An Open Model for Probabilistic Seismic Hazard Assessment on the Indian Subcontinent

N. Ackerley[1,2]

[1]Istituto Universitario di Studi Superiori, Pavia, Italy
[2]Université Joseph Fourier, Grenoble, France

March 31, 2016

Abstract

Open models enable peer review and collaboration; open models can be combined and improved upon. This report details the implementation and verification of the PSHA of Nath and Thingbaijam (2012) for the Indian subcontinent within the OpenQuake framework. The electronic supplement of Nath and Thingbaijam (2012) does not completely describe the models. For example it was necessary to infer the correct mapping of tectonic region types to areal zones from geological maps and cited references, and smoothed-gridded models had to be interpreted as total seismicity rather than seismicity rates. The logic tree describing epistemic uncertainty in earthquake frequency-magnitude distributions (FMDs) of the areal zones was found to be intractably complex if taken literally. A method was developed to "collapse" FMD uncertainty which gives exact results for the mean hazard and reduces the computational complexity of full enumeration of the logic tree by approximately one million googols. Recommendations are made for future work regarding improvements to the GMPE logic tree and source modelling. Implementation of these improvements is beyond the scope of this project, but it is hoped that NRML input and output model files published at the OpenQuake hazard wiki will provide a sound basis for future work. Hazard has been verified within ±26% for cities with high hazard and within ±0.09 g for the rest. Pending verification of the full hazard map, this model of seismic hazard on the Indian subcontinent is ready for incorporation into the GEM Hazard Input Models Database.

Contents

List of Figures

List of Tables

1 Introduction

In this study the seismic hazard model for peninsular India proposed by Nath and Thingbaijam (2012) is implemented within the OpenQuake (Pagani et al., 2014; Crowley et al., 2015) platform.

This report is intended to be archived with the input and output files necessary to replicate the results at `https://hazardwiki.openquake.org/`. References to file names in the electronic data are shown in `typewriter` font, as are keywords specific to OpenQuake, such as `bGRRelative`.

1.1 Seismic hazard in the Indian subcontinent

The study of seismic hazard in India has been progressing steadily, from deterministic studies (Bureau of Indian Standards, 2002) to probabilistic seismic hazard assessment (PSHA) and from site-specific towards larger regional studies. Ashish et al. (2016) gives an up-to-date overview of the importance and history of this work. Of particular note is the fact that the Bureau of Indian Standards has not updated their seismic hazard zonation since 2002 (Bureau of Indian Standards, 2002). Nath and Thingbaijam (2012) summarize concerns with this standard (currently in force), including underestimation of hazard, application of single zone factor to regions with very different hazard, and lack of treatment of uncertainty.

Some studies have focused on the extreme hazard of the Himalayas (Bilham et al., 2001) in the north-east, including the Shillong plateau, (Das et al., 2006) and north-west (Mahajan et al., 2009). Other studies have focused on regions of lesser but nonetheless high hazard such as Gujarat (Yadav et al., 2008) or considered the whole of stable "peninsular India" (Jaiswal and Sinha, 2007; Ashish et al., 2016). Only Bhatia et al. (1999) considered the whole of India, but as Ashish et al. (2016) points out, since it was part of a global hazard mapping project (GSHAP) it only included "only a few sources for Peninsular India focusing on the inter-plate region along the Himalayan belt".

Nath and Thingbaijam (2012) is thus distinguished from previous work in providing a detailed probabilistic hazard assessment for the whole of India, including neighbouring states such as Bangladesh and Nepal. It is the culmination of several previous works, some unpublished, involving the same group of authors. These works include development of a uniform catalogue (Nath et al., 2010), development of ground-motion prediction equations (GMPEs) specific to the Shillong region (Nath et al., 2012), evaluation of a suite of GMPEs applicable to India (Nath and Thingbaijam, 2011) and development of smoothed-gridded and areal seismicity models (Thingbaijam and Nath, 2011). Although there are inevitably some limitations, as we shall see later, this work represents the current state-of-the-art as far as PSHA in the Indian subcontinent.

1.2 Open science and OpenQuake

The seismological research community is a collegial one: researchers generally share data, models, software and results freely. However it is becoming generally recognized that scientific computation is falling short of expectations in terms of reproducibility (Fomel and Claerbout, 2009; Donoho et al., 2009). As computing power grows, so do models and their complexity; it seems that our ability to describe these models is not keeping pace. In other disciplines, the components of a properly-documented experiment are well-known and widely practised. Scientific computing is a relative newcomer, and presents new challenges, such as the constant evolution of programming languages.

Reproducibility is one of the fundamental tenets of science. In the context of scientific computing reproducibility requires, at a minimum, a complete description of model, software versioning and results, open source code, and access to sufficient computing power (Hinsen, 2011).

Nath and Thingbaijam (2012) provides the majority of the model description and results as an electronic supplement. Unfortunately this description is incomplete, and worse, the software used to run the simulations is not freely available. The consequence is that results cannot be verified, errors cannot be corrected and improvements cannot be made.

OpenQuake (Pagani et al., 2014) is a fully-featured suite of software for the modelling of seismic hazard and risk. It is based on the OpenSHA framework (Field et al., 2003) but developed in the Python programming language. The source code is open, freely distributable and modifiable, and version-controlled at `https://github.com/gem/`. Input and output files are encoded using and XML schema called the Natural Hazard Risk Markup Language (NRML) and which is both human- and machine-readable. NRML input files are standardized and can be combined between projects; NRML hazard output files can become the input to subsequent risk analyses. OpenQuake-engine is an ideal platform for development of PSHA models. In fact, there is an ongoing effort to build a Global Earthquake Model (Pagani et al., 2014) based on OpenQuake.

1.3 Overview

In Section 2 the process followed to translate the model of Nath and Thingbaijam (2012) for OpenQuake-engine is detailed.

Section 2.1 uses the GMPE logic tree as an introduction to the tectonic subregions and associated GMPEs. Potential improvements in terms of subregion and GMPE selection are reserved for Appendix A.

The source model is described in Section 2.2. In particular issues relating to tectonic region assignments and interpretation smoothed-gridded seismicity data files are discussed. Recommendations for an improved smoothed-gridded model are made in Appendix B. Improvements to source modelling are proposed in Appendix C.

In Section 2.3 issues encountered in implementing ground motion prediction equations (GMPEs) are discussed.

The modelling of source frequency-magnitude distribution (FMD) uncertainty described in Nath and Thingbaijam (2012) turned out to be unimplementable in the strictest sense in OpenQuake and possibly on any platform, so compromises made are described in Section 2.4.

Section 3.1 verifies the current results against those of Nath and Thingbaijam (2012). In particular, for selected cities, hazard curves and tables of ground motion with various probabilities of exceedance are presented and evaluated. Inconsistencies between the figures and electronic supplement of Nath and Thingbaijam (2012) are discussed.

Section 3.2 summarizes the results and directions for future work while reserving more detailed discussion for the appendices.

Finally Appendix D gives an overview of model files and related source code.

2 Implementation

2.1 Ground motion prediction logic tree

The GMPE logic tree conveys some of the complexity of predicting earthquake hazard in peninsular India, and provides a starting point for discussion of both tectonic subregions and the GMPEs associated with them. The logic tree diagrammed in Nath and Thingbaijam (2012, Figure 3) is redrawn for clarity in Figure 1. The tectonic region names and GMPEs listed differ slightly from Nath and Thingbaijam (2012) but are exactly as found in the NRML model input files (e.g. source models mapped in Figure 2 and Figure 3) and the OpenQuake-engine source code.

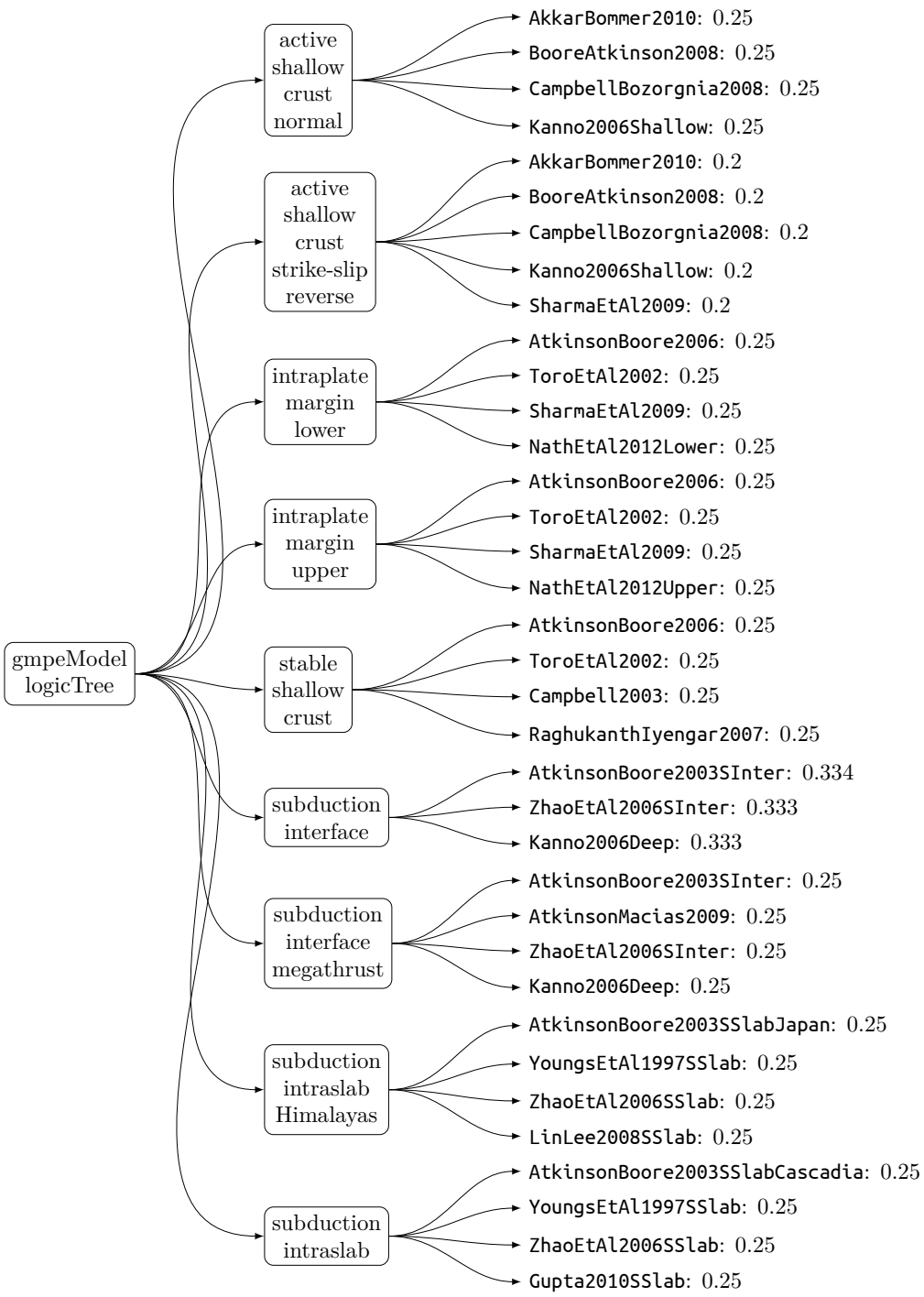

Figure 1: GMPE logic tree of Nath and Thingbaijam (2012), as encoded in `gmpe_logic_tree.xml`. Middle column selects tectonic region types mapped in Figure 2. OpenQuake GMPE class names and assigned weights are given on the right side. GMPE characteristics are are summarized in Table 3

Nath and Thingbaijam (2012, Figure 3) show Sharma et al. (2009) being used for normal faulting in the active shallow crust, but normal faulting is the only kind of faulting *not* supported by Sharma et al. (2009). This was assumed to be an error made in the drawing of the logic tree rather than the actual implementation. Thus Figure 1 shows Sharma et al. (2009) given weight only in zones where strike-slip or reverse faulting are predominant.

Although model input files may be "human-readable" no textual format of a directed graph will be easy for a human to read. This sort of error highlights the need for ways of visually diagramming logic trees directly from the model input files. A script for converting NRML logic trees to LaTeX was developed for this study.

Assignment of specific zones to tectonic subregions is treated further in Section 2.2.2. In particular methods are discussed for distinguishing between dominant fault mechanisms in the shallow crust and selecting GMPEs for intraslab subduction in deeper layers.

Delavaud et al. (2009) point out that macroseismic intensity observations are more abundant than instrumental recording, and go on to demonstrate that they can be used almost interchangeably for the purpose of quantitative assessment of GMPE efficacy. This is particularly important in areas of low seismicity or sparse instrumentation macroseismic intensity, such as India. Nath and Thingbaijam (2011) have made good use of this fact, but Nath and Thingbaijam (2012) appear to utilize their efficacy assessments imperfectly. This and other issues which could be addressed in future work in this area are addressed in Appendix A.

2.2 Seismogenic sources

The electronic supplement of Nath and Thingbaijam (2012) provides most but not all of the information required to generate a complete source model, even when supplemented by the earlier unpublished work of Thingbaijam and Nath (2011) which focuses specifically on source modelling. This section thus focuses on bridging the gaps to construct a complete source model.

Nath and Thingbaijam (2012) proposed three source models: a single set of areal seismogenic source zones, and two smoothed-gridded point source models. Combining these models using a logic tree (as discussed in Section 2.4 and diagrammed in Figure 5) allows the benefits of each model to be combined. All models are derived from the catalogue of Nath et al. (2010) for sub-catalogues with different minimum magnitudes and depth ranges.

2.2.1 Model layers

Thingbaijam and Nath (2011) divide their model into four layers as summarized in Table 1 and Figure 15. Crustal thicknesses vary significantly across the region of study, but the convenience of constant model layer thicknesses turns out to be not entirely unrealistic. The continental crust is 75-80 km thick beneath the Himalayas where the tectonics can be divided into shallow crust and interface (Thingbaijam and Nath, 2011). Similarly in the Shillong plateau of Northeast India the crust is quite thick and significant variation of stress drop with depth has been noted, with devastating "pop-up" type events (Bilham and England, 2001) being generated in the lower crust (Nath et al., 2012). In stable continental regions the crustal thickness is a more usual 35-45 km, with seismicity concentrated in the uppermost 25 km. The preceding seismotectonic features can be represented reasonably well using two seismogenic layers: 0-25 km and 25-70 km.

Intra-slab subduction occurs in three or four broad zones: the Hindu-Kush and Pamir ranges in the north-west, the eastern Himalayas and Indo-Myanmar subduction zones in the north-east and the Sumatra-Andaman subduction zone in the south east. Deep-seated seismicity only occurs in the first and last region. The tectonics of the Indo-Myanmar region are a combination of oblique subduction, accretion and collision (Wang et al., 2014). These tectonic zones are represented by two deeper seismogenic layers: 70-150 km and 150-300 km.

Table 1: Summary of layer characteristics used for source models. Completeness magnitudes and years used in generating original smoothed-gridded seismicity models are from Table 1 of Thingbaijam and Nath (2011). Layer identifiers used throughout this report are indicated. Tops and bottoms of layers have been taken as seismogenic depth limits. Hypocentral depths listed are at mid-layer.

minimum magnitude				4		4.5		5.5	
layer	depth (km)			start	end	start	end	start	end
	min.	max.	hypo.						
1	0	25	12.5	1994	2008	1964	2008	1903	2008
2	25	70	47.5	1990	2008	1964	2008	1902	2008
3	70	180	125	1996	2008	1964	2008	1914	2008
4	180	300	240	1970	2008	1984	2008	1912	2008

This stack of depth-limited seismogenic zones can crudely represent the fact that subduction events are generally spread over a dipping plane (see Figure 16). The four-layer structure furthermore captures the fact that there are 4 clear modes in the distribution of depths (see Figure 15).

2.2.2 Areal zones

Areal source models are appropriate when source mechanisms and seismicity rates are relatively uniform across a given area. They can provide a sound basis for regional assessment of b-value, maximum magnitude and other key parameters of a frequency-magnitude distribution, as shown in Thingbaijam and Nath (2011).

Selection of GMPEs (and thus the implementation of GMPE logic trees, see Section 2.1) depends on correct assignment of tectonic region types. The main difficulty in implementing the areal source model of Nath and Thingbaijam (2012) was that although the authors' intentions were generally clear, tectonic region assignments were not made explicit.

In layer 1, the shallow crust, a first distinction was made between active and stable regions according to seismicity. In active regions, assignments were made for this study using a combination of the representative focal mechanisms reported by Nath and Thingbaijam (2012) and fault maps such as the HimaTibetMap database (Styron et al., 2010). Zones obviously dominated by subduction faults were assigned "subduction interface"; for the rest the representative rake was used to distinguish between fault mechanisms, as is customary in GMPE implementations.

$$\text{mechanism} = \begin{cases} \text{reverse} & \text{if threshold} < \text{rake} < 180^\circ - \text{threshold} \\ \text{normal} & \text{if threshold} < -\text{rake} < 180^\circ - \text{threshold} \\ \text{strike-slip} & \text{otherwise} \end{cases}$$

A threshold of 30° was chosen, consistent with the OpenQuake implementations of Boore and Atkinson (2008); Campbell and Bozorgnia (2008); Sharma et al. (2009) but not Zhao et al. (2006) which uses 45°. Since the representative focal mechanism was computed as the average of the moment tensors reported in the GCMT database weighted by magnitude it is biased in favour of the larger earthquakes (Thingbaijam and Nath, 2011).

In layer 2, the deep crust, most zones are assumed to be dominated by interface subduction, except those in the stable continental part of peninsular India.

Intraslab subduction is expected to be dominant in layers 3 and 4. Nath and Thingbaijam (2011) show that the efficacy of GMPEs varies greatly between the Pamirs in the north-west and the Indo-burman subduction zone in the north-east. Unfortunately, Nath and Thingbaijam (2012) gives no hints as to how to treat intraslab subduction for sources in Andaman-Sumatra and in the eastern Himalayas. In the end it was decided to treat (Nath and Thingbaijam, 2012, Figure 3) as if "Indo-Myanmar" was intended to include the Andaman-Sumatra subduction as well. Thus one group of GMPEs is used for "subuction Himalayas" while another is used for "subduction" (i.e. everywhere else).

The lack of clear indications of how to assign various tectonic regions is a major shortcoming of the electronic supplement of Nath and Thingbaijam (2012). The tectonic region type assignments which were made for this study are summarized in Figure 2. Among these assignments, the following could be problematic:

- Zone 17 at the edge of the Pamir ranges is arguably "stable shallow crust" but was assigned "subduction interface".
- Zone 21 in the Himalayas has been assigned "subduction interface" because of the presence of the main Himalayan thrust fault. However events in this zone were included in Sharma et al. (2009) and thus could possibly be modelled better using "stable shallow crust strike-slip reverse".
- Zones 1, 2, 4, 5, 13–17, 22, 910 and 911 of layer 1 have been assigned "subduction interface" because of the presence of thrust faults but could arguably be better treated as part of "stable shallow crust strike-slip reverse" although events in these zones are not part of Sharma et al. (2009). Note that Nath and Thingbaijam (2011) only evaluated interface subduction models for events below 25 km depth.
- Zone 906 in the Great Himalayas just north of the Shillong plateau was assigned "active shallow crust strike-slip reverse" even though the main Himalayan thrust runs through it, because the representative focal mechanism is strike-slip.
- Zones 903 and 915-918 are predominantly oceanic crust but have been assigned "active shallow crust" or "subduction interface" according to the dominant focal mechanism and fault types. For example zone 903 includes the Murray Ridge and so exhibits predominantly normal faulting as expected for a spreading ridge. It is classified for the purpose of GMPE selection as "active shallow crust normal", but is likely in reality to produce ground motions distinct from an active continental crust.
- Zones 71, 86 on layer 1 and zones 9031, 9081, 9131, 9151 and 9171 on layer 2 have have a values of zero and so were assigned a "no seismicity" type and omitted from the areal source model.
- Zones 169, 170 and 172 on layer 4 capture seismicity at 180-300 km depth, but only Youngs et al. (1997) and Kanno et al. (2006) support depths below 180 km (see Table 3).

Zones numbered as "9xx" in Nath and Thingbaijam (2012) represent the amalgamation of several zones from Thingbaijam and Nath (2011). In some cases this was done because of similarity of source mechanisms and statistics while in others it was necessary because the amount of seismicity in one of the zones was insufficient for FMD characterisation. Furthermore, zones numbered "9xx1" in layer 2 have effectively had their seismicity transferred to the corresponding zones 9XX in layer 1. For example zones 32 and 115 in Thingbaijam and Nath (2011) become zones 908 and 9081 in Nath and Thingbaijam (2012), where zone 9081 has no seismicity.

Note that on layers 3 and 4 two distinct tectonic region types are defined for intraslab subduction (Nath and Thingbaijam, 2012, p. 137). Specifically, the "Indo-Myanmar and Andaman-Sumatra subduction zones" are assigned "intraslab" while the "Himalayas and northwest India-Eurasia convergence" are assigned "intraslab Himalayas". Different GMPEs are applied in these

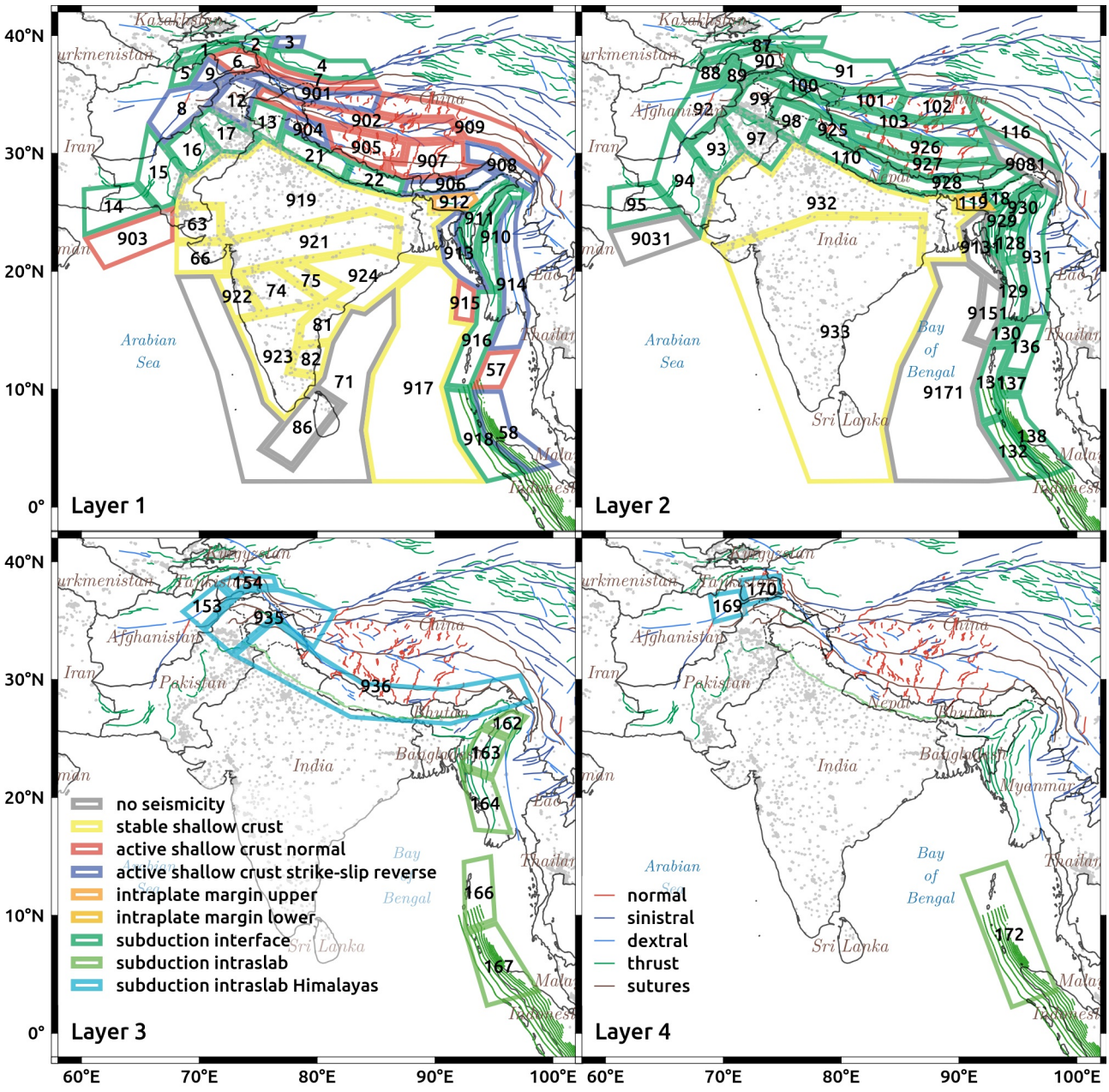

Figure 2: Areal source model tectonic region assignments used in GMPE logic tree. The areal source models is encoded in `areal_source_model.xml`. Zone identification numbers from Nath and Thingbaijam (2012) are indicated. Fault traces are from HimaTibetMap-1.0 (Styron et al., 2010) except the Sumatran subduction fault which is from SLAB 1.0 (Hayes et al., 2012). Fault data from the stable regions of India is lacking. Urban areas, "contiguous patches of built-up land greater than 1 km^2" (Schneider et al., 2009), are indicated in darker grey.

regions, as described in Section 2.1, in particular the Japan and Cascadia adjustments of Atkinson and Boore (2003) are applied, respectively.

Magnitude-scaling relations are used in PSHA to determine the actual rupture dimensions once a magnitude has been drawn from a frequency-magnitude distribution. These were relatively straightforward to select once the tectonic region assignments were made, since "Wells and Coppersmith (1994) for crustal events and those given by Strasser et al. (2010) for the subduction earthquakes" (Nath and Thingbaijam, 2012, p. 140). It was inferred that for interface and intraslab regions `StrasserInterface` and `StrasserIntraslab` should be used, respectively. The comment that "the fault-rupture area estimated from the magnitude is constrained by a factor of 2" (Nath and Thingbaijam, 2012, p. 140) was similarly interpreted as a width/depth aspect ratio of 2.

Since it is not explicitly stated in Nath and Thingbaijam (2012) the seismogenic depth was assumed to be midway between the minimum and maximum for each layer. Potential refinements to this setup are discussed in Appendix C.

The supplementary information required to generate the fully specified areal source model from the electronic supplement files `polygonlay%d.txt` and `seismicitylay%d.txt` files in the is contained in `auxiliary data.csv`.

2.2.3 Smoothed-gridded points

Smoothed-gridded seismicity models aim to replicate geographic variations of activity rates in a catalogue-driven way. Typically a smoothing kernel is used which enforces a correlation distance and limits the resolution. The electronic supplement of Nath and Thingbaijam (2012) includes smoothed-gridded values of ν for a set of latitudes and longitudes spanning the indian Subcontinent. Unfortunately neither the paper nor the data files themselves state whether ν is an annual activity rate, the total number of events, or something else.

After some discussion with K. Thingbaijam it was decided that although the models are described as "spatially varying annual activity rates" (Nath and Thingbaijam, 2012, p. 140) the electronic supplement actually contains spatially smoothed total seismicity, i.e. number of events (per cell). In order to convert this information to activity rates, i.e. number of events per year (per cell), it was necessary to obtain the duration of each sub-catalogue. Fortunately this missing ingredient is summarized in (Thingbaijam and Nath, 2011, Table 1) and reproduced in Table 1.

Given the total seismicity N and the length in years of the relevant catalogue T (see Table 1) the annual rate ν for a given model is obtained using:

$$\nu = N/T \tag{1}$$

In OpenQuake-engine a smoothed-gridded seismicity model is handled as a set of point sources with specified frequency-magnitude distributions: at a minimum a, b and $m_{\max}$ must be specified. Nath and Thingbaijam (2012) indicate that "b-value and $m_{\max}$ remain fixed within the source zone". Thus in the present study for the smoothed seismicity model the parameters b and $m_{\max}$ of the truncated Gutenberg-Richter magnitude-frequency distributions are inferred from the areal source model zonation. For points inside zones with non-zero a values in the areal source model this is trivial; for points outside these zones the zone with the shortest perpendicular distance to the point was chosen.

A gridded point source model also requires specification of tectonic region type and source mechanism for the selection and implementation of GMPEs, as well as the uncertainty in the FMDs. Thus the same procedure was used to assign tectonic subregion, rake, dip, strike, magnitude scaling relations, σ_b and $\sigma_{m_{\max}}$. For example the tectonic subregion assignments are shown for the smoothed-gridded model with $m_min = 4.5$ in Figure 3.

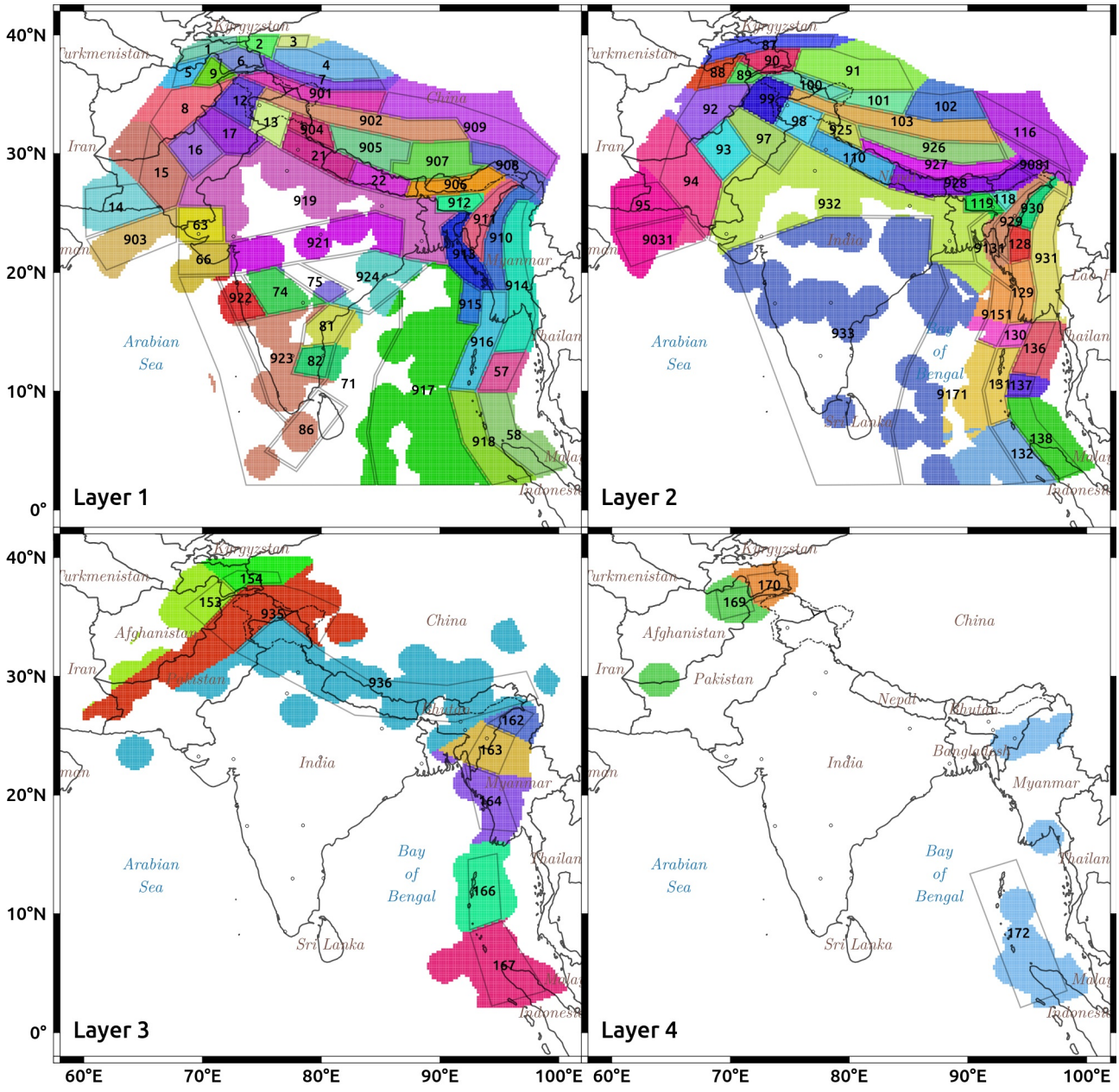

Figure 3: Areal zone associations for smoothed-gridded seismicity point source model with $m_{\min} = 4.5$. While each point source has its own activity rate, other properties of the frequency-magnitude distribution, such as b and $m_{\max}$, and the tectonic subregion used for GMPE selection are taken from the associated areal zone. Note that although the model is on a 0.1° grid, only the points on a 0.2° grid are plotted here. Nominal smoothed seismicity source models are encoded in `smoothed_source_model_mmin4.5.xml` and `smoothed_source_model_mmin5.5.xml`

The truncated Gutenberg-Richter magnitude-frequency distribution in OpenQuake-engine implements

$$\lambda(M \geq m) = 10^{a-bm} = e^{\alpha-\beta m}$$

Ignoring events below some threshold $m_{\min}$, the annual rate becomes

$$\lambda(M \geq m_{\min}) = e^{\alpha-\beta m_{\min}} e^{-\beta(m-m_{\min})} = \nu e^{-\beta(m-m_{\min})}$$

Thus to compute the a value for a point source from the activity rate ν for a given magnitude threshold, we take into account the b value for the zone as follows:

$$a = \log_{10}(\nu) + bm_{\min}$$

Similarly to compute the activity rate for an areal source we can use

$$\nu = 10^{a-bm_{\min}} \tag{2}$$

Since the data files for the smoothed seismicity models were neither clearly labeled nor described in detail in Nath and Thingbaijam (2012) there was some doubt as to their proper interpretation. In order verify the assumptions made above it was decided to attempt to reconstruct the smoothed-gridded point source model from the events labeled as mainshocks in the original catalogue of Nath et al. (2010). Some details of the smoothing are contained in the unpublished Thingbaijam and Nath (2011). The smoothing methodology of Frankel (1995) was used. The years over which the catalogue was treated as complete are given (see Table 1 and Appendix B). The smoothing kernel was Gaussian, with correlation distances of 65 and 85 km for $m_{\min}$ of 4.5 and 5.5 respectively.

Figure 4 shows the original and reconstructed gridded-smoothed seismicity models. The recomputed model was obtained using the implementation of Frankel (1995) in the OpenQuake Hazard Modeller's Toolkit (Weatherill, 2014). Isotropic Gaussian smoothing kernels were used with correlation distances as as described above. The correspondence is good but not perfect.

Discrepancies between the original and reconstructed models can arise in many ways. One possibility is different handling of catalogue events with unspecified depths (they were assumed to be in the shallowest layer) and depths below layer 4 (they were assumed to be in layer 4). Doubts are raised as to the quality of declustering in Appendix B; it may be that the events labeled as "mainshocks" in Nath et al. (2010) were not the ones used by Thingbaijam and Nath (2011). Edge effects may be important in that the geographic limits of the catalogue and smoothed-gridded model are not identical. In any case the reconstructed model shown was trimmed to contain only points specified in the original model. Finally, it may be that the completeness intervals and correlation distances given in the preiminary study of Thingbaijam and Nath (2011) were not those used in the final work of Nath and Thingbaijam (2012). The correspondence is not perfect, but good enough to validate the assumptions made as to the interpretation of the electronic supplement data.

In order to verify that the smoothed and areal models are approximately equivalent to each other and the catalogue, annual activity rates were computed for each. Areal activity rates were computed using (2). Smoothed model activity rates were computed by summing the seismicity for all points and then applying (1). Catalogue activity rates were computed by querying the catalogue of Nath et al. (2010) with appropriate minimum magnitudes and within the bounds of the areal source model. Note also that events are only counted if the epicentre is within one of the zones of the areal model. This was done on a layer-by-layer basis as well as over the whole model.

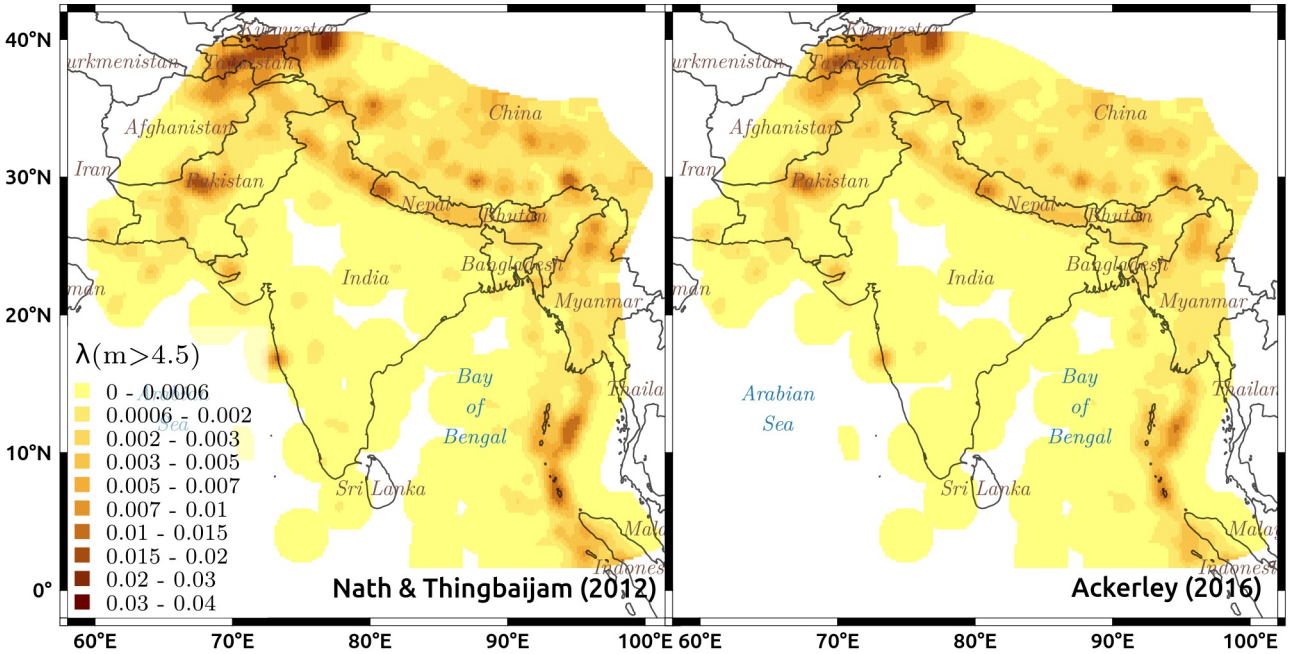

Figure 4: Smoothed-gridded annual seismicity rates for layer 1 and $m_{\min} = 4.5$. Original model of Nath and Thingbaijam (2012) is on left, while on the right is a model reconstructed from the catalogue of Nath et al. (2010).

Table 2: Comparison of annual seismicity rates for areal model, smoothed-gridded seismicity model and catalogue. In each case the value shown is the average or expected number of events per year ν above the given minimum magnitude. Catalogue events and smoothed-gridded point sources are only counted if the epicentre is within one of the zones of the areal model.

$m_{\min}$		4.5			5.5	
source	areal	smoothed	catalogue	areal	smoothed	catalogue
layer						
1	80	130	54	8.4	4.1	3.1
2	68	174	78	10.4	3.6	3.8
3	36	89	40	2.9	1.7	1.6
4	12	43	10	1.6	1.2	1.2
total	194	435	182	23.3	10.6	9.7

The results are tabulated in Table 2. Both the areal and smoothed models tend to overestimate the seismicity in the catalogue. Discrepancies between the areal model and the catalogue are likely an artefact of taking the total seismicity for a given zone, computing a frequency-magnitude distribution, and applying that FMD uniformly over the zone. Discrepancies between the smoothed model and the catalogue cannot be explained by the smearing effect of the smoothing kernel, because this should result in smoothed seismicity rates lower than the catalogue rates when computed over the same area, whereas we observe smoothed seismicity rates which are higher.

Improvements to the smoothed seismicity model are proposed in Appendix C.

Other issues of note:

- Zones 9031, 9081, 9131, 9151 and 9171 on layer 2 have $m_{\max}$ values values of zero. These zones all the smoothed seismicity points in or nearest to these zones on layer 2 were assigned the $m_{\max}$ values from the corresponding zones on layer 1, namely zones 903, 908, 913, 915 and 917.
- Given that the Japan/Cascadia regional adjustments are used for intraslab subduction, it is not clear why they are not also applicable for interface subduction.
- Although the hazard maps in the electronic supplement are at 0.2° and the paper says the smoothed-gridded models are also at 0.2° they are in fact at 0.1°. Figure 3 shows the model at just 0.2° for convenience.

2.3 Ground-motion prediction equations

In order to evaluate seismic hazard across the Indian subcontinent it is necessary to consider a wide range of tectonic and crustal propagation regimes, including some which may be unique to the region. In all, Nath and Thingbaijam (2012) uses 21 GMPEs from 17 references, summarized in Table 3.

Nine of these GMPEs were new to OpenQuake-engine. These were implemented following the quality assurance procedures described in Pagani et al. (2014). In this section the focus is on the necessity and appropriateness of these new GMPEs, as well as concerns which arose during their implementation.

In the process of the development of a GMPE specific to the Indian Himalayas, an area of rapid urbanization and elevated hazard (Sharma et al., 2009) excluded shallow India-Bangladesh and deep India-Burma border events from their database on the basis that PGA has different distance scaling. This observation points to the necessity of different GMPEs for these types of events, in the former case a GMPE specific to the Shillong plateau (Nath et al., 2012) and in the latter case, one specific to Indoburman subduction (Gupta, 2010).

The Shillong plateau (zones 118 and 912 in Figure 2) is an example of a tectonic regime specific to India. Situated in between the Himalayan and Indoburman subduction zones it would be considered a stable crustal region were it not for the massive normal faulting events known to occur there. The great Assam earthquake of 1897 destroyed buildings within several hundred km. The two main structures involved, the Dauki and Oldham faults, are capable of M $>$ 8 plateau-building events with a recurrence interval of 3-8 kyr each (Bilham and England, 2001). Nath et al. (2012) notes stress drop apparently increasing with depth and models κ using a database of recent and minor but well-recorded earthquakes, and uses this information to develop stochastic models for events in the upper and lower crust. The simulations are of vertical rather than horizontal motion at a hard-rock site.

The GMPE of Sharma et al. (2009) is intended for the Indian Himalayas but is based on data from both Zagros Mountains in Iran and the Himalayas. The database included only a small number of events (see Table 3), of which only a few were in the Himalayas, and none a

Table 3: Ground motion prediction equations used in this study. "N" indicates that models were newly implemented in OpenQuake for the current study. "S" indicates that the model has since been superseded by an equivalent model from the same authors. Among the databases used, "ENA" stands for eastern North America, and "NGA" stands for next generation attenuation. The tectonic region "Type" uses the following abbreviations: "active" shallow crust, "intraplate" margin, "stable" continental crust, "interface" subduction and "intraslab" subduction. N_E and N_R are the number of earthquakes and records in the database, respectively. H, M and R are the ranges of depth, magnitude and distance over which the GMPE is considered by the authors to be valid. The component "C" for which the GMPE is defined can be "H" for unspecified horizontal, "R" for random horizontal, "A" for average of horizontals, "M" for median of horizontals "G" for geometric mean of horizontals rotated into most adverse (GMRotI50) (Boore et al., 2006), "S" for peak of square root of sum of squares of horizontals or "V" for vertical.

OpenQuake class	Reference	N	S	Database	Type	N_E	N_R	H [km]		M		R [km]		C
`ToroEtAl2002`	Toro (2002)			ENA	stable	simulation				5.0	8.0		1000	A
`Campbell2003`	Campbell (2003)			ENA	stable	hybrid				5.0	8.2	0	1000	A
`AtkinsonBoore2006`	Atkinson and Boore (2006)			ENA	stable	simulation		2	30	5.0	8.3		1000	H
`RaghukanthIyengar2007`	Raghukanth and Iyengar (2007)	✓		peninsular India	stable	simulation		5	15	4.0	8.0		300	A
`BooreAtkinson2008`	Boore and Atkinson (2008)		✓	NGA-West1	active	58	1574			5.0	8.0	0	200	G
`CampbellBozorgnia2008`	Campbell and Bozorgnia (2008)		✓	NGA-West1	active	72	942			4.0	8.0	0	200	G
`SharmaEtAl2009`	Sharma et al. (2009)	✓		Himalayas & Zagros	active	16	201			5.0	7.0		100	A
`AkkarBommer2010`	Akkar and Bommer (2010)		✓	Europe & Middle East	active	131	532			5.0	7.6	0	100	A
`NathEtAl2012Upper`	Nath et al. (2012)	✓		Shillong	intraplate	simulation		0	25	4.8	7.6	10	100	V
`NathEtAl2012Lower`		✓		plateau	margin			25	40	4.8	8.1			
`AtkinsonBoore2003SInter`	Atkinson and Boore (2003)			global	interface	80	1155	20	50	5.0	8.3	10	550	R
`ZhaoEtAl2006SInter`	Zhao et al. (2006)		✓	Japan	interface	269	1520	25	50	5.0	8.3		300	A
`AtkinsonMacias2009`	Atkinson and Macias (2009)			Cascadia	interface	simulation				7.5	9.0		400	R
`Kanno2006Shallow`	Kanno et al. (2006)	✓		Japan	active *or* interface	83	3769	0	30	5.5	8.2		450	S
`Kanno2006Deep`	Kanno et al. (2006)	✓		Japan	intraslab	111	8150	30	200	5.5	8.2		450	S
`YoungsEtAl1997SSlab`	Youngs et al. (1997)			global	intraslab	164	480	50	229	5.0	7.8	10	500	A
`AtkinsonBoore2003SSlabJapan` `AtkinsonBoore2003SSlabCascadia`	Atkinson and Boore (2003)	✓		global	intraslab	80	1155	50	100	5.0	8.3	30	550	R
`ZhaoEtAl2006SSlab`	Zhao et al. (2006)		✓	Japan	intraslab	269	1725	50	120	5.0	8.3		300	A
`LinLee2008SSlab`	Lin and Lee (2008)			Northeast Taiwan	intraslab	54	4823	39	161	4.1	6.7	40	600	A
`Gupta2010SSlab`	Gupta (2010)	✓		Indoburman Arc	intraslab	3	56	91	148	6.3	7.2		375	M

result of normal faulting. Compared to the other models available for active regions it is the only one based on data from India. It is furthermore a valuable addition to a logic tree in the Shillong plateau because unlike the other models available for that region, it is not based on stochastic simulation. The GMPE lacks a M^2 term and so Cotton et al. (2006) would counsel against its inclusion in a logic tree, but it is retained for lack of an alternative for the region. During implementation it was observed that it does not actually define coefficients for PGA so they were assumed to be the same as for the spectral acceleration at 0.04 s.

The GMPE of Gupta (2010) is essentially a regionalization of Atkinson and Boore (2003) for intraslab subduction. On the basis of a database of just three events, the constant term of Atkinson and Boore (2003) was recalculated, leaving the distance, depth, magnitude and site amplification terms unchanged.

The GMPE of Raghukanth and Iyengar (2007) was developed for the stable shallow crust of peninsular India using stochastic simulation. Raghukanth and Iyengar (2007) actually describe three models based on regional variations in Q, for Koyna-Warna, southern India and western-central India, plus a model for all of peninsular India obtained by sampling the regional models in proportion to their landmass. It has been assumed that Nath and Thingbaijam (2012) did not use the regional models. In implementing Raghukanth and Iyengar (2007) typographical errors were identified in the coefficient Tables 2, 3 and 5 by comparing results obtained with the smoother published curves in Figures 3 and 5. The grossest error in Table 2(b) was fixed while 3 other errors causing a maximum error of approximately 10% error were not fixed (see `http://docs.openquake.org/oq-hazardlib/master/gsim/raghukanth_iyengar_2007.html`).

Kanno et al. (2006) specifies two models, for shallow and deep events, based on data predominantly from Japan. Rather than distinguishing between seismotectonic regimes, this GMPE gives appropriate scaling relations based on depth alone. Thus "both crustal and subduction interface events fall into the category of shallow events" (p. 883) where "shallow" is defined as a "focal depth of 30 km or less" (p. 883). This flexibility allows the GMPE to be used for many tectonic regimes, although as discussed in Appendix A it should in future work be restricted to regions where it demonstrates good efficacy.

Table 3 shows that layer 4 (180-300 km) is significantly deeper than deepest events used in regression for Atkinson and Boore (2003, 100 km), Lin and Lee (2008, 161 km), Zhao et al. (2006, 120 km) and Gupta (2010, 148 km). Of the GMPEs used for interface subduction in layer 4 only Youngs et al. (1997, 229 km) includes events in the correct depth range. In future work it would be beneficial to identify GMPEs appropriate for use in the depth range of layer 4 (see Appendix A.

Oddly, Kanno et al. (2006) is specified to 200 km depth, but is only used for interface events (layer 2). Given the poor efficacy of Kanno et al. (2006) in the Hidukush/Pamirs it makes sense that it should be omitted in the Himalayas, but not in the Indo-Myanmar subduction zone.

It should be noted, finally, that Kanno et al. (2006) is defined for the "peak square root of the sum of squares of two orthogonal horizontal components in the time domain" (p. 880). Since the peak value is taken after computing the vectorial sum of the horizontals this is different from the (similarly rare) "vectorial addition" of Douglas (2003) where the sum is taken after peak values are located in the time domain. This ground motion intensity measure component is more conservative than choosing a random horizontal component or the average of the peak horizontal components, but it is less conservative than the aforementioned vectorial addition. Table 3 shows that GMPEs for many different ground motion components are mixed in Nath and Thingbaijam (2012); this practice is not unusual but if left uncorrected errors do propagate through to hazard curves, and the resulting aleatory uncertainty is under-estimated (Beyer and Bommer, 2006). OpenQuake tracks the type of horizontal component measured in its base `GroundShakingIntensityModel` class, but does not currently make the necessary corrections to

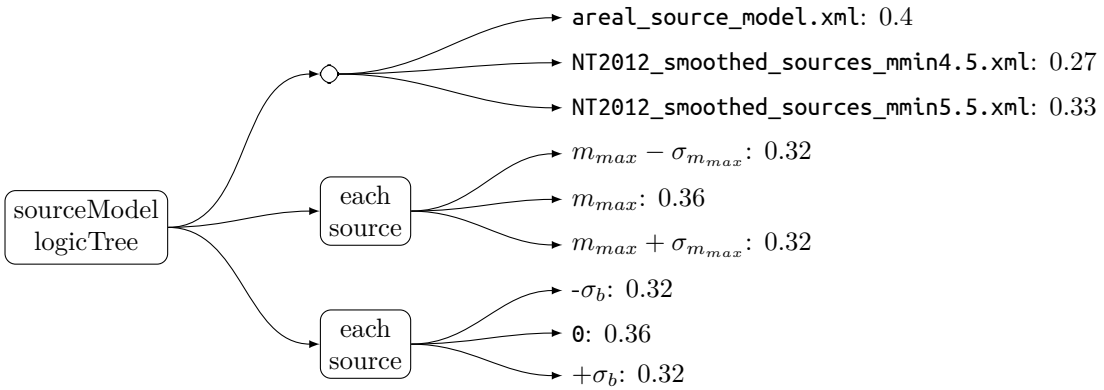

Figure 5: Symbolic source model logic tree of Nath and Thingbaijam (2012).

the mean or standard deviation of the ground motion.

2.4 Source model logic tree

Nath and Thingbaijam (2012) accounts for the epistemic uncertainty in seismicity model parameters by estimating the standard deviations of b and $m_{\max}$ in each source zone and assigning weights to ± 1 standard deviation for each source.

In reviewing the data, it was observed that `seismicitylay2.txt` listed standard deviation of b in zone 93 as 0.9, and this results in an absurd amount of variability in the FMD for this zone. The rest of the standard deviations for b vary from 0.03 to 0.16. It was decided that this must be a simple typographical error, and that the intended value must have been 0.09 instead.

The source modelling part of the logic tree of Nath and Thingbaijam (2012, Figure 4) is shown in Figure 5.

Note that although Figure 4 of Nath and Thingbaijam (2012) shows the activity rate ν (and by implication a) varying with b. No values for the standard deviation of a or ν are provided in the electronic supplement of Thingbaijam and Nath (2011), thus it was assumed that the intent is to recalculate a whenever b is varied in order to maintain constant total moment rate. This is the default behaviour in OpenQuake-engine when using the uncertainty type `bGRRelative`.

2.4.1 Full or partial enumeration

Figure 5 is the source model logic tree in *symbolic* form. If this is actually expanded to encompass all source zones the result is a source model logic tree far too large to represent on a page; just a portion of it is shown in Figure 6.

The fact that Figure 6 has to be truncated is not simply a lack of page space. Although it is common practice to diagram logic trees in parallel, full enumeration of such logic trees actually takes place in series. Thus, rather than just 3×223 branches there are in fact $3^{223} \approx 10^{106}$ (i.e. approximately one million googol) terminal branches implied in a full enumeration of Figure 6. Full enumeration of the logic tree is clearly out of the question. It may even be computationally prohibitive simply to calculate the weights on the terminal branches, a necessary precursor to partial enumeration by Monte Carlo sampling. (In any case it was not possible to complete even an analysis by partial enumeration on the current version of OpenQuake.)

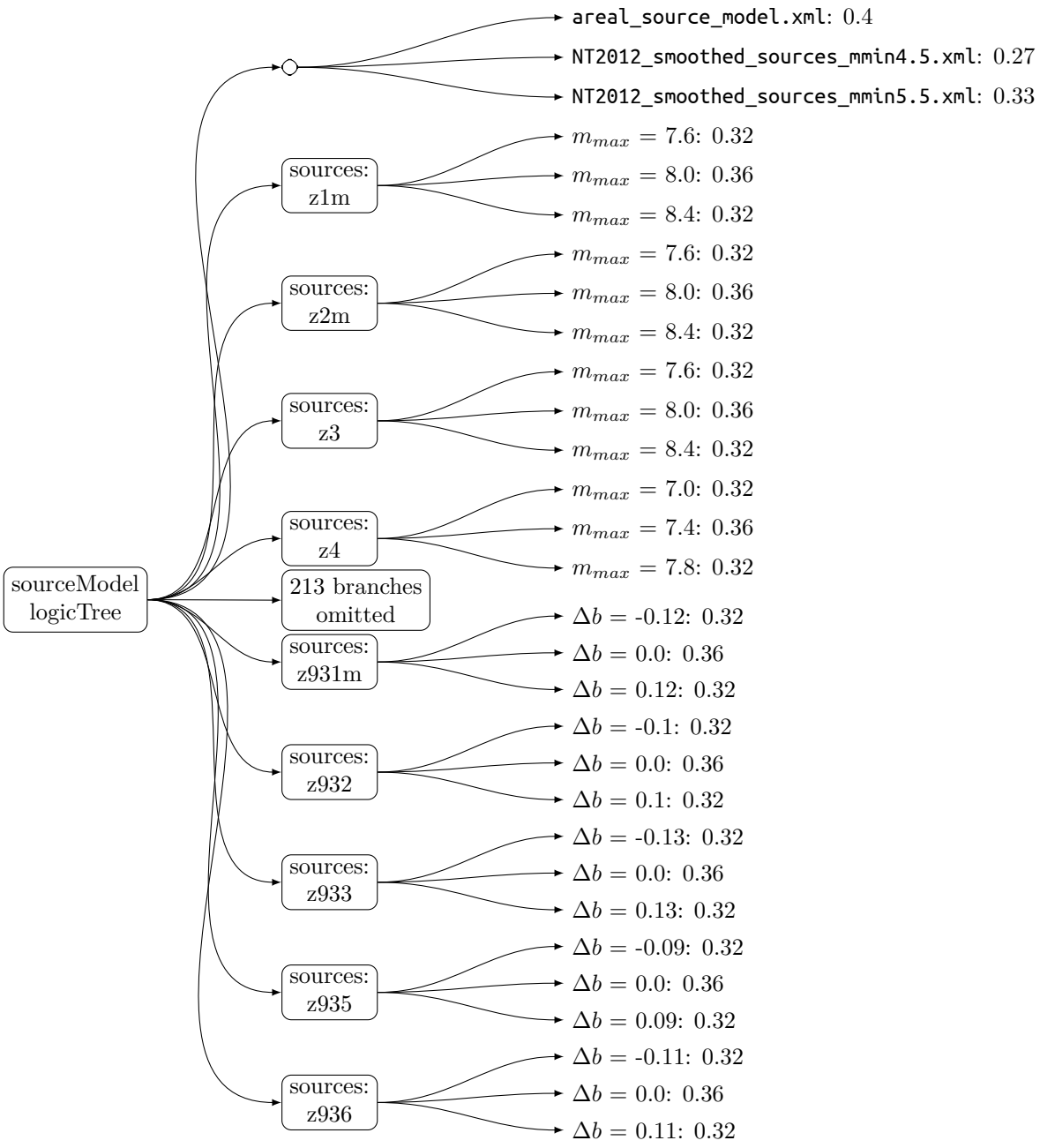

Figure 6: Partial source model logic tree of Nath and Thingbaijam (2012). The full model is encoded in `source_model_logic_tree.xml`

Given that sources are only considered when located within a certain radius of any given site (Nath and Thingbaijam, 2012, 200 km), an algorithm could be developed to find all subsets of the sites with a common subset of sources within that radius. Of the cities listed in Nath and Thingbaijam (2012, Table 3), Shillong and Imphal have the most zones within a 200 km radius, at 14, and Guwahati is next at 12. If for each site only the nearest 14 zones are considered, then the symbolic logic tree of Figure 5 expands to a more tractable $3^{29} \approx 7 \times 10^{13}$ branches. For a mapping application where many sites are located on a grid those sites could be sorted into groups of sites having common contributing source zones; each group would then have a distinct set of logic tree realizations to enumerate or be sampled from. The scheme described above is effectively a means of generating a logic tree without its geographically irrelevant branches (n.b. not "pruning" which implies branch exists before it is removed). This mode of logic tree enumeration is not currently available in OpenQuake nor is it mentioned in Nath and Thingbaijam (2012).

Note when a source model consists of a grid of point sources, it is neither possible nor advisable to try to apply full enumeration of a symbolic logic tree such as Figure 5. It is not possible because when the number of sources is on the order of 400000, as the case of Nath and Thingbaijam (2012), the number of terminal branches is not even representable as a 64-bit float. It is not advisable because if it makes sense to vary b and $m_{\max}$ at all it is because these properties vary on a sub-regional scale. For example an $m_{\max}$ exists because of the actual physical sizes of the faults, and a b value exists because of stress regimes. Using the site grouping and distance limit as described above it should be possible to fully enumerate a smoothed-gridded source model, as long as the variability b and $m_{\max}$ is applied only on groups of point sources, the most natural grouping being one devised according to an areal source model.

Given the above discussion, it is unlikely that Nath and Thingbaijam (2012) performed full or even partial enumeration of their model of source uncertainty.

Another possibility is that they applied the positive and negative deviations to the mean values of b and $m_{\max}$ for all areal and point sources simultaneously. In this case there would be a very manageable number of terminal branches, $3^2 = 27$, each with a different FMD for every zone. This possibility must be dismissed as grossly over-weighting the possibility, for example, that all sources jointly have the highest possible $m_{\max}$ and lowest possible b.

Nath and Thingbaijam (2012) only publish mean hazard curves and maps. Nowhere are median or quantile hazards discussed; effectively there is no discussion of the effect of aleatory and epistemic uncertainty in the models on the uncertainty of the results.

2.4.2 Collapse of frequency-magnitude distributions

In classical PSHA the hazard integral gives an estimate of rate at which the ground motion Y at a site of interest will exceed some value y. For a set of N_S sources generating ruptures of N_M magnitudes m_j at N_R distances r_k using discrete summations (Baker, 2008):

$$\lambda(Y > y) = \sum_{i=1}^{N_S} \lambda(M_i > m_{\min}) \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P_i(Y > y | m_j, r_k) P_i(M_i = m_j) P_i(R_i = r_k) \qquad (3)$$

where for source i:

- $\lambda_i(M_i > m_{\min})$ is the rate of earthquakes greater than $m_{\min}$
- $P_i(Y > y | m_j, r_k)$ is the ground motion prediction equation (GMPE) and incorporates aleatory uncertainty in the ground motion
- $P_i(M_i = m_j)$ is the frequency-magnitude distribution (FMD)
- $P_i(R_i = r_k)$ is the distribution of distance measures from points within the source to the site of interest.

For a single point source, neglecting finite source effects, this simplifies to:

$$\lambda(Y > y) = \lambda(M > m_{\min}) \sum_{j=1}^{N_M} P(Y > y|m_j, R) P(M = m_j)$$

Now suppose that certain epistemic uncertainties have been identified. In order to estimate the effect of this lack of knowledge on the distribution of ground motions suppose N_G GMPEs $P_\ell(Y > y|m_j, R)$ have been assigned weights w_ℓ and N_F FMDs $P_m(M = m_j)$ have been assigned weights w_m. This is accomplished via

$$\lambda(Y > y) = \lambda(M > m_{\min}) \sum_{m=1}^{N_F} w_m \sum_{\ell=1}^{N_G} w_\ell \sum_{j=1}^{N_M} P_\ell(Y > y|m_j, R) P_m(M = m_j)$$

but we can reorder the summations to obtain

$$\lambda(Y > y) = \lambda(M > m_{\min}) \sum_{\ell=1}^{N_G} w_\ell \sum_{j=1}^{N_M} P_\ell(Y > y|m_j, R) \sum_{m=1}^{N_F} w_m P_m(M = m_j)$$

By pre-computing the "collapsed" FMD for the source

$$P_C(M = m) \triangleq \sum_{m=1}^{N_F} w_m P_m(M = m)$$

we can simplify the hazard summation to

$$\lambda(Y > y) = \lambda(M > m_{\min}) \sum_{\ell=1}^{N_G} w_\ell \sum_{j=1}^{N_M} P_\ell(Y > y|m_j, R) P_C(M = m_j)$$

At this point it is critical to note that a further "collapse" of GMPE uncertainty is not possible because the ground motion is conditional upon the magnitude.

Returning to the case of multiple sources in (3), we can include ground-motion and frequency-magnitude epistemic uncertainty by writing

$$\lambda(Y > y) = \sum_{i=1}^{N_S} \lambda_i(M_i > m_{\min}) \sum_{m=1}^{N_F} w_{mi} \sum_{\ell=1}^{N_G} w_{\ell,i} \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P_{\ell i}(Y > y|m_j, r_k) P_{mi}(M_i = m_j) P_i(R_i = r_k)$$

and we can still reorder the summations, as long as we recognize that the collapsed FMD is different for each source

$$\lambda(Y > y) = \sum_{i=1}^{N_S} \lambda(M_i > m_{\min}) \sum_{\ell=1}^{N_G} w_{\ell i} \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P_{\ell i}(Y > y|m_j, r_k) P_i(R_i = r_k) \sum_{m=1}^{N_F} w_{mi} P_{mi}(M_i = m_j)$$

Thus for a rate-based formulation we can pre-compute the collapsed FMD for each source

$$P_{Ci}(M_i = m) \triangleq \sum_{m=1}^{N_F} w_{mi} P_{mi}(M_i = m) \tag{4}$$

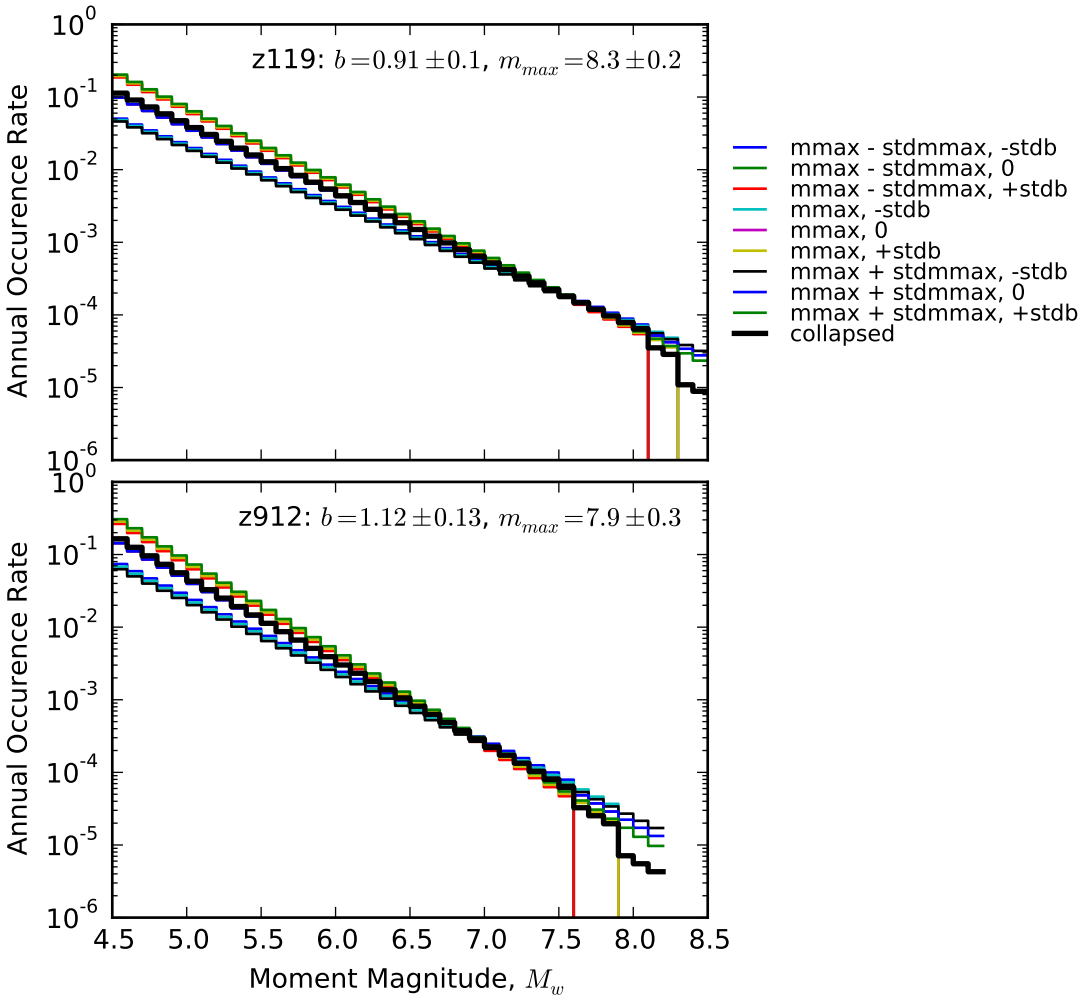

Figure 7: Frequency-magnitude distributions corresponding to various branches of the logic tree of Nath and Thingbaijam (2012). The zones chosen are the two closest to Guwahati, India. Zone 119 at 25-50 km depth is the zone capable of producing a devastating "pop-up" type event (Bilham et al., 2001; Nath et al., 2012). The "collapsed" FMD uncertainty computed using (4) is shown as a heavy black line.

And compute the mean hazard using

$$\lambda(Y > y) = \sum_{i=1}^{N_S} \lambda(M_i > m_{\min}) \sum_{\ell=1}^{N_G} w_{\ell i} \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P_{\ell i}(Y > y | m_j, r_k) P_i(R_i = r_k) P_{Ci}(M_i = m_j) \quad (5)$$

An example of the process of pre-computing collapsed FMDs as described in equation (4) is shown in Figure 7.

The preceding analysis shows that collapsed FMDs should give an exact result for the mean exceedance rate computed using a rate-based formulation. Figure 8 shows that in practice using OpenQuake, the collapsing of FMDs gives a very good approximation of the mean hazard for low probabilities of exceedance but not for high. The difference is due to the fact that OpenQuake-engine computes PSHA not using the classical rate-based formulation but using the probability-based formulation of OpenSHA (Field et al., 2003; Pagani et al., 2014). This issue will be revisited at the end of this section.

The important point here is that the "collapsed" FMD can be pre-computed for each source. In an application with N_S sources, if there are 3^2 branches to the FMD logic tree this means 3^{2N_S} branches can be eliminated in all. In a simple example with a single site, two areal sources and a moderately complex GMPE logic tree (see Figure 8) computation time on a quad-core laptop dropped from 36 minutes to 55 seconds when FMD uncertainty was collapsed. In mapping type applications with gridded sites and hundreds of areal sources, each with FMD uncertainty, the computational savings can make an intractable problem tractable.

The hazard integral (3) only estimates the mean hazard, not the median or quantiles. Aleatory variability is modelled in most if not all GMPEs using a log-normal distribution so the ability to compute an arbitrary quantile of ground-motion exceedance is built-in.

In OpenQuake, the effect of epistemic variability on quantiles is handled as follows. First, as mentioned in Section 2.4.1, terminal branches or "realizations" are enumerated, and weights assigned to each. In some cases, with large numbers of sources, each having FMD uncertainty, OpenQuake-engine founders simply in enumerating realizations. For each realization and intensity measure type and level the probabilities of exceedance are computed. Finally, these probabilities are then sorted quantiles are located with reference to the weighting scheme. As a simplistic example, if there are 3 FMD branches for a source, then whichever produces the largest probability of exceedance at a given level of ground motion will determine the probability of exceedance for any percentile from the 67^{th} through the 100^{th}. Furthermore, if there are no other logic trees, for GMPEs or source models, then in fact the 67^{th} will be the same as the 100^{th} percentile hazard curve.

With the FMD logic trees collapsed the ground motions for different realizations are no longer computed and cannot be sorted, as required to find a quantile of interest. The result is just plain wrong, farther from the true result than if no FMD uncertainty is considered at all, as shown in Figure 9. Given this understanding it is clear that "collapse" of FMDs cannot produce correct results for quantile hazard curves, or at least it cannot convey properly the part of the uncertainty due to epistemic uncertainty in the FMDs.

A slightly more general case than that given by Baker (2008) is to treat the probability of an event at a given distance from a site as dependent on magnitude. In this case for a non-point source (Field et al., 2003, equation A2)

$$\lambda(Y > y) = \lambda(M > m_{\min}) \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P(Y > y | m_j, r_k) P(R = r_k | m_j) P(M = m_j)$$

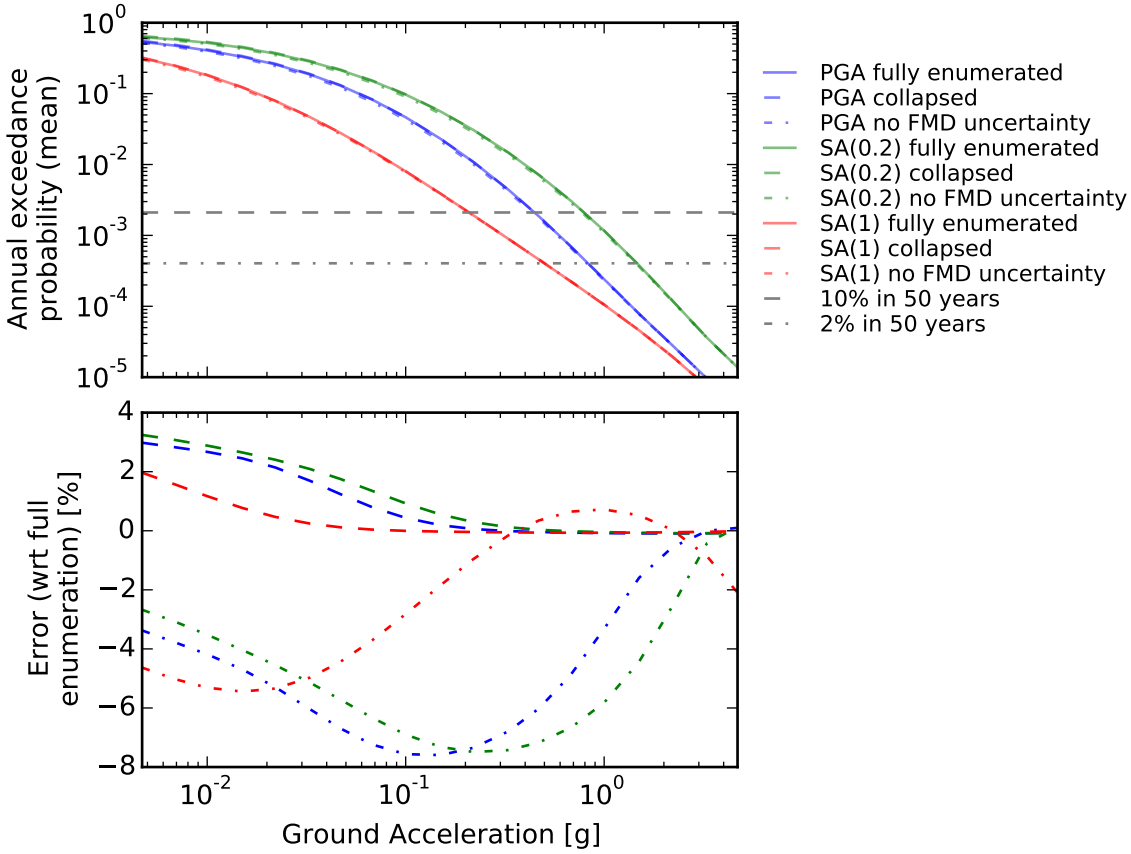

Figure 8: Mean hazard curves computed using various levels of FMD uncertainty. The site of interest is the city of Guwahati. The source model consists only of zones 119 and 912 of Nath and Thingbaijam (2012). The full GMPE logic tree of Nath and Thingbaijam (2012) is used. The "fully enumerated" result implements the FMD logic tree described in Nath and Thingbaijam (2012) while "collapsed" implements (4) and "no FMD uncertainty" models only the FMD logic tree branches with the largest weights.

To assess the effect of epistemic uncertainties on the mean we must compute

$$\lambda(Y > y) = \lambda(M > m_{\min}) \sum_{m=1}^{N_F} w_m \sum_{\ell=1}^{N_G} w_\ell \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P_\ell(Y > y | m_j, r_k) P(R = r_k | m_j) P_m(M = m_j) \quad (6)$$

but in many cases it will be reasonable to treat the distribution of site-source distances as independent of the frequency-magnitude distributions, so that the summations can still be reordered

$$\lambda(Y > y) = \lambda(M > m_{\min}) \sum_{\ell=1}^{N_G} w_\ell \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P_\ell(Y > y | m_j, r_k) P(R = r_k | m_j) \sum_{m=1}^{N_F} w_m P_m(M = m_j)$$

and the FMDs can be collapsed.

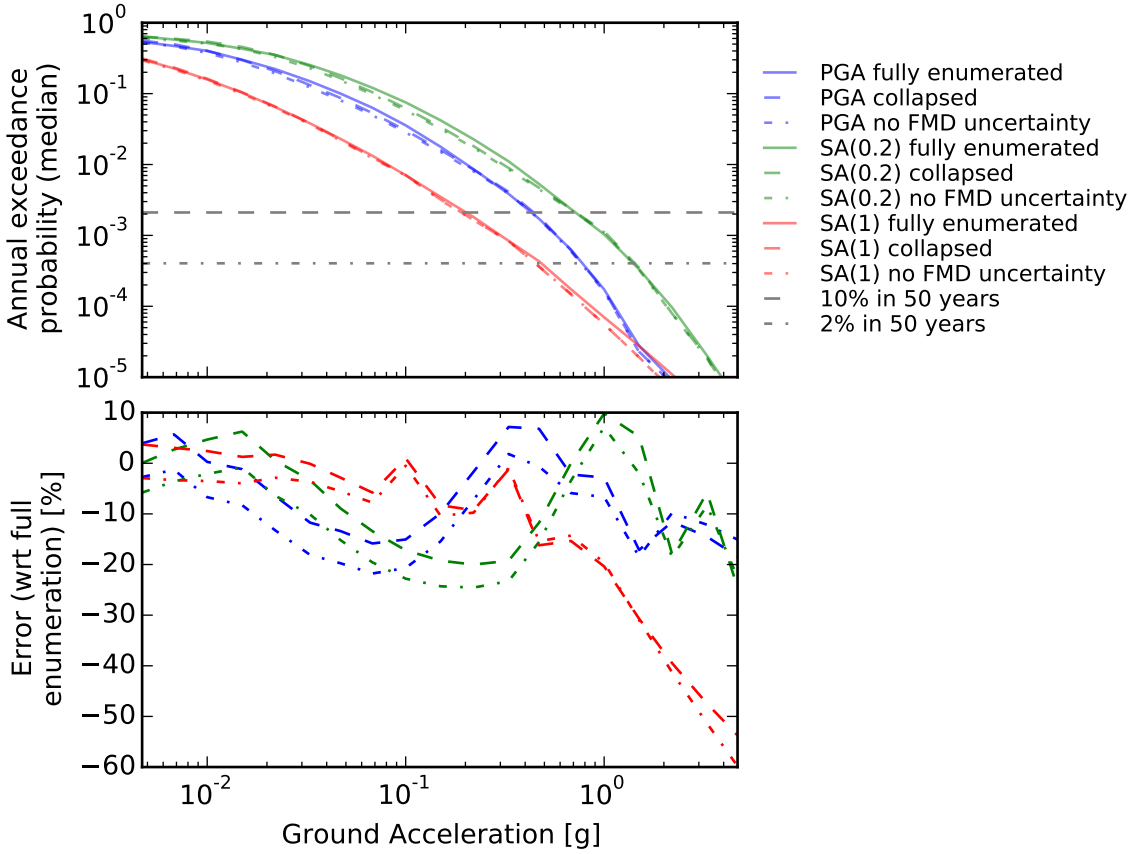

Figure 9: Median hazard curves computed using various levels of FMD uncertainty. For detailed description see Figure 8.

The hazard for each source must then be summed to compute the hazard for a given site. There may be instances where uncertainty regarding $m_{\max}$ is connected to uncertainty regarding the physical extent of a source. This sort of epistemic uncertainty would have to be handled differently, as weighted alternative source models.

Finally we return to the issue of the inaccuracy of the results obtained using OpenQuake at high probability of exceedance. The formulation of the hazard integral (3) implemented in OpenQuake-engine assumes a Poisson model for earthquake occurrence, and furthermore that the probability of two or more occurrences of each source is zero in the time interval of interest (Field et al., 2003; Pagani et al., 2014). Rather than the rate of exceedance, the probability of exceedance in a given time interval is computed. Similarly, rather than depending on the rate of events over a given minimum $\lambda(M > m_{\min})$, it is written in terms of the probability of a single occurrence of an event (with a minimum magnitude) in a given time interval $P(\text{occur}_i)$. Rewriting the formulation of Field et al. (2003, equation A9):

$$P(Y > y) = 1 - \prod_{i=1}^{N_S} \left[1 - \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P(\text{occur}_i) P(Y > y | m_{ij}, r_{ijk}) P(R = r_{ijk} | m_{ij}) P(M = m_{ij}) \right]$$

Whereas when working with rates epistemic uncertainty is treated by computing a weighted sum of predicted weights, with probabilities it is treated by computing one minus the product of the probabilities of non-occurrence.

The incorporation of epistemic uncertainty in the selection of GMPEs and FMDs is implemented in OpenQuake-engine by computing a weighted sum of probabilities (Graeme Weatherill, personal communication). Neglecting the possibility of dependence of the distance distribution on magnitude, the summation is then:

$$P(Y > y) = \sum_{m=1}^{N_F} w_m \sum_{\ell=1}^{N_G} w_\ell \left\{ 1 - \prod_{i=1}^{N_S} \left[1 - \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P(\text{occur}_i) P_\ell(Y > y | m_{ij}, r_{ijk}) P(R = r_{ijk}) P_m(M = m_{ij}) \right] \right\}$$

As long as the probability of exceedance for any given realization m, ℓ is much smaller than unity, the product can be rewritten as a summation by discarding higher-order terms. In this case we get

$$P(Y > y) = \sum_{m=1}^{N_F} w_{mi} \sum_{\ell=1}^{N_G} w_{\ell i} \sum_{i=1}^{N_S} \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P(\text{occur}_i) P_\ell(Y > y | m_{ij}, r_{ijk}) P(R = r_{ijk}) P_m(M = m_{ij})$$

and proceed as from (6) to arrive at

$$P(Y > y) = \sum_{\ell=1}^{N_G} w_\ell \sum_{i=1}^{N_S} \sum_{j=1}^{N_M} \sum_{k=1}^{N_R} P(\text{occur}_i) P_\ell(Y > y | m_{ij}, r_{ijk}) P(R = r_{ijk}) P_{Ci}(M_i = m_j)$$

where the collapsed FMD for source i, $P_{Ci}(M_i = m_j)$ is computed as for the rate-based formulation using equation (4)

In conclusion, the collapse of FMD logic trees by weighted summation of branches is perfectly safe for the classical formulation of the hazard integral with handling of epistemic uncertainty by summation of rates. When mean hazard is computed using the OpenSHA formulation of Field et al. (2003) and epistemic uncertainty is handled by summation of probabilities, for example using OpenQuake, care must be taken that the probabilities of exceedance of the individual realizations are all small. In many cases inaccuracy at low-probability of exceedance is effectively a computational artefact, a curiosity of no engineering significance. It is always possible, however, to reduce the error without computational cost by reducing the investigation time. For example, in the above example reducing the investigation time from 1 year to 1 month reduced the maximum error in the mean hazard from 3% to 0.25%.

This method gives a significant reduction in computational complexity, but can only be used to compute the mean hazard, not the median or quantiles. In many cases, such as the present one, an intractable problem become tractable. In Section 3 the hazard results presented were obtained via the collapsing of FMD uncertainty. In collapsing the FMD branches of the source model logic tree of Figure 5 these branches are effectively deleted, leaving only the branch set giving weights to the areal and smoothed source models.

3 Hazard results

Validation of PSHA results would require comparison against observed hazard. This in turn would require a catalogue many times longer than the significant return periods, something which won't be available for centuries. In this work we are merely verifying that the current model gives result close to those of Nath and Thingbaijam (2012).

Results in this section were obtained using the model files in Appendix D and OpenQuake Release 1.9.

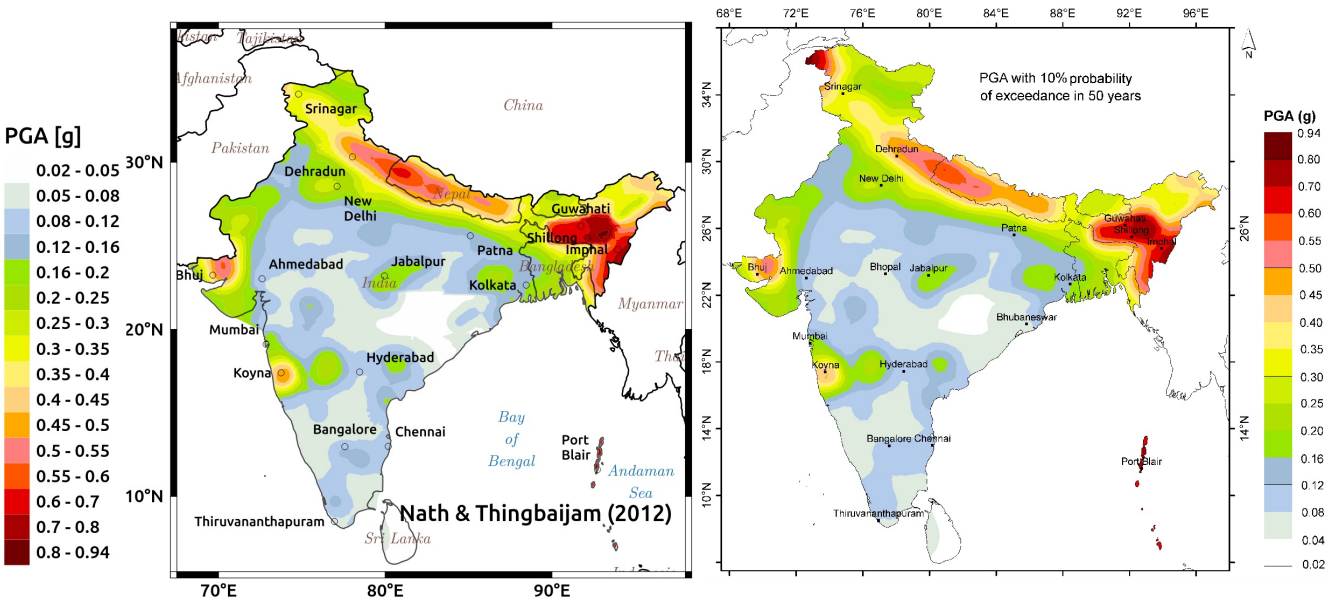

Figure 10: Comparison of original hazard maps for PGA with 10% probability of exceedance (POE) in 50 years. On the right is Figure S1 A from the electronic supplement, which is a colour version of the top left panel of Figure 7 in Nath and Thingbaijam (2012). On the left is data from the relevant column of `India_pga.csv`, contoured using the Contour plugin version 1.3.5 of the QGIS software mapping package (QGIS Development Team, 2016).

3.1 Verification

The electronic supplement of Nath and Thingbaijam (2012) contains probabilities of exceedance on an 0.2° grid of latitude and longitude for several intensity measures. Unfortunately this data does not correspond to the hazard maps shown in Figures 7 and S1 of Nath and Thingbaijam (2012), as shown in Figure 10.

There are many points of dissimilarity between the published figures and data for hazard results. One of the most confounding is the "ridge" of rapid change in PGA at about 22°N between 80°E and 85°E. The ridge does not correspond to the boundary of any areal source. Similar features are seen in the data files for all intensity measures and probabilities of exceedance. For these reasons the electronic supplement result data files are considered to be in error. Future verification of hazard should be done exclusively against the published figures.

Unfortunately, due to the size of the model and problems with the current version of the OpenQuake-engine (Release 1.9) it has not been possible to generate a full hazard map. For now we must content ourselves instead with verification of hazard at selected cities.

The current model is compared to Table 3 of Nath and Thingbaijam (2012) in Table 4. Similarly, the hazard curves in Figure 6 of Nath and Thingbaijam (2012) are compared to the current study in Figure 11.

Expressed as a percentage Table 4 shows significant difference between current results and published values. At 10% POE in 50 years there is a single site with PGA 26% too high (Sringar) and many sites with PGA up to 48% too low. However, there is a trend to larger error where the hazard was lower. It turns out the absolute error is never greater than 0.18 g, and among the cities with high hazard the error is never greater than ±26%.

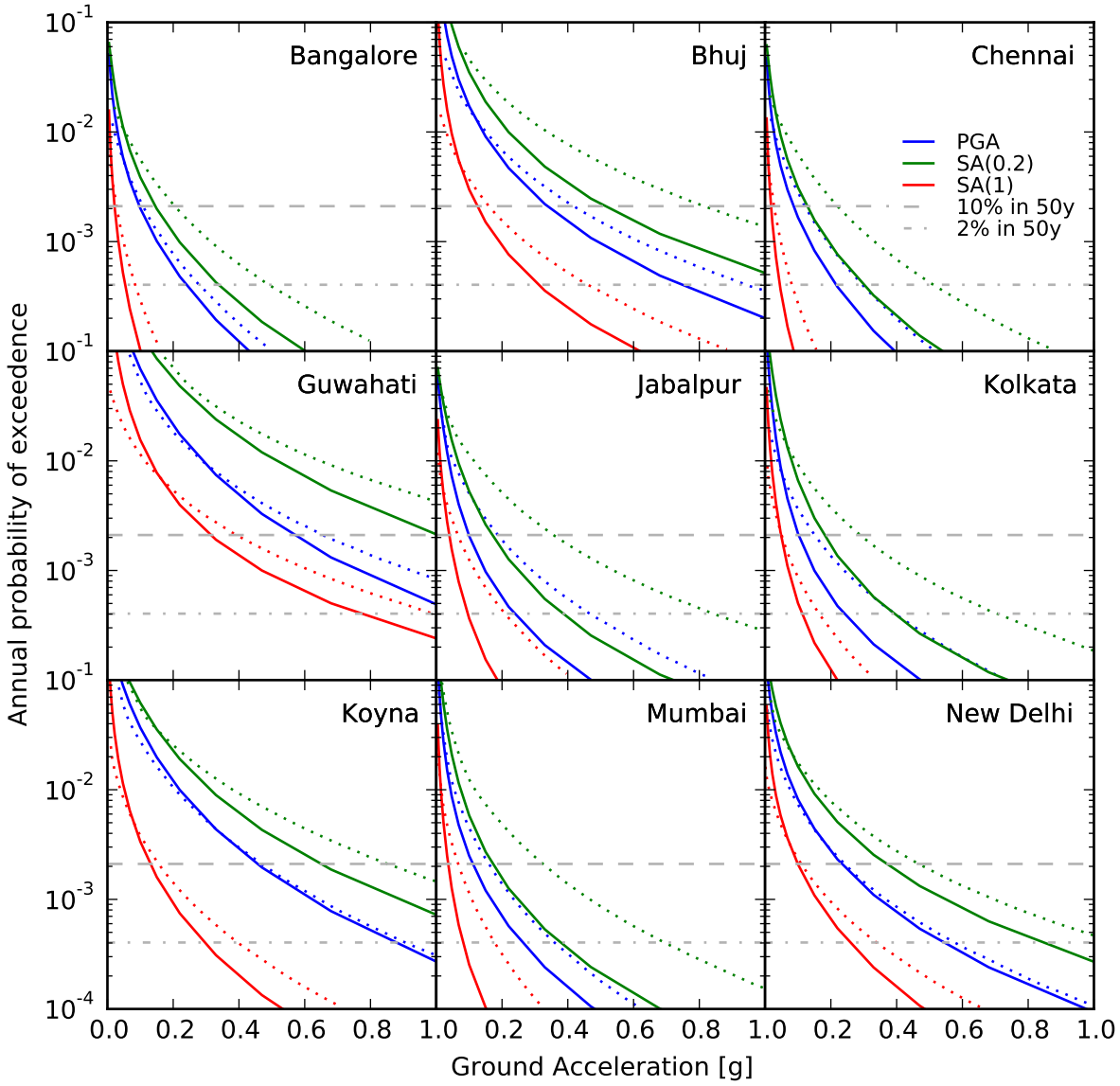

Figure 11: Mean hazard curves for selected cities. Solid lines are the results of the current study. Dashed lines results digitized from Figure 6 of Nath and Thingbaijam (2012).

Table 4: Mean PGA in selected cities. Cities are located on the map of Figure 10. 'NT2012" is the peak ground acceleration at 10% probability of exceedance in 50 years from Table 3 of Nath and Thingbaijam (2012). "A2016" is the same measure of hazard from the current study. The error is that of A2016 relative to NT2012.

City	Latitude [°N]	Longitude [°E]	NT2012 [g]	A2016 [g]	Error [g]	Error [%]
Srinagar	34.08	74.80	0.33	0.417	0.09	26
Koyna	17.40	73.75	0.47	0.455	-0.02	-3
New Delhi	28.56	77.11	0.24	0.232	-0.01	-3
Imphal	24.78	93.94	0.68	0.636	-0.04	-6
Bangalore	12.98	77.58	0.11	0.097	-0.01	-12
Guwahati	26.18	91.73	0.66	0.562	-0.10	-15
Dehradun	30.33	78.04	0.47	0.393	-0.08	-16
Port Blair	11.61	92.72	0.71	0.583	-0.13	-18
Bhuj	23.25	69.66	0.42	0.337	-0.08	-20
Shillong	25.48	92.11	0.72	0.538	-0.18	-25
Chennai	13.00	80.18	0.12	0.086	-0.03	-29
Mumbai	19.11	72.85	0.16	0.110	-0.05	-31
Kolkata	22.65	88.45	0.15	0.102	-0.05	-32
Hyderabad	17.45	78.46	0.09	0.061	-0.03	-32
Patna	25.60	85.10	0.13	0.085	-0.04	-35
Ahmedabad	23.03	72.61	0.11	0.062	-0.05	-43
Thiruvananthapuram	8.50	76.95	0.07	0.037	-0.03	-47
Jabalpur	23.20	79.95	0.19	0.099	-0.09	-48

Figure 11 shows that in the cities with highest hazard (e.g. Bhuj, Guwahati, Koyna and New Delhi) the current model gives lower hazard at low probability of exceedance and higher hazard at high probability of exceedance. In other cities, those with lower hazard, the current model gives lower hazard at all probabilities.

3.2 Discussion

Although better agreement was hoped for, agreement within ±26% in the cities with high hazard is a good result for the first revision of the new model, especially given the incomplete documentation of the reference model.

More work is needed to determine where the remaining discrepancies lie. First it will be important to plot the hazard map for the new model, and particularly the difference between the new map and the reference map. This will give a sense of exactly how the error is regionalized. For example, the remaining error may be associated with one or more zones or tectonic region types, or it may track the peaks of the smoothed seismicity model.

A hazard error map such as those used to compare 2008 and 2014 models in Petersen et al. (2014) would be quite diagnostic; without it we can really only speculate as to likely sources of discrepancies.

One possibility is that the subduction interface tectonic region types should not be used in layer 1 because Nath and Thingbaijam (2011) only evaluated interface subduction GMPEs for events below 25 km depth. It may not be correct to use GMPEs for active shallow crust where subduction faults come to the surface, but at the outset we are just trying to reproduce the

results of Nath and Thingbaijam (2012). There could in fact be many more incorrectly assigned tectonic region types; it should be possible to locate and correct these as well using a hazard error map.

Another possibility is that the smoothed seismicity model files could still be being interpreted incorrectly. In this case switching to regenerated smoothed seismicity models as discussed in Appendix B will bring us no closer to verification.

It is possible that no correspondence between hazard map errors and areal or smoothed seismicity models will be observed. The conclusion would be that the rupture forecast (i.e. distributions of rupture areas and hypocenters) generated by OpenQuake is significantly different from the closed-source software of Nath and Thingbaijam (2012). In this case there is nothing we can do to better verify the model, but at the same time there would be nothing wrong with it and it could safely be used as the basis for future work.

A final comment. Bommer and Scherbaum (2008) notes that "PSHA should always be done together with a disaggregation ... [and] ... setting up a logic tree should always be accompanied by a sensitivity study". This certainly holds true in the current case. Disaggregation jobs for some of the high hazard cities, starting with Guwahati, have been set up but are currently failing to complete. Disaggregation could help to explain discrepancies between the current model and the reference. Perhaps more importantly it can guide future efforts towards the source zones, magnitude ranges and GMPEs which dominate hazard.

Similarly it was hoped that this report could show the relative importance of the newly and updated GMPEs (columns 'N' and 'S' in in Table 3) to hazard results via a sensitivity analysis. This would help guide improvements to logic trees in future work.

4 Conclusions

This study has on the one hand been a cautionary tale about the reproducibility of PSHA and on the other an attempt to completely document an open model for seismic hazard the Indian subcontinent.

Besides being based on closed-source software, the PSHA of Nath and Thingbaijam (2012) lacks explicit documentation of tectonic region types and the proper interpretation of smoothed-gridded model files. There are furthermore a few minor but obvious errors, one in the specification of Sharma et al. (2009) for normal faulting, another in the standard deviation of b value given for zone 93. Aside from these defects the documentation of Nath and Thingbaijam (2012) is quite complete.

The model of Nath and Thingbaijam (2012) is furthermore very well thought-out, and quite close to the state of the art. Potential improvements are discussed in the appendices, among them upgrading the GMPEs selected, revising upwards the completeness magnitude used in generating smoothed-gridded models and incorporating more geological and paleoseismological constraints in modelling certain faults. Even without these improvements this model is ready for incorporation into the GEM Hazard Input Models Database, pending a few more verification exercises.

The most important outstanding verification is the generation of a hazard map. More particularly a hazard error map like those used to show differences between 2008 and 2014 models in Petersen et al. (2014) is essential to isolate mis-assigned tectonic region types and/or identify problems in the interpretation of smoothed-gridded model files. At the time of this writing this additional work is being held up by server scheduling but also problems with the way OpenQuake Release 1.9 partitions large jobs. It is hoped that before publication at `https://hazardwiki.openquake.org/` this verification can be completed.

Acknowledgement

Thanks to Marco Pagani and Graeme Weatherill for setting an interesting problem and particularly to Graeme for managing my recalcitrant jobs on the server. Thanks to Kiran Thingbaijam for clarifications and engaging discussion. Thanks to Amanda for your unfailing support.

References

Abrahamson, N., Gregor, N., and Addo, K. (2016). Bc hydro ground motion prediction equations for subduction earthquakes. *Earthquake Spectra*, 32(1):23–44. (Cited on page 37.)

Akkar, S. and Bommer, J. J. (2010). Empirical equations for the prediction of PGA, PGV, and spectral accelerations in Europe, the Mediterranean region, and the Middle East. *Seismological Research Letters*, 81(2):195–206. (Cited on page 15.)

Anbazhagan, P., Sreenivas, M., Ketan, B., Moustafa, S. S., and Nassir, S. A.-A. (2015). Selection of ground motion prediction equations for seismic hazard analysis of peninsular India. *Journal of Earthquake Engineering*, (just-accepted). (Cited on pages 36 and 37.)

Ashish, Lindholm, C., Parvez, I. A., and Kühn, D. (2016). Probabilistic earthquake hazard assessment for peninsular India. *Journal of Seismology*, pages 1–25. (Cited on pages 3, 42, and 43.)

Atkinson, G. M. and Boore, D. M. (2003). Empirical ground-motion relations for subduction-zone earthquakes and their application to Cascadia and other regions. *Bulletin of the Seismological Society of America*, 93(4):1703–1729. (Cited on pages 10, 15, and 16.)

Atkinson, G. M. and Boore, D. M. (2006). Earthquake ground-motion prediction equations for eastern North America. *Bulletin of the seismological society of America*, 96(6):2181–2205. (Cited on pages 15 and 36.)

Atkinson, G. M. and Macias, M. (2009). Predicted ground motions for great interface earthquakes in the Cascadia subduction zone. *Bulletin of the Seismological Society of America*, 99(3):1552–1578. (Cited on pages 15 and 35.)

Baker, J. W. (2008). An introduction to probabilistic seismic hazard analysis. *Report for the US Nuclear Regulatory Commission, page Version*, 1. (Cited on pages 19 and 22.)

Berryman, K., Ries, W., and Litchfield, N. (2014). The Himalayan frontal thrust: attributes for seismic hazard, version 1.0. Technical report. `http://www.nexus.globalquakemodel.org/gem-faulted-earth/` Accessed 2015-12-04. (Cited on page 42.)

Beyer, K. and Bommer, J. J. (2006). Relationships between median values and between aleatory variabilities for different definitions of the horizontal component of motion. *Bulletin of the Seismological Society of America*, 96(4A):1512–1522. (Cited on page 16.)

Bhatia, S. C., Kumar, M. R., and Gupta, H. K. (1999). A probabilistic seismic hazard map of india and adjoining regions. *Annals of Geophysics*, 42(6). (Cited on pages 3 and 42.)

Bilham, R. and England, P. (2001). Plateau 'pop-up'in the great 1897 Assam earthquake. *Nature*, 410(6830):806–809. (Cited on pages 6, 14, and 42.)

Bilham, R., Gaur, V. K., and Molnar, P. (2001). Himalayan seismic hazard. *Science*, 293(5534):1442–4. (Cited on pages 3 and 21.)

Bommer, J. J. and Scherbaum, F. (2008). The use and misuse of logic trees in probabilistic seismic hazard analysis. *Earthquake Spectra*, 24(4):997–1009. (Cited on page 29.)

Boore, D. M. and Atkinson, G. M. (2008). Ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods between 0.01 s and 10.0 s. *Earthquake Spectra*, 24(1):99–138. (Cited on pages 7 and 15.)

Boore, D. M., Watson-Lamprey, J., and Abrahamson, N. A. (2006). Orientation-independent measures of ground motion. *Bulletin of the Seismological Society of America*, 96(4A):1502–1511. (Cited on page 15.)

Bozorgnia, Y., Abrahamson, N. A., Atik, L. A., Ancheta, T. D., Atkinson, G. M., Baker, J. W., Baltay, A., Boore, D. M., Campbell, K. W., Chiou, B. S.-J., et al. (2014). Nga-west2 research project. *Earthquake Spectra*, 30(3):973–987. (Cited on page 35.)

Bureau of Indian Standards (2002). Criteria for earthquake resistant design of structures, part 1 - general provisions and buildings. Number 1893–2002 in IS. New Delhi. (Cited on page 3.)

Campbell, K. W. (2003). Prediction of strong ground motion using the hybrid empirical method and its use in the development of ground-motion (attenuation) relations in eastern North America. *Bulletin of the Seismological Society of America*, 93(3):1012–1033. (Cited on pages 15 and 36.)

Campbell, K. W. and Bozorgnia, Y. (2008). Nga ground motion model for the geometric mean horizontal component of PGA, PGV, PGD and 5% damped linear elastic response spectra for periods ranging from 0.01 to 10 s. *Earthquake Spectra*, 24(1):139–171. (Cited on pages 7 and 15.)

CEUS–SSCn (2012). Central and Eastern United States seismic source characterization for nuclear facilities. Technical report, EPRI, U.S. DOE, and U.S. NRC. `http://www.ceus-ssc.com/Report/Downloads.html` Accessed 2016-03-01. (Cited on page 42.)

Cotton, F., Scherbaum, F., Bommer, J. J., and Bungum, H. (2006). Criteria for selecting and adjusting ground-motion models for specific target regions: Application to central Europe and rock sites. *Journal of Seismology*, 10(2):137–156. (Cited on pages 16 and 35.)

Crowley, H., Monelli, D., Pagani, M., Silva, V., Weatherill, G., and Rao, A. (2015). *OpenQuake Engine User Instruction Manual Version 1.5*. Global Earthquake Model (GEM). `http://www.globalquakemodel.org/media/cms_page_media/432/oq-manual-15.pdf` Accessed 2015-10-15. (Cited on page 3.)

Das, S., Gupta, I. D., and Gupta, V. K. (2006). A Probabilistic Seismic Hazard Analysis of Northeast India. *Earthquake Spectra*, 22(1):1–27. (Cited on pages 3 and 42.)

Delavaud, E., Scherbaum, F., Kuehn, N., and Allen, T. (2012). Testing the global applicability of ground-motion prediction equations for active shallow crustal regions. *Bulletin of the Seismological Society of America*, 102(2):707–721. (Cited on pages 36 and 37.)

Delavaud, E., Scherbaum, F., Kuehn, N., and Riggelsen, C. (2009). Information-theoretic selection of ground-motion prediction equations for seismic hazard analysis: An applicability study using Californian data. *Bulletin of the Seismological Society of America*, 99(6):3248–3263. (Cited on pages 6 and 35.)

Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M., and Stodden, V. (2009). Reproducible research in computational harmonic analysis. *Computing in Science & Engineering*, 11(1):8–18. (Cited on page 3.)

Douglas, J. (2003). Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. *Earth-Science Reviews*, 61(1):43–104. (Cited on page 16.)

Field, E. H., Jordan, T. H., and Cornell, C. A. (2003). OpenSHA: A developing community-modeling environment for seismic hazard analysis. *Seismological Research Letters*, 74(4):406–419. (Cited on pages 4, 22, 24, and 25.)

Fomel, S. and Claerbout, J. F. (2009). Reproducible research. *Computing in Science & Engineering*, 11(1):5–7. (Cited on page 3.)

Frankel, A. (1995). Mapping seismic hazard in the central and eastern united states. *Seismological Research Letters*, 66(4):8–21. (Cited on pages 12 and 38.)

Gardner, J. and Knopoff, L. (1974). Is the sequence of earthquakes in southern california, with aftershocks removed, poissonian? *Bull. Seismol. Soc. Am*, 64(5):1363–1367. (Cited on page 41.)

Gupta, I. (2010). Response spectral attenuation relations for in-slab earthquakes in Indo-Burmese subduction zone. *Soil Dynamics and Earthquake Engineering*, 30(5):368–377. (Cited on pages 14, 15, and 16.)

Hayes, G. P., Wald, D. J., and Johnson, R. L. (2012). Slab1.0: A three-dimensional model of global subduction zone geometries. *Journal of Geophysical Research: Solid Earth (1978–2012)*, 117(B1). (Cited on pages 9, 42, and 43.)

Hinsen, K. (2011). A data and code model for reproducible research and executable papers. *Procedia Computer Science*, 4:579–588. (Cited on page 3.)

Jaiswal, K. and Sinha, R. (2007). Probabilistic Seismic-Hazard Estimation for Peninsular India. *Bulletin of the Seismological Society of America*, 97(1B):318–330. (Cited on pages 3 and 42.)

Kanno, T., Narita, A., Morikawa, N., Fujiwara, H., and Fukushima, Y. (2006). A new attenuation relation for strong ground motion in Japan based on recorded data. *Bulletin of the Seismological Society of America*, 96(3):879–897. (Cited on pages 8, 15, 16, and 36.)

Lin, P.-S. and Lee, C.-T. (2008). Ground-motion attenuation relationships for subduction-zone earthquakes in northeastern Taiwan. *Bulletin of the Seismological Society of America*, 98(1):220–240. (Cited on pages 15, 16, and 36.)

Mahajan, A. K., Thakur, V. C., Sharma, M. L., and Chauhan, M. (2009). Probabilistic seismic hazard map of NW Himalaya and its adjoining area, India. *Natural Hazards*, 53(3):443–457. (Cited on page 3.)

Nath, S. K. and Thingbaijam, K. K. S. (2011). Peak ground motion predictions in India: an appraisal for rock sites. *Journal of Seismology*, 15(2):295–315. (Cited on pages 3, 6, 8, 28, 35, and 36.)

Nath, S. K. and Thingbaijam, K. K. S. (2012). Probabilistic seismic hazard assessment of India. *Seismological Research Letters*, 83(1):135–149. (Cited on pages 1, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 16, 17, 18, 19, 21, 23, 25, 26, 27, 28, 29, 35, 36, 37, 39, 40, and 42.)

Nath, S. K., Thingbaijam, K. K. S., and Ghosh, S. K. (2010). Earthquake catalogue of South Asia – a generic m_w scale framework. `http://www.earthqhaz.net/sacat/` Accessed 2015-12-01. v2. (Cited on pages 3, 6, 12, 13, 37, 38, 39, 40, 41, and 43.)

Nath, S. K., Thingbaijam, K. K. S., Maiti, S. K., and Nayak, A. (2012). Ground-motion predictions in Shillong region, northeast India. *Journal of Seismology*, 16(3):475–488. (Cited on pages 3, 6, 14, 15, and 21.)

Pagani, M., Monelli, D., Weatherill, G., Danciu, L., Crowley, H., Silva, V., Henshaw, P., Butler, L., Nastasi, M., Panzeri, L., et al. (2014). Openquake engine: an open hazard (and risk) software for the global earthquake model. *Seismological Research Letters*, 85(3):692–702. (Cited on pages 3, 4, 14, 22, and 24.)

Petersen, M. D., Moschetti, M. P., Powers, P. M., Mueller, C. S., Haller, K. M., Frankel, A. D., Zeng, Y., Rezaeian, S., Harmsen, S. C., Boyd, O. S., et al. (2014). Documentation for the 2014 update of the United States national seismic hazard maps. Technical report, US Geological Survey. (Cited on pages 28, 29, and 42.)

QGIS Development Team (2016). *QGIS User Guide Version 2.8*. Open Source Geospatial Foundation Project. `http://docs.qgis.org/2.8/pdf/en/QGIS-2.8-UserGuide-en.pdf` Accessed 2016-03-15. (Cited on page 26.)

Raghukanth, S. and Iyengar, R. (2007). Estimation of seismic spectral acceleration in peninsular India. *Journal of Earth System Science*, 116(3):199–214. (Cited on pages 15, 16, and 36.)

Scherbaum, F., Delavaud, E., and Riggelsen, C. (2009). Model selection in seismic hazard analysis: An information-theoretic perspective. *Bulletin of the Seismological Society of America*, 99(6):3234–3247. (Cited on page 36.)

Scherbaum, F. and Kuehn, N. M. (2011). Logic tree branch weights and probabilities: summing up to one is not enough. *Earthquake Spectra*, 27(4):1237–1251. (Cited on page 37.)

Schneider, A., Friedl, M. A., and Potere, D. (2009). A new map of global urban extent from MODIS satellite data. *Environmental Research Letters*, 4(4):044003. (Cited on page 9.)

Sharma, M. L., Douglas, J., Bungum, H., and Kotadia, J. (2009). Ground-motion prediction equations based on data from the Himalayan and Zagros regions. *Journal of Earthquake Engineering*, 13(8):1191–1210. (Cited on pages 6, 7, 8, 14, 15, and 29.)

Stepp, J. (1972). Analysis of completeness of the earthquake sample in the puget sound area and its effect on statistical estimates of earthquake hazard. In *Proc. of the 1st Int. Conf. on Microzonazion, Seattle*, volume 2, pages 897–910. (Cited on pages 38, 40, and 41.)

Strasser, F. O., Arango, M., and Bommer, J. J. (2010). Scaling of the source dimensions of interface and intraslab subduction-zone earthquakes with moment magnitude. *Seismological Research Letters*, 81(6):941–950. (Cited on page 10.)

Styron, R., Taylor, M., and Okoronkwo, K. (2010). Database of active structures from the Indo-Asian collision. *Eos, Transactions American Geophysical Union*, 91(20):181–182. (Cited on pages 7, 9, 42, and 43.)

Thingbaijam, K. K. S. and Nath, S. K. (2011). A seismogenic source framework for the Indian subcontinent. Unpublished manuscript. (Cited on pages 3, 6, 7, 8, 10, 12, 17, 37, and 38.)

Toro, G. R. (2002). Modification of the Toro et al.(1997) attenuation equations for large magnitudes and short distances. *Risk Engineering, Boulder, Colorado*. (Cited on pages 15, 36, and 37.)

Uhrhammer, R. (1986). Characteristics of northern and central california seismicity. *Earthquake Notes*, 57(1):21. (Cited on page 41.)

Wang, Y., Sieh, K., Tun, S. T., Lai, K.-Y., and Myint, T. (2014). Active tectonics and earthquake potential of the Myanmar region. *Journal of Geophysical Research: Solid Earth*, 119(4):3767–3822. (Cited on page 6.)

Weatherill, G. A. (2014). *OpenQuake Hazard Modeller's Toolkit - User Guide*. Global Earthquake Model (GEM). (Cited on page 12.)

Wells, D. L. and Coppersmith, K. J. (1994). New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the seismological Society of America*, 84(4):974–1002. (Cited on page 10.)

Woessner, J., Laurentiu, D., Giardini, D., Crowley, H., Cotton, F., Grünthal, G., Valensise, G., Arvidsson, R., Basili, R., Demircioglu, M. B., et al. (2015). The 2013 european seismic hazard model: key components and results. *Bulletin of Earthquake Engineering*, 13(12):3553–3596. (Cited on page 42.)

Yadav, R. B. S., Tripathi, J. N., Rastogi, B. K., and Chopra, S. (2008). Probabilistic Assessment of Earthquake Hazard in Gujarat and Adjoining Region of India. *Pure and Applied Geophysics*, 165(9-10):1813–1833. (Cited on pages 3 and 42.)

Youngs, R., Chiou, S.-J., Silva, W., and Humphrey, J. (1997). Strong ground motion attenuation relationships for subduction zone earthquakes. *Seismological Research Letters*, 68(1):58–73. (Cited on pages 8, 15, 16, and 36.)

Zhao, J. X., Zhang, J., Asano, A., Ohno, Y., Oouchi, T., Takahashi, T., Ogawa, H., Irikura, K., Thio, H. K., Somerville, P. G., et al. (2006). Attenuation relations of strong ground motion in Japan using site classification based on predominant period. *Bulletin of the Seismological Society of America*, 96(3):898–913. (Cited on pages 7, 15, and 16.)

Appendix A Alternative GMPE logic tree

In this section, possible improvements to the GMPE logic tree of Nath and Thingbaijam (2012) in future work are discussed.

An obvious upgrade to the GMPE logic tree of Nath and Thingbaijam (2012) would be to use updated models, whenever they exist, as per the recommendations of Cotton et al. (2006). For example, the NGA-West1 models of 2008 were superseded in 2014 by NGA-West2, and so the newer models should be used Bozorgnia et al. (2014). Models which have been superseded and should be updated are indicated in Table 3. This is trivial to implement since the newer models have already been implemented in OpenQuake.

A less straightforward but nonetheless important improvement would be to select GMPEs and possibly also to assign weights using measures of GMPE efficacy. Delavaud et al. (2009) point out that macroseismic intensity observations are more abundant than instrumental recording and go on to demonstrate that the two can be used almost interchangeably for the purpose of quantitative assessment of GMPE efficacy. This is particularly important in areas of low seismicity or sparse instrumentation, such as India. Nath and Thingbaijam (2011) have made good use of this fact, but Nath and Thingbaijam (2012) appear to utilize their efficacy assessments only imperfectly.

For example in Nath and Thingbaijam (2012, Figure 3) there is a branch for megathrust earthquakes, because the GMPE of Atkinson and Macias (2009) demands it, but no regional distinctions are made. Yet Table 5 shows clear differences in GMPE efficacy between interface subduction in the Himalayas and in the Andaman-Sumatra subduction zone.

Table 5: Relative efficacy of GMPEs for interface subduction in the Indian subcontinent. Negative average sample log likelihood (LLH) scores are from (Nath and Thingbaijam, 2011, Table 5) while weights are computed using Delavaud et al. (2012).

(a) Himalayas

Model	LLH	weight	DSI
KAN06	2.4190	0.19	12.2
NATH09	2.4280	0.18	11.5
ATBO03	2.5733	0.17	0.8
ZHAO06	2.6512	0.16	-4.5
LILE08	2.6789	0.15	-6.3
YOU97	2.7117	0.15	-8.4

(b) Andaman-Sumatra

Model	LLH	weight	DSI
ATMA09	2.5644	0.30	1.4
MEPA10	3.3970	0.17	-43.0
ATBO03	3.4345	0.16	-44.5
PETE04	3.5942	0.15	-50.3
ZHAO06	3.7918	0.13	-56.7
KAN06	4.2216	0.09	-67.8

It appears as though GMPEs for all megathrust earthquakes were chosen by taking the top four ranking GMPEs (by LLH) for events in the Himalayas. Many authors (Scherbaum et al., 2009; Nath and Thingbaijam, 2011; Delavaud et al., 2012; Anbazhagan et al., 2015) are quite interested in "ranking", i.e. constructing an ordered list of GMPEs, but few clearly explain how rankings are to be used. This is not a horse race; we are not betting that a GMPE will "win, place or show".

Scherbaum et al. (2009) suggest a way to turn an LLH score into a logic-tree weight and Delavaud et al. (2012) developed the concept of "data support index". Using these measures Table 5 shows that in the Himalayas the data support the models more-or-less equally, and so it would make no sense to omit Youngs et al. (1997) or Lin and Lee (2008) on this basis.

A better method than applying an arbitrary LLH or ranking cutoff for pruning logic tree branches would be to apply the principles of mutual exclusivity and collective exhaustiveness (Scherbaum et al., 2009). Models should be excluded if they are very similar in their predictions, particularly if the methodology for producing them is similar. As an example, note that among the models applied in the stable shallow crust, all but Raghukanth and Iyengar (2007) were developed for eastern North America, and all but Campbell (2003) are based on stochastic simulations. Since Atkinson and Boore (2006) and Toro (2002) have similar methodologies and Nath and Thingbaijam (2011) show that they have similar LLH scores they are not mutually exclusive and in future work it would make sense to omit one, likely the latter since it has a slightly higher LLH. By the same token, the addition of a fully-empirical model (if it exists) would bring the set of GMPEs closer to being collectively exhaustive, as would improved models specific to peninsular India or its subregions.

Whereas a distinction is made in Nath and Thingbaijam (2012) between intraslab subduction in the Himalayas and Indo-Burman subduction zones, none is made for interface subduction. Table 5 suggests that model efficacy does differ greatly between the two regions. While many models are equally supported for the data in the Himalayas, several, notably Kanno et al. (2006), are used by Nath and Thingbaijam (2012) where they are not well-supported by the data. Therefore, in future work it would be appropriate to select subduction interface GMPEs differently in the Himalayas, the Indo-Burman subduction zone and Andaman-Sumatra.

Care must be taken when using efficacy measures to assess GMPEs in seismically stable regions. For example, it is dangerous to recommend a GMPE for a regions on the basis of a single event. One study which does just this is Anbazhagan et al. (2015). In an extreme case they propose different logic tree weights for Anjar and Bhuj even though the epicentres and depths were very close together. In contrast Nath and Thingbaijam (2011) compute LLH for 7

regions (using 38 events total) and state that, "individual events do not have significant number of observations to support a viable ranking basis." Anbazhagan et al. (2015) furthermore seem to misuse the concept of data support index (DSI) by simply setting weights to zero when the DSI is negative. In contrast Delavaud et al. (2012) insist that the difference between DSIs is more diagnostic than the sign of a given DSI.

The collective exhaustiveness requirement is trickier to implement. It is this requirement which pushes hazard modellers to seek out models which are complementary. Thus models with broad data support from other regions complement models with poor data support from the target region. Stochastic models supplement data-driven models. Models with different functional forms, distance or magnitude ranges can complement each other.

An example of a model based on an improved global database is Abrahamson et al. (2016). It should be considered for inclusion, although it is recommended for use only to 120 km depth so more work is needed to identify GMPEs appropriate for layer 4 of this model.

Once the set of relevant GMPEs has been determined by successive cycles of growth and pruning.

Scherbaum and Kuehn (2011) give specific recommendations as to the proper use of rankings, proposing a "stick-breaking" method of assigning weights. In this method the GMPEs are considered in order by rank. Each is assigned a weight, where the weight is understood as the probability that the model is the true model given that previously considered models are known not to be true. In this model equal weights can still be arrived at but the procedure is as follows. Given three models, they are ranked, perhaps by LLH, and the first is assigned a 33⅓% weight. The second is assigned 50% of the remaining 67⅔% or 33⅓%, and the third is assigned 100% of the remaining 33⅓% because it must be true.

In future work it may be instructive to apply the stick-breaking method of assigning weights.

The process of developing a logic tree to assess epistemic uncertainty is clearly a dialectical one. Mutual exclusivity and collective exhaustiveness comprise opposing forces which must be exerted alternately and in tandem.

Bearing this in mind, in future work it is recommended to:

1. Replace models which have been superseded, in particular update NGA-West1 models to NGA-West2.
2. Split both intra-slab and interface subduction into Himalayan, Indo-Burman and Andaman-Sumatran tectonic subregions.
3. Incorporate models using different methodologies and/or databases, e.g. the BC Hydro subduction model of Abrahamson et al. (2016).
4. Incorporate models using more region-specific models.
5. Prune models which are quite similar to those already used, e.g. Toro (2002).
6. Use efficacy measures to rank GMPEs and prune obviously inapplicable ones.
7. Assign weights by stick-breaking as per Scherbaum and Kuehn (2011)

Appendix B Catalogue evaluation

This section is a brief review of the catalogue of Nath et al. (2010). This is the catalogue used by Thingbaijam and Nath (2011) and subsequently Nath and Thingbaijam (2012) to generate frequency-magnitude distributions for the areal model and smoothed-gridded activity rates for the point-source models. It is found that declustering is incomplete and the magnitude of completeness has been underestimated. The magnitude of completeness is critical for the accuracy of activity rates using methods which don't automatically compensate for incomplete catalogues

Table 6: Details of two largest mainshocks in catalogue of Nath et al. (2010).

event ID	agency	date	magnitude	layer ID	depth [km]
32897	EHB	2004-12-26 00:58:52.280	9.1	1	22
32898	GCMT	2004-12-26 01:01:09.000	9.1	2	29

such as Frankel (1995). In particular when the magnitude of completeness is underestimated, event counts, activity rates and overall hazard will also be underestimated.

Figure 12 gives an overview of the events labelled as mainshocks. This figure raises serious doubts about the time spans treated as complete by Thingbaijam and Nath (2011). Consider for example the assertion completeness at magnitude 5.5 to 1906; clearly the rate is constant from 1964 on, but before that there is variation over time of the completeness. Similarly at magnitude 4.5 the catalogue is definitely complete back to 1987, possibly back to 1978, but definitely not back to 1964.

In order to investigate completeness more rigorously, the method of Stepp (1972) was applied to estimate the completeness as a function of time. The results are shown in Figure 13. Whereas a sharp knee is expected when the catalogue goes from complete to incomplete as the observation time is extended further and further back in time, Figure 13 shows gentle curves. Furthermore, in layers 1 and 2 in particular, there are no time periods where the activity rate has a $1/\sqrt{T}$ slope expected of a constant occurrence rate.

Nath et al. (2010) do not specify the declustering method used, but in any case it was imperfect, as a cursory review reveals events with nearly identical characteristics. Conspicuously there are two magnitude 9+ events in 2004, one each in layers 1 and 2.

The duplication of the Andaman-Sumatra megathrust of 2004 and other events suggests that part of the problem in assessing the completeness of the catalogue events labelled as "mainshocks" is the declustering method used.

Figure 14 shows the results of declustering the raw catalogue anew. The duplicate Andaman-Sumatra megathrust was successfully removed, as expected. Similar results are obtained if declustering is performed using only the mainshocks of the original catalogue as input. With the results of Figure 14 various completeness tables are now easy to formulate. For example in layer 2, the catalogue is clearly complete at magnitude 4.5 back to 1991 (not 1964) and at magnitude 5.5 back to 1960 (not 1905).

It is recommended that for future work that smoothed-gridded seismicity models be generated using completeness intervals selected from Figure 14. This will ensure that activity rates and thus hazard are not underestimated.

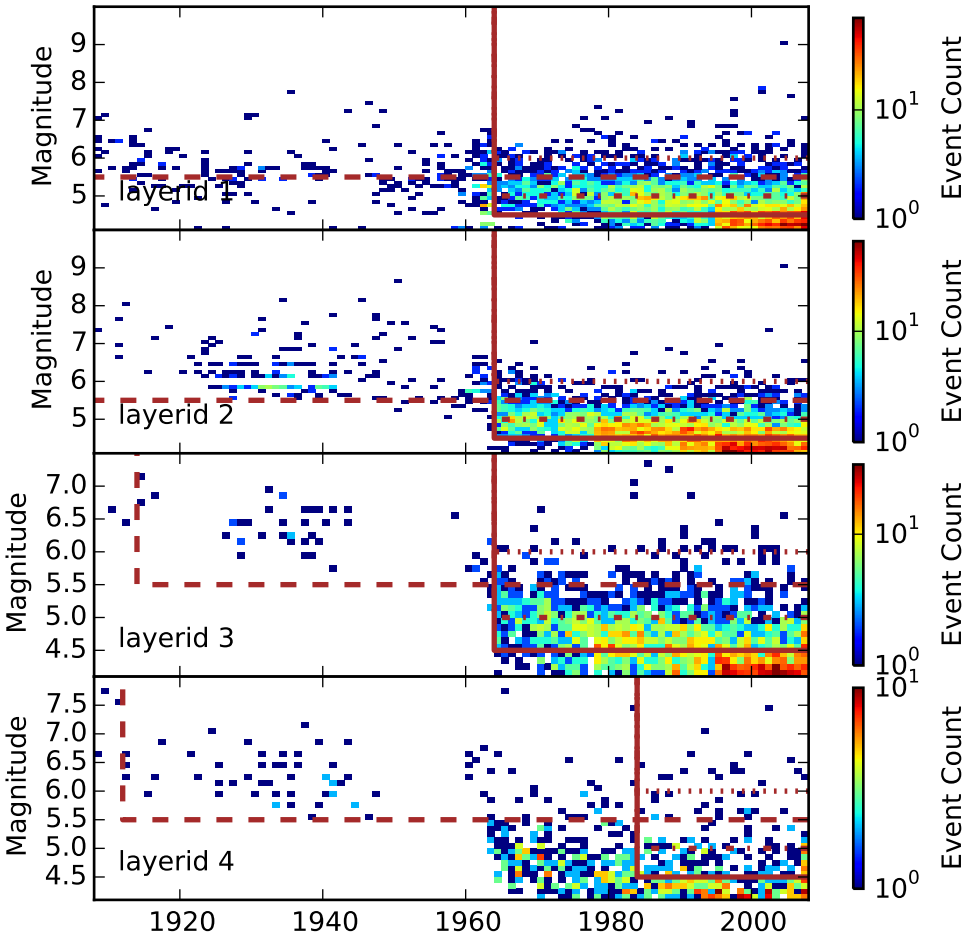

Figure 12: Magnitude-time density plot for mainshocks. Mainshock identification is that of Nath et al. (2010). Layer depth limits are from Table 1. Overlaid in brown are completeness intervals from Table 1 used by Nath and Thingbaijam (2012) in generating gridded-smoothed seismicity models.

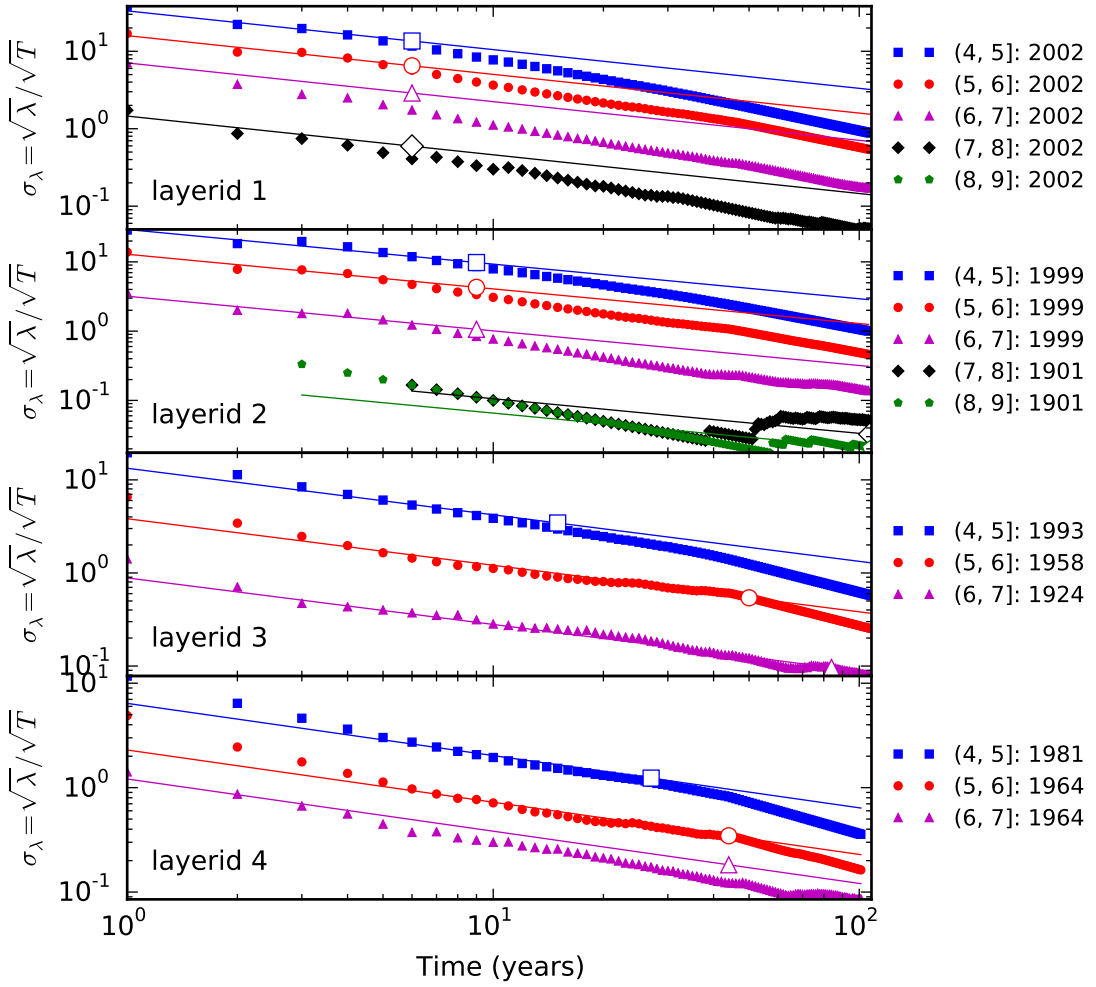

Figure 13: Completeness analysis of mainshocks identified by Nath and Thingbaijam (2012) in the catalogue of Nath et al. (2010) using the method of Stepp (1972). Closed symbols show the square root of the ratio of the activity rate λ to the time period T for various time periods and cut-off magnitudes. For a perfectly uniform occurrence rate the slope is $1/\sqrt{T}$ with a standard deviation of $1/\sqrt{T}$. Open symbols are estimates of the completeness period obtained by detecting a change in slope. Solid lines are reference curves with a slope of $1/\sqrt{T}$ and passing through the estimated completeness period.

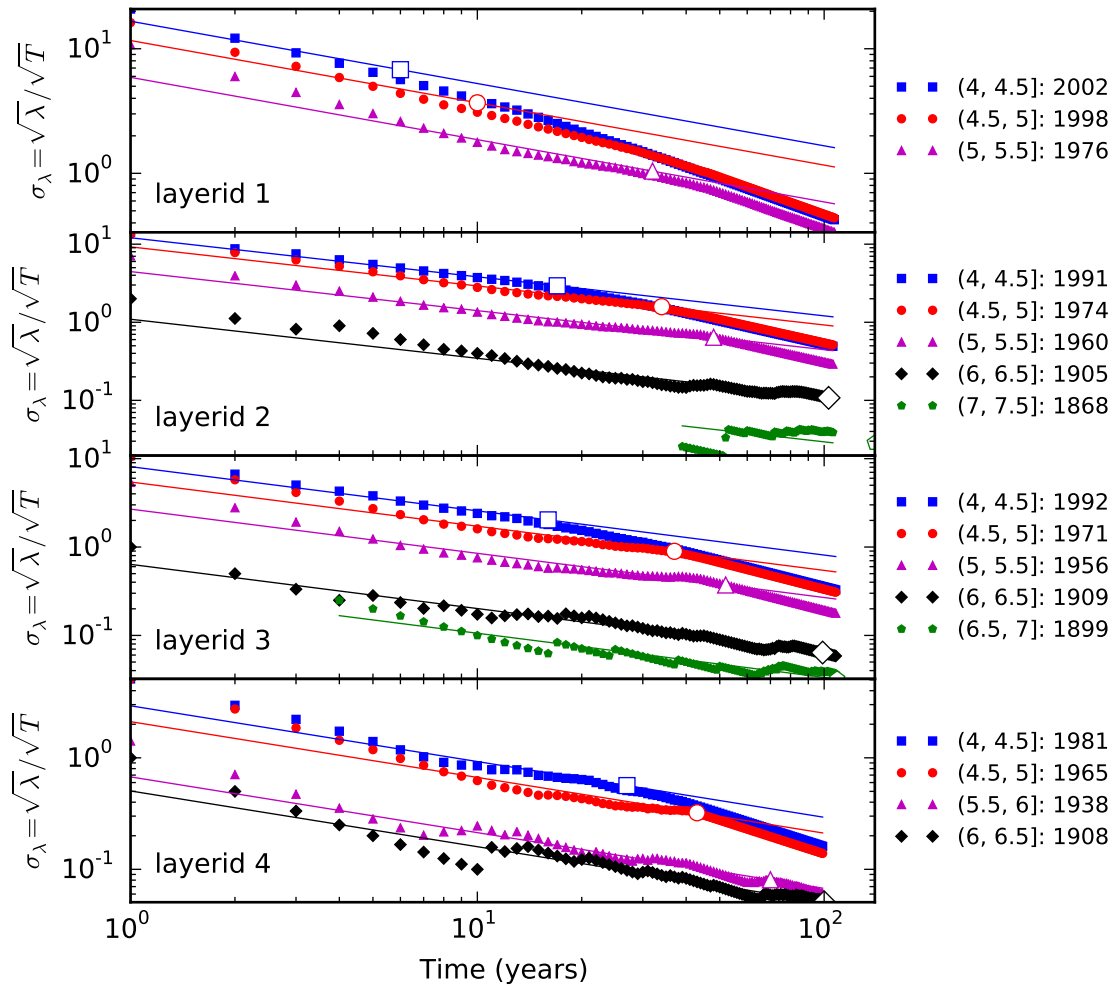

Figure 14: Completeness analysis of mainshocks selected by declustering the catalogue of Nath et al. (2010) using the method of Gardner and Knopoff (1974) with an Uhrhammer (1986) time window and equal fore- and after-shock windows. See caption of Figure 13 for explanation of symbols and overview of the Stepp (1972) method.

Appendix C Source model improvements

There are several aspects of the source model of Nath and Thingbaijam (2012) which could be improved upon in future work.

Without altering the source modelling framework there are two improvements which could be made. The first is motivated the catalogue study in Appendix B. The catalogue is simply not complete to the magnitude assumed by Nath and Thingbaijam (2012). The smoothed-gridded models should be revised using either a magnitude of completeness approximately 0.5 units higher, over significantly reduced time spans, or using a time-varying magnitude of completeness.

Second, For the areal and smoothed-gridded seismicity models the hypocentral depth was assumed to be midway between the layer boundaries. This is a crude assumption made due to a lack of indication of what was actually done by Nath and Thingbaijam (2012). A clear improvement would be to distribute hypocentral depths in a way which corresponds to the depths observed in the catalogue. This could either be done on the basis of the whole catalogue, as shown in Figure 15 or for each areal zone.

The use of smoothed-gridded model in conjunction with areal zones to model background seismicity in a catalogue-driven way is a sound and widely applied practice. This approach is the only one available when knowledge of geology is sparse, as in peninsular India.

The current state of the art, however, is to include more detailed modelling of faults where possible (e.g. Woessner et al., 2015; Petersen et al., 2014). While earlier investigations in the Indian subcontinent (Bhatia et al., 1999; Das et al., 2006; Yadav et al., 2008; Jaiswal and Sinha, 2007) relied on areal seismogenic source zonation, Nath and Thingbaijam (2012) adds smoothed-gridded point sources. Ashish et al. (2016) adds fault-modelling but is limited to the stable-continental regions.

Knowledge of the location of a surface trace of a fault is not sufficient; this must be augmented by geodetic slip rates and paleoseismic estimates of the size and timing of past earthquakes. Recently published fault maps (Styron et al., 2010; Berryman et al., 2014) and fault-specific studies (Bilham and England, 2001; Hayes et al., 2012) represent significant progress in this direction (see Figure 2) and this knowledge should be incorporated into future seismic source models.

Berryman et al. (2014) make very clear recommendations for modelling the main frontal thrust of the Himalayas. The fault is divided into three segments based on historic and paleoseismic observations. The segments are characterised by different lengths, slip rates and therefore $m_{\max}$. All segments have a relatively shallow dip of just 10°. The authors ultimately conclude that a $m_{\max} = 9.1$ event may occur anywhere on the fault, and thus it can be modelled as a single simple fault. Berryman et al. (2014) also recommend enforcing $m_{\min} = 7.0$ since events below this magnitude likely do not significantly contribute to the slip rate budget. The existing areal zones which overlap the main frontal thrust (zones 13, 21, 22, 906) should thus be left in place, with their $m_{\max}$ reduced to 7.0 in order to avoid "double-counting" (CEUS–SSCn, 2012). Note that the fault is believed to extend from the surface to just 17-20 km depth. Thus zones 98, 110 and 928 on layer 2 and 936 on layer 3 should not be affected the addition of a fault model for the main frontal thrust.

Bilham and England (2001) is an example of a paleoseismic study which gives us sufficient information to fully characterise a fault source for a significant fault which doesn't even extend to the surface. In particular the necessary geometry and slip rates are provided for both the Oldham and Dauki faults beneath the Shillong plateau. The best-fitting solution of Bilham and England (2001) for the M_w8.1 1897 Assam earthquake was for slip extending from 9 to 45 km, overlapping both layers 1 and 2 of the current source model. Thus both zones 118 and 912 should have their $m_{\max}$ adjusted downward after the incorporation of simple fault models to

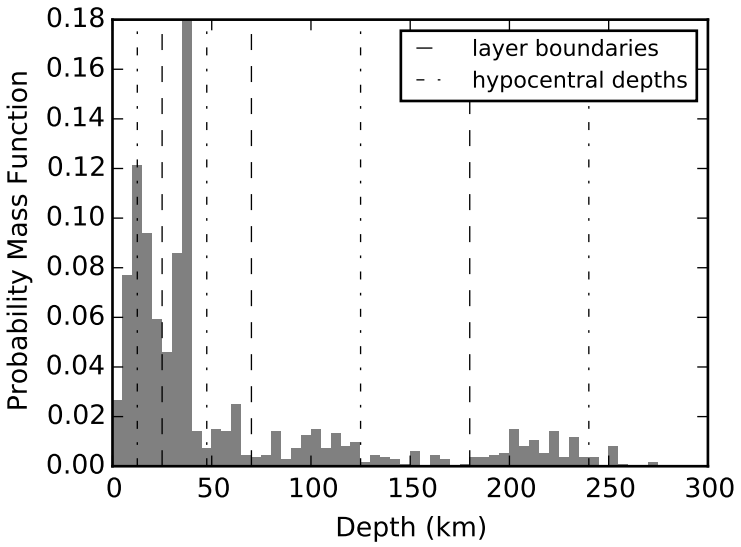

Figure 15: Depth histogram for mainshocks over magnitude 5.5. Mainshock identification is that of Nath et al. (2010). Seismogenic layer boundaries and hypocentral depths used in the current implementation are indicated as dashed and dash-dotted lines respectively.

avoid double-counting. It is therefore recommended in future work that both of these faults be modelled as simple faults.

Modelling of the Sumatra-Andaman fault following Hayes et al. (2012) would improve hazard modelling in the southeast, although it is not going to be very important for risk in peninsular India.

Other potentially significant faults with surface traces defined in Styron et al. (2010) include:

- Makran thrust belt in zones 15 and 16
- main Pamir thrust in zones 1 and 2
- Sagaing Fault in zone 914

In each of these cases more research is needed to obtain the geodetic and paleoseismic data needed to constrain fault models.

It should be noted, finally, that while Ashish et al. (2016) explicitly models 10 faults in peninsular India, seismicity is modelled using a Gutenberg-Richter truncated at an $m_{\max}$ which is not constrained by slip rates. Further, rather than modelling the dip of the faults, the fault is effectively modelled as an areal source with an outline which is offset by 0.2° from the surface trace. This approach only concentrates the seismicity near known faults; it does not incorporate geodetic or paleoseismological constraints.

The addition of fault models will greatly improve modelling of the expected distance distribution for sites near the faults. For deeper source regions, such as intraslab subduction under the Pamir range and the Andaman-Sumatra islands, existing areal zones probably already model the actual distribution of hypocenters (see Figure 16) well enough.

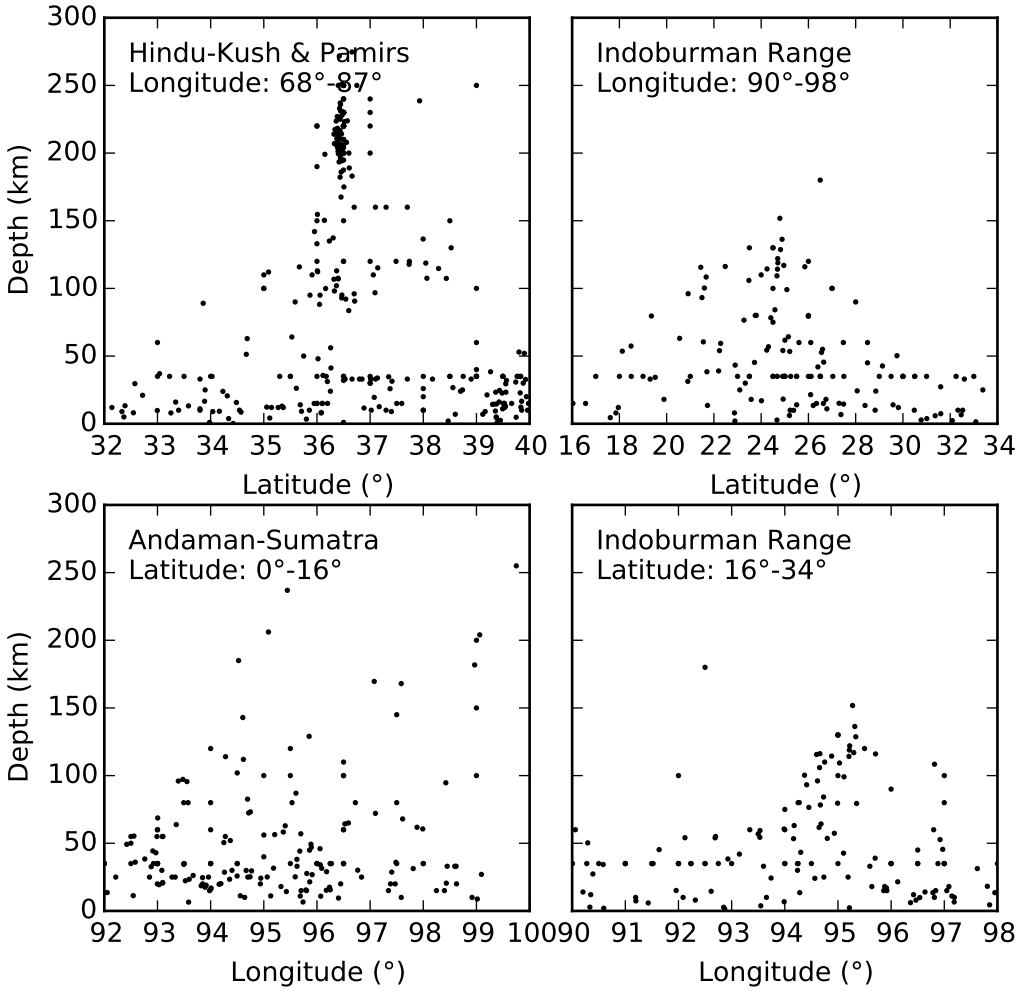

Figure 16: Depth vs. distance for mainshocks in regions with deep events. Subregions are indicated on each map; top left is the Hindu-Kush and Pamir ranges in the northwest of India viewed from the east, top right is the Andoman-Sumatran subduction zone viewed from the south while bottom left and right are beneath the Indoburman range and viewed from the east and south respectively. Sub-catalogues were selected for events over magnitude 5.5 within a rectangular box of latitude and longitude as indicated on each individual plot. Horizontal and vertical axes are plotted at different scales.

Appendix D Summary of electronic data

This section summarizes the input and output files relating to this model. It is expected to need revision prior to publication at `https://hazardwiki.openquake.org/`.

Electronic supplement corrections:

- `seismicitylay2.txt` (corrected σ_b for zone 93)

Areal zone assignments for tectonic region type, magnitude scaling relation and aspect ratio:

- `auxiliary data.csv`

Site coordinates:

- `NT2012_Table_3_lon_lat.csv`
- `NT2012_Figure_7_Indian_subcontinent_lon_lat.csv`

Full enumeration of epistemic uncertainty:

- `full_enumeration.ini`
- `gmpe_logic_tree.xml`
- `source_model_logic_tree.xml`
- `areal_source_model.xml`
- `NT2012_smoothed_sources_mmin4.5.xml`
- `NT2012_smoothed_sources_mmin5.5.xml`

Collapsed FMD uncertainty:

- `collapsed.ini`
- `collapsed_map.ini`
- `source_logic_tree_collapsed.xml`
- `areal_collapsed.xml`
- `NT2012_smoothed_collapsed_mmin4.5.xml`
- `NT2012_smoothed_collapsed_mmin5.5.xml`

No FMD uncertainty:

- `no_fmd_uncertainty.ini`
- `no_fmd_uncertainty.xml`

New GMPEs omitted:

- `no_fmd_uncertainty_standard_gmpes.ini`
- `gmpe_logic_tree_omit_new.xml`