderate magnitude events may be available for some smaller events in certain regions where high-quality networks are available, but otherwise this information is absent for most earthquakes. Removal of records lacking this data would ultimately eliminate most available data for comparison against GMMs, while limiting the selection of GMMs only to those not requiring such detail would the eliminate most models. Instead, it is necessary to attempt to define "reasonable" values of these finite rupture parameters from the information available. Fortunately, when the earthquake is small magnitude (M < 5) the finite rupture properties can be approximated by their point-source equivalents without excessive bias. In this case top-of-rupture depths and bottom of rupture depths can be considered equivalent to hypocentral depth (potentially with a small correction of < ± 1 km), finite fault properties (e.g. length, width, strike, dip) can be approximated by a small vertical plan of unit area, while basic rule of thumb equivalences in distance can be applied (e.g. $R_{JB} \approx R_{EPI}, R_{RUP} \approx R_{HYP}, R_X \approx -R_{EPI}, R_{y0} \approx R_{EPI}$). For moderate magnitudes (e.g., $5.0 \leq M_W \leq 6.5$), we recommend the construction of idealised rupture planes, whose sizes scale in proportion to magnitude according to existing magnitude scaling relations, and for which the hypocentre is positioned in the centroid of the plane. If information to constrain the orientation is available, such as a focal mechanism or stress axes, this can be used to align the plane. Tools to assist in the process of characterising finite ruptures and the corresponding metadata are available in PyPSHATest.

Finally, the last metadata parameters that may require additional effort to constrain relate to the site properties. For many recording networks information on the V_{S30} may be available for a large proportion of stations, but seldomly all. For other parameters such as basin depths (Z1.0 and Z2.5) their availability is less common. If using stations from weak motion networks, it is uncommon to undertake engineering seismological investigation into the recording sites, and most are simply classified according to the surface geology. If V_{S30} values are not available for the site then we recommend using those from national or regional scale compilations, otherwise from the United States Geological Survey (USGS) Global V_{S30} database (Wald & Allen, 2007). In this case vs30measured should be set to False. Basin depth parameters may be estimated from the directly measured V_{S30} or its equivalent proxy using the empirical V_{S30} to $Z_{1.0}$ and V_{S30} to $Z_{2.5}$ conversion relations of Chiou & Youngs (2014) and Campbell & Bozorgnia (2014), which are currently implemented in PyPSHATest. Other conversions can be found in the literature and may be added in due course.

It should be emphasised that Table 1 defines the minimum set of attributes needed, and where additional information is available these can also be supplied in the flatfile. As additional attributes are not needed by the GMMs they may not influence the residual calculation, but they may still assist with managing and filtering the flatfiles prior to the residual analysis.

3.2. Compiling Ground Motion Data for PSHA Comparisons: The Leaky Data Pipeline

In compiling ground motion data for comparison against probabilistic seismic hazard curves, we must acknowledge that our objectives differ from other applications of ground motion records. Whether the data is collected from accelerometer networks, broadband velocity sensors, or both, we are looking to assemble a complete archive of ground motions of relevance at the station. This contrasts with many seismological applications such as tomographic studies or even GMM development, which require that the set of records is of a high enough quality to be usable for the intended purpose. In that sense,

completeness of the history as a station(s) is not a key objective and few, if any, existing compilations of ground motion records found in flatfiles or other databases will have been constructed.

To illustrate how and why we do not necessarily produce a complete archive of observations, even when stations are in operation in a region, consider the illustration of the "leaky data pipeline" in Figure 15. An earthquake takes place in an idealised, well-instrumented region and this earthquake produces shaking at the locations of 100 sensors. The numbers are selected as purely illustrative and do not necessarily represent the exact proportions of records lost. From those 100 stations a proportion of sensors do not record the strong ground motion. This can happen for a variety of reasons, but among the most common are that the sensor may be offline, the environment might happen to be unusually noisy (e.g., rainstorm, anthropogenic noise etc.) or simply the motion too weak for the STA/LTA trigger.

The majority of sensors that *should* have recorded the motion usually do so, but the next leaks are in the data archiving pipeline itself. It is important to recall that unless the data are being used for real-time impact assessment, there is usually a gap in time between the event occurring and the compilation of the waveform archive. During this time some recordings may be removed by the primary (local) network operator, which may be due to known problems with the instrument, noise and potentially errors in the timing of the instrument that may result in it being misattributed to another event. These issues are less common and can sometimes be resolved at later dates after manual investigation by the operator. Assuming that the waveform record is archived by the primary operator, there can also be a loss of observations *at the time of compilation* if the compilation is being made via a secondary archive of webservice. EIDA/ESM/RRSM is an example of a secondary archive that relies on data feeds from the contributing networks. In some cases, however, this second step of data archive and retrieval can also result in loss of observations. Possible reasons can include decisions by the primary network to restrict access to certain stations or simply not make them available to the secondary archive, embargoed data, downtime of the data feeds from the local network to the federated service at the time of retrieval, or again, incorrect information in the waveform metadata that means the waveform is not included in the data collection.

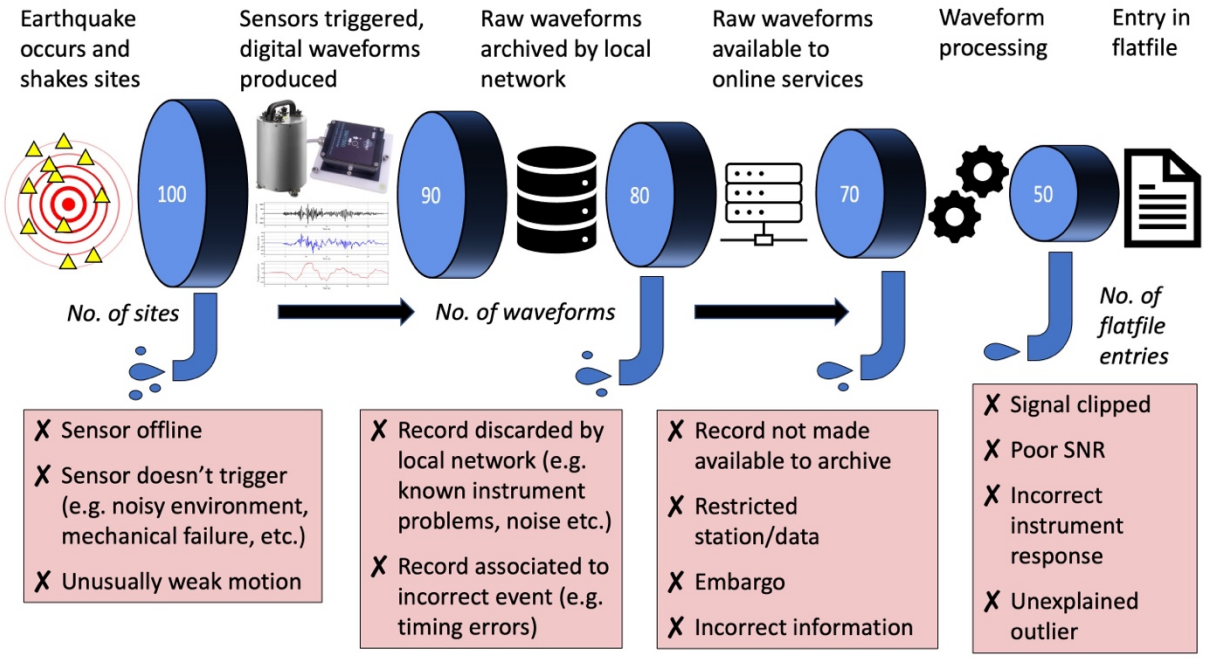


Figure 15: The observation processing chain and causes of loss of observations from the processed database. The numbers in the circles indicate the relative proportion of records retained at each step (starting with 100); however, these are illustrative and vary from case to case.

The final, and potentially largest, "leak" in the observation pipeline comes at the waveform processing stage. Here we encounter problems with a proportion of remaining records that may prevent us from retrieving an accurate measure of the required ground motion parameters. These problems include

clipping of the waveforms (more prevalent when using records from weak motion broadband sensors), poor signal to noise ratio (SNR) across the required frequency range, identified errors in instrument response (e.g., incorrect gain), contamination of the record from soil structure response and/or housing, or simply that the record is an unexplained outlier.

At the end of this pipeline, we may be left with a suite of usable records from which we can retrieve accurate ground motion parameters; however, this is no longer a complete archive of the history of shaking in a region. The extent of the loss of data will naturally vary from region to region and network to network, and when considering ground motions of engineering significance, it is less likely that stations will lose data through, for example, not triggering or poor SNR. However, we must acknowledge that expanding our data set to consider ground motions from both weak and strong motion recording networks that these problems may contaminate a greater proportion of the data set and overall completeness becomes lower.

With the leaky data pipeline it is not easy to know exactly what proportion of the expect observations are missing in the data set and when. This is not trivial for our purpose because for quantitative comparisons of expected and observed rates of exceedances at stations in a region our results may be at best somewhat biased, and at worst completely misleading, in suggesting the expected hazard is overestimating the observed rates of exceedance, if in our observed data set we record only a small to moderate proportion of the actual exceedances that occurred.

If we are aware of the problems of loss of observations, what options are available to help quantify this and somehow correct for this when making the comparisons?

1) Limit the stations under consideration to only those known to have the longest, and/or most complete, period of operation.

2) If we are aware of periods of station downtime during its history, then retain the observations but re-scale the duration of operation at the station, t_s.

3) Use known information about earthquakes that occurred to make inferences of the corresponding expected motion at a site if no observation is found in the archive.

The first of the options may be the simplest to apply, and the resulting archive may be considered more confidently as a complete history of strong shaking at the sites. There are several significant downsides, however: i) we are likely to discard a large proportion of the available data, leaving a smaller number of available sites from which statistical inferences are weaker, ii) we rely on prior information about station downtime and still can't necessarily know when and how certain records a lost, iii) we may still lose a significant proportion of records in the waveform processing stage. Also, stations that have been in operation over the longest time spans are likely to have seen several changes instrumentation, potentially even in location and housing too. Relying on this subset of stations may also introduce other problems.

The second option of reducing the operation time of the station, t_s, if the station downtime periods are known is appealing. In Europe, the EIDA webservices have developed particularly useful tools for automatically retrieving waveform metadata for its contributing stations via the EIDA webservice waveform catalogue (EIDAWS-WFCatalog) (http://www.orfeus-eu.org/data/eida/webservices/). This information feeds the EIDA Data Availability web tool (http://www.orfeus-eu.org/data/eida/quality/availability/), and allows us to create a day-by-day archive of data availability by station and sensor for most of the stations in EIDA. In developing the ground motion database of strong and weak motions we also extracted daily station operation archives for each station and instrument, which give a percentage of data availability. This revealed some particularly critical insights into the availability of ground motion data in Europe, in which we see that the "real" duration of operation of a station (which we define as the total number of days in which the data availability at a station is 90 % or more) is highly variable by network, site and sensor. In almost all cases this real duration is less than the operation period indicated for the site by the EIDA station book, and the gaps may be significant (on the order of months or years).

While the EIDA waveform metadata services are an invaluable tool for our purposes and *may* be able to identify periods of significant station/sensor downtime, it still does not solve all the problems in the leaky pipeline. Specifically, it cannot necessarily identify when stations do not trigger because of high

noise, nor can it help identify the proportion of records lost in the waveform processing stage (though this could potentially be worked out with some laborious cross-checking of the flatfile against the original data source).

Compared to the first and second option, the third option redirects the problem away from identifying and correcting for incompleteness and instead toward making the archive itself complete. Here, we require first a reference homogeneous earthquake catalogue for a region that we consider complete above the magnitude of interest for PSHA. This catalogue is taken as the true record of earthquakes that have occurred in a region, and from the times, magnitudes and locations of the events in the catalogue we can use existing GMMs to predict the distribution of ground motions at the stations where they either i) would be expected but are not found in the flatfile, or 2) the stations were not yet in (or have gone out of) operation. Where the ground motion record for the event is found then this record is taken as an observation without uncertainty, otherwise the "observation" of ground motion takes the form of a normal distribution with median and standard deviation predicted by the selected GMM(s).

The third approach can be termed *data imputation by regression* as we are using existing models of GMM to substitute for missing data. To a certain extent there are precedents for this approach in the literature. Ward (1995) uses existing ground motion models for California (thus calibrated primarily on Californian data) to predict the expected ground motions across a regular grid of sites from a catalogue of known events. In their case they do not use observed records directly, but the concept is similar. A clearer precedent can be found in Tasan et al. (2014) who apply data imputation by regression to fill in gaps in observed data from France and Turkey when comparing against PSHA models. Imputation by regression has some specific advantages over other methods. Firstly, as we do not need to know why an expected observation is not present in the flatfile, we can compensate for all potential causes of data loss and do not necessarily need to change the duration window. In fact, we are able to do the opposite, which is to pad out the observation periods for any station, when necessary, such the duration periods for all stations are normalised to the same t_s. This can have some particularly useful consequences for the statistics of the comparisons being made in section 4.

Of course, however, we must also account now for the uncertainty in the imputed ground motion and treat the "observed" history now describes a distribution of exceedances (in most cases). Similarly, the use of an existing GMM to impute observations also blurs the line between model prediction and observation for testing, as the observations may to a greater or lesser extent represent the predictions of a not necessarily independent model and may introduce their own biases. The introduction of new biases due to the GMM can be moderated, if not avoided entirely, by selection of a GMM that is shown to fit well to the observed ground motion data for a region. "Fit well" in this sense would mean to have a minimal demonstrable bias the scaling of ground motion residuals with respect to the most critical source, path and site properties. If multiple GMMs were considered adequate for data imputation in a region then one might wish to select one or more GMMs that were not used in the considered PSHA models, in order to reduce the possibility of certain PSHA models being favoured by virtue of the common GMM in the model and the data prediction. In some cases, this may be unavoidable, however, and a more reasonable approach might be to repeat the imputation process using several GMMs and demonstrate the outcomes of the quantitative hazard model to data comparisons are reasonably robust to the choice of model.

3.3. Filling in the Gaps: Data Imputation via Mixed Effects Regression

In the previous sub-section we have identified the *data imputation via regression* approach as the preferred option for addressing the issue of incompleteness in our observed ground motion data. In their previous application of this approach, Tasan et al. (2014) use the GMMs in their original form for the purpose of imputation, meaning that the distribution of ground motions predicted by the GMM corresponds to that of the entire ergodic aleatory uncertainty. However, while this approach has the aforementioned benefits, it discards information that we do have from the observed ground motion data, and this information can be used to help reduce the uncertainty in the imputed observations when available. We illustrate this concept with two cases in Figure 16 and consider a ground motion model with the form:

$$\ln Y_{es} = f(M, R, \boldsymbol{\theta}) + \delta_T \quad 3.1$$

Where $\ln Y_{es}$ is the logarithm of the observed ground motion at site s from event e, $f(M, R, \boldsymbol{\theta})$ is the expected median ground motion from the GMM as a function of event magnitude, M, source-to-site distance R, and set of other predictive parameters $\boldsymbol{\theta}$ depending on the GMM in question. δ_T is the total residual (the difference between the observation and prediction) and is a normally distributed variate with mean of 0 and standard deviation σ.

On the left image in Figure 16 we see a single event occurring in Northern Italy, which is recorded by 77 stations in our data file (mixing strong and weak motion). These stations are not all the stations that would have been expected to have recorded the earthquake, so some observations from this event are missing. However, with enough observations of the event we can make an inference as to whether the event is more or less energetic than expected by the ground motion model, implying higher or lower stress drop respectively. We quantify this by the between-event residual δB_e and is normally distributed with a mean of 0 and a standard deviation τ. The between event residual is routinely determined in the regression of a GMM using mixed effects regression, such that the GMM in 3.1 becomes:

$$\ln Y_{es} = f(M, R, \boldsymbol{\theta}) + \delta B_e + \delta W \quad 3.2$$

where δW is the ergodic within-event residual.

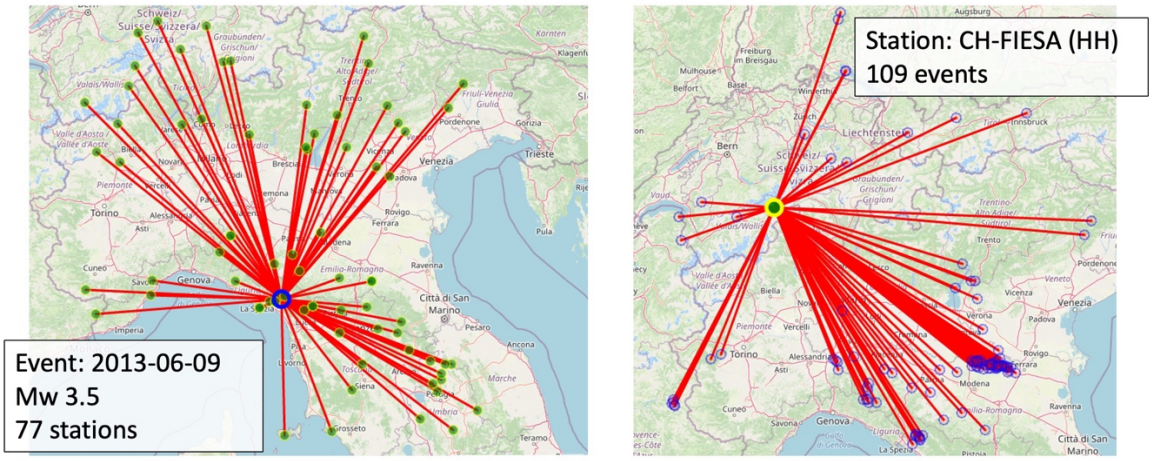


Figure 16: Illustrations of typical cases of numbers of stations recording an event in the ground motion database (left) and number of events recorded by a single station in the database (right)

On the right image in Figure 16 we consider a single recording station in Switzerland (CH-FIESA) for which observations of ground motion from 109 events are found in the data file. We do not know whether this is the complete history of observations at the site but even if it is not, it is a sufficient number of observations for us to determine whether the ground motion at the site is systematically higher or lower compared to the centre of the full distribution of site-to-site variability implied by the GMM. This is the site-to-site (or between-station) residual, $\delta S2S_s$, and is a normally distributed variate with a mean of 0 and a standard deviation ϕ_{s2s}. The above ergodic within-event residual can be separated into the site-to-site component $\delta S2S_s$, and the remaining site-corrected within-event component, δW_{es}:

$$\ln Y_{es} = f(M, R, \boldsymbol{\theta}) + \delta B_e + \delta S2S_s + \delta W_{es} \quad 3.3$$

Where δW_{es} is a normally distributed variate with mean of 0 and standard deviation ϕ_0. With this separation of the ground motion residuals into its between-event, site-to-site and site-corrected within-event components (random effects) then the total ergodic aleatory variability of the ground motion at a site is given by:

$$\sigma = \sqrt{\tau^2 + \phi_{s2s}^2 + \phi_0^2} \quad 3.4$$

The decomposition of residuals through mixed effects regression is commonplace in ground motion modelling. For a single site with repeated observations (such as that of CH-FIESA in the example) part of the ergodic aleatory variability can be removed from the total standard deviation to produce a *partially non-ergodic* prediction of ground motion. Where data are sufficient to do so further repeatable random effects such as source-region to source-region variability ($\delta L2L_l$) and path-to-path variability ($\delta P2P$) (Al Atik et al., 2010). Such GMMs would be considered *fully non-ergodic* models, and we shall return to the relevance of these for probabilistic seismic hazard analysis later in Section 4.

With a ground motion data set sufficiently large to constrain δB_e and δW_{es} for thousands of earthquakes and recording stations we can capitalise on this information within the data imputation process to make better predictions of the missing data and reduce uncertainty on the observation in the process. In this sense we are using *data imputation by mixed effects regression*. The general workflow for this is illustrated in Figure 17. Initially we identify (or construct) a reference catalogue of earthquakes for the region and time period of interest, which should be harmonised in terms of moment magnitude. We then associate the magnitudes in the catalogue to their corresponding events in the flatfile and filling in additional metadata where necessary (see Section 3.2) we undertake a comparison of the ground motion data against selected GMMs for a range of intensity measures/spectral periods of interest. The comparisons are undertaken by deconstructing the residual $\delta = \ln Y - f(M, R, \boldsymbol{\theta})$ into its random effects components via mixed effects regression (Bates et al., 2015). This returns a set of δB_evalues for many events in the catalogue, and a set of $\delta S2S_s$ values for some stations. To make a comparison between the data and a specific model, we also need to know the minimum magnitude considered for the model itself, M_{min}, and the maximum source-to-site distance R_{max} (the *integration distance* in OpenQuake's terminology).

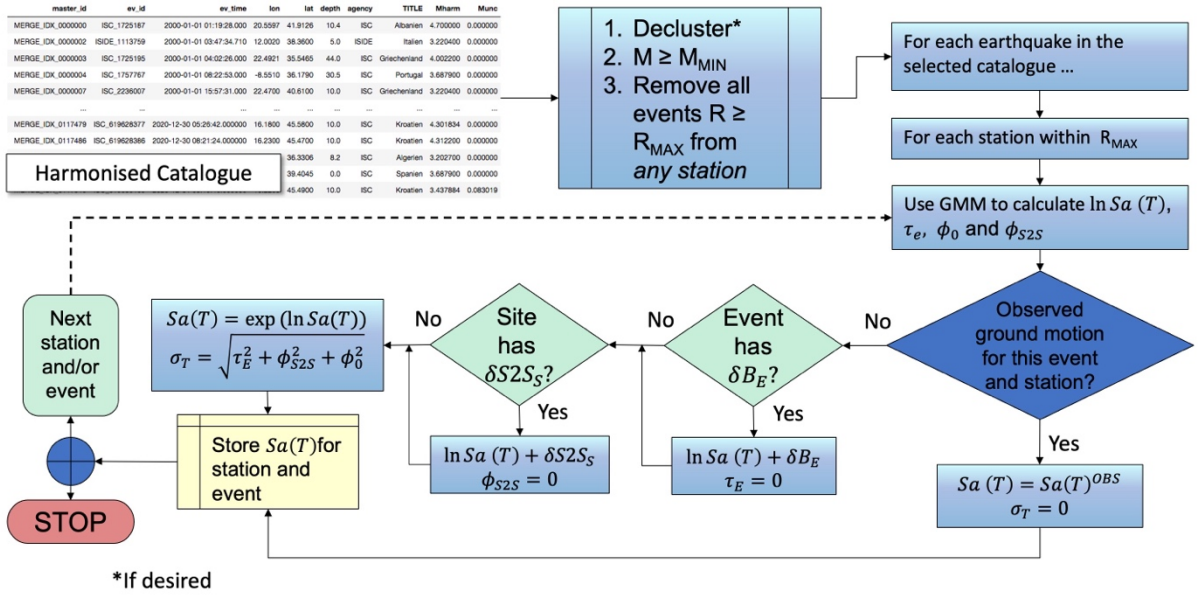


Figure 17: Workflow for data imputation via mixed effects regression

The workflow in Figure 17 evaluates the catalogue event by event above the minimum magnitude, and for each event identifies the stations within R_{max}. For the selected GMM, the expected ground motion at the station $\ln Sa\,(T)$ is calculated from a GMM ($\overline{\ln Sa(T)}$) along with the between-event, site-corrected within-event and between-site standard deviations (τ_e, ϕ_0, and ϕ_{s2s} respectively). If an observed ground motion for the specific event and site is available this is taken as the observed value $Sa(T)^{OBS}$ without uncertainty and the algorithm proceeds to the next station/event. If no observed ground motion is available, the algorithm will then check to see if a *valid* δB_e for the event is present. If so, then this is added to $\overline{\ln Sa(T)}$ and the corresponding τ_e set to 0, otherwise δB_e is set to 0 and τ_e from the GMM is retained. The algorithm will then check if a *valid* $\delta S2S_s$ is present for the site in question. If so, then this too is added to $\overline{\ln Sa(T)}$ and the corresponding ϕ_{S2S} set to 0, otherwise $\delta S2S_s$ is set to 0 ϕ_{s2s} from the GMM is retained. The imputed ground motion at the site is then determined from $\mathrm{Sa(T)^{IMP}} = \exp(\overline{\ln Sa(T)} + \delta B_e + \delta S2S_s)$ and the resulting log-normal standard deviation as defined

in equation 3.4. In this formulation, where no observed ground motion from a specific earthquake at a specific site if found but would be expected, the best case is that a valid δB_e and $\delta S2S_s$ are available, in which case the imputed ground motion is adjusted using these linear terms and the resulting uncertainty reduces to just the site-corrected between-event standard deviation ϕ_0. The standard deviation then increases depending on which information is available, until one reaches the case that no valid δB_e and $\delta S2S_s$ is available, wherein the total ergodic standard deviation σ_T is used.

This approach of *data imputation via mixed effects* regression effectively develops on that adopted by Tasan et al. (2014), but with the novelty that where *some* observations are available for a given earthquake and a given site then these can be used to calibrate the imputed data and reduce uncertainty with respect to that of a fully ergodic GMM. In the description above we have used the term *valid* to describe both a δB_e and a $\delta S2S_s$, and what this effectively means is an estimation constrained by enough observations. The exact number of observations for sufficient constraint may be somewhat subjective, but for the current purposes we define a δB_e or $\delta S2S_s$ are being "well-constrained" if it is based on 10 or more observations and "unconstrained" if it is based on less than three observations. From a purely statistical perspective, uncertainty on δB_e and $\delta S2S_s$ could be propagated into the imputation process by setting τ_e and/or ϕ_{S2S} equivalent to the measurement error on their respective residual terms. This is not currently supported in the tools but may be added in future.

3.4. Setting up a PSHA Testing Case Study: France

In Section 2 we focused on the example of France as a test case study for comparing distributions of source models between the Drouet et al. (2020) model [FR2020] and the 2020 European Seismic Hazard Model [ESHM20], and then looked at the France-Germany border region for comparing hazard distributions. With these same models we will illustrate the use of PyPSHATest on a case study application for comparing probabilistic seismic hazard maps against observations of ground motion for metropolitan France. The details of the data compilation process are summarised here. As a region of low-to-moderate seismicity, France provides a particular challenge for quantitative comparison of PSHA against observations, with relatively few earthquakes and few recorded ground motions of engineering significance against which we can compare out models. As such it is a region that is not abundant in usable records of shaking from strong, and those histories of ground motions that are present are likely to be incomplete. In this case study, however, we will show how the data available can still be used for quantitative comparisons against PSHA models.

3.4.1. Building the Database of Ground Motion Observations

With the PSHA models for France (FR2020 and ESHM20) already selected and implemented in OpenQuake, our first major challenge is to set up a database of ground motion observations for France and the surrounding regions. Despite its low-to-moderate seismicity, France boasts a high-quality seismological recording network operated by Réseau Seismologique et Géodésique Français (RESIF), with the vast majority of recorded data made available via the EPOS European Integrated Data Archive (EPOS-EIDA). This includes both the RESIF-RAP Accelerometric Permanent Network (in operation since 1995) and the RESIF Broadband network in operation since 1962. These networks along may provide a large body of data from which to make our comparisons; however, we can also gain important insights into ground motions from earthquakes occurring in and around metropolitan France by also accessing data from neighbouring countries, including Spain, Germany, Switzerland and Italy. With many of the seismological recording networks in these respective countries providing access to their data via the EPOS-EIDA portal we begin by throwing the net wide and compiling a large-scale database of both weak and strong ground motion observations across a broader region of western Europe to create the northwest European ground motion flatfile.

To build the flatfile we use a bulk data download and automatic record processing software called Stream2Segment (Zaccarelli et al., 2019), which allows us to access ground motions made available via both the EPOS-EIDA and Incorporated Research Institutions for Seismology (IRIS) webservices. The specifications of the download are as follows:

Time period: 01/01/2000 00:00:00.0 (start) to 01/01/2021 00:00:00.0 (end) (21 full calendar years of data)

Region: Earthquakes should fall within -6°E to 19°E (west to east) and 40°N – 55°N, which we limit to a depth range of 0 – 50 km hypocentral depth

Magnitudes: M ≥ 3.5 in *any* scale, according to the International Seismological Centre event service.

Stations: All stations within 3.5° of the epicentre location

Channels: Selected channels are HN (accelerometric), HG (accelerometric), HH (broadband velocity), EH (broadband velocity) and HL (broadband velocity). Channel depth must be less than 20 m.

Waveforms: Data are downloaded for a total of 4 minutes, beginning one minute before the theoretical P-wave arrival time computed considering a global velocity model, and continuing for 3 minutes subsequent to the expected P-wave arrival.

The downloaded waveforms (segments) are automatically processed with the following selection criteria applied:

1. Clipped signals are removed based on a threshold of 80 % of the maximum number of counts for a 24-bit digitizer
2. Instrumental response is then removed, with different pre-filters according to the magnitude
3. Signal to noise ratios (SNR) were computed in the Fourier domain for 13 different ranges of bandwidths, spanning f_{min} = 0.5 Hz (M ≤ 3.0), 0.3 Hz (3.0 < M ≤ 6.0) or 0.08 Hz (M ≥ 6.0) to $f_{max} = \max(40.0, 0.9 \cdot f_{Nyquist})$. Records with SNR < 3 in the frequency range 0.15 Hz to 6.3 Hz were removed.
4. Multiple arrivals are identified based on the occurrence of significant peaks in the second derivative of cumulative squared acceleration filtered over the bandwidth $0.5 \leq f(Hz) \leq 15$.
5. Records from the broadband velocity channels HH and EH were removed when M ≥ 5.5 at distances R_{EPI}≤ 50 km.
6. Records are excluded if the ratio of PGV for the two horizontal components was less than 0.1 or greater than 10.0.

For all the resulting records the Fourier amplitude spectra (FAS) and response spectra are calculated. Finally, as a consistency check the PGA and PGV were compared to those expected from the Bindi et al. (2014) GMM and records removed if the between-event and between-station residuals were significant outliers.

After application of the full processing chain and selection criteria our complete database contains 177,816 individual spectra (more than 85,000 horizontal pairs) from 1,823 events recorded at 2,810 stations (operated by 77 networks). This includes both weak and strong motion records *and* both permanent and temporary network deployments. While the full flatfile spans a large region including all of France, Switzerland, Belgium and Germany, and much of Spain, Italy and the United Kingdom, the events and stations in the flatfile relevant to metropolitan France are shown in Figure 18. As an automatically processed flatfile, and one that contains both weak and strong motion, there are inevitable limitations in relation to the waveform quality. There exists the possibility in some cases that errors, incorrect station/instrument information and other problems may exist in the data. We will see in due course that there are some tools in PyPSHATest that may help to explore and identify stations that may have problems and to exclude these from the analysis. It will also be seen that while we are confronted with a large file of observations, once we apply the various testing processes it is likely that a smaller number of sites with repeated observations may be the most relevant. An initial end-to-end testing application can help identify the most important sites and/or events that are influencing the comparisons with the hazard, which can then be inspected more carefully to identify potential errors or problems that may influence the results.

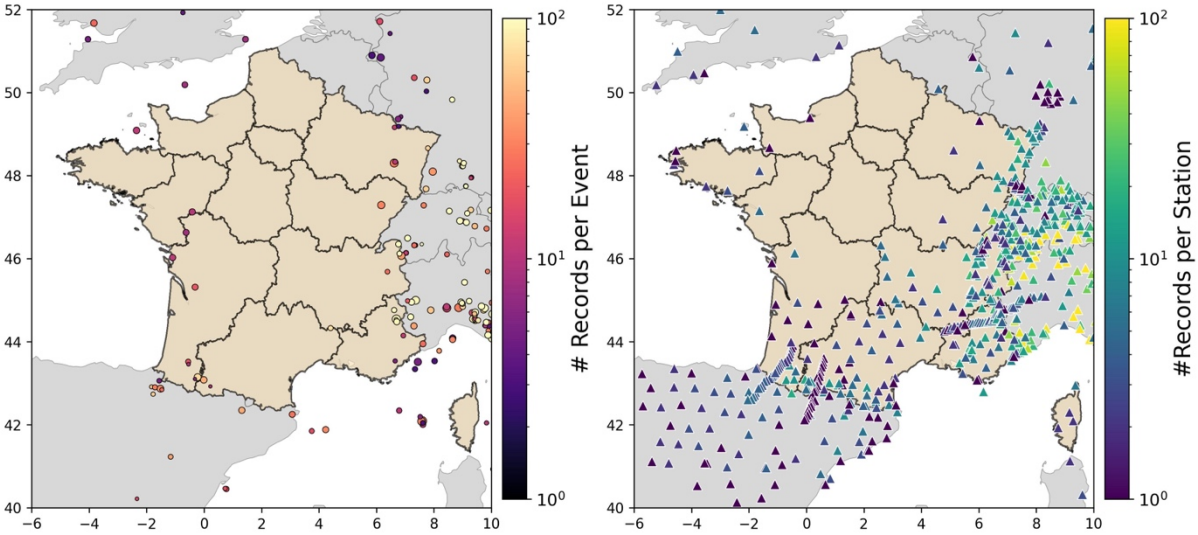


Figure 18: Flatfile of ground motions for northwest Europe in relation to the target region of France: Earthquakes (left) and Stations (Right)

3.4.2. The reference earthquake catalogue

In the data imputation process, we see that there are two points where a harmonised earthquake catalogue is necessary for testing PSHA. The first is to help us identify when and where observations are missing from the dataset by providing a "true" record of the earthquake history, from which we can identify when observations should be expected but are missing in the flatfile. When the observation is missing, the imputation process will determine a distribution of ground motion according to a GMM. Once again, the harmonised catalogue is necessary to provide a (usually) moment magnitude M_W or equivalently-scaling proxy (M_W^*) for input into the GMM to predict the ground motion at a site and, if available, to provide a δB_e value that is constrained from observations at other stations. For general usage of PyPSHATest there is no specific requirement to use a particular catalogue, nor any restriction that would prevent a non-harmonised catalogue being used (i.e., one in which the magnitudes may refer to a mixture of reported scales). The obvious limitation of using a non-harmonised catalogue is in the determination of δB_e, which will display increased variability owing to the heterogeneity in magnitude definitions and may result in apparent regional trends owing to country-by-country or network-by-network differences is magnitude estimation.

As the region we are considering spans multiple countries, we need to retrieve the earthquake locations and magnitudes from a harmonised European earthquake catalogue covering the period 2000 to 2020 inclusive. Ideally, we would have adopted the Euro-Mediterranean Earthquake Catalogue (EMEC) (Grünthal & Wahlström, 2012), which was updated for use in ESHM20 (Danciu et al., 2021). However, this catalogue is not usable for our purpose as it ends at the end of 2014 and it imposes a threshold magnitude of M_W^* 3.5, which may be slightly too high for some of the ground motion calculations. Instead, we create a new harmonised catalogue that integrates data from multiple networks in the region of interest, including many that form part of the input data for EMEC. These are harmonised using a process that is close to that of the EMEC compilation procedure, with some minor adjustments for regional biases. The input seismicity bulletins are taken from:

1. The harmonised SiHex instrument catalogue for France updated to the end of 2020 (https://gitlab.com/eost/bulletins/-/tree/master/sismicite_1962_2020)
2. The bulletin of the Swiss Seismological Service (http://seismo.ethz.ch/en/home/)
3. The Italian Instrumental Paramteric Database (ISIDE) (https://terremoti.ingv.it/en/iside)
4. The parametric earthquake catalogue of the Bundesanstalt für Geowissenschaften und Rohrstoffe (BGR) Germany (https://www.bgr.bund.de/EN/Themen/Seismologie/Erdbebenauswertung_en/Erdbebenkataloge_en/historische_Kataloge/germany_en.html)

5. The parametric earthquake catalogue of the Royal Observatory of Belgium (ROB) (https://www.geo.be/catalog/)
6. The parametric earthquake catalogue of the British Geological Survey (BGS) (http://www.earthquakes.bgs.ac.uk/)
7. The pan-European earthquake catalogue produced by GFZ GEOFON (https://geofon.gfz-potsdam.de/)
8. The bulletin of the International Seismological Centre (http://www.isc.ac.uk/)

The respective bulletins and the polygons used in the EMEC harmonisation process are shown in Figure 19, while the harmonised catalogue itself is shown in Figure 20. Altogether the harmonised instrumental catalogue contains 19,159 events with $M_W^* \geq 3.5$ spanning the period 2020 to 2021 (from a total of 117,543 in the whole catalogue).

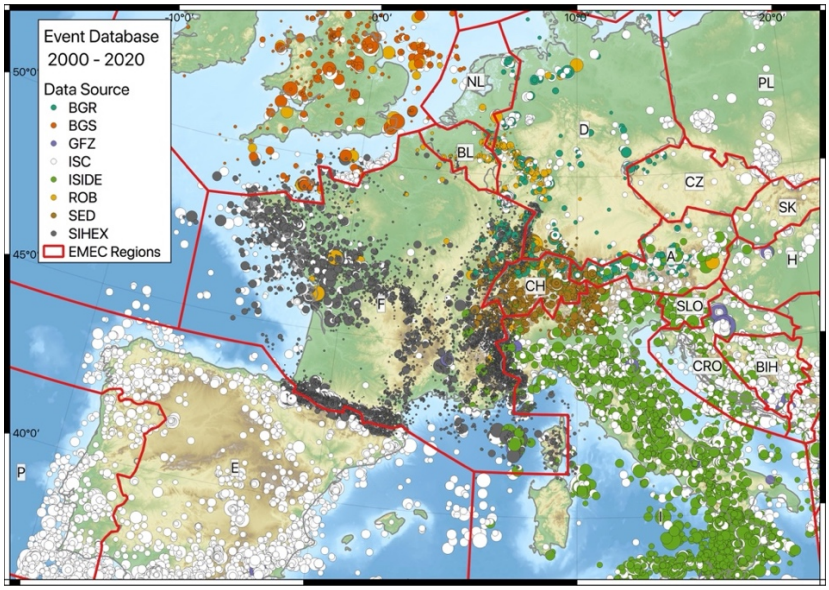


Figure 19: Earthquake and their data sources used in the compilation of the harmonised catalogue. EMEC harmonisation regions shown by red polygons

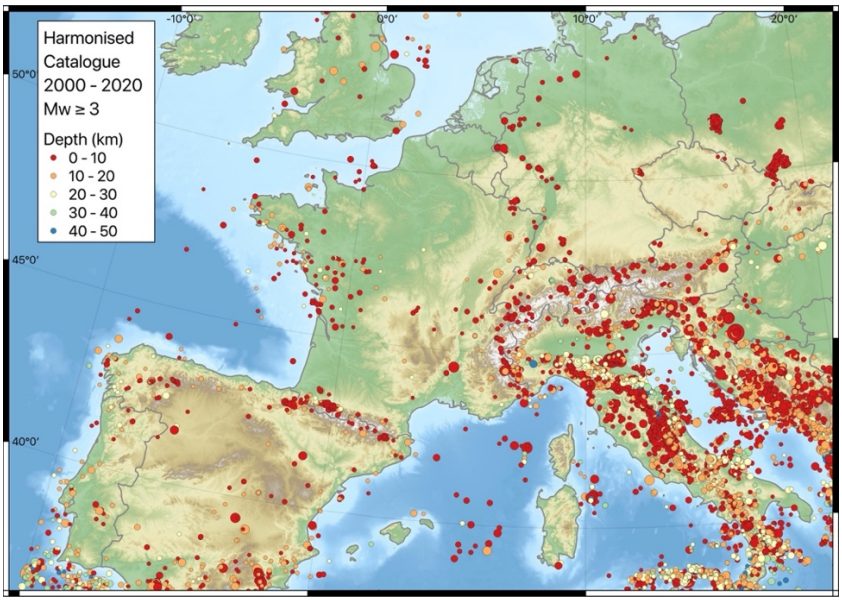


Figure 20: Harmonised earthquake catalogue with $M_W \geq 3$ for northwest Europe for the period 2000/01/01 to 2020/12/31

Once again, we must acknowledge that there are caveats still in the process pertaining to the accuracy of the catalogue, which are relevant when we consider that this catalogue forms the basis for the "true" history of earthquakes to have occurred in the target region and period. Errors in location and harmonised magnitude are possible, and while harmonisation should go some way to correcting for regional biases in the ground motion residuals, there may still be biases due to the magnitude conversion scales and other information used. We can state, however, that δB_e was determined prior to harmonisation and after it, and the harmonisation did appear to correct both magnitude- and region-dependent trends in the between event-residuals. So, we consider this to be adequate for the current purpose.

3.4.3. Harmonising the metadata

While the harmonised earthquake catalogue provides a common magnitude for all the events in the flatfile, there is still a need to add additional data that is required for some (or all) of the GMM under consideration. For the sites themselves we need an estimate of V_{S30}, which we take as either the measured V_{S30} if it is available for the station, or else we adopt the topographically-inferred proxy V_{S30} compiled as part of the 2020 European Seismic Risk Model (Weatherill et al., 2023c). For the measured V_{S30}, this information along with the operational start date, end date and status is taken from the EPOS EIDA station book (http://orfeus-eu.org/stationbook/). As several GMMs adopt different coefficients for the site scaling term and/or variability when the V_{S30} is either measured or inferred, this information too is added to the flatfile.

For the source and path distances we are confronted with a requirement that many GMMs require parameters derived from the finite fault source representation of an event. This is available for only a few earthquakes found in the catalogue, so we need to find a way to manage these parameters in the absence of a finite fault. While the vast majority of the events are small to moderate magnitude ($M_W \leq 5.0$) the finite fault can be adequately approximated by a point source such that $Z_{TOR} = Z_{BOR} = h_D$, $R_{JB} = R_{EPI}$, $R_{RUP} = R_{HYPO}$, $R_{y0} = R_{EPI}$ and $R_X = -R_{EPI}$ (this effectively turns off hanging wall scaling, albeit that for small magnitude events hanging wall scaling will generally taper to 0 in those models that have it). If a focal mechanism is available then strike, dip and rake are sampled randomly from the two planes, while if no focal mechanism is available then values of dip = 90°, rake = 0° and strike = 0° are assumed. For larger events a focal mechanism is always available, so to calculate the finite fault properties we generate a random surface by selecting the nodal plane randomly from the two available and placing the hypocentre in the centroid of the fault plane. Fault plane length and width are calculated based on the rupture expected from the Wells & Coppersmith (1994) magnitude-area scaling relation, with an assumed aspect ratio of 1.25.

3.5. Using PyPSHATest for Building the Station Database

Once we have the flatfile and the harmonised earthquake catalogue, we can undertake the ground motion residual analysis that will contribute to the database of station information that will be used in the data imputation process in the following section. This database forms the core of PyPSHATest and can allow us to make comparisions between models and data, and to explore the ground motion data for our test region in more depth.

The workflow for this process is broken down into several steps: loading of the flatfile, merging the harmonised catalogue, setting up the data for determination of the GMM residuals, execution of the GMM residuals and storage into an hdf5 database (separate from the main station database). Several library imports are needed: 1) Numercial Python (Numpy) and Pandas, 2) OpenQuake's functions for creating the GMMs and the intensity measures, and 3) from PyPSHATest we need the catalogue parser for a generic csv catalogue file ("harmonised_catalogue_csv_parser"), the a tool to load and manage the flatfile ("FlatfileFormatter") and tools for executing the residual analysis ("Residuals") and for visualising the results ("ResidualVisualizations). These can be imported as illustrated below:

```python
# Useful Python Libraries
import numpy as np
import pandas as pd

# OpenQuake libraries
from openquake.hazardlib import valid
from openquake.hazardlib.imt import SA, PGA, PGV

# PSHA Testing Tools (imported in order of use)
from pshatest.catalogue_tools import harmonised_catalogue_csv_parser
from pshatest.flatfile_handler import FlatfileFormatter
from pshatest.gmm_residual_analysis import Residuals, ResidualVisualizations
```

In the first step we need to load the flatfile. The flatfile has been constructed in this case initially as a pandas DataFrame, which can be exported to several different formats. Our preferred choice is usually either as a Pandas HDF5Store (an hdf5 binary formatted to store dataframes and their metadata) or as a comma separated value (csv) file. These can be loaded as follows:

```python
# Load in the data from hdf5
flatfile_raw = pd.read_hdf(
    ".path/to/flatfile.hdf5",
    key="complete"  # Name of data table in hdf5
)

# ... or from csv
flatfile_raw = pd.read_csv(
    "./path/to/flatfile.csv",
    sep=",",
)
flatfile = FlatfileFormatter(
    flatfile_raw,
    "A Sample Flatfile Name",  # Flatfile name or description
    station_id_type="network-station-channel",  # How to define the station ID
)
```

One important parameter here is the "station_id_type", which defines how the station is identified in the residual analysis. Two options can be input here: 1) "network-station", which will create a unique ID combining just the network and station name, and in doing so will then group together all observations from all channels recording at the station and not separate broadband from accelerometrically recorded ground motions, or 2) "network-station-channel", which will create a unique ID for each channel at the site, allowing for observations to be separated out by sensor type and thus create a separate $\delta S2S$ for each sensor, albeit each sensor may be constrained by fewer observations. In this case, we choose to separate the station IDs per channel and thus calculated residuals separately. The flatfile *must* contain the information and headers indicated in Table 1.

In the next step we can load the harmonised catalogue and merge it to the flatfile (this need only be done once, and the additional information stored with the flatfile for future use). An example of how to do this is as follows:

```python
# Load in the harmonised catalogue from csv
harm_cat = harmonised_catalogue_csv_parser(
    "./path/to/the/harmonized_catalogue.csv",
    "NWEurope",  # Name for the catalogue
    mmin=3.0, mmax=np.inf,  # Minimum and maximum magnitude
    start_time="2000-01-01T00:00:00.0",  # Start time (UTC) in ISO format
    end_time="2022-01-01T00:00:00.0",  # End time (UTC) in ISO format
    is_master=True,  # This is a magnitude-harmonised catalogue
)
# This returns a pshatest.catalogue_tools.Catalogue object,
# which contains the catalogue dataframe as an attribute
# .catalogue
flatfile.merge_harmonised_catalogue(
    harm_cat.catalogue,  # Catalogue
    time_window=60, # Maximum time difference (in s)
    distance_window=30.0,  # Maximum distance difference (in km)
)
# Fixes other metadata
flatfile.cleanup()
```

In this example the catalogue is stored in a csv file with the required headers: "ev_id" (the unique event ID), "ev_time" (the event time as a datetime string, "lon" (event longitude °E), "lat" (event latitude °N),

"depth" (hypocentral depth, km), "M" (harmonised magnitude). Filtering of the catalogue can be done during the loading process as shown in the example, where we limit it to $M_W^* \geq 3.0$ and to the end of 2022. With the catalogue loaded, this is merged with the flatfile. As the event time and location in the flatfile may not correspond to that in the harmonised catalogue, we must merge via spatio-temporal association. In the above example the event in the harmonised catalogue is associated to that in the flatfile if it falls within ± 60 s of the flatfile event time, and within 30 km of the flatfile location. For most modern instrumental catalogues, a smaller temporal window could be assumed, though we find $\delta t \pm 30$ to 60 seconds usually sufficient in this case.

With the flatfile now ready and harmonised, the next step is to setup the data ready for the residual calculations. This is done as follows:

```
# Define a set of filters, e.g.
# Depth < 50 km
# M > 3.0
# RJB <= 350 km
# 150 <= Vs30 (m/s) <= 2500
filter_set = {
    "ev_depth": ("range", 0.0, 50.0),
    "mag": ("range", 3.0, np.inf),
    "r_jb": ("range", 0.0, 350.0),
    "vs30": ("range", 150.0, 2500.0),
}

# Returns the ground motion database for input
# into the residual calculations
gmdb_strong, gmdb_weak = flatfile.setup_gmvs(
    filters=filter_set,  # Filters to be applied
    split_strong_weak_motion=True,  # Split the strong/weak motions
)
```

To select the data for input into the residual calculations we can define a set of filters for any attribute in the flatfile. In the above example we choose to limit the data to $h_D \leq 50$ km, $M \geq 3.0$, $R_{JB} \leq 350$ km and $150 \leq V_{S30}(\mathrm{m/s}) \leq 2500$. Filters are set as tuples of "('range', low, high)" or "('equality', value)". The filters will also automatically remove any ground motions for which the selected filter attribute is missing, so the filters can be useful even if opting to keep as many of the records as possible. With the filters define then we export our flatfile into two Dictionaries, one containing all the ground motion values and metadata for the strong motion records, the other for the weak motion records. This made possible with the argument "split_strong_weak_motion", which will split the database if set to True, or keep all records together otherwise.

The data is ready for calculation of the residuals. First, we define the list of IMTs and GMMs for which we plan to calculated the residuals. Here we need the OpenQuake tools imported previously, which allow us to define the IMTs as PGA, PGV or and SA(T) in terms of OpenQuake's "IMT" object. Then the GMMs should be defined as a list of tuples, with each tuple containing a compact shortform name of the GMM and a valid instance of the corresponding OpenQuake GMM object:

```
# Define the selected intensity measured as
# OpenQuake IMT objects
imts = [SA(0.01), SA(0.025), SA(0.04), SA(0.05),
        ..., SA(9.0), SA(10.0)]

# Provide a list of the GMMs and corresponding
# labels for their storage in the database
gmms_labels = [
    ("Akk2014", valid.gsim("AkkarEtAlRjb2014")),
    ("Ameri14", valid.gsim("Ameri2014Rjb")),
    ("ASK2014", valid.gsim("AbrahamsonEtAl2014")),
    ...
    ("ESHM20", valid.gsim("KothaEtAl2020ESHM20")),
]
```

Finally, we are now ready to run the residual calculations. This is done in two steps following the example below. In the first step we setup the class ready for calculation, then in the second the calculations are

run. Depending on the size of the ground motion database, the number of GMMs and the number of IMTs, calculating the residuals may take a while to run (anything from a couple of minutes to a couple of hours). We recommend storing these results to file straight away, which we can do via ".to_hdf".

```
# Setup the residuals for the strong motions
resids_strong = Residuals(
    gmdb_strong,  # GM data ready for calculation
    gmms_labels,  # List of GMMs and their labels
    imts,  # List of IMTs
)
# Calculate the residuals
resids_strong.run()

# Store the residual calculations to a binary file
resids_strong.to_hdf(
    "./path/to/strong_motion_residuals.hdf5"
)
```

The above step is then repeated for the weak motion data.

If one wishes to retrieve results from previous calculations without re-running, there is also the option to create the object from the datastore:

```
# Create a Residuals object with results
# from this (or other) analysis
resids_previous = Residuals.from_hdf5(
    "./path/to/strong_motion_residuals.hdf5",
    gmms_labels,  # Need to manually specify the GMMs and labels
)
```

Once the residuals are calculated we will want to inspect the results for trends with respect to different predictor variables. These can help identify when GMMs are systematically over-/under-estimating the ground motions and/or where they may not capture elements of source, path and site scaling well. To facilitate this, we have included a set of flexible visualisation tools, which can be found in the object "ResidualVisualizations". Below provides an example of how, for a given GMM and IMT, we can create a plot of δB_e against magnitude, $\delta S2S_S$ against V_{S30} and δW_{es} against Joyner-Boore distance:

```python
# Setup tool for visualizing results
visualizer = ResidualVisualizations(resids)

# Plot the between-event residuals (dBe)
# with respect to magnitude
visualizer.between_event(
    "ASK2014", # GMM (label)
    "SA(0.2)", # IMT (as string)
    "mag", # Variable to plot on x-axis
    xlim=(3.5, 7.5), # Limits of the y-axis
    ylim=(-3.0, 3.0), # Limits of the x-axis
    bins=np.arange(3.5, 7.0, 0.5),  # Bins for violin plot
)

# Plot the between-station residuals (dS2S)
# with respect to Vs30
visualizer.between_station(
    "ASK2014",
    "SA(0.2)",
    "vs30",
    pred_var_units="m/s", # Units of the predictor variable
    xlim=(150, 1000),
    ylim=(-3.0, 3.0),
    bins=np.arange(200.0, 1000.0, 100.0),
)

# Plot the within-event residuals (dWes)
# with respect to Joyner-Boore distance
visualizer.within_event(
    "ASK2014",
    "SA(0.2)",
    "r_jb",
    pred_var_units="km",
    xlim=(0, 350),
    ylim=(-3.0, 3.0),
    bins=np.arange(0, 375.0, 25),
)
```

Several more optional arguments can be passed to the function, and the user is referred to the API documentation for a full description. To export these figures to file, one can add the optional keyword "filename" and supply the path to the file in which the figure should be saved.

Returning to our ground motion database for northwest Europe, we can see examples of the residuals for both the strong and weak motion subsets in Figure 21 (using the Abrahamson et al., 2014 GMM with Sa (0.01 s)) and Figure 22 (using the Kotha et al. (2020) GMM adapted for the ESHM20 with Sa (0.2s))

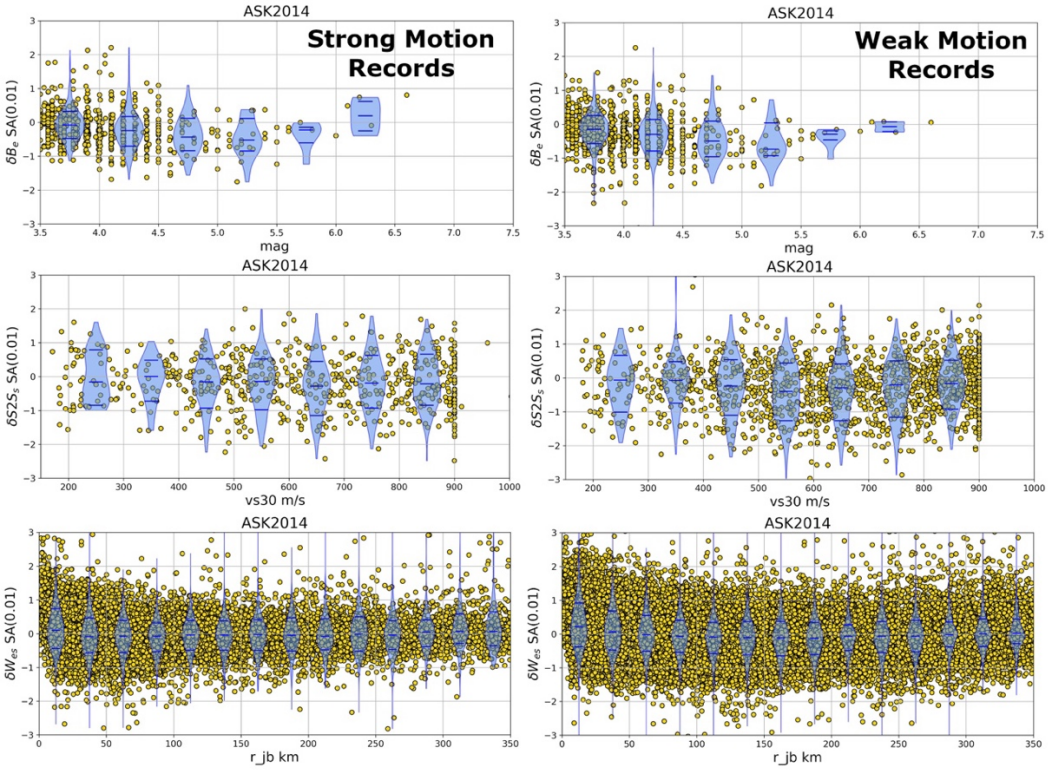


Figure 21: Between event (top row), between-station (middle row) and site-corrected within event residuals for Sa (0.01 s) using the Abrahamson et al. (2014) GMM. Values from strong motion records only are shown in the left column and from weak motion in the right column.

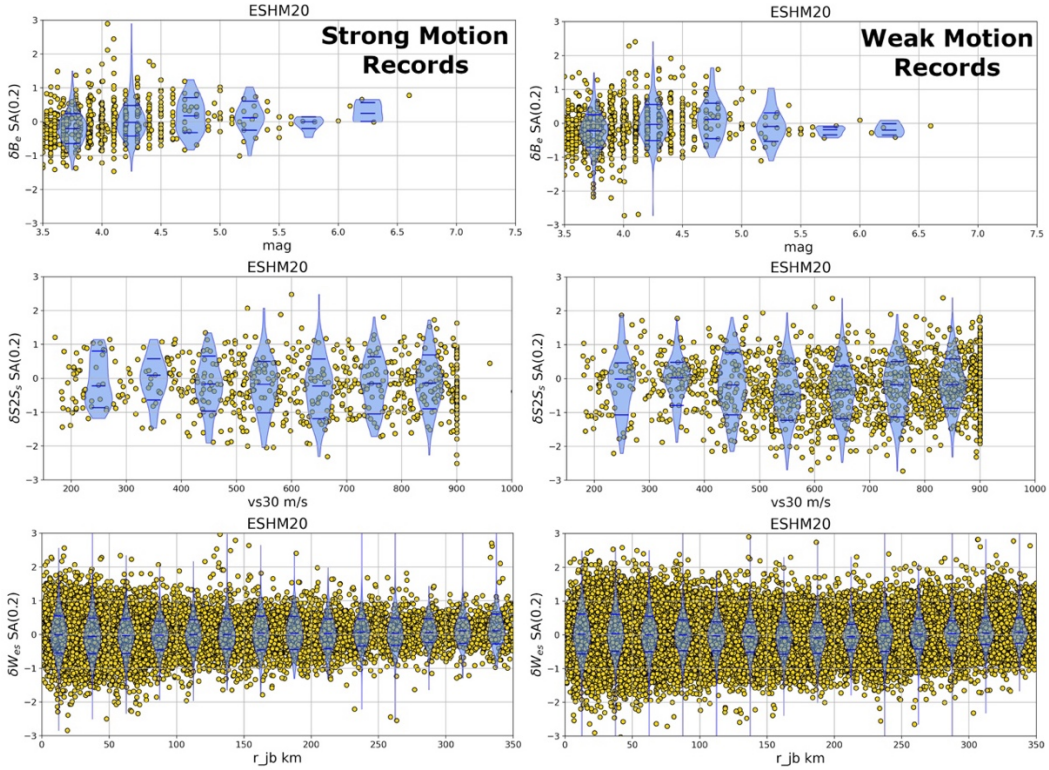


Figure 22: As Figure 21 but for the ESHM20 GMM and Sa (0.2 s).

3.5.1. Constructing the Complete Station and Hazard Database

With the above steps run, we now have all the information necessary to construct the complete database that was shown in Figure 2. The minimum amount of information required is a harmonized earthquake catalogue and a complete set of ground motion residual calculations, the outputs of which store the relevant flatfile, ground motion observations and metadata. To construct and allow analysis of the database, we have a specific class in the tools called "StationDatabase", which is imported via:

```
from pshatest.station_database import StationDatabase
```

We assume initially that one wishes to create the database; however, the "StationDatabase" class is a point of interaction with any database, including existing ones, so one does not need to re-create the database on subsequent usage. In the following code snippet, we will create a database and load in the station information via the ground motion residuals alongside the harmonised earthquake catalogue:

```
dbname = "./path/to/ground_motion_hazard_database.hdf5"

# Create the database
db1 = StationDatabase(
    dbname,
    start_date="2000-01-01", # Start date/time of the database
    end_date="2020-12-31",  # End date/time
)

# Add in the ground motion residual information and catalogue
db1.setup(
    # add the strong motion residuals
    strong_motion_residual_file="./path/to/strong_motion_residuals.hdf5",
    # add the weak motion residuals (optional)
    weak_motion_residual_file="./path/to/weak_motion_residuals.hdf5",
    # add the harmonized earthquake catalogue
    earthquake_catalogue_file="./path/to/harmonized_earthquake_catalogue.csv"
)
```

A comprehensive overview of all the attributes and functions of the StationDatabase class is available in the online documentation. Of note, however, are the attributes "stations", which produces a summary table of the station information:

```
db1.stations
```

	sids	lons	lats	N	net	name	vs30	vs30measured	start_eida	end_eida	ndays	HN	HG	HH	EH	geometry
1N-AIGO	100000	6.877168	44.787328	1	1N	AIGO	794.246704	False	NA	NA	NaN	False	False	True	False	POINT (6.87717 44.78733)
1N-VIL1	100001	0.122391	49.398194	1	1N	VIL1	675.000000	False	NA	NA	NaN	False	False	True	False	POINT (0.12239 49.39819)
3A-MZ01	100000	13.270813	42.671575	71	3A	MZ01	538.388123	False	NA	NA	NaN	True	False	False	True	POINT (13.27081 42.67157)
3A-MZ02	100001	13.285812	42.669947	8	3A	MZ02	636.115906	False	NA	NA	NaN	True	False	False	True	POINT (13.28581 42.66995)
3A-MZ03	100002	13.288322	42.677017	8	3A	MZ03	809.518433	False	NA	NA	NaN	True	False	False	True	POINT (13.28832 42.67702)
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

"catalogue", which produces a single dataframe showing the harmonised catalogue information:

```
db1.catalogue
```

	master_id	ev_id	agency	ev_time	lon	lat	depth	M	sigma_m
0	MERGE_IDX_0000000	ISC_1725187	ISC	2000-01-01 01:19:28.000	20.5597	41.9126	10.40	4.7000	0.0
1	MERGE_IDX_0000001	SHX-19151	SIHEX	2000-01-01 03:20:57.700	-2.9910	48.1490	1.90	2.4000	0.1
2	MERGE_IDX_0000002	ISIDE_1113759	ISIDE	2000-01-01 03:47:34.710	12.0020	38.3600	5.00	3.2204	0.0
3	MERGE_IDX_0000003	ISC_1725195	ISC	2000-01-01 04:02:26.000	22.4921	35.5465	44.00	4.0022	0.0
4	MERGE_IDX_0000004	ISC_1757767	ISC	2000-01-01 08:22:53.000	-8.5510	36.1790	30.50	3.6879	0.0
...	...	...	...	...	...	...	...	...	...

"metadata", which produces the summary table of the ground motion values and metadata used in the GMM residual calculations:

```
1 db1.metadata
```

	z_tor	f_length	sta	r_x	r_jb	ev_lon	r_rup	M_CAT_ID	dip	ev_depth	...	pSA_5.49451	pSA_5.98802	pSA_6
83	21.990809	2.818383	FR-SAOF-HN	115.909509	115.909509	8.4483	118.247936	MERGE_IDX_0002475	90.0	23.4	...	2.684985e-05	1.737821e-05	1.2834
84	21.990809	2.818383	RA-STET-HN	135.163107	135.163107	8.4483	137.173706	MERGE_IDX_0002475	90.0	23.4	...	1.959065e-05	1.690958e-05	1.1681
115	9.583232	0.833536	FR-SAOF-HN	74.165397	74.165397	7.3430	74.836530	Not in harmonised catalogue	90.0	10.0	...	4.398853e-06	3.997694e-06	3.2328
156	15.567310	2.065380	FR-SAOF-HN	53.747955	53.747955	7.4413	56.253024	MERGE_IDX_0004324	90.0	16.6	...	1.055918e-05	8.946461e-06	7.0399
157	15.567310	2.065380	RA-STET-HN	92.836263	92.836263	7.4413	94.308705	MERGE_IDX_0004324	90.0	16.6	...	6.241589e-06	5.103085e-06	4.5817
...	...	...	...	...	...	...	...	...	...	...	...	...	...	

There are also available some functions to plot information extracted from the data, and, in particular, to plot the station-to-station residual $\delta S2S$ for a given station with respect to different GMM *and* for multiple channels if available for the station:

```
1 db1.plot_ds2s_for_site(
2     "FR-SAOF",  # Station Name
3     ["ASK2014", "ESHM20"],  # Choice of GMMs
4     ["HH", "HN"],  # Channel Choice
5     ylim=(-3.0, 3.0),  # Limits on the y-axis
6 )
```

Which can produce a plot such as that shown in Figure 23.

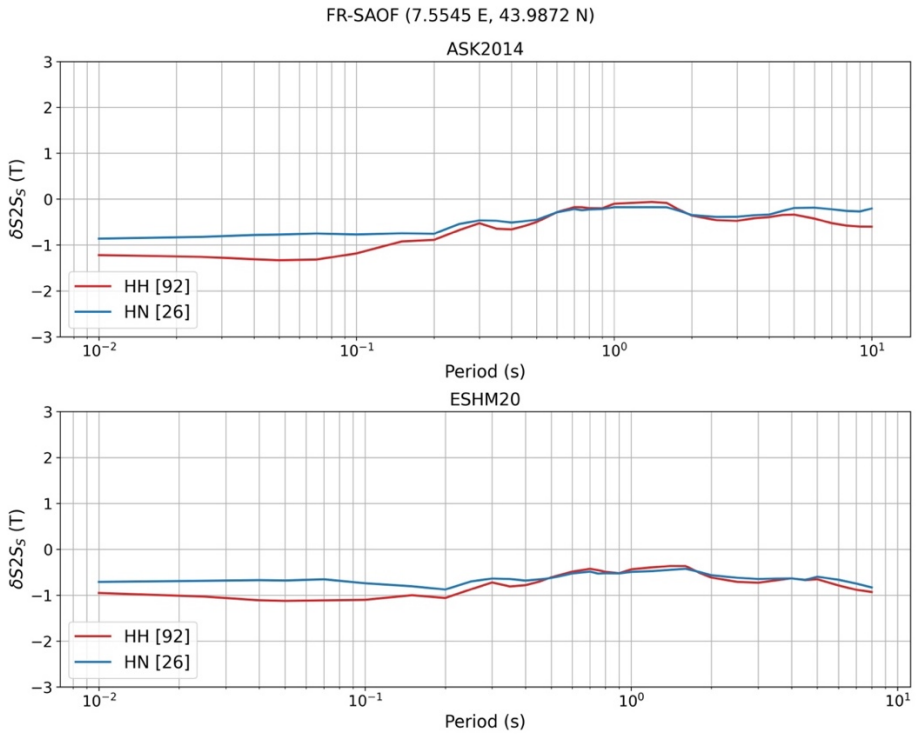


Figure 23: Example $\delta S2S(T)$ for a given station in the database (FR-SAOF) using two different GMMs (ASK2014 & ESHM20) and two different channels (HH [weak] and HN [strong]), numbers in square brackets indicate the number of records per channel

With the database of stations created the next step is to calculate seismic hazard at the stations. PyPSHATest itself does not wrap or execute commands to run OpenQuake or setup hazard models, so

we assume that if the user execute the PSHA calculations separately. If the PSHA model is intended to be run in OpenQuake, however, the "StationDatabase" class has a function to export the relevant station information into an OpenQuake site model file:

```
db1.sites_to_openquake_site_model(
    "./path/to/openquake_site_model.xml",
    sites=[...],  # List of sites (if not specified then all will be exported)
     # If choosing to run on a reference site condition then ...
    reference_vs30 = 800.0,  # Vs30 = 800 m/s
    reference_vs30_measured = True,  # Is a measured site condition
    reference_z1pt0 = ...,
    ...
)
```

With this function the user can choose which sites to export and can control whether to set the site conditions to the values found in the flatfile, or whether to assign them to a reference site condition in alignment with the typical hazard map usage. The properties of the reference site are specified via the additional arguments.

If the PSHA for the sites has been executed successfully then the results of the calculations should be compiled together using the "OpenQuakeHazardOutputManager" as illustrated in section 2. Once again, while this current deliverable assumes that OpenQuake will be used for the PSHA calculation, results from any other software can also be used equivalently if they can be translated into the hdf5 datastore format shown. If the PSHA results have been successfully stored in the hdf5 format generated by the OpenQuakeHazardOutputManager then these can be added into the main StationDatabase via:

```
# Add in some hazard results (from model 1)
db1.add_hazard_from_oq_manager(
    "./path/to/hazard_model_datastore.hdf5",
    models = ["model 1", "model 2"],  # Choice of models to import
    )
```

Here one simply needs to specify the path to the hazard model datastore created via the OpenQuakeHazardOutputManager, and the choice of which models from the datastore to import into the StationDatabase (if not specified then all the models will be imported).

By this point the complete set of information for testing PSHA against data is stored in the database. If we reflect on all the steps necessary for this process it is clear that a large amount of effort is needed to prepare the models *and* the observed site data needed for comparisons. One would hope that in the course of a seismic hazard model development many of the steps taken here would actually be undertaken as part of the development process, e.g., compilation of a harmonised earthquake catalogue, compilation of the flatfile, exploration of GMMs and calculation of residuals etc. It is possible that several functions implemented in PyPSHATest could server wider purposes beyond the testing context in which they are being applied here. At the very least, one the information has been gathered following the steps above, PyPSHATest allows for a lot of flexibility and opportunity to explore and understand the hazard models and data available. In the next section, however, we will focus on how to compare models of observed ground motions and illustrate the features and interpretations using the example ground motion and hazard data from metropolitan France.

4. Comparing Seismic Hazard Models with Observed Ground Motions

4.1. Statistics of Exceedances

For time-independent PSHA the expected number of exceedances of a given level of ground motion, A, at a site s within a given time period, t, is determined from the rate of exceedance output from a chosen hazard model $\mathcal{h}$:

$$E_s[N(a \geq A|t, \mathcal{h})] \quad = \quad \lambda_s(a \geq A|t, \mathcal{h}) \cdot t \tag{4.1}$$

If the recording history of the site in question can be assumed to contain a *complete* archive of all exceedances of ground motion level A during a (now site-specific) operation period t_s then the probability of observing n exceedances at the site is described from a Poisson distribution:

$$P(n) = \frac{\left(\lambda_{s|\mathcal{h}} \cdot t_s\right)^n e^{-\lambda_{s|\mathcal{h}} \cdot t_s}}{n!} \tag{4.2}$$

Where $\lambda_{s|h} = \lambda_s(a \geq A|t, \mathcal{h})$. Stirling & Gerstenberger (2010) propose a simple criterion for testing observations against a model at a given site, which is simply whether or not the observed number of exceedances of A falls within the 2.5 and 97.5 percentile of the distribution. As discussed previously, however, for an individual site the observation durations t_s are too short, and corresponding number of exceedances of accelerations of engineering significant usually too few, to make a valid statistical inference.

If a set of S observation sites are chosen such that probabilities of exceedance are independent, then the expected number of exceedances can be determined from a sum of independent Poisson processes, which is itself Poissonian:

$$E[N(a \geq A|t\mathcal{h})] = \sum_{s=1}^{S} \lambda_s(a \geq A|t_s, \mathcal{h}) \cdot t_s \tag{4.3}$$

Where S is the total number of observation sites.

Expanding our consideration to multiple sites means that we are afforded several different ways of describing the aggregated seismic hazard curve: either through the total number of exceedances of ground motion over S sites, or through the total number of sites of one or more exceedance of ground motion. Mak & Schorlemmer (2016) present a clear overview of four different forms of aggregated seismic hazard curve, each of which allows for the application of different statistical models. We summarise these here:

Form 1: Total number of exceedances of a fixed level of ground motion over multiple sites

For S sites, each site s has a N_s records of observed ground motion in its *complete* duration period t_s years. The i^{th} observation of ground motion from N_S records at site s is given by O_{Si}. For a given level of ground motion A the "observed" hazard curves is given by:

$$h_{obs}(A) = \sum_{s=1}^{S} \sum_{i}^{N_S} I(O_{Si} \geq A) \tag{4.4}$$

Where $I(\cdot) = 1$ if the logical term $(\cdot)$ is true, or 0 otherwise.

The forecasted hazard from the $\mathcal{h}^{th}$ model is then taken as the sum of Poisson random variables, $\mathcal{P}$, each with their respective mean corresponding to the expected number of exceedances from model $\mathcal{h}$ given the duration of observation at the site:

$$H^{(1)}(\hbar) = \sum_{s=1}^{S} \mathcal{P}(\lambda_s(a \geq A|t_s, \hbar) \cdot t_s) \quad 4.5$$

Which has the following as the expected hazard:

$$h_{\exp}^{(2)}(A|\hbar) = \sum_{s=1}^{S} \lambda_s(a \geq A|t_s, \hbar) \cdot t_s \quad 4.6$$

In the case of time independent PSHA, $h_{\exp}(A|\hbar)$ is Poisson distributed as it refers to the sum of Poisson random variables.

Form 2: Total number of sites with one or more exceedances of a fixed ground motion level

In this form the observed hazard curve is described by the number of sites exceeding a given ground motion level A:

$$h_{obs}(A) = \sum_{s=1}^{S} I(\exists i: O_{Si} \geq A) \quad 4.7$$

Where $\exists i: O_{Si} \geq A$ indicates that at least one or more observations from N_S observations at the site s exceeds A. In this case the forecast hazard is a sum of heterogeneous Bernoulli random variables, $B_{P(x)}$ (where $P(x)$ is the probability of the binary variable x yielding a successful outcome). This therefore corresponds to the sum of the probabilities of one of more exceedances of ground motions at each site:

$$H^2(A|\hbar) = \sum_{s=1}^{S} B_{P(X_{\lambda_{s|h,t_s}>0})} \quad 4.8$$

Where $P\left(X_{\lambda_{s|h}t_s} > 0\right)$ is the probability of observing at least one exceedance of A at the site according to the model.

The expected hazard from the $\hbar^{th}$ model is:

$$h_{exp}^2(A|\hbar) = \sum_{s=1}^{S} P\left(X_{\lambda_{s|h,t_s}} > 0\right) \quad 4.9$$

In this notation X_x denotes a Poisson random variable with mean x. $h_{exp}(A|\hbar)$ is the sum of heterogeneous Bernoulli random variables (heterogenous because the probability changes per site), meaning that it is modelled as a Poisson-binomial random variable.

Form 3: Total number of exceedances of ground motion with a fixed annual rate of exceedance across multiple sites

If $A_{s|\hbar}$ is the level of ground motion with a fixed annual rate of exceedance according to model $\hbar$, then the observed number of exceedances of ground motion corresponding to a fixed annual rate of exceedance is model-dependent:

$$h_{obs}(A|\hbar) = \sum_{s=1}^{S} \sum_{i}^{N_S} I\left(O_{Si} \geq A_{S|\hbar}\right) \quad 4.10$$

The forecasted hazard is now model independent and is simply the sum of Poisson random variables with mean $\lambda_r \cdot t_s$, where λ_r is the annual rate of exceedance according to a given return period:

$$H^3 = \sum_{s=1}^{S} \mathcal{P}(\lambda_r \cdot t_s) \quad 4.11$$

While the expected hazard is then:

$$h_{exp}^{3} = \sum_{s=1}^{S} \lambda_r \cdot t_s \tag{4.12}$$

As with form 1, H^3 is Poissonian as it is the sum of Poisson random variables.

Form 4: Total number of sites with at least one exceedance of ground motion with a fixed rate of exceedance

This is the adaptation of form 2, which rather than fixing the ground motion threshold instead sets a site- and model-dependent threshold ground motion corresponding to a given rate of exceedance, $A_{S|\hbar}$:

$$h_{obs}(A|\hbar) = \sum_{s=1}^{S} I\left(\exists i: O_{Si} \geq A_{S|\hbar}\right) \tag{4.13}$$

Where $I\left(\exists i: O_{Si} \geq A_{S|\hbar}\right)$ takes the value of 1 if any observed ground motion from those observed at the site exceeds the model-dependent threshold value.

The forecasted hazard no longer depends on either the model or the ground motion level as it corresponds to a sum of heterogeneous Bernoulli variables whose individual probabilities of success depend only on the specified annual rate of exceedances and the duration of operation of the site:

$$H^{(4)} = \sum_{s=1}^{S} B_{P\left(X_{\lambda_{s|h},t_S} > 0\right)} \tag{4.14}$$

With expected hazard:

$$h_{exp}^{(4)} = \sum_{s=1}^{S} P\left(X_{\lambda_s|h,t_S} > 0\right) \tag{4.15}$$

This last form of the model makes H^4 a Poisson-binomial distributed random variable as it is the sum of heterogeneous Bernoulli random variable. However, if t_s were to be constant for all sites then the distribution is no longer heterogeneous and can instead be modelled by a binomial distribution, $\mathcal{B}$.

The four forms of aggregated seismic hazard curve outlined by Mak & Schorlemmer (2016) provide a clear statistical framework to describe how we can make comparisons between models and data over multiple sites. There are some important considerations within these comparisons, however. The first is that these comparisons assume that the individual Poisson and/or Bernoulli processes at each site are statistically independent, and therefore by extension exchangeable. For a given seismic recording network the proximity of stations if often such that several may be located sufficiently close to one another that the probabilities of exceedance cannot be said to be independent. We will look in more detail it this issue in the subsequent section.

The second consideration from the frameworks presented here is that in the case of forms 2 and 4 of the hazard curves the forecasted hazard, H, is represented by sum of independent Bernoulli trials with heterogeneous probabilities. The expected number of sites exceeding the ground motion level, be it a fixed ground motion level or one dependent on the annual probability of exceedance, is therefore described by a Poisson-Binomial distribution. The only exception, though a particularly relevant one, is form 4 in the case that the observation times t_s are equal, such that H is a sum of Bernoulli trials with equal probabilities, in which case $H^{(4)}$ is Binomially distributed. Given that the Poisson-Binomial distribution is so prevalent in these considerations, a further look into this distribution and its usage is warranted.

4.1.1. The Poisson-Binomial Distribution

In the second and fourth form of aggregated seismic hazard curves the forecast hazard is given in terms of the number of sites, N, out of a total of S sites with one or more exceedances of a given level of ground motion in their operation time t_s. The exceedance of a given level of ground motion at each site, s, is described by a Bernoulli random variable with probability P of a successful outcome (success = "exceedance of the ground motion within the operational time t_s". This is denoted previously as

$P\left(X_{\lambda_{s|h}t_s} > 0\right)$, which we will refer to now as P_{Xs} for short. The probability mass function (pmf) describing the number of N successful outcomes from S independent Bernoulli experiments each with individual probability p_i is given by:

$$P(k = K) = \sum_{M \in F_k} \prod_{j \in M} p_j \prod_{j \in M^c} (1 - p_j) \quad 4.16$$

Where F_k is the set of all subsets of k integers that can be selected from $i = 1, 2, \ldots, S$ (Wang, 1993), e.g. for $N = 3$: $F_1 = \{M = \{1\},\ M = \{2\}, M = \{3\}\}, F_2 = \{M = \{1,2\},\ M = \{1,3\},\ M = \{2,3\}\}$, and $F_3 = \{M = \{1,2,3\}\}$. M^c is the complement of M. The corresponding cumulative density function (cdf) is then given as:

$$F_N(k) = \sum_{l=0}^{k} \left[\sum_{M \in F_l} \prod_{j \in M} p_j \prod_{j \in M^c} (1 - p_j) \right] \quad 4.17$$

The Poisson-binomial distribution is somewhat complicated to apply in practice, firstly because it has no closed-form expression for the distribution function, and secondly because the number of subsets of k integers, F_k, scales by:

$$|F_k| = \frac{S!}{(S-k)! \cdot k!} \quad 4.18$$

This means that for even moderate S, evaluation of the pmf and cdf is computationally demanding. This latter problem has been addressed by Hong (2013), who presents a computationally efficient method for approximation of F_k.

As a sum of Bernoulli variables, the first two moments of the Poisson-binomial distribution are simply defined by:

$$\mu = \sum_{s=1}^{S} P_s \quad 4.19$$

And

$$\sigma^2 = \sum_{s=1}^{S} (1 - P_s) \cdot P_S \quad 4.20$$

Returning to our general definition of number of sites, the mean and variance of the number of sites N from a total S at which one or more exceedances of a given level of ground motion should be observed is predicted by a given hazard model, $\hbar$, as:

$$\mu(S|\hbar) = \sum_{s=1}^{S} P(a \geq A|\hbar, t_s) \quad 4.21$$

And:

$$\sigma(S|\hbar) = \sqrt{\sum_{s=1}^{S} P(a \geq A|\hbar, t_s) \cdot [1.0 - P(a \geq A|\hbar, t_s)]} \quad 4.22$$

This latter form of the model corresponds to the "Counting Method" described by Albarello & D'Amico (2008; 2015), who also argue that as S becomes large then from Central Limit Theorem we can define data-to-model differences as normalised residuals from a Gaussian distribution:

$$Z_\hbar = \frac{N - \mu(S|\hbar)}{\sigma(S|\hbar)} \quad 4.23$$

from which more conventional hypothesis tests can be applied. We should note, however, that the Gaussian approximation to the cdf of the Poisson-binomial distribution proposed by Albarello & D'Amico (2008), though emerging from the Central Limit Theorem, does differ slightly from the approximate representations described by Hong (2013). In their comparison of exact and approximate methods for representing the Poisson-binomial cdf, they consider approximation by both a normal distribution (with a continuous correction):

$$F_N(k) \approx \Phi(z'), \quad k = 1, 2, \dots, S \tag{4.24}$$

where $Z' = \frac{k+0.5-\mu}{\sigma}$ and $\Phi(x)$ the cdf of the standard normal distribution, along with a "refined normal approximation":

$$F_N(k) \approx \Phi(z') + \frac{\gamma \cdot (1 - z'^2) \cdot \phi(z')}{6} \tag{4.25}$$

Where $\phi(x)$ is the pdf of the standard normal distribution and γ the skewness of the Poisson-binomial distribution:

$$\gamma = \sigma^{-3} \sum_{s=1}^{S} P_S \cdot (1 - P_S) \cdot (1 - 2P_S) \tag{4.26}$$

In computational tests against the exact representation of $F_N(k)$ the "refined normal approximation" is found to be adequate for both small and large S.

4.2. Log-Likelihood

An alternative means of quantifying model-to-data fit is considered by Albarello & D'Amico (2008; 2015), which emerges from the log-likelihood of the Poisson-binomial distribution. Given an observed set of ground motion exceedances at N out of S sites, which we denote as E, the probability that a given model $\hbar$ attributes to the specific set of exceedances is:

$$P(\hbar|E) = c_\hbar \cdot P(E|\hbar) = c_\hbar \cdot L_\hbar \tag{4.27}$$

Where $c_\hbar$ is a constant and $L_\hbar$ the model likelihood of the Poisson-binomial distribution defined by:

$$L_h = \prod_{s=1}^{N_\hbar} P(a \geq A_S|t_s, \hbar) \prod_{s=N_\hbar+1}^{S} 1 - P(a \geq A_S|t_s, \hbar) \tag{4.28}$$

Where $N_\hbar$ is the predicted number of sites with ground motion exceeding A_S according to model $\hbar$. The log-likelihood function, $l_\hbar$, is then:

$$l(\hbar) = \sum_{s=1}^{N_\hbar} \log[P(a \geq A_S|t_s, \hbar)] + \sum_{s=N_\hbar+1}^{S} \log[1 - P(a \geq A_S|t_s, \hbar)] \tag{4.29}$$

Albarello & D'Amico (2008) suggest that the overall performance of a model $\hbar$ with respect to the data can be assessed using a Monte Carlo approach. From the model itself a set of V realisations of exceedances for the respective sites can be generated by drawing a uniformly distributed sample $x = \mathcal{U}(0, 1)$ for each site within each realisation $v = 1, 2, \dots, V$. If $x < P(a \geq A_s|t_s, \hbar)$ for the site in question then the site is placed in the set $s = [1, N_\hbar]$ indicating exceedance, otherwise it is in the set $s = [N_\hbar + 1, S]$. For each sample realisation v the log-likelihood $l^v(\hbar)$ is calculated using Equation 4.29, and the mean and standard deviation of $l^v(\hbar)$ calculated from the V samples: $\mu(l^v(\hbar))$ and $\sigma(l^v(\hbar))$. The log-likelihood is then determined for the observed set of exceedances, $l^0(\hbar)$, which can be quantified relative to the model by:

$$\tau(\hbar) = \frac{l^0(\hbar) - \mu(l^v(\hbar))}{\sigma(l^v(\hbar))} \tag{4.30}$$

Where $\tau(\hbar)$ is distributed according to a Student's t-distribution, which allows for the definition of a prediction interval within which $\tau(\hbar)$ should be found if the model is assumed to be supported by the

observations. Albarello & D'Amico (2015) suggest that a simple criterion may be that the model is not supported by the observation if $|\tau(\hbar)| > 2$.

4.3. Adapting the Statistical Tests for Uncertain Observations

The statistical frameworks for comparing models and data presented in sections 4.1 and 4.2 assume that the exceedance of a level of ground motion (be it fixed *a priori* or hazard-model dependent) at a given site is known with certainty, and thus has a binary outcome. From section 3, however, in order to account for completeness of the observations at a site we allow for ground motions to be predicted according to a distribution in the case that shaking is expected but not present in our observed ground motion data. In this case we can no longer describe exceedance in terms of $I(O_{Si} \geq A)$ (as Equation 4.4) nor the total number of sites with one or more exceedances of ground motion as we have done in Equation 4.13, but rather we can only describe them in terms of a *probability* that the ground motion was exceeded in the observation time window. This change in the definition of observed hazard can be accommodated by adapting the formulation proposed by Mak & Schorlemmer (2016) for counting exceedances of ground motion when considering observations in terms of macroseismic intensity, which must be converted to ground motion accelerations along with their corresponding uncertainty. Here, if we define our observed hazard in terms of the number of stations exceeding a given level of acceleration (be the level fixed or hazard-dependent) then when the exceedance of the ground motion level at station s from each individual earthquake i of N_S earthquakes expected to produce shaking at the station the resulting hazard is defined as:

$$h_{obs}(A) = \sum_{s=1}^{S}\left[1 - \sum_{i=1}^{N_s} P\left(a < A \middle| \mu_{si}^{imp}, \sigma_{si}^{imp}\right)\right] \tag{4.31}$$

Where $P\left(a < A \middle| \mu_{si}^{imp}, \sigma_{si}^{imp}\right)$ is the probability the ground motion at site s from event i did not produce an exceedance of A given the expected ground motion level and its uncertainty from the imputation $(\mu_{si}^{imp}, \sigma_{si}^{imp})$. For the cases when an observation is available for the site then μ_{si}^{imp} is simply the observation and $\sigma_{si}^{imp} = 0$ accordingly. The Type 2 and Type 4 forms of aggregated seismic hazard curve from Mak & Schorlemmer (2016) adapt easily to this formulation of uncertain observation. In this formulation $h_{obs}(A)$ does not take integer values at the number of stations exceeding levels of ground motion, but rather a continuous values reflecting the sum of the probabilities of exceedance at each station.

To illustrate this change in definition we can consider the case of our observations for France. In Figure 24 we consider four increasing levels of ground motion intensity (0.01g, 0.02 g, 0.05 g and 0.1 g), and from our observations of ground motion in the 21 year period for which our flatfile is constructed (2000 – 2021 inclusive) we show the sites for which the respective intensity measure level is exceeded. From 115 potential stations the number exceeding the threshold decreases from 27 when $A = 0.01g$, to 13 when $A = 0.02\ g$, then 4 when $A = 0.05\ g$ and just 2 when $A = 0.1\ g$. By contrast Figure 25 shows the definition once the uncertainty from the data imputation process is incorporated. Here we can now consider all 115 sites for which we have observations of ground motion (it will be seen in due course how we arrive at 115 sites), and for each site we now have a corresponding probability of exceeding the respective intensity measure levels.

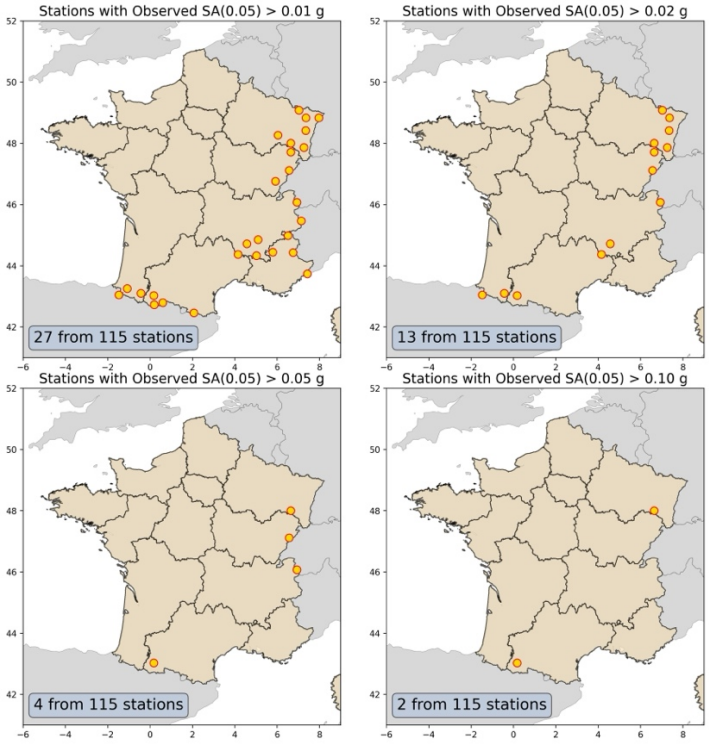


Figure 24: Stations in our ground motion database for which threshold accelerations (0.01 g, 0.02 g, 0.05 g and 0.1 g) have been exceeded

In addition to correcting for missing observations, the data imputation approach affords us another benefit that assists in the statistical comparisons between observed and predicted exceedances, which is that we can normalise the "observation" period to be equal for all stations regardless of their actual duration of operation. Thus, for every station we produce a 21 year pseudo-history that mixes together observed ground motions *and* distributions of ground motion for missing cases. The benefit comes in the Type 4 aggregated hazard curves in which the expected hazard is a sum of Bernoulli variables. If the duration of operation at the site changes from one site to another then the probabilities of exceeding the ground motion level associated to a particular return period, $A_{s|\hbar,t_S}$, is site dependent as t_S varies from site to site. The Bernoulli variables describing $h_{\exp}$ are therefore heterogeneous and the resulting probability distribution is Poisson-binomial, which, as we have seen, has no closed form density function. However, as imputation is allowing us to normalise the effective station history to the same temporal duration t_S is constant for all sites, and the expected hazard is then the sum of a set of homogenous Bernoulli variables, which can be modelled with a conventional Binomial distribution. As the Binomial distribution does have closed form density functions then we can define the confidence intervals directly, greatly simplifying the comparisons. Otherwise, when a Poisson-binomial distribution is assumed then the confidence intervals may need to be defined by Monte Carlo sampling *or* from the refined normal approximation (equations 4.24 – 4.26).

It is not only the aggregated hazard curve definitions that require adaptation for the uncertain observations, to fully be able to integrate the data imputation approaches into the testing framework we should also adapt the log-likelihood formulation. In this case it is possible to do so via Monte Carlo simulation such that $\ell^0(\hbar)$ is no longer a single observation but rather now becomes $\ell^\psi(\hbar)$ where $\psi = 1, 2, \dots, V$ samples of the observed hazard with uncertainty from imputation. Note, however, that this results in $\ell^\psi(\hbar) = \mathcal{N}\left(\mu^\psi\left(\ell^\psi(\hbar)\right), \sigma^\psi\left(\ell^\psi(\hbar)\right)\right)$, meaning that one can no longer quantify the fit of the observation to the model via a Student's t-distribution. Instead, we can make the comparison using an unequal variances t-test (or Welch's t-test), which allows us to test the hypothesis that two samples have equal means in the case the samples have unequal population variances. Alternative, and arguably more robustly, the null hypothesis that the distribution observed log-likelihood

$\mathcal{N}\left(\mu^{\psi}\left(\ell^{\psi}(\hbar)\right), \sigma^{\psi}\left(\ell^{\psi}(\hbar)\right)\right)$ is the same as that of the model log-likelihood $\mathcal{N}\left(\mu(\ell^{V}(\hbar)), \sigma\left(\ell^{V}(\hbar)\right)\right)$ could be tested via a Brunner-Munzel test (Brunner & Munzel, 2000) or a 2-sample Kolmogorov-Smirnov test.

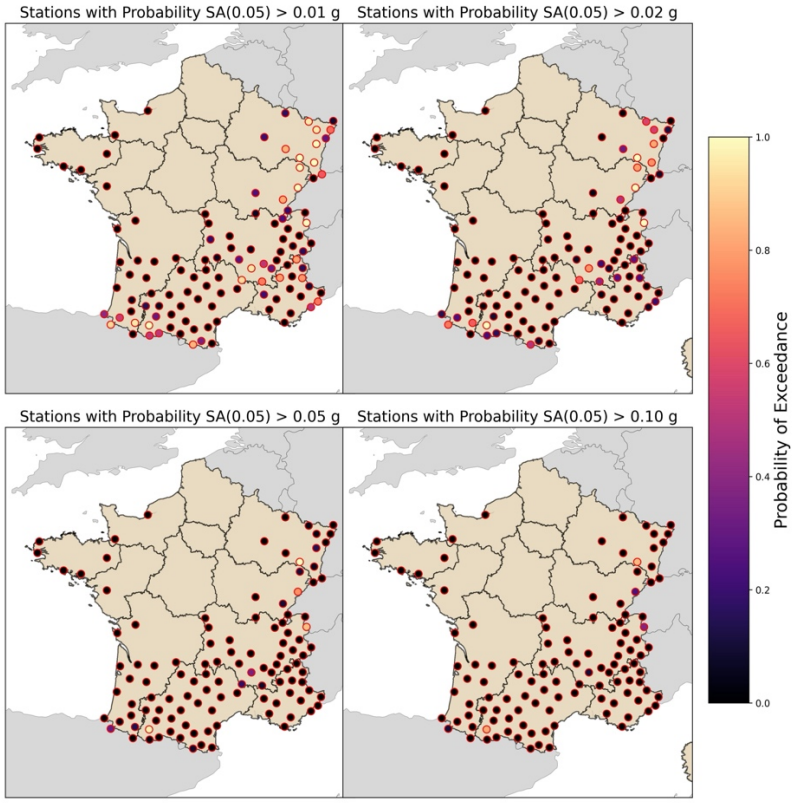


Figure 25: Probability that acceleration levels 0.01 g, 0.02 g, 0.05 g and 0.1 g have been exceeded at the respective stations between 2000 – 2020 inclusive.

One last point to be made regarding the use of the data imputation process and the necessary adaptations to the testing process to allow for uncertain observations of ground motion is that, in theory, this could also be extended to include observations from the pre-instrumental era and/or recent databases of macroseismic intensity. In theory, a historical earthquake catalogue could be used as the basis for data imputation, while observed macroseismic intensities can themselves be converted ground motions using an appropriate ground motion intensity conversion equation (GMICE). This is effectively a hybrid of one of the earliest testing approaches proposed by Ward (1995), supplemented by observation where available. Both will inevitably yield distributions of ground motion at target sites of interest, which can be treated in the manner described here. Incorporating older events would inevitably mean greater uncertainty in the "true" history of shaking at a site, as the locations and magnitudes would become more uncertain the further one looks back in time. Similarly, it would not be unexpected to observe potential changes in observed rates of occurrence at the point at which the macroseismic intensity-based observations gives way to instrumental ones. These and other limitations of macroseismic intensity for PSHA testing have been a motivation for focussing exclusively on observed ground motions here, but it is important to emphasise that the two approaches can be integrated into the same framework adopted by PyPSHATest if desired.

4.4. Running the PSHA Testing Tools

In the following example we will use the ground motion data, the earthquake catalogue and the residual analysis undertaken so far to demonstrate a complete end-to-end workflow for testing the two PSHA models of interest for France (Drouet et al., 2020) and ESHM20 (Danciu et al., 2021). In the previous section the relevant information had been combined into the Station and Hazard Database, which will form our starting point here. The tools for testing PSHA against observed ground motions can be found in the class "PSHATests", which is imported via:

```
from pshatest.psha_testing_tools import PSHATests
```

We begin our comparison by first connecting to the database we constructed at the end of Section 3.

```
# Define the database name
dbname = "./path/to/ground_motion_hazard_database.hdf5"

# Instantiate the database class
db1 = StationDatabase(dbname,
                      start_date="2000-01-01",  # Start date of database
                      end_date="2020-12-31",  # End-date
                     )
# Connect to the database
db1.open()
```

4.4.1. Declustering

In our PSHA testing process we have the option to decide whether we consider all observations (and therefore all events in the harmonized catalogue) for comparison against the PSHA models, or whether to consider only the mainshocks, which we assume a Poissonian. There is some debate in the scientific literature as to which approach is preferable. PyPSHATest leaves this option to the user; however, the authors of this deliverable are of the view that declustering the observations would be the preferable option to ensure fairer comparison of the tests. There are a wide variety of declustering algorithms in the literature and we would encourage the PSHA testers to explore the impact of different options, including some that are available inside OpenQuake itself. Any declustering algorithm could be considered provided that it can, for the earthquake catalogue of length N_{EQ} events, provide two outputs:

vcl: A vector of length N_{EQ} indicating (via numerical integer) to which cluster an earthquake belongs, or 0 if the event is not assigned to any cluster.

flagvector: A vector of length N_{EQ} indicating whether an event is a Poisson mainshock (indicated with 0) or a transient event (indicated with a non-zero). When executed in the testing process only ground motions from those events marked as Poissonian mainshocks will be used to generate the "observed" hazard curve.

Although we favour no specific declustering algorithm, for users who might be interested to understand the impact of declustering on the comparisons we have included a simple declustering algorithm into the tools, which is based on the Gardner & Knopoff (1974) approach but with the scaling of spatio-temporal aftershock windows fixed to the European calibrated adjustment factors adopted by Grünthal et al. (2018). This declustering tool can be imported via:

```
from pshatest.declustering_tools import simple_declustering
```

When using the harmonised catalogue integrated into the Station and Hazard Database, the results of a declustering can be added to the database and labelled, so that different declustered catalogues could be selected in the subsequent tests:

```
# Get the catalogue from the database ...
catalogue = db1.catalogue.copy()
# and make sure that the event time is a datetime object
catalogue['ev_time'] = catalogue["ev_time"].astype(np.datetime64)

# Run the declustering algorithm
# vcl = Vector indicating to which cluster each event belongs
# flagvector = Vector classifying each event as a foreshock (-1)
#              mainshock (0) or aftershock (1).
vcl, flagvector = simple_declustering(catalogue)

# Add the declustering results to the database
db1.add_declustering_results(
    vcl,
    flagvector,
    event_id = catalogue["master_id"].to_numpy(),  # Set the event IDs
    # Add identifier for declustering results in database
    decluster_model_name="gruenthal_declustering1",
    # Add any other description of interest (just for book-keeping, not used)
    info={"description": "Gruenthal type declustering (descending magnitude)"}
)
```

Here the database now stores the results of this declustering under the header "gruental_declustering1".

4.4.2. Selecting Minimum Site Spacing

In the first step of our analysis, we need to select the sites from the database that will be used for comparison against PSHA. Recalling that our database stores the station information as a dataframe under the attribute "StationDatabase.stations", we can use this information to select our sites of interest. In the current example, for our target region of metropolitan France (plus a 20 km buffer), we have in our database 357 individual stations (with some stations having both weak and strong motion channels), illustrated as shown below:

```
db1.stations
```

	sids	lons	lats	N	net	name	vs30	vs30measured	start_eida	end_eida	ndays	HN	HG	HH	EH	geometry
1N-AIGO	100000	6.877168	44.787328	1	1N	AIGO	794.246704	False	NA	NA	NaN	False	False	True	False	POINT (6.87717 44.78733)
1N-VIL1	100001	0.122391	49.398194	1	1N	VIL1	675.000000	False	NA	NA	NaN	False	False	True	False	POINT (0.12239 49.39819)
4H-R0220	100081	7.459959	47.630631	3	4H	R0220	531.066162	False	NA	NA	NaN	False	False	False	True	POINT (7.45996 47.63063)
4H-R0320	100082	7.450140	47.657658	4	4H	R0320	644.646179	False	NA	NA	NaN	False	False	False	True	POINT (7.45014 47.65766)
4H-R30FE	100083	7.294344	47.684685	2	4H	R30FE	890.013306	False	NA	NA	NaN	False	False	False	True	POINT (7.29434 47.68468)
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

In our presentation of the forms of aggregated seismic hazard curves and the Poisson-binomial distribution, one important assumption that underpins the statistics is that each Bernoulli variable is independent. This means that both the observed and expected rates of exceedance at each of the stations under consideration cannot contain spatial dependencies. This is challenging to achieve in practice, as for stations located in close geographical proximity their history of ground motions will likely contain many common events. Similarly, their expected hazard will contain contributions from many of the same seismogenic sources. In many PSHA testing applications considered previously (e.g., Albarello & D'Amico, 2008; Stirling & Gerstenberger, 2010; Tasan et al., 2014; Mak & Schorlemmer, 2016), the requirement of independence is fulfilled by selecting stations with a minimum spacing from one another on the order of several tens of kilometres. Exactly what minimum spacing is required seems to vary from study to study, with Tasan et al. (2014) adopting 10 km while Mak & Schorlemmer (2016) propose minimum spacings of 25 km (for California) and 50 km (for the eastern United States). Given the subjective judgement required here, we add in a function to select stations from all those that are available while enforcing a user defined minimum spacing. In contrast to Mak & Schorlemmer (2016), who propose different strategies for random sampling with minimum inter-site spacing, we choose an approach to site selection that favours sites with more recordings. We do so by sorting the sites in descending order of number of records and beginning with the most well recorded site we consider each station in turn, adding to the selection if it is more than R km from *any* other site among those already selected, or removing it otherwise. This function is called "select_sites_minimum_spacing", which is imported via:

```
from pshatest.station_database import select_sites_minimim_spacing
```

And can be applied to the station dataset via:

```
france_stations_30km = select_sites_minimim_spacing(
    db1.stations,  # Station dataframe
    30.0, # Minimum spacing in km
)
```

The in the current example we will use a minimum spacing of 30 km, which when applied to our station set for France yields 115 sites for consideration in the PSHA tests.

4.4.3. Running the Data Imputation

The next step is to run the data imputation process to generate the set of "observed" histories for these sites. This is illustrated in the example below in which we will use the Abrahamson et al. (2014) GMM to run the data imputation, we will consider only mainshocks above M 4.5 (the minimum magnitude used in both FR2020 and ESHM20) and will only consider observed ground motions for stations within an epicentral distance of 250 km of the event. Our data imputation time is limited to 1 January 2000 to 31 December 2020 (i.e., 21 years) and we will get imputed accelerations for 12 selected spectral periods between T = 0.05 s and T = 3.0 s.

```
# Select the GMM to use and get the OpenQuake instance
gsim = valid.gsim("AbrahamsonEtAl2014")

# Set up the data imputation tool
observation_handler = ObservedHazardDataImputation(
    dbname,  # The path to the station and hazard database
    mmin=4.5,  # The minimum magnitude to consider
    rmax=250.0,  # The maximum epicentral distance
    # The list of stations for imputation
    region=france_stations_30km.index.to_list(),
    # (optional) The preferred declustering results
    decluster_method="gruenthal_declustering1",
)

# Define the specific spectral periods for which we want
# the imputed accelerations
selected_periods = np.array([0.05, 0.075, 0.1, 0.15, 0.2, 0.25,
                             0.3, 0.5, 0.75, 1.0, 2.0, 3.0])

# Run the data imputation
observed_results = observation_handler.get_accelerations_for_stations(
    periods = selected_periods,
    gmm="ASK2014",  # Label for the GMM model
    gmm_model=gsim,  # OpenQuake instance of the model
    start_date="2000-01-01",
    end_date="2020-12-31"
)
```

The output from the imputation process is another class (StationSetResults), which stores the observed and imputed ground motions and their respective uncertainties for all the stations considered in the analysis. This class has some useful features to explore and inspect the imputation results, including plotting individual station pseudo-histories such as that shown in Figure 26 for the station FR-SAOF. As a container for the results of the imputation, the StationSetResults object is needed for the PSHA tests.

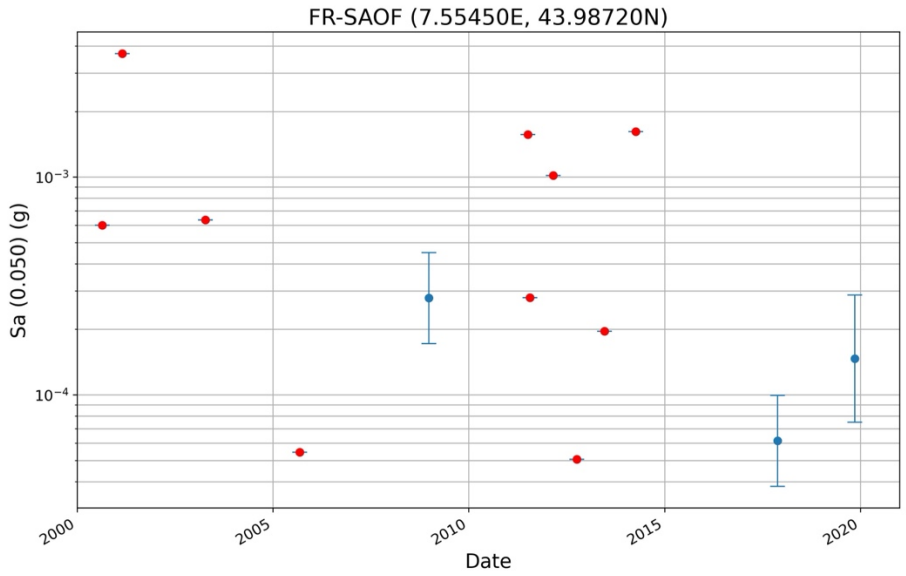


Figure 26: Station pseudo-history for Sa (0.05 s) at site FR-SAOF, with observated motions marked by red dots and the mean and 5 – 95 % confidence intervals of the imputed motions shown by the blue dots and error bars.

4.4.4. Comparing Hazard Curves against Observations

With the StationSetResults containing the observed and imputed ground motions for the set of stations of interest, we need to retrieve the corresponding seismic hazard curves from the PSHA models. For this we require that the selected sites are among those for which the PSHA calculations have been run. The full set of seismic hazard curves from a given PSHA calculation is retrieved from the Station and Hazard Database via:

```
# Selected ground motion intensity measures (as strings)
selected_imts = [
    "SA(0.05)", "SA(0.1)", "SA(0.15)", "SA(0.2)", "SA(0.3)",
    "SA(0.5)", "SA(0.75)", "SA(1.0)", "SA(2.0)", "SA(3.0)"
]

# Get the hazard curves for the model
curves_model_1 = db1.get_model_hazard_curves_stations(
    "HAZARD_MODEL_1",
    imts = selected_imts,
)
```

To make our model to data comparisons we will produce plots using the Type 2 and Type 4 aggregated curves. Recall that the Type 2 aggregated curve takes as the ordinate the number of sites with at least one exceedance, and as the abscissa the ground motion level. The Type 4 aggregated curve has the same ordinate but takes as the abscissa the probability of exceedance (or return period, but we prefer PoE in this case). For our comparisons we need to define a vector of target intensity measure levels (for the Type 2 curves) and a vector of target probabilities of exceedance for the Type 4 curves. Note that probabilities of exceedance here will refer to the probabilities of exceedance for an investigation time corresponding to the *duration of the observation period*, and the necessary conversions from the calculated probabilities of exceedance with respect to their original investigation time are made by the tools themselves. Examples of there are shown here:

```
# Set our list of intensity measure type
imts = ["SA(0.05)", "SA(0.1)", "SA(0.2)",
        "SA(0.5)", "SA(1.0)", "SA(2.0)"]

# Define target intensity measure levels
target_imls = np.array([
    0.0001, 0.0005, 0.001, 0.002, 0.005,
    0.0075, 0.01, 0.02, 0.03, 0.05, 0.075,
    0.1, 0.15, 0.2, 0.5
])

# Apply these for all the intensity measure types to create the
# target intensity measure types and levels
target_imtls = dict([(imt, target_imls) for imt in imts])

# Need to define the target probabilities of exceedance in Nyrs
target_probs = np.array([
    0.5, 0.4, 0.3, 0.2, 0.15, 0.1,
    0.075, 0.05, 0.03, 0.02, 0.01
])
```

In the above example we define intensity measure levels for the Type 2 curves from 1.0E-4 g to 0.5 g and PoEs in 21 years for the Type 4 curves ranging from 0.5 to 0.01 (corresponding to a range of return periods of ≈30 to ≈ 2000 years).

To setup the tests we instantiate the PSHATest class, for which we will need the outputs of the imputation process, the curves for the sites, and the start and end date:

```
# Instantiate the testing tool
psha_test = PSHATests(
    observations = observed_results,  # (Imputed) Observations
    models = curves_model_1,  # PSHA model results
    start_date = "2000-01-01T00:00:00.0",
    end_date = "2021-01-01T00:00:00.0",
    use_stats = True,
)
```

The last argument passed here "use_stats" indicates whether the comparisons should be made against the mean and quantiles of the logic tree (True) or against all the curves (False). For the examples here we will show the comparisons only for the mean and quantiles.

To compare the models and data using the Type 2 form of aggregated curve we need the observed rates of exceedance of the target ground motion levels and the model probabilities of exceedance:

```
observed_exceedence_by_imt, model_prob_exceedance = psha_test.get_all_exceedance_imls(
    target_imtls,
    use_stats = True
)
```

These outputs can then be input into a plotting function to produce comparisons of models to observations in the Type 2 aggregated curve form:

```
psha_test.plot_nsites_exceeding_imls(
    target_imtls,  # Intensity measure types and levels
    observed_exceedence_by_imt,  # Observed exceedances of the IMLs per IMT
    model_prob_exceedance,  # Model Probabilities of exceedance
    imt = "SA(0.05)",  # Intensity measure type
    title="A Figure Title, e.g. Type 2 Curves",
    ylim=(0, 120), # y-axis limits
    filename="./path/to/output_image_file.jpg",
    filetype="jpg",
)
```

Figure 27 shows the Type 2 curve comparisons for the FR2020 and ESHM20 models against the observed ground motion data for France. The observed number of sites exceeding the given intensity measure level is shown in black, which is contrasted against the mean and quantiles of the PSHA models. The black line falls below the mean curve and the middle quantiles in both cases, suggesting that the observed rates of exceedances are lower than those in the centre of the model. In both cases the observations do fall within the 5th to 95th percentile, and in the case of the ESHM20 within the 16th to 84th percentile. The *tentative* interpretation here might be that the hazard models are overestimating with respect to the observed rate of exceedance (given the observation period considered), albeit that for the ESHM20 the agreement is closer.

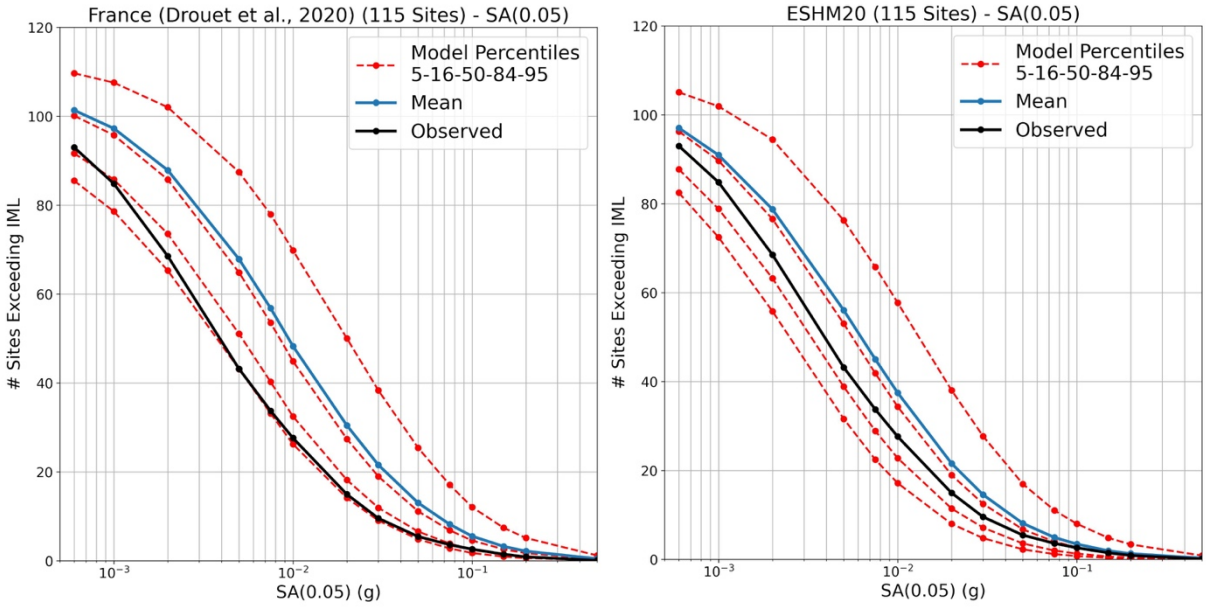


Figure 27: Comparison of the Type 2 aggregated curves between the observations and epistemic uncertainty range of the FR2020 PSHA model (left) and ESHM20 (right)

In the case of the Type 2 curves the underlying distribution of the observed exceedances is Poisson-binomial, so in the initial version of the toolkit we do not determine the confidence intervals. This will be added in future versions.

To plot the Type 4 curves, we need to determine the observed occurrences of ground motions exceeding the intensity measure levels corresponding to the specific probabilities of exceedance:

```
# Get the model IMLs corresponding to the fixes PoEs and the
# observed exceedances for these probabilities
obs_exceeded_probs, model_imls = psha_test.get_all_exceedance_poes(
    target_probs = target_probs,
    imts = imts,
    use_stats = True
)
```

Which are then input into the plotting functionality to return the comparison of Type 4 aggregated seismic hazard curves:

```
psha_test.plot_nsites_exceeding_poes(
    target_probs=target_probs,
    obs_exceed_poes=obs_exceeded_probs,
    ylim=(0, 70),
    imt="SA(1.0)",
    title="A Figure Type, e.g. Type 4 Curves",
    filename="./path/to/output_image_file.jpg",
    filetype="jpg",
)
```

The resulting Type 4 aggregated curves are shown for the FR2020 and ESHM20 models for Sa (0.05 s) in Figure 28. Compared to the Type 2 curves there are two significant differences. The first is that in this case the observations are modelled using a Binomial distribution, meaning that we can easily retrieve the confidence intervals for the observed number of stations exceeding the IML with a given PoE in 21 years. These are shown in the shaded region around the observed curve (in black). The second difference is that the observed curve appears above the mean. This is not a contradictory result to the Type 2 curves, but a manifestation of the fact that if the model is "overestimating" hazard then the IMLs for the give PoE will be higher and the number of sites exceeding their corresponding IML will be lower, and vice versa. So in this case the image is reversed but the actual trend is consistent with the Type 2 curves in Figure 27, which is that the observations seem to align with the lower quantiles (still in the 5- 95 percentile range) but that in the case of the ESHM20 a closer agreement with the centre of the hazard model distribution is found, particularly for the lower PoEs that get closer to the range of engineering significance.

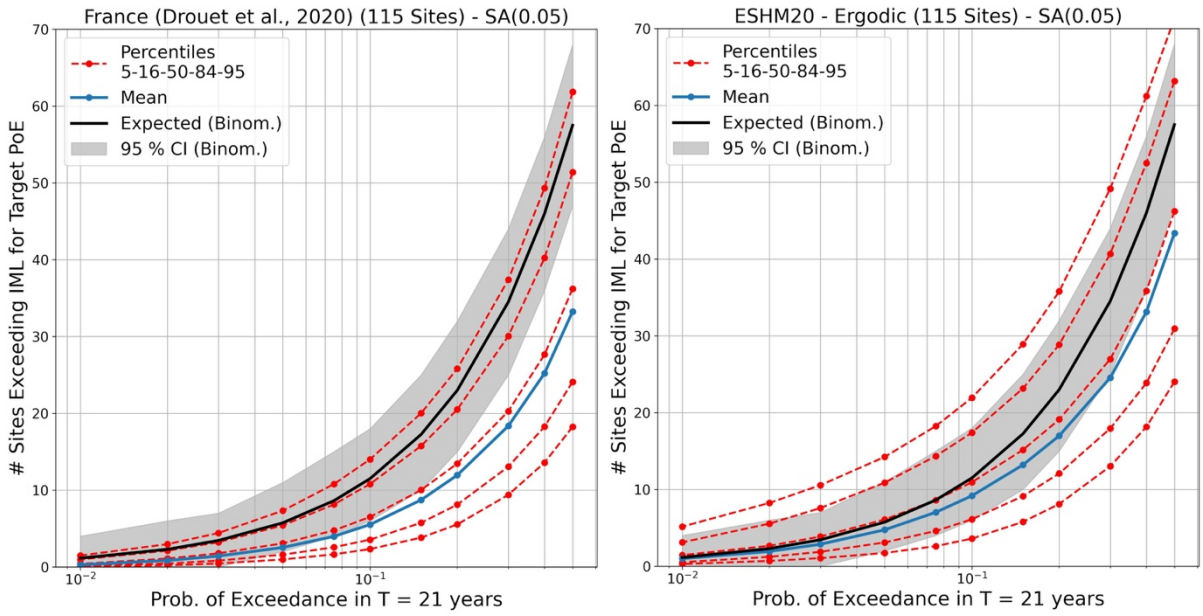


Figure 28: Type 4 aggregated hazard curves comparing observed numbers of sites exceeding their respective hazard level (blue line for the mean hazard and red dashed lines for the quantiles) and the number of stations exceeding their threshold level according to their set probability and the Binomial distribution with expectation (black line) and 5 – 95 % confidence intervals (grey shaded region)

We can also undertake a log-likelihood comparison of the models using the method of Albarello & D'Amico (2008). In this case we need to apply Monte Carlo sampling to both the observed and modelled log-likelihood. Similar to the Mak & Schorlemmer (2016) aggregated curves, the log-likelihood analysis can be run for with respect to exceedance of fixed intensity measure levels or with respect to exceedance of IMLs for fixed probabilities of exceedance. These two options are supported in the functions "get_likelihoods_imls" and "get_likelihoods_poes", respectively. Here we will focus only on the latter. To run this, we need to select appropriate probabilities of exceedance, which we will set here as the 0.2 and 0.0432 PoE in 21 years (≈ 95-year and 500-year return period respectively), and specify the number of samples to draw (here we use 1000). From the same PSHATests class we can retrieve the distribution of model and observed log-likelihoods for each probability of exceedance and each hazard curve. Once again, we can undertake this comparison on a curve-by-curve basis, albeit that this becomes a larger volume of data to manage (particularly for larger logic trees). Instead, we follow suit from the previous analysis and make the comparisons only against the mean and quantile curves. To retrieve the sample distribution of model and observation log-likelihoods we can run:

```
# Define target probabilities of exeedance in T = 21 years
# 0.2 ≈ 95 year RP, 0.0432 ≈ 500 year RP
target_probs = np.array([0.2, 0.0432])
# Get the distributions of log-likelihoods
model_llhs, obs_llhs = psha_test.get_likelihoods_poes(
    target_poes = target_probs,
    imt = "SA(0.05)",  # Choice of IMT
    nobs_samp=1000,  # Number of samples of observed llh
    nsamp=1000,  # Number of samples of model llh
    use_stats=True,  # Run only for the mean and quantiles
)
```

The outputs of this function ("model_llhs" and "obs_llhs" in the code snipped above) are both 3D arrays of log-likelihood values with dimension $[N_{POES}, N_{SAMP}, N_{CURVES}]$. The observed log-likelihood is dependent on the hazard curve at it relates to exceedance of the IML for a given PoE. For each PoE and hazard curve we are left with our N_{SAMP} values of log-likelihood, for which we leave the user to apply the appropriate hypotheses tests, which can be implemented easily for those available in the Scientific Python (Scipy) statistical tests library (https://docs.scipy.org/doc/scipy/reference/stats.html), or similar software. For a basic visualisation and interpretation, however, we do include tools to compare the respective histograms of log-likelihood function:

```
psha_test.plot_likelihood_stats(
    model_stats=model_llhs,  # Model log-likelihood
    obs_stats=obs_llhs,  # Observation log-likelihood
    poes = target_probs,  # All target PoEs
    selected_poe=0.0432,  # Which PoE to plot (from those in target probs.)
    llh_specs = (-30.0, 0.0, 1.5),  # (low, high, interval) for histograms
    ylim=(0.0, 0.5),
    xlim=(-30.0, 0.0),
    filename="./path/to/image_file.jpg",
    filetype="jpg"
)
```

For our two models (FR2020 and ESHM20) the above command produces a set of histogram comparisons (one per curve for the mean and quantiles), which are shown for the 0.2 PoE in 21 years case (≈ 95-year return period) in Figure 29 and for the 0.0432 PoE in 21 years case (≈ 500-year return period) in Figure 30. The distributions of the histograms largely convey the same narrative as shown in the Type 2 and Type 4 aggregated hazard curves. For the higher PoE the distribution of observed ℓ values is centred to the left of those for the model when considering the mean, median and upper quantiles. For FR2020 the greatest overlap in distributions is for the 16th percentile, while for the ESHM20 it is closer to the 50th percentile, suggesting observations are closer to the centre of the distribution for the ESHM20 case. For the lower PoE agreement is better overall for both cases and the centre of the observed ℓ aligns well with that of the mean and median curves for the ESHM20.

This demonstration outlines several of the comparison tools that are available and shows how we can use them to visualise and interpret the results. With further usage of the outputs from these functionalities one can readily design and implement hypothesis tests that would yield more binary

pass/fail criteria, but this is something that we stop short of doing in the current tools. We can, however, use the available tools to explore deeper into the models and, with careful usage of the PSHA calculations, take a look at certain research questions.

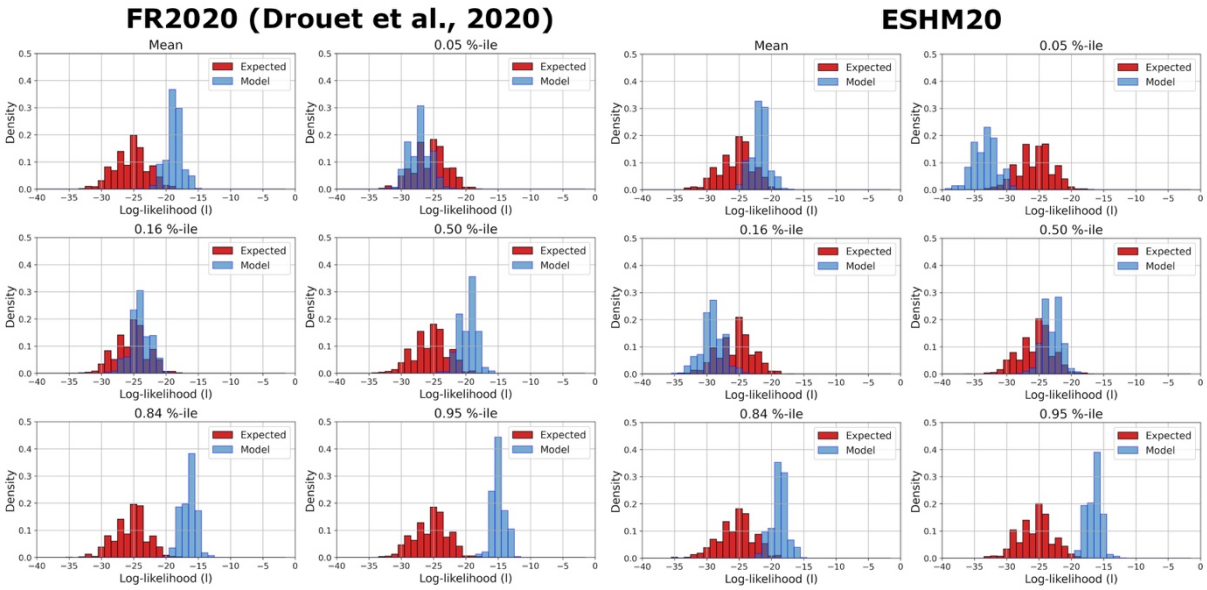


Figure 29: Comparison of the ℓ distributions for observations and hazard model (mean and quantiles) for the 0.2 PoE in 21 years for the FR2020 model (left) and ESHM20 (right)

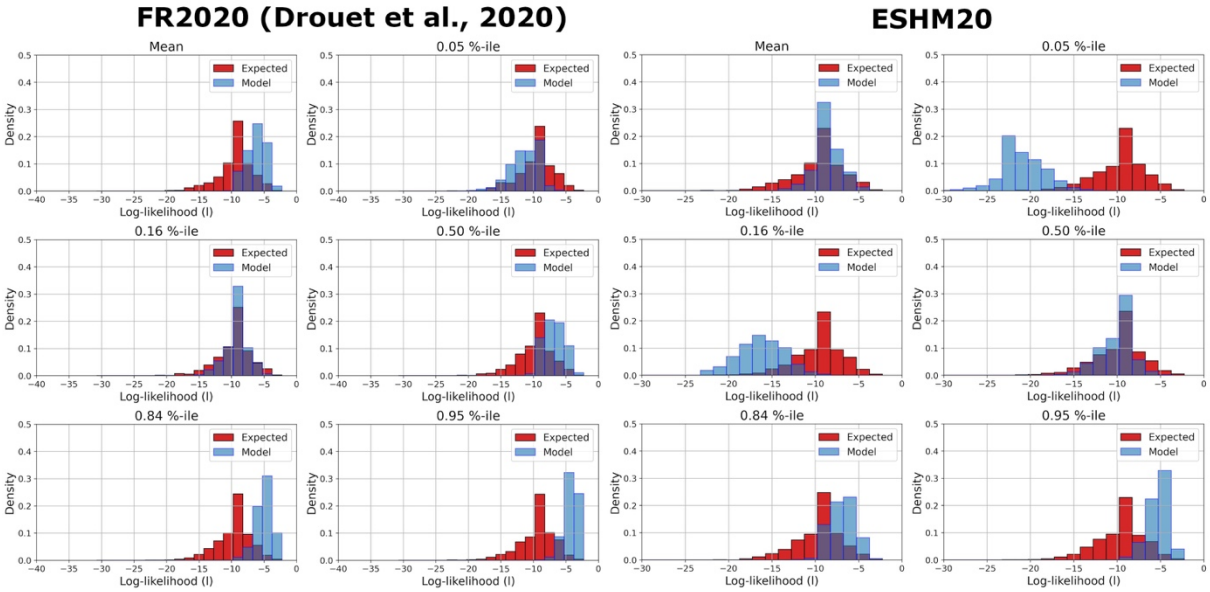


Figure 30: As Figure 29 for the 0.0432 PoE in 21 years.

4.5. Does (Partially) Non-Ergodic PSHA yield a different result when compared against ground motions?

The development of the data imputation by mixed effects regression described in Section 3 and implemented in Section 4 inadvertently highlights a critical issue in the PSHA to ground motion data comparisons. When we compare seismic hazard curves against multiple stations each will multiple observations of ground motion at a site, we are challenging the assumption of ergodicity in the models. In the development of GMMs, the median model is fit to observations of ground motion data pooled from a larger area (or from many tectonically analogous regions). For a given earthquake scenario

($M, R, \boldsymbol{\theta}$) the variability, σ, in the GMM will reflect the total variability from all source, path and site effects across the larger area from which the records are taken. This is the ergodic assumption.

In the case of a single site, for example, and assuming a GMM that adopts as its predictors M, R and V_{S30} the site-to-site variability ϕ_{S2S} of the ergodic GMM will reflect the variability in site amplification scaling across all sites with similar V_{S30}, therefore representing the amplification effects implicit from many different soil profiles. With repeated observations at a site, we can begin to constrain $\delta S2S_S$ specific to the site in question, with uncertainty that should decrease as the number of observations increases. With this information we can understand that the "real" variability in ground motion at the site in question is smaller than the ergodic variability as the site will tend to persistently produce ground motions amplifications greater or lower than that of the centre of the distribution implied from the full suite of observations used to constrain the GMM. In site-specific PSHA, obtaining multiple recordings at a site, even if from smaller or moderate magnitude events, allows us to constrain $\delta S2S_S$ and remove the site-to-site variability ϕ_{S2S} from the model (or at least transfer the uncertainty out of the GMM σ and into the reducible epistemic uncertainty). If one has available repeated observations of ground shaking for a given locality from the same (or nearby) seismogenic sources and/or covering similar travel paths then these erogodic components of source-to-source and path-to-path variability can be gradually partitioned out of the GMM σ, reducing it until the point it should converge toward the "true" non-ergodic variability at a site. For a more detailed explanation the reader is referred to a dedicated chapter on this topic in the state-of-the-art seismic hazard and risk textbook of Baker et al. (2021).

For the seismic hazard curves themselves the gradual reduction in σ from moving toward more partially or even fully non-ergodic GMMs has an important influence in steepening the seismic hazard curve, in most cases reducing the seismic hazard at lower probabilities of exceedance. In the case of site-to-site variability ($\delta S2S_S$), when we consider multiple sites with constrained $\delta S2S_S$ we would expect the median ground motion to increase ($\delta S2S_S > 0$) and to decrease ($\delta S2S_S < 0$) in roughly equal proportion. The absolute change in median ground motion will effectively translate the curve left and right along the axis of ground motion intensity level. But if in all cases σ is reduced, as $\phi_{S2S} \rightarrow 0$, then the curves will steepen, and the seismic hazard will overall tend toward a decrease across multiple sites. Add in more repeatable effects and the trend continues until one reaches the maximum possible reduction in σ from constraint of all possible repeatable source, path and site effects.

Why is this relevant for PSHA to observation comparison here? Our ground motion observations are a manifestation of the non-ergodic process at each specific site and thus $P(a \geq A|M, R, \theta)$ will reflect a variability that is smaller than the variability implied from the ergodic GMM. In most PSHA testing studies undertaken to date, stations with long time histories and multiple observations of ground motion are selected, meaning that the actual observations of ground motion exceedances will reflect more closely the non-ergodic hazard across the sites. We have made a similar selection bias in the tools here by specifically preferring stations with higher numbers of records when selecting stations with minimum spacing. The PSHA models we have considered so far, however, (Drouet et al., 2020; ESHM20) have applied specifically the ergodic form of the ground motion models. This is not an oversight or example of poor practice, but rather a necessity because regional models must provide estimates of seismic hazard across *all* sites of the target region, the vast majority of which will have no observations of ground motion, nor be located sufficiently close to well-instrumented areas to infer repeatable site effects. It is not necessarily a coincidence, nor evidence of systematic failures of PSHA (as some have claimed), that most PSHA testing studies return the result that the regional hazard model is overestimating the rate of exceedance with respect to observations. In these cases, ergodic PSHA, with its higher σ, is being compared against the manifestation of a non-ergodic process.

If the above assertion is true, how can we demonstrate it quantitatively? In the ideal case, we could want to implement PSHA for a *fully non-ergodic* GMM, i.e., one with fully constrained source, path and site effects. Recent developments in ground motion modelling have aimed to calibrate such models, which contain GMM coefficients that allow for variation in the stress parameter, attenuation and linear site amplification term across a high-resolution grid of cells. Such models are possible to constrain with a sufficient density of ground motion observations and earthquakes, and examples can be found for California (Landwehr et al., 2016; Abrahamson et al., 2019), Italy (Lanzano et al., 2021; Keuhn, 2023) and for Fourier Amplitude Spectra in France (Sung et al., 2022). Unfortunately, such models have not been included yet as part of regional scale PSHA (though they may soon do so) and are not yet supported by the PSHA software available for this study (OpenQuake). However, we do have two

possible options that, while not necessarily conclusive in quantifying exactly the impact of fully non-ergodic PSHA on the comparisons, may allow us to understand the general trend.

The first option is currently available in the ESHM20 model and its OpenQuake implementation, and that is to run a partially non-ergodic PSHA such that the median GMM is increased by $\delta S2S_S$ where available and the ϕ_{S2S} of the ground motion model is removed, such that $\sigma^2 = \tau^2 + \phi_0^2$. ESHM20 is one of the first regional scale models to adopt the backbone GMM strategy for epistemic uncertainty modelling. Here the GMM of Kotha et al. (2020; 2022) is selected as the core ("backbone") model. This GMM is itself an example of an intermediate point in non-ergodic GMM modelling, as it is constructed around regionalisations of ground motion observations in Europe (based on geology and tectonics) that allow for repeatable source region ($\delta L2L_l = \mathcal{N}(0, \tau_{L2L})$) and path/site region attenuation ($\delta c_3 = \mathcal{N}(0, \tau_{c3})$) differences to be quantified. The region-to-region variability in repeatable source-region and attenuation region effects is then used to constrain the epistemic uncertainty, represented by a distribution median ground motions (Weatherill et al., 2020). As per current practice, $\delta S2S_S$ was also determined for more than 1,100 stations across Europe. Conceptually, the fully ergodic form of the model is represented as:

$$\ln Y = f(M, R, \delta c_3) + \delta L2L_l + \delta B_e^0 + \delta S2S_s + \varepsilon \quad 4.32$$

$$\sigma^2 = \tau_{L2L}^2 + \tau_{c3}^2 + \tau_0^2 + \phi_{S2S}^2 + \phi_0^2 \quad 4.33$$

Where δB_e^0 is the source-region corrected between-event variability ($\delta B_e^0 = \mathcal{N}(0, \tau_0)$) and $\varepsilon = \mathcal{N}(0, \phi_0)$ the site corrected within-event variability. In the development of the logic tree both τ_{L2L} and τ_{c3} are used to constrain the source and attenuation variability in the backbone GMM logic tree, and are thus removed from σ. Its original form, presented in Danciu et al. (2021), the hazard calculation adopts $\sigma^2 = \tau_0^2 + \phi_0^2 + \phi_{S2S}^2$, which is what we have compared so far. In OpenQuake we also have the possibility to set $\delta S2S_S$ explicitly for the sites we have data available, and thus $\phi_{S2S} \approx 0$. This is a *partially non-ergodic* implementation of the model; partial because we have not necessarily been able to remove repeatable source and path effects to the extent done by *fully* non-ergodic models.

The second option available to us is to adopt the ergodic GMMs in the respective models without adjustment, but instead replace their original ergodic σ with a value that reflects the greatest reduction in σ that might be achievable were the data sufficient to constrain a fully ergodic model. Effectively, we use the ergodic GMMs with a (close to) fully non-ergodic σ. This should not be seen as a fully accurate approach, but it can be indicative of the scale of the reduction in variability that is achievable if we assume that the current centre of the distribution of ground motions across the target sites is representative of the centre of the eventual non-ergodic model were the data available. Basically, we are reducing σ but cannot systematically shift the median ground motion to higher or lower values across our region. We refer to this as the *fully* non-ergodic hazard, albeit it contains the caveats indicated above. For this purpose, we consulted the non-ergodic GMMs compared for Italy by Kuehn (2023), which yield non-ergodic σ in the range 0.25 – 0.43. We select a mid-case option of $\sigma = 0.36$, which is applied to all periods.

We will not repeat all of the steps shown in the usage of PyPSHATest, we have kept the same data set of observations from weak and strong motions for the same sites in France shown previously. The imputation process also remains the same, and the Abrahamson et al. (2014) GMM is used for this purpose as it was found be largely unbiased in trends of δB_e, δW_{ES} and $\delta S2S_S$. The Kotha et al. (2020) model with the ESHM20 adjustments was found to perform similarly well; however, given that the ESHM20 GMM logic tree is based entirely on this model (for shallow seismicity) it is more arguable that we might be skewing our data to favour the ESHM20 model if we used it for the imputation (this was actually tested and found not to be true, but we will proceed with Abrahamson et al. (2014) in the examples here). For the same set of 357 weak and strong motion target sites in France, we run five different PSHA calculations:

1) ESHM20 source and ground motion model with the fully ergodic site-to-site variability (as run for the ESHM20 itself)

2) ESHM20 source and ground motion model with partially non-ergodic site-to-site variability using $\delta S2S_S$ where available and setting ϕ_{S2S} to 0.

3) ESHM20 source and ground motion model with "fully" non-ergodic variability ($\sigma = 0.36$) for all periods

4) FR2020 source and ground motion model with fully ergodic variability

5) FR2020 source and ground motion model with full non-ergodic variability

As Figures 27 to 30 effectively indicated the same narrative, we show only the results of these comparisons for the Type 4 aggregated curves, though it should be easy to see that the full suite of results can be retrieved following the examples illustrated previously. The comparison of the five different hazard models for same 115 selected target sites (from all 357 available) is shown for Sa (0.05 s), Sa (0.2 s) and Sa (1.0 s) in Figures 31, 32 and 33 respectively.

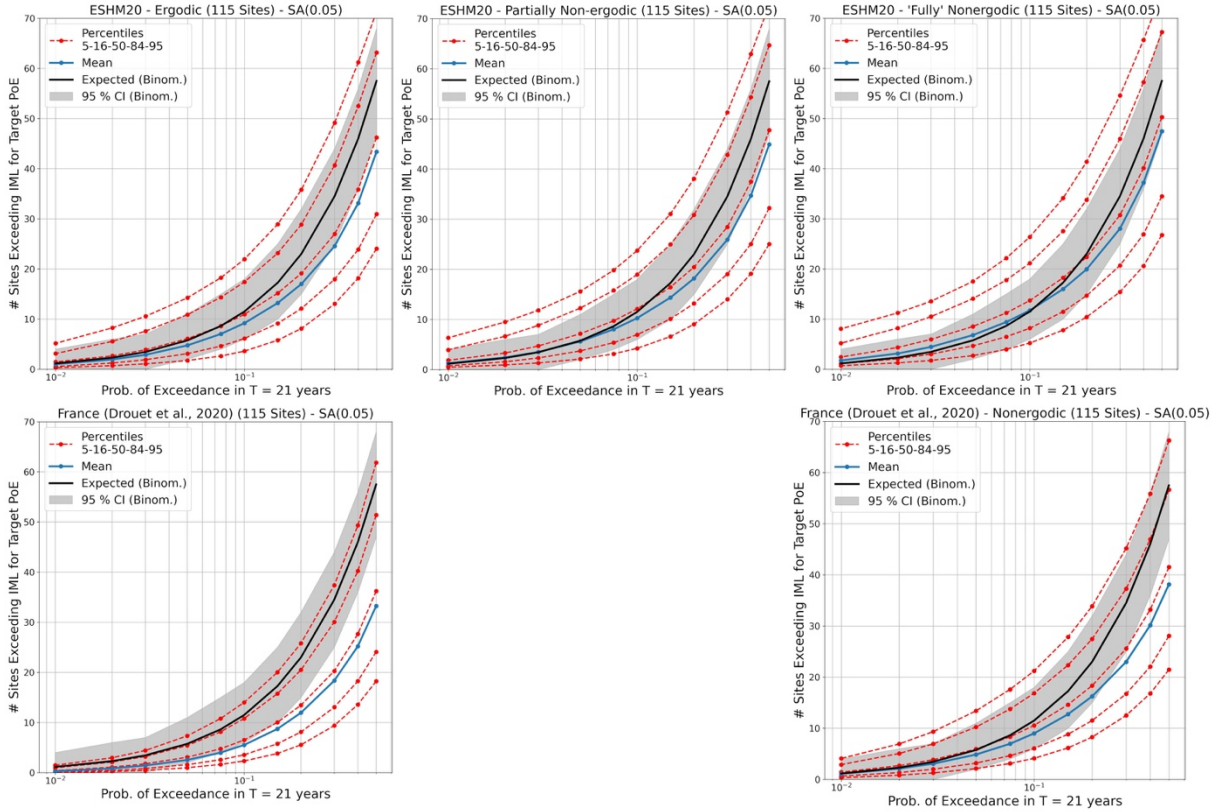


Figure 31: Comparison of Type 4 aggregated seismic hazard curves and data for Sa (0.05) ESHM20 (top row) and FR2020 (bottom row) using ergodic PSHA (left columns), partially non-ergodic PSHA (centre column, ESHM20 only) and "fully" non-ergodic σ (right)

The general trend of the Type 4 curves compared in Figures 31 - 33 bear out our assumption that as one transitions from fully ergodic to non-ergodic, we generally see a shift in where the observed exceedance curves sit with respect to the quantiles of the model. Looking at the results for the short period Sa (0.05 s) curves we see that for probabilities of exceedance of engineering relevance (in the range 0.03×10^{-2} to 0.1 in 21 years) in both cases the observed exceedance curves come into closer agreement with the mean and median curves of the hazard models. For higher probabilities of exceedance, the observed hazard curves sit above the mean and median hazard; however, for the ESHM20 model with fully non-ergodic σ both the mean and median hazard curve fall within the 5 – 95 % confidence intervals of the observed exceedance. Based on the comparisons shown here we cannot necessarily argue that there is no systemic bias toward overestimation or underestimation of ground motion exceedance in the models, but we see that as we remove the ergodic components of variability the tendency toward systematic overestimation reduces. This has important implications, which we will revisit in the concluding section of the report.

A more perplexing trend does seem to be apparent, however, in comparing Type 4 curves for short period motion Sa (T ≤ 0.2 s) to those for longer period motion (Sa (T = 1 s)). The same trend toward

closer convergence between model and observation when considering fully non-ergodic PSHA is apparent, but there appears to be more systemic overestimation of exceedance rates by the models for longer spectral periods. The exact cause of this is unclear and will warrant further investigation following this report. That this degree of overestimation is not present for shorter periods would seem to bring into question the idea that this indicative of a systematic overestimation in the hazard, but the specific cause(s) may be hard to discern. One possibility is that the use of weak motion to constrain δB_e and $\delta S2S_S$ may be introducing a systematic bias that is reducing the number of records with strong longer period components. Comparison of the ground motion residuals for weak and strong records does not necessarily indicate such a systematic problem. The only trend of note appears to be a slight tendency toward negative δB_e at larger M for the Abrahamson et al. (2014) GMM, but this same tendency is present in both the strong and weak motion data. Likewise, site effects may play a role; however, of the 115 stations that are retained in this comparing 73 are on stiff soil or rock sites with $V_{S30} \geq 600$ m/s. So, it would seem unlikely that local site issues can explain the full discrepancy. Further research is ongoing to explore the robustness of this observation to different modelling assumptions.

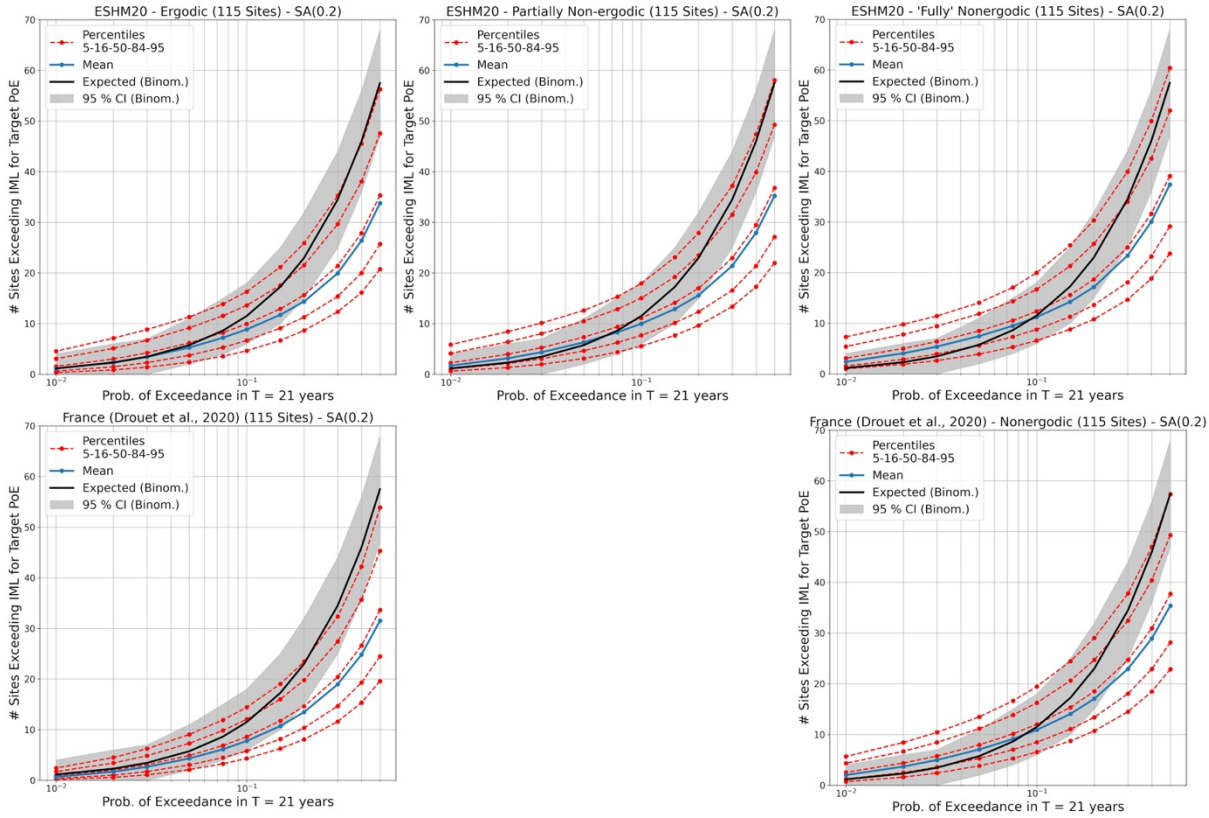


Figure 32: As Figure 31 for Sa (0.2 s)

This analysis does not provide a conclusive answer to the question of whether non-ergodic PSHA results in differences in the model to data comparisons, but it does support the notion. This is not unexpected given the entire premise on which non-ergodic PSHA is based; however, both the hazard models and the testing tools give us a means of demonstrating it. While the data and models themselves can be challenging and time consuming to set up and run, the workflow for comparison of the models shown here is a culmination of a process that can be implemented in a simple computational framework building on the tools developed within PyPSHATest. This can be (and in fact was) implemented in a single Jupyter Notebook or Python script that allows the comparisons to be run in a matter of seconds, affording the modellers control over specific assumptions in order to understand their impact.

The analysis presented here also did not set out with the objective of "testing" the different models for the purposes of commenting on the suitability of either ESHM20 or FR2020 for France. Additionally, we do not overlook the clear fact that for the range of PoEs of engineering significance, the number of sites expected to be exceeded is small (less than 20) and that the confidence intervals of the observed exceedance curve span much of the predicted hazard range. Taken at face value, one might easily claim that this analysis verifies that the models agree with observations for the probabilities of exceedance of

engineering interest, but this should be approached with caution given the limited time period of observation and the low overall level of seismicity. With the possible exception of the results at Sa (1.0 s), which clearly require further investigation, it is arguable that none of the comparisons here would necessarily suggest any model should be rejected for consideration in applications when considering both the full epistemic uncertainty range of the model *and* the uncertainty in the observations themselves. This brief demonstration does, however, indicate that the degree of ergodicity in the GMM variability can have a significant impact in reconciling observations and predictions of exceedance of ground motions. This is an important consideration when looking at what actions might be made on the basis of such comparisons.

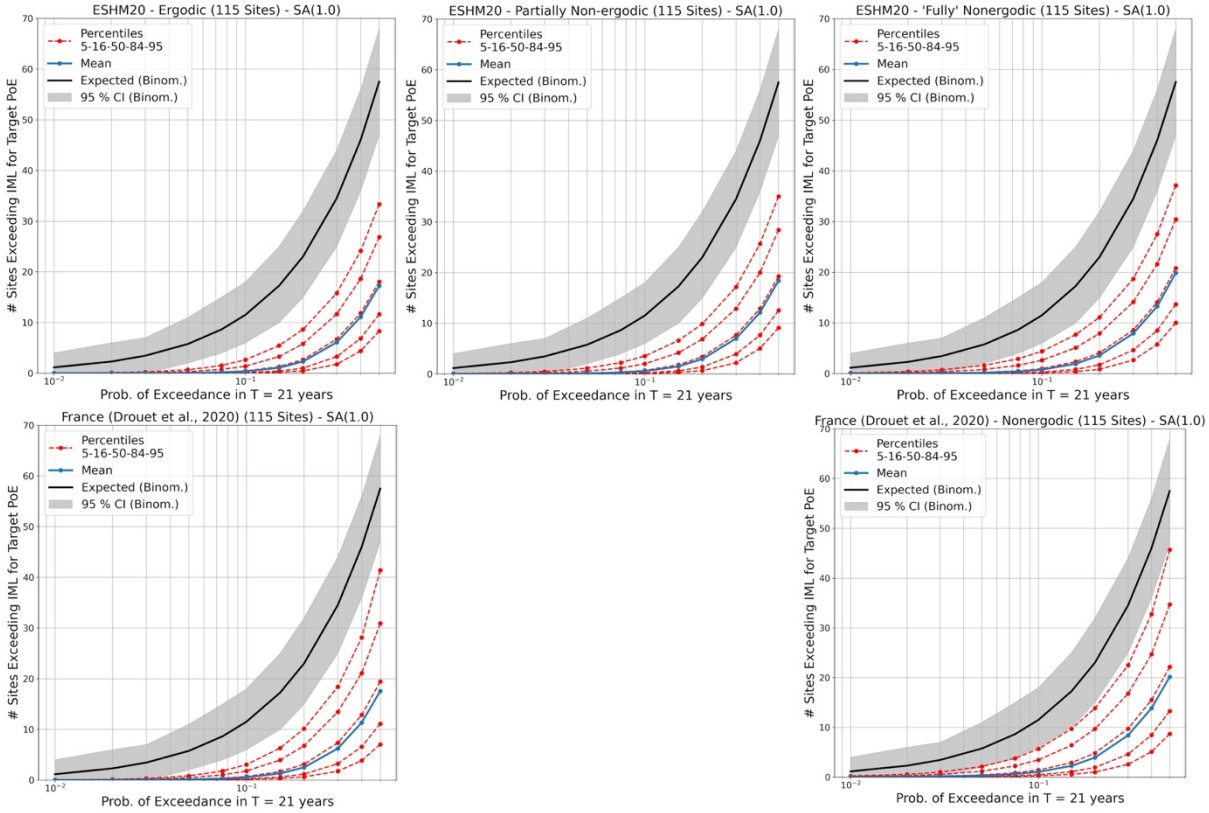


Figure 33: As Figure 31 for Sa (1.0 s)

5. Hazard to Data Comparisons: Considerations for Application and Future Developments

When we began the process of developing tools for "testing" PSHA against observed ground motions, we did so with the awareness of the limitations of such comparisons (e.g., Beauval et al., 2008; Mak et al., 2014). Despite the occurrence of the term "test" in the name of the toolkit, we have stopped short of implementing simple functions that compare hazard curves and observations from a simple file and return a binary pass/fail outcome. While inconsistency with observations is something that should be addressed carefully by the hazard modeller, there are many limitations in direct comparison with ground motion that need to be addressed and corrected for. The scope of PyPSHATest therefore expands outside of the features needed to reproduce some of the existing studies in the literature, at the cost of not (yet) including some that might have been adopted in similar testing studies. There is work ongoing to look at some specific issues that fell within the scope of the work in the METIS project but have not been completed to satisfaction within the time available. This work will continue, and we hope to keep expanding the functionality and robustness of PyPSHATest in the future.

Based on the experiences learned in constructing the tools within PyPSHATest and applying them in the examples here (and elsewhere), we take the opportunity to reflect on what we aimed to do in this process, the extent to which this has been achieved, and how the outputs of the tools should be used and interpreted by the modeller. The developments have been undertaken with these questions in mind, and so we conclude this report by seeking to provide insights, clarification, and guidance for hazard modellers in the future.

If PyPSHATest does not provide tools for statistical testing (e.g., hypothesis testing) of PSHA models, then what is its purpose?

The direct testing/validation of PSHA curves using observed ground motion data, which is implied by this question, contains a series of assumptions and uncertainties that cannot necessarily be reduced into a simple pass/fail framework, and it may be misleading to do so. The primary objective of PyPSHATest has been to delve deeply into the process how the comparisons are undertaken, what is certain and what may be unconstrained, and to then provide tools to allow reproduction of the process from end-to-end. In this sense the greatest effort in PSHA to data comparison is in setting up both the models and the data to the extent that the fairest comparisons between the two can be undertaken. This requires transparency and reproducibility in the PSHA calculations, something that is greatly facilitated with the use of OpenQuake and the integration of OpenQuake's own functionality into PyPSHA test. It also, however, requires transparency in the data set used for testing, the way it is processed, the inherent biases present in the data and how we correct them. The toolkit therefore contains many tools to facilitate this process too, especially in relation to the use of data imputation to correct of the inherent bias of the incompleteness of observations in a data set.

So what can a hazard modeller do with PyPSHATest? If we look back through the workflow, we can see that there are many areas where model development and model testing are more closely interlinked, and we have aimed to make this process easy and reproducible. Examples include:

1) Comparing seismicity models from two different PSHA models, which can be easily done from models supplied in an OpenQuake seismogenic source model logic tree structure using the quantitative comparison tools presented in section 2.

2) Different PSHA models for the same region can be compared quantitatively via a variety of metrics, including some that can help identify spatial variation in the differences. The use of the datastore here can make customised analysis and enquiries of the models fast and efficient.

3) Comparing seismogenic source models models against observed seismicity. Though not done directly in the toolkit, PyPSHATest can rapidly render a source model (or suite of source models) into a seismicity rate grid format and export them directly into a GriddedForecast for application

with PyCSEP. In this sense, we have not re-invented the wheel in terms of adding seismicity forecast testing functionality but rather we allow for direct integration with this powerful and growing software, leveraging on the range of tests on offer and the expertise of the existing seismicity forecasting community.

4) Comparing GMMs against data and assessing trends/biases in source, path, and site scaling. This can be done with the ground motion residual tools that were developed as part of the data imputation process in Section 3. This does not include a model "scoring" functionality that would select/reject or up-weight/down-weight a given model(s) (though such scoring metrics could be easily added using the outcomes of the residual analysis), but it does provide a means of making deeper insights into the agreement between models and data to give the modeller the capability of assessing suitability themselves.

5) Analysis of station to station, and even channel-to-channel, variability at observed recording stations. Application of the residual calculations in Section 3 will, for a given flatfile, yield a database of site-to-site residual terms $\delta S2S_S$ for all sites in the file. Where data are sufficient, this can help identify locations of particularly high or low amplification, and with careful analysis of $\delta S2S_S$ and the observed motions could even be useful to identify potential problems at stations with co-located accelerometer and broadband velocity recording instruments.

6) With a ground motion flatfile and a harmonised earthquake catalogue (two data sets that are commonly developed in any PSHA calculation), the tools can quickly provide an apparent earthquake history for the stations in the database and allows the user to decide how to fill in the gaps using inferences from available data.

7) Parameters to control both the data processing and the model to data comparisons are easily configured by the user. This means it is possible to assess rapidly the influence of, for example, the choice of GMM for the data imputation, the influence of declustering, the minimum magnitude and/or maximum distance, etc.

8) To the author's knowledge, we provide the first open-source tools that specifically implement existing hazard model comparisons strategies described in previous publications (e.g., Albrarello & D'Amico, 2008; Mak & Schorlemmer, 2016), and do so in a framework compatible with state-of-the-art PSHA software. We have also adapted some of these methods to account for the uncertainty in the observed rates of ground motion exceedance.

Is it still possible to apply statistical hypothesis tests to compare models and data, based on the functionalities here?

One cannot readily anticipate all potential comparisons and tests that one may wish to undertake; however, once the user is familiar with the input and outputs of the functions in PyPSHATest they may have the capability to use functions available among existing Python packages for statistics and machine learning, such as Scipy.stats, Scikit-Learn or StatsModels. Further extensions to the current library to allow greater interoperability with these additional packages would be anticipated for the future, albeit we would wish to solicit feedback from users as to the nature of their workflows to best understand how to meet their needs. The toolkit is released under a Git development platform, which gives users an easy means of providing feedback, reporting problems, or making requests (with explanations) for new features.

Can I use weak motion data and strong motion data for comparison against PSHA model and can I access existing databases of ground motions?

PyPSHATest not only makes the process of integrating weak motion data into the comparisons possible, or at the very least it can be agnostic about the type of instrument used to record the motion, it also provides tools to interrogate the data further and make comparisons. In the example application shown in Sections 3 and 4 we have used such a database of weak and strong motion, which was compiled using open data: EPOS EIDA (http://www.orfeus-eu.org/data/eida/) and IRIS (https://service.iris.edu/). Download and automatic processing of the ground motion data was undertaken using Stream2Segment (Zaccarelli et al., 2019; https://github.com/rizac/stream2segment), and open-source tool developed for this purpose. In this sense the entirety of the workflow is reproducible, and the tools are capable of undertaking analysis using any processed database of ground motions. An interested user may wish to

use instead existing flatfiles or databases of ground motions if they wish. One limitation; however, is that widening the number of available stations and earthquakes being used will inevitably introduce problems of recording quality and missing metadata. Some quality checks can be implemented in the automatic processing with Stream2Segment, while others may require custom code and workflows developed by the user. Further examples may be added to the available documentation to demonstrate how these quality checks and metadata issues were addressed in the present case.

Can PyPSHATest be used to prune or down-weight branches of the logic tree that are incompatible with observations?

PyPSHATest offers the ability to compare observations and data on a branch-by-branch basis. Should a user wish to define a statistical basis for compatibility in terms of a hypothesis test or similar criterion then this could be implemented combining outputs of PyPSHATest's functionality with existing tools. But while this practice is certainly possible, it should be approached with caution. There are several reasons why this may not necessarily be advisable. The first is the limitation of the data set used for the comparison and the potential for over-interpreting inferences gained from such limited data. The application shown in Section 4 illustrates the degree of uncertainty we can have on our observed hazard curve because of the completeness. When factoring in the recognition that the window of observation of instrumental ground motions is inevitably short compared to the time-span of the earthquake process, and that inferences based on high probability (short return period) hazard do not necessarily provide insights into the processes controlling the lower probability (long return period) hazard, pruning the branches on the basis of direct comparison with the resulting hazard may be dangerous.

The second issue is that in a suite of hazard curves from a logic tree one cannot always identify a one-to-one relation between the high/low outlier branches and the specific factor in the component models that produce them. There can be trade-offs between the relative impacts of the source model and ground motion model that could produce branches of hazard curves that are not incompatible with direct observations of shaking, even though neither source nor ground motion model branch is itself incompatible with the underlying data. For example, a source branch yielding low seismicity rates (say -1σ), combining with a branch that produces systematically low ground motions (such as a low stress drop branch in a backbone GMM logic tree), may produce a particularly low PoE hazard curve that would be inconsistent with observations. Their combination of probabilities *should* result in a curve with a low weight but given that neither source model nor GMM branch is itself unlikely, and both are independent, then is there no sound basis for removal of the branch compared to accepting it with its current low weight. Conversely, two contrasting outlier branches (e.g., low seismicity rate source model with high stress drop GMM, or vice versa) may be individually highly unlikely and incompatible with their own respective observations of seismicity/ground motion, yet still combine to produce a hazard branch that *is* compatible with observations.

The third reason is effectively illustrated by the comparison of ergodic and non-ergodic PSHA models in Section 4.5. Both models (FR2020 and ESHM20) had logic trees on the order of hundreds or thousands of branches, and their distribution in terms of hazard curves could be seen and compared in Section 2 and Section 4. When comparing the ergodic PSHA models against observations we saw that the observations agreed well with the lower quantiles (5 % and 16 %), that that mean and median seemed to fall in the 5 – 95 % confidence intervals of the observed rate of exceedance (so not rejected, but not necessarily favoured) and that the upper quantiles would be outside the confidence intervals and thus "rejected". Once we begun to remove the ergodic components of the ground motion variability, these trends weakened, and the observed curve tended to align more closely with the centre of the distribution, and the upper branches no longer fell outside the confidence interval. Perhaps this is an unusually favourable example, but had one pruned the logic tree on the basis of the ergodic model then one would have biased the ergodic hazard to match the non-ergodic observations, potentially eliminating important branches that capture the possibility of, for example, higher than expected ground motions or seismicity, and thus skewing the resulting hazard toward the low end of the distribution with potentially serious consequences for safety in seismic design.

The comparison of the ESHM20 provides a clear case-in-point illustration of the danger of this approach. In the GMM logic tree of the ESHM20 there are 15 branches, five for source-region stress parameter uncertainty combined with three for residual attenuation region uncertainty (Weatherill et al., 2020). The source region stress parameter uncertainty is quantified using source-region to source-region

variability in the Kotha et al. (2020) model (τ_{L2L}), which is combined with an additional large magnitude uncertainty but that is not influencing here. The distribution $\mathcal{N}(0, \tau_{L2L})$ was calibrated directly on observed ground motion data, which we could see in the region-to-region variability $\delta L2L_l$, and mapped into five branches to capture the primary moments of the distribution. As a direct up/down scaling factor on the median ground motion the τ_{L2L} uncertainty was a dominant one in the full logic tree, with many of the lower and/or higher curves corresponding directly to the low/high τ_{L2L} branches. Had we undertaken the comparisons shown in Figures 31 - 33, we would have concluded that the low τ_{L2L} branches fit the direct observations better and upweighted these, while down-weighting (or removing entirely) the high τ_{L2L} branches. Yet if we had looked in detail at the $\delta L2L_l$ inferred for regions in France we would have seen slightly positive values, suggesting that the ground motions themselves would have been better aligned with the upper τ_{L2L} branches. Were we to have made the comparisons based on the partially or fully non-ergodic analysis, however, we would then have seen that the observed exceedances for PoEs of engineering relevance fit well the mean curve and even skewed toward the higher branches. This would have been in better alignment with what is seen directly in the ground motion data.

Where the need for pruning is deemed necessary, we would largely advocate that this is done based on the component level comparisons, such as the seismicity model forecasts or the ground motion residual distributions. Several of these comparisons are possible with, or facilitated via, PyPSHATest, and we would be keen to continue to add others in the future.

Do I need to have a minimum spacing between stations or could I use all stations to increase the number of observations?

This question relates to an interesting problem that has been the discussion of several papers (Albarello & Peruzza, 2017; Iervolino et al. 2017), which is that of the assumption of independence between the stations. The assumption of independence is necessary for both the calculation of seismic hazard and the statistical model to describe the probabilities of exceedance of ground motion at multiple sites. On the seismic hazard side $P[a \geq A|M, R, \boldsymbol{\theta}]$ is calculated independently for each site, and while each rupture in the ERF will inevitably produce ground motion at relevance of multiple sites, the probabilities of exceedance at a given site is determined conditionally independent of that of any neighbouring sites. Observations of ground motion records from well recorded events demonstrate clearly that ground motion variability is correlated between two or more sites then those sides are located close together with small separation distance, h, and that this correlation decreases as h increases. Empirical models of describing the correlation in ground motion variability with distance and with spectral period have been in the literature for nearly two decades (e.g., Boore et al., 2003; Goda & Hong, 2008; Jayaram & Baker, 2009; Esposito & Iervolino, 2012). Such models are usually based on fit of semi-variograms, which describe the correlation according to an exponential model $\rho(h) = 1 - \exp\left(-\frac{3h}{r}\right)$, where r is the *range* of the correlation (or correlation length), i.e. the distance h where the $\rho(h)$ decays to the point of practical triviality (usually $\rho(h) = 0.05$). The actual correlation length is dependent on the spectral period of ground motion, the choice of GMM used to determine the residuals, and the region in question. While the shape of the curves and the correlation lengths do differ, for the spectral period range being considered here most models to converge toward $\rho(h) < 0.1$ at distances beyond 30 km. This was the basis for selecting a minimum spacing of 30 km in the example shown here, as this would seem to be a distance at which conditional independence in ground motion exceedance in the hazard may be a reasonable assumption.

The assumption of conditional independence in exceedances across multiple sites is not just an issue for the hazard calculation, but also for the probability distributions assumed in making the statistical comparisons. Most critically, the probability distributions of aggregated exceedance curves are based on the assumption that exceedance of ground motion across S sites represents the outcome of S independent Bernoulli trials, which results in a Binomial distribution if P_S are equal and a Poisson-binomial distribution if they are not. When sites are located close together, however, our observed exceedance can no longer be represented by a set of independent Bernoulli trials as the probability of exceedance at one site is not independent of that at a neighbouring one. This dependence has the impact of increasing the variance σ of the corresponding probability distributions, but the degree of increase depends on the spatial configuration of the stations in question and the seismicity in a region.

Iervolino et al. (2017) utilise a stochastic event based hazard calculation approach to generate hazard curves across multiple strong motion sites in Italy, incorporating spatial correlation into the sampling of ε from the ground motion model to allow conditional dependence of probabilities of exceedance at closely situated sites. Their analysis demonstrated that the variance σ increased, meaning that that rejection of models because of their position in the distribution tails was less likely. Albrarello & Peruzza (2017) propose a method to calibrate the correction factor $\sigma' = c \cdot \sigma^u$ that should be applied to the Poisson-binomial distribution modelling the number of exceedances of ground motion. Their method requires modelling the change in σ for different GMMs and spectral periods by simulating ground motions at the stations from an observed catalogue of earthquakes *a priori*. Other approaches to the problem can be attempted by basing comparison statistics on probability distributions that do not require independence in the Bernoulli trials.

OpenQuake and PyPSHATest allow for some exploration of this topic and further research is ongoing. From its stochastic event based PSHA calculator OpenQuake can generate seismic hazard curves that incorporate spatial correlation between sites, and this is easily configurable from the available models used in the literature. Thus, it can reproduce the approach of Iervolno et al. (2017) without the need for modification and the resulting seismic hazard outputs can be analysed with PyPSHATest in the manner shown here. Similarly, the data imputation processes within PyPSHATest can be adapted to implement the method proposed by Albarello & Peruzza (2017) to constrain σ'. The focus of remaining activities in METIS on this specific topic is to extend capability of PyPSHATest to incorporate probability distributions that can allow for associated Bernoulli trials with exchangeability. Until the new features are further understood and implemented, we would recommend maintaining the approach of using a minimum station spacing in the site selection. However, interested users may wish to explore different spacings to understand the potential impact on the comparisons.

Can I use macroseismic intensity to test the seismic hazard?

The limitations of macroseismic intensity for this purpose were discussed in section 1. However, the adaptation of the procedure for constraining observed exceedance in order to allow uncertainty, does allow for compatibility between macroseismic intensity and the data imputation. Moderate extension to the data imputation tools could allow this to be integrated with minimal cost; however, certain challenges remain. Arguably the greatest of these relates to the completeness of the macroseismic intensity observations and, by extension, to the completeness of the harmonised earthquake catalogue. We would encourage users to outline a case study application here and proposed the most suitable strategies to compensate for the significant spatio-temporal variation in earthquake catalogue completeness that accompanies the use of historical (pre-instrumental) seismicity.

Can I contribute to the development of PyPSHATest?

As a final point, PyPSHATest is an open-source tools, which is available from (https://gitlab.pam-retd.fr/openmetis/pypshatest) at the time of preparation of the report. We strongly encourage users to test the code and use the "issues" of the code repository to communicate feedback and make requests for new features. The code is free to download and distributed under a GNU GPL v3.0 licence. Users are welcome to contribute code at any time, though it will be subject to review, and contributions are managed via Git.

6. Bibliography

Abrahamson, N. A., Silva, W. J., & Kamai, R. (2014). Summary of the ASK14 Ground Motion Relation for Active Crustal Regions. Earthquake Spectra, 30(3), 1025–1055. https://doi.org/10.1193/070913EQS198M

Abrahamson, N., Kuehn, N., Walling, M., & Landwehr, N. (2019). Probabilistic Seismic Hazard Analysis in California Using Nonergodic Ground Motion Models. Bulletin of the Seismological Society of America, 109(4), 1235–1249. https://doi.org/10.1785/0120190030

Albarello, D., & D'Amico, V. (2008). Testing probabilistic seismic hazard estimates by comparison with observations: An example in Italy. Geophysical Journal International, 175(3), 1088–1094. https://doi.org/10.1111/j.1365-246X.2008.03928.x

Albarello, D., & D'Amico, V. (2015). Scoring and Testing Procedures Devoted to Probabilistic Seismic Hazard Assessment. Surveys in Geophysics, 36(2), 269–293. https://doi.org/10.1007/s10712-015-9316-4

Albarello, D., & Peruzza, L. (2017). Accounting for spatial correlation in the empirical scoring of probabilistic seismic hazard estimates. Bulletin of Earthquake Engineering, 15(6), 2571–2585. https://doi.org/10.1007/s10518-016-9961-0

Albarello, D., Peruzza, L., & D'Amico, V. (2015). A scoring test on probabilistic seismic hazard estimates in Italy. Natural Hazards and Earth System Sciences, 15(1), 171–186. https://doi.org/10.5194/nhess-15-171-2015

Anderson, J. G., Brune, J. N., Biasi, G., Anooshehpoor, A., & Purvance, M. (2011). Workshop Report: Applications of Precarious Rocks and Related Fragile Geological Features to U.S. National Hazard Maps. Seismological Research Letters, 82(3), 431–441. https://doi.org/10.1785/gssrl.82.3.431

Atik, L. A., Abrahamson, N., Bommer, J. J., Scherbaum, F., Cotton, F., & Kuehn, N. (2010). The Variability of Ground-Motion Prediction Models and Its Components. Seismological Research Letters, 81(5), 794–801. https://doi.org/10.1785/gssrl.81.5.794

Baker, J. W., Abrahamson, N. A., Whitney, J. W., Board, M. P., & Hanks, T. C. (2013). Use of Fragile Geologic Structures as Indicators of Unexceeded Ground Motions and Direct Constraints on Probabilistic Seismic Hazard Analysis. Bulletin of the Seismological Society of America, 103(3), 1898–1911. https://doi.org/10.1785/0120120202

Baker, J. W., Bradley, B. A., & Stafford, P. J. (2021). Seismic Hazard and Risk Analysis. Cambridge University Press.

Beauval, C., Bard, P.-Y., Hainzl, S., & Gueguen, P. (2008). Can Strong-Motion Observations be Used to Constrain Probabilistic Seismic-Hazard Estimates? Bulletin of the Seismological Society of America, 98(2), 509–520. https://doi.org/10.1785/0120070006

Bindi, D., Massa, M., Luzi, L., Ameri, G., Pacor, F., Puglia, R., & Augliera, P. (2014). Pan-European ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5 %-damped PSA at spectral periods up to 3.0 s using the RESORCE dataset. Bulletin of Earthquake Engineering, 12(1), 391–430. https://doi.org/10.1007/s10518-013-9525-5

Boore, D. M. (2010). Orientation-Independent, Nongeometric-Mean Measures of Seismic Intensity from Two Horizontal Components of Motion. Bulletin of the Seismological Society of America, 100(4), 1830–1835. https://doi.org/10.1785/0120090400

Boore, D. M., Gibbs, J. F., Joyner, W. B., Tinsley, J. C., & Ponti, D. J. (2003). Estimated Ground Motion From the 1994 Northridge, California, Earthquake at the Site of the Interstate 10 and La Cienega Boulevard Bridge Collapse, West Los Angeles, California. Bulletin of the Seismological Society of America, 93(6), 2737–2751. https://doi.org/10.1785/0120020197

Boore, D. M., Stewart, J. P., Seyhan, E., & Atkinson, G. M. (2014). NGA-West2 Equations for Predicting PGA, PGV, and 5% Damped PSA for Shallow Crustal Earthquakes. Earthquake Spectra, 30(3), 1057–1085. https://doi.org/10.1193/070113EQS184M

Brune, J. N. (1999). Precarious Rocks along the Mojave Section of the San Andreas Fault, California: Constraints on Ground Motion from Great Earthquakes. Seismological Research Letters, 70(1), 29–33.

Brune, J. N. (2002). Precarious-Rock Constraints on Ground Motion from Historic and Recent Earthquakes in Southern California. Bulletin of the Seismological Society of America, 92(7), 2602–2611. https://doi.org/10.1785/0120000606

Brunner, E., & Munzel, U. (2000). The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. Biometrical Journal, 42(1), 17–25.

Campbell, K. W., & Bozorgnia, Y. (2014). NGA-West2 Ground Motion Model for the Average Horizontal Components of PGA, PGV, and 5% Damped Linear Acceleration Response Spectra. Earthquake Spectra, 30(3), 1087–1115. https://doi.org/10.1193/062913EQS175M

Cauzzi, C., Faccioli, E., Vanini, M., & Bianchini, A. (2015). Updated predictive equations for broadband (0.01–10 s) horizontal response spectra and peak ground motions, based on a global dataset of digital acceleration records. Bulletin of Earthquake Engineering, 13(6), 1587–1612. https://doi.org/10.1007/s10518-014-9685-y

Chiou, B. S.-J., & Youngs, R. R. (2014). Update of the Chiou and Youngs NGA Model for the Average Horizontal Component of Peak Ground Motion and Response Spectra. Earthquake Spectra, 30(3), 1117–1153. https://doi.org/10.1193/072813EQS219M

Committee on the Safety of Nuclear Installations (CSNI), Nuclear Energy Agency. (2015). Workshop on Testing Probabilistic Seismic Hazard Analysis Results and the Benefits of Bayesian Techniques.

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

Danciu, L., Nandan, S., Reyes, C., Basili, R., Weatherill, G., Beauval, C., Rovida, A., Vilanova, S., Sesetyan, K., Bard, P. Y., Cotton, F., Wiemer, S., & Giardini, D. (2021). The 2020 update of the European Seismic Hazard Models—ESHM20: Model Overview [EFEHR Technical Report 001 v1.0.0]. https://doi.org/10.12686/a15

Drouet, S., Ameri, G., Le Dortz, K., Secanell, R., & Senfaute, G. (2020). A probabilistic seismic hazard map for the metropolitan France. Bulletin of Earthquake Engineering, 18(5), 1865–1898. https://doi.org/10.1007/s10518-020-00790-7

Esposito, S., & Iervolino, I. (2012). Spatial Correlation of Spectral Acceleration in European Data. Bulletin of the Seismological Society of America, 102(6), 2781–2788.

Esteva, L. (1968). Regionalizacíon sísmica de México para fines de Ingenhería [PhD Thesis]. School of Engineering, National Autonomous University of Mexico (UNAM).

Field, E. H., Jordan, T. H., & Cornell, C. A. (2003). OpenSHA: A Developeing Community-modeling Environment for Seismic Hazard Analysis. Seismological Research Letters, 74(4), 406–419.

Fujiwara, H., Morikawa, N., Ishikawa, Y., Okumura, T., Miyakoshi, J., Nojima, N., & Fukushima, Y. (2009). Statistical Comparison of National Probabilistic Seismic Hazard Maps and Frequency of Recorded JMA Seismic Intensities from the K-NET Strong-motion Observation Network in Japan during 1997-2006. Seismological Research Letters, 80(3), 458–464. https://doi.org/10.1785/gssrl.80.3.458

Goda, K., & Hong, H. P. (2008). Spatial Correlation of Peak Ground Motions and Response Spectra. Bulletin of the Seismological Society of America, 98(1), 354–365. https://doi.org/10.1785/0120070078

Grünthal, G., Stromeyer, D., Bosse, C., Cotton, F., & Bindi, D. (2018). The probabilistic seismic hazard assessment of Germany—Version 2016, considering the range of epistemic uncertainties and aleatory variability. Bulletin of Earthquake Engineering, 16(10), 4339–4395. https://doi.org/10.1007/s10518-018-0315-y

Grünthal, G., & Wahlström, R. (2012). The European-Mediterranean Earthquake Catalogue (EMEC) for the last millennium. Journal of Seismology, 16(3), 535–570. https://doi.org/10.1007/s10950-012-9302-y

Hale, C., Abrahamson, N., & Bozorgnia, Y. (2018). Probabilistic Seismic Hazard Analysis Code Verification (PEER Report No. 2018/03; pp. 1–139). Pacific Earthquake Engineering Research Center.

Hong, Y. (2013). On computing the distribution function for the Poisson Binomial Distribution. Computational Statistics and Data Analysis, 59(C), 41–51.

Iervolino, I., Chioccarelli, E., & Cito, P. (2023). Testing three seismic hazard models for Italy via multi-site observations. PLOS ONE, 18(4), e0284909. https://doi.org/10.1371/journal.pone.0284909

Iervolino, I., Giorgio, M., & Cito, P. (2017). The effect of spatial dependence on hazard validation. Geophysical Journal International, 209(3), 1363–1368. https://doi.org/10.1093/gji/ggx090

Iervolino, I., Vitale, A., & Cito, P. (2021). Empirical assessment of seismic design hazard's exceedance area. Scientific Reports, 11(1), 18803. https://doi.org/10.1038/s41598-021-98388-9

Jayaram, N., & Baker, J. W. (2009). Correlation model for spatially distributed ground-motion intensities. Earthquake Engineering & Structural Dynamics, 38(15), 1687–1708. https://doi.org/10.1002/eqe.922

Kotha, S. R., Weatherill, G., Bindi, D., & Cotton, F. (2020). A regionally-adaptable ground-motion model for shallow crustal earthquakes in Europe. Bulletin of Earthquake Engineering, 18(9), 4091–4125. https://doi.org/10.1007/s10518-020-00869-1

Kotha, S. R., Weatherill, G., Bindi, D., & Cotton, F. (2022). Near-source magnitude scaling of spectral accelerations: Analysis and update of Kotha et al. (2020) model. Bulletin of Earthquake Engineering, 20(3), 1343–1370. https://doi.org/10.1007/s10518-021-01308-5

Kuehn, N. (2023). A comparison of nonergodic ground-motion models based on geographically weighted regression and the integrated nested laplace approximation. Bulletin of Earthquake Engineering, 21(1), 27–52. https://doi.org/10.1007/s10518-022-01443-7

Landwehr, N., Kuehn, N. M., Scheffer, T., & Abrahamson, N. (2016). A Nonergodic Ground-Motion Model for California with Spatially Varying Coefficients. Bulletin of the Seismological Society of America, 106(6), 2574–2583. https://doi.org/10.1785/0120160118

Lanzano, G., Sgobba, S., Caramenti, L., & Menafoglio, A. (2021). Ground-Motion Model for Crustal Events in Italy by Applying the Multisource Geographically Weighted Regression (MS-GWR) Method. Bulletin of the Seismological Society of America, 111(6), 3297–3313. https://doi.org/10.1785/0120210044

Lanzano, G., Sgobba, S., Luzi, L., Puglia, R., Pacor, F., Felicetta, C., D'Amico, M., Cotton, F., & Bindi, D. (2019). The pan-European Engineering Strong Motion (ESM) flatfile: Compilation criteria and data statistics. Bulletin of Earthquake Engineering, 17(2), 561–582. https://doi.org/10.1007/s10518-018-0480-z

Mak, S., Clements, R. A., & Schorlemmer, D. (2014). The Statistical Power of Testing Probabilistic Seismic-Hazard Assessments. Seismological Research Letters, 85(4), 781–783. https://doi.org/10.1785/0220140012

Mak, S., & Schorlemmer, D. (2016). A Comparison between the Forecast by the United States National Seismic Hazard Maps with Recent Ground-Motion Records. Bulletin of the Seismological Society of America, 106(4), 1817–1831. https://doi.org/10.1785/0120150323

Marzocchi, W., & Jordan, T. H. (2014). Testing for ontological errors in probabilistic forecasting models of natural systems. Proceedings of the National Academy of Sciences, 111(33), 11973–11978. https://doi.org/10.1073/pnas.1410183111

Marzocchi, W., & Jordan, T. H. (2018). Experimental concepts for testing probabilistic earthquake forecasting and seismic hazard models. Geophysical Journal International, 215(2), 780–798. https://doi.org/10.1093/gji/ggy276

McGuire, R. K. (1976). FORTRAN computer program for seismic risk analysis (pp. 1–90) [U. S. Geological Survey Open File Report 76-67].

Meletti, C., Marzocchi, W., D'Amico, V., Lanzano, G., Luzi, L., Martinelli, F., Pace, B., Rovida, A., Taroni, M., Visini, F., & Group, M. W. (2021). The new Italian seismic hazard model (MPS19). Annals of Geophysics, 64(1), 6. https://doi.org/10.4401/ag-8579

Mezcua, J., Rueda, J., & Garcia Blanco, R. M. (2013). Observed and Calculated Intensities as a Test of a Probabilistic Seismic-Hazard Analysis of Spain. Seismological Research Letters, 84(5), 772–780. https://doi.org/10.1785/0220130020

Milner, K. R., Shaw, B. E., Goulet, C. A., Richards-Dinger, K. B., Callaghan, S., Jordan, T. H., Dieterich, J. H., & Field, E. H. (2021). Toward Physics-Based Nonergodic PSHA: A Prototype Fully Deterministic Seismic Hazard Model for Southern California. Bulletin of the Seismological Society of America, 111(2), 898–915. https://doi.org/10.1785/0120200216

Mosca, I., Sargeant, S., Baptie, B., Musson, R. M. W., & Pharaoh, T. C. (2022). The 2020 national seismic hazard model for the United Kingdom. Bulletin of Earthquake Engineering, 20(2), 633–675. https://doi.org/10.1007/s10518-021-01281-z

Ordaz, M., & Reyes, C. (1999). Earthquake Hazard in Mexico City: Observations versus Computations. Bulletin of the Seismological Society of America, 89(5), 1379–1383.

Pagani, M., Monelli, D., Weatherill, G., Danciu, L., Crowley, H., Silva, V., Henshaw, P., Butler, L., Nastasi, M., Panzeri, L., Simionato, M., & Vigano, D. (2014). OpenQuake Engine: An Open Hazard (and Risk) Software for the Global Earthquake Model. Seismological Research Letters, 85(3), 692–702. https://doi.org/10.1785/0220130087

Rey, J., Beauval, C., & Douglas, J. (2018). Do French macroseismic intensity observations agree with expectations from the European Seismic Hazard Model 2013? Journal of Seismology, 22(3), 589–604. https://doi.org/10.1007/s10950-017-9724-7

Rota, M., & Rosti, A. (2017). Comparison of PSH results with historical macroseismic observations at different scales. Part 1: Methodology. Bulletin of Earthquake Engineering, 15(11), 4585–4607. https://doi.org/10.1007/s10518-017-0157-z

Savran, W., Werner, M., Schorlemmer, D., & Maechling, P. (2022). pyCSEP: A Python Toolkit For Earthquake Forecast Developers. Journal of Open Source Software, 7(69), 3658. https://doi.org/10.21105/joss.03658

Schorlemmer, D., Werner, M. J., Marzocchi, W., Jordan, T. H., Ogata, Y., Jackson, D. D., Mak, S., Rhoades, D. A., Gerstenberger, M. C., Hirata, N., Liukis, M., Maechling, P. J., Strader, A., Taroni, M., Wiemer, S., Zechar, J. D., & Zhuang, J. (2018). The Collaboratory for the Study of Earthquake Predictability: Achievements and Priorities. Seismological Research Letters, 89(4), 1305–1313. https://doi.org/10.1785/0220180053

Stein, S., Geller, R., & Liu, M. (2011). Bad Assumptions of Bad Luck: Why Earthquake Hazard Maps Need Objective Testing. Seismological Research Letters, 82(5), 623–626.

Stirling, M., & Gerstenberger, M. (2010). Ground Motion-Based Testing of Seismic Hazard Models in New Zealand. Bulletin of the Seismological Society of America, 100(4), 1407–1414. https://doi.org/10.1785/0120090336

Stirling, M., & Petersen, M. (2006). Comparison of the Historical Record of Earthquake Hazard with Seismic- Hazard Models for New Zealand and the Continental United States. Bulletin of the Seismological Society of America, 96(6), 1978–1994. https://doi.org/10.1785/0120050176

Stirling, M. W. (2012). Earthquake Hazard Maps and Objective Testing: The Hazard Mapper's Point of View. Seismological Research Letters, 83(2), 231–232. https://doi.org/10.1785/gssrl.83.2.231

Stirling, M. W., Gerstenberger, M. C., Manea, E. F., & Bora, S. (2022). Testing and Evaluation of the New Zealand National Seismic Hazard Model (University of Otago, Dunedin (NZ) Technical Report).

Sung, C.-H., Abrahamson, N. A., Kuehn, N. M., Traversa, P., & Zentner, I. (2022). A non-ergodic ground-motion model of Fourier amplitude spectra for France. Bulletin of Earthquake Engineering. https://doi.org/10.1007/s10518-022-01403-1

Tasan, H., Beauval, C., Helmstetter, A., Sandikkaya, A., & Guéguen, P. (2014). Testing probabilistic seismic hazard estimates against accelerometric data in two countries: France and Turkey. Geophysical Journal International, 198(3), 1554–1571. https://doi.org/10.1093/gji/ggu191

Thomas, P., Wong, I., & Abrahamson, N. (2010). Verification of Probabilistic Seismic Hazard Analysis Computer Programs (pp. 1–176) [PEER Report No. 2010/06]. Pacific Earthquake Engineering Research Center.

Wald, D. J., & Allen, T. I. (2007). Topographic Slope as a Proxy for Seismic Site Conditions and Amplification. Bulletin of the Seismological Society of America, 97(5), 1379–1395. https://doi.org/10.1785/0120060267

Wang, Y. H. (1993). On the number of successes in independent trials. Statistica Sinica, 3, 295–312.

Ward, S. N. (1995). Area-Based Tests of Long-Term Seismic Hazard Predictions. Bulletin of the Seismological Society of America, 85(5), 1285–1298.

Weatherill, G. A., Pagani, M., & Garcia, J. (2016). Exploring earthquake databases for the creation of magnitude-homogeneous catalogues: Tools for application on a regional and global scale. Geophysical Journal International, 206(3), 1652–1676. https://doi.org/10.1093/gji/ggw232

Weatherill, G., Cotton, F., Daniel, G., & Zentner, I. (2023a). Implementation of the Drouet et al. (2020) PSHA for France in OpenQuake: Comparisons and Modelling Issues (Project Deliverable No. SIGMA2-XXXX-YY-ZZ; SIGMA 2: Research and Development Program on Seismic Ground Motion). EDF.

Weatherill, G., Cotton, F., Daniel, G., Zentner, I., Iturrieta, P., & Bosse, C. (2023b). Strategies for Comparison of Probabilistic Seismic Hazard Models and Insights from Application in the Germany and France Border Region. Natural Hazards and Earth System Sciences, submitted.

Weatherill, G., Crowley, H., Roullé, A., Tourlière, B., Lemoine, A., Gracianne, C., Kotha, S. R., & Cotton, F. (2023c). Modelling site response at regional scale for the 2020 European Seismic Risk Model (ESRM20). Bulletin of Earthquake Engineering, 21(2), 665–714. https://doi.org/10.1007/s10518-022-01526-5

Weatherill, G., Kotha, S. R., & Cotton, F. (2020). A regionally-adaptable "scaled backbone" ground motion logic tree for shallow seismicity in Europe: Application to the 2020 European seismic hazard model. Bulletin of Earthquake Engineering, 18(11), 5087–5117. https://doi.org/10.1007/s10518-020-00899-9

Wiemer, S., Danciu, L., Edwards, B., Marti, M., Fäh, D., Hiemer, S., Wössner, J., Cauzzi, C., Kästli, P., & Kremer, K. (2016). Seismic Hazard Model (2015) for Switzerlan (SUIhaz2015) (p. 164). Swiss Seismological Service, ETH Zurich. 10.12686/a2

Zaccarelli, R., Bindi, D., Strollo, A., Quinteros, J., & Cotton, F. (2019). Stream2segment: An Open-Source Tool for Downloading, Processing, and Visualizing Massive Event-Based Seismic Waveform Datasets. Seismological Research Letters. https://doi.org/10.1785/0220180314