



Originally published as:

Chen, Y.-S., Weatherill, G., Pagani, M., Cotton, F. (2018): A transparent and data-driven global tectonic regionalisation model for seismic hazard assessment. - *Geophysical Journal International*, 213, 2, pp. 1263—1280.

DOI: <http://doi.org/10.1093/gji/ggy005>

# A transparent and data-driven global tectonic regionalization model for seismic hazard assessment

Yen-Shin Chen,<sup>1,2</sup> Graeme Weatherill,<sup>1,2</sup> Marco Pagani<sup>1</sup> and Fabrice Cotton<sup>2,3</sup>

<sup>1</sup>Global Earthquake Model Foundation, I-27100 Pavia, Italy. E-mail: [yenshin@gfz-potsdam.de](mailto:yenshin@gfz-potsdam.de)

<sup>2</sup>GFZ German Research Centre For Geosciences, D-14473 Potsdam, Germany

<sup>3</sup>Institute for Earth and Environmental Sciences, University of Potsdam, D-14469 Potsdam, Germany

Accepted 2018 January 7. Received 2018 January 2; in original form 2017 August 17

## SUMMARY

A key concept that is common to many assumptions inherent within seismic hazard assessment is that of tectonic similarity. This recognizes that certain regions of the globe may display similar geophysical characteristics, such as in the attenuation of seismic waves, the magnitude scaling properties of seismogenic sources or the seismic coupling of the lithosphere. Previous attempts at tectonic regionalization, particularly within a seismic hazard assessment context, have often been based on expert judgements; in most of these cases, the process for delineating tectonic regions is neither reproducible nor consistent from location to location. In this work, the regionalization process is implemented in a scheme that is reproducible, comprehensible from a geophysical rationale, and revisable when new relevant data are published. A spatial classification-scheme is developed based on fuzzy logic, enabling the quantification of concepts that are approximate rather than precise. Using the proposed methodology, we obtain a transparent and data-driven global tectonic regionalization model for seismic hazard applications as well as the subjective probabilities (e.g. degree of being active/degree of being cratonic) that indicate the degree to which a site belongs in a tectonic category.

**Key words:** Fuzzy logic; Earthquake hazards; Seismicity and tectonic.

## 1 INTRODUCTION

The assessment of seismic hazard is a fundamental step in the evaluation and mitigation of the risk that earthquakes pose to society. At present, observations of seismicity and its resulting ground motion at any given site are insufficient to characterize the long-term recurrence of specified levels of shaking. It is therefore necessary to aggregate seismological observations over a given spatial or tectonic domain in order to derive models to compute the expected levels of shaking and their uncertainties, given a set of variables describing the fundamental characteristics of the earthquake process. Inherent within this process are several key assumptions indicating that certain properties of the earthquake process can be regionalized on the basis of geophysical and/or tectonic similarity. These classifications may apply to the properties of the seismogenic source, the attenuation of seismic waves within the crust, or to the effects of surface geology on ground motion.

In seismic hazard assessment, tectonic regionalizations are often used for the selection of the ground motion model (or models) and the areas to which they should be applied (Abrahamson & Shedlock 1997; García *et al.* 2012; Delavaud *et al.* 2012). In the development of ground motion prediction equations, regionalization has become an important topic, as the increasing availability of data has led ground motion model developers to identify key differences in the

behaviour of models for different regional subsets of ground motion records, even amongst those records classified as belonging to the same tectonic regime.

A legacy of geographical and tectonic regionalizations for the purpose of earthquake classification extends back in the scientific literature for over half a century, when the establishment of the Worldwide Standard Seismograph Network brought with it large volumes of seismological and geological data for routine processing and assessment. In response, Flinn & Engdahl (1965) proposed, and later revised (Flinn *et al.* 1974; Young *et al.* 1996), a division of the Earth into 51 zones based on both seismotectonic and geographical considerations. The Flinn–Engdahl regionalization remains in widespread use as many organizations, including the International Seismological Centre (ISC) and United States Geological Survey (USGS), continue to use it to associate seismological data to given regions of the Earth. It also forms the basis upon which the USGS Shakemap system selects appropriate ground motion models for their real-time estimates of ground shaking, the tectonic classification of which is seen in Fig. 1. Other regionalization schemes have focussed specifically on the seismic wave properties in order to understand the large-scale tectonic variation in tectonic material (Toksöz & Anderson 1966; Kanamori 1970; Dziewonski 1971; Jacob 1972; Jordan 1981; Okal 1977; Mitchell 1995; Mitchell *et al.* 1997; Gudmundsson & Sambridge 1998; Mitchell *et al.* 2008;

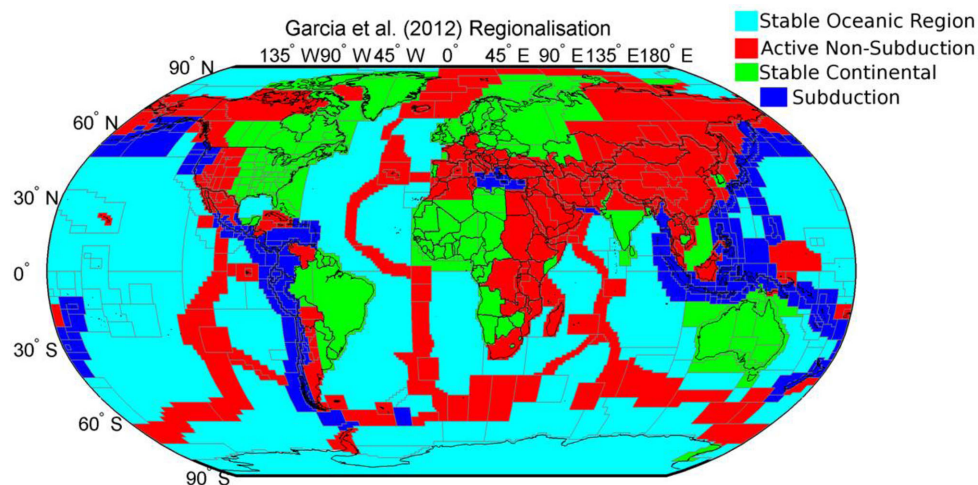


Figure 1. The tectonic regionalization developed by García *et al.* (2012) in the use of Global ShakeMap.

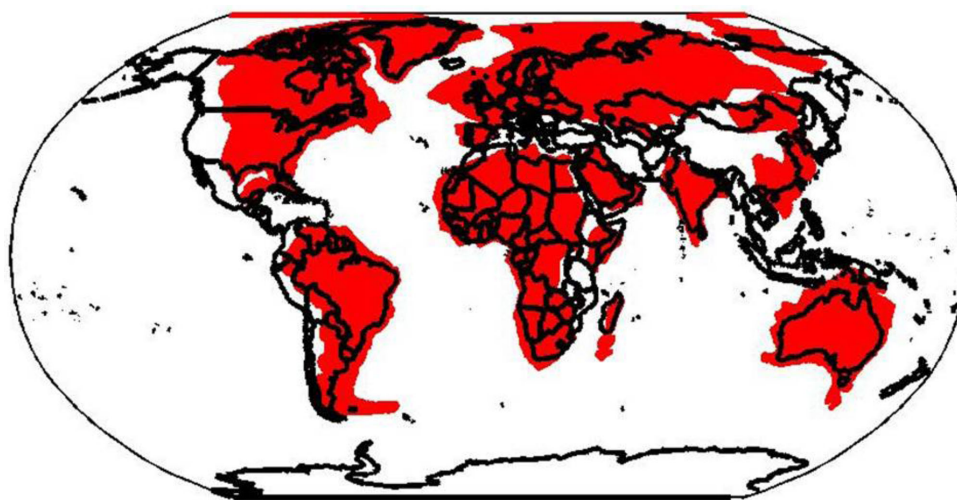


Figure 2. Stable continental regions (in red) defined in Johnston *et al.* (1994).

Khan *et al.* 2011; Lekic & Romanowicz 2011; Heeszel *et al.* 2013). Generally, these models partition the Earth's surface into several major divisions according to the material properties (e.g. geological era, attenuation, and seismic velocity); hence their utility in a seismic hazard assessment context is less clear.

A different regionalization model proposed by Kagan *et al.* (2010) was used to classify earthquakes according to tectonic regime such for similar purposes of analysing the statistical properties of recurrence within each. The Kagan *et al.* (2010) approach is particularly pertinent for this work as it describes an algorithmic process for classification in which expert judgement is minimized, and is applicable on a global scale. It is illustrative of an approach that can be reproducible in areas of generally high seismicity. It found application in global earthquake forecasting where regionally derived parameters of the earthquake recurrence models are applied to predict the spatial variation in the rates of events above a given magnitude (Kagan & Jackson 2012; Bird *et al.* 2015), and would therefore be illustrative of how an automated regionalization process can be informative in seismogenic source modelling.

Other regionalizations have been defined for a broader range of objectives. The aforementioned stable continental regions model of Johnston *et al.* (1994, see Fig. 2) is designed for the purpose of describing earthquake recurrence properties in regions of low

seismicity by considering data from all tectonically analogous regions. The exploration of variability in geophysical properties within the stable continental domain, with a particular focus on eastern North America, has also been revisited from a seismogenic source modelling perspective within the Central & Eastern US Seismic Source Characterization (CEUS-SSC) for Nuclear Facilities Project (Coppersmith *et al.* 2012), and subsequently from a ground motion attenuation perspective in the Next Generation Attenuation (NGA) East Project (Dreiling *et al.* 2014). The CEUS-SSC project revisited the separation of extended and non-extended crust proposed by Johnston *et al.* (1994) as distinct regimes in terms of recurrence and maximum magnitude. The new proposed classification separates between those regions belonging to extended Mesozoic and younger domains and those belonging to older extended and non-extended domains, the former of these believed to be capable of producing slightly larger earthquakes than the latter. In contrast, the regionalization study conducted within the NGA East Project focuses specifically on regional differences in attenuation in the Eastern US, identifying a significant difference in attenuation characteristics in the Gulf Coast and Mississippi Embayment compared to other regions of eastern North America.

Although it is evident that tectonic regionalization does play a role in many areas of seismic hazard analysis, and that attempts at

global regionalization have already been made, few regionalizations have been designed specifically in this context that can be applied on a global scale. Recent national and regional scale probabilistic seismic hazard models have critical regionalizations built within. For example, it has been long established that a national boundary between ‘active’ and ‘stable’ regions in the United States National Seismic Hazard model (Petersen *et al.* 2015) is approximately delineated by the eastern limit of the Rocky Mountains. In Europe and the Middle East, the tectonic picture is highly complex, with regions of subduction, continental convergence, extensional basins and transcurrent plate boundaries all situated in this compact area. Even the stable areas of northern Europe display characteristics that make it somewhat distinct from similar stable regions in North America and other parts of the world (Delavaud *et al.* 2012). Tectonic regionalizations built within recent hazard models, such as the 2013 European Seismic Hazard Model (Woessner *et al.* 2015) and 2014 Earthquake Model of the Middle East (Danciu *et al.* 2016), have relied on expert judgement regarding both the types of zones and their spatial extent.

The precedents for tectonic regionalization in hazard, whilst serving their purpose, suffer from several major limitations that make comparison between models difficult. The first limitation is the lack of reproducibility and clarity regarding the definition of the zone itself. In the ground motion modelling domain regionalization is based around a set of non-specific criteria that does not specify any clear spatial extent but only acts to classify subsets of strong motion records within a database. Subregionalization within a tectonic regime in varying the ground motion model coefficients is often according to political boundaries (e.g. Boore *et al.* 2014; Kotha *et al.* 2016) rather than geophysical. Little explicit consideration is given to the geophysical properties of the regions that account for the regional differences. The second limitation in the current approaches is that classifications are ‘hard’, in the sense that a location (be it source or site) is unambiguously placed into a single category. Given the inherent lack of precision in the classification of the tectonic regimes, it is counter-intuitive that the boundaries between tectonic regimes are generally treated as discrete conditions. In reality, the boundary between ‘active’ and ‘stable’ crustal regions, for the purposes of seismic hazard assessment, represents a continuum that transitions gradually over a distance.

To address the limitations described, this work presents a methodology for application to the problem of tectonic regionalization specifically for a seismic hazard analysis context. The methodology adopts a fuzzy inference process with criteria (here considered as ‘rules’) that are based on the interpretation from global geophysical data sets whose properties are related to the description of the earthquake source and path effects that seismic hazard analysis attempts to capture. It will be illustrated how this principally data-driven approach can still be combined with expert judgement albeit in a manner that is both quantifiable and replicable. Being predicated on a fuzzy system the main outputs of this process are geographical distributions of ‘memberships’, that is, quantified measures indicating the degree to which a site belongs to a tectonic category. These can be used to subsequently categorize a site to a regime according to a user-decision of an acceptable level or membership, or alternatively they could inform the assignment of weights within an analysis of epistemic uncertainty.

In order to test the methodology proposed, we classify sources and site locations found in the existing strong motion databases whose tectonic region type is assumed *a priori*, such as the NGA West 2 (Ancheta *et al.* 2014) and NGA East (Goulet *et al.* 2014) and we discuss the consistency of between the predicted regionalization

memberships and those already in use for probabilistic seismic hazard assessment (PSHA).

## 2 MAJOR TECTONIC CLASSES CONSIDERED IN THIS STUDY

The major tectonic regions used in the vast majority of hazard studies are Active Shallow Region, Subduction, and Stable Continental Region (Abrahamson & Shedlock 1997). For each of these three tectonic regions, a further subclassification has been suggested in some cases, for example, Akkar *et al.* (2012), Delavaud *et al.* (2012) and Stewart *et al.* (2013).

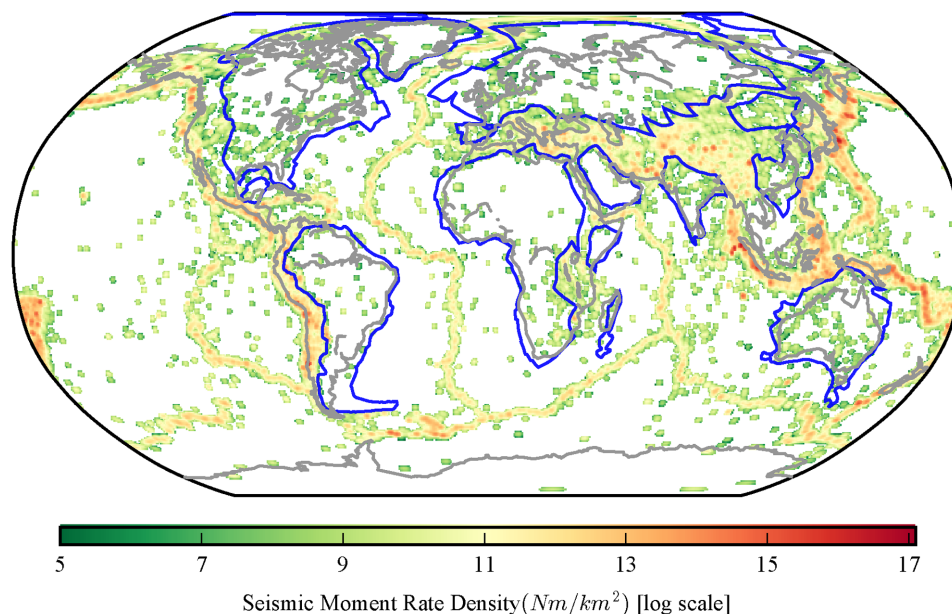
Within *Stable Continental Region*, it is recommended to consider separate ground-motion prediction equations (GMPEs) for cratonic regions (characterized by low deformation with slow attenuation), and non-cratonic regions (low deformation but faster attenuation; Somerville *et al.* 2009; Akkar *et al.* 2012) because of the significant difference in crust material properties. Inside subduction zones, it is considered important to differentiate between earthquakes generated along the interface and within the slab due to sensibly different propagation paths. Additionally, the earthquakes with mainly oceanic travel paths exhibit different behaviours from continental regions (Somerville *et al.* 2009; Akkar *et al.* 2012; Delavaud *et al.* 2012).

Taking into consideration the major differences mentioned above, the first step is to classify the active region from stable region, since the major tectonic regimes mentioned above can be broadly separated into active and stable categories. This binary classification, whilst often recognized in many applications as a practical division, is not representative of the physics of the lithosphere in which the deformation occurs across a continuum. We therefore consider the concepts of ‘active’ and ‘stable’ within a subjective probabilistic description (degree of being active), with the physical definition of activity or stability indicative of higher or lower long-term seismicity and strain, noting that a ‘stable’ region does not imply the absence of either.

The relevant global seismicity data and geophysics data sets have been appraised for integration into a data-driven fuzzy classification scheme. Instead of assigning a region as being active region (True) or not (False), in this study, the degree of membership to an active region is provided for each target site on the globe. Adopting a similar strategy, using relevant crustal property data sets, the degree of membership to a cratonic region for each target site is provided in the study. It should be emphasized at this point that although oceanic regions are given some consideration here, a regionalization for the purposes of seismic hazard assessment is primarily focused on continental regions. Further categorizations can be made using this result together with additional digital relevant data sources such as the Slab 1.0 model for subduction classification. In our study, seven target tectonic regimes are defined: (1) Active Shallow Region (2) Active Oceanic Region (3) Subduction-Interface (4) Subduction-Intraplate (5) Stable Continental Region-Craton (6) Stable Continental Region-Non Craton (7) Stable Oceanic Region.

## 3 DESCRIPTION OF THE DATA SET USED FOR TECTONIC REGIONALIZATION

Input to a probabilistic seismic hazard analysis consists of both seismic source characterization and the ground motion characterization, so various global seismotectonic and geophysical data sets are appraised not only for homogeneity, quality or spatial coverage,



**Figure 3.** A global map of the smoothed seismic moment rate density ( $\text{Nm km}^{-2}$ ) derived from the Weatherill *et al.* (2016) earthquake catalogue using the smoothing methodology proposed by Kagan & Jackson (1994) with a smoothing kernel width of 100 km. The stable continental region polygons (in blue) defined by EPRI (Johnston *et al.* 1994) are superposed on the map.

but also their relevance to the prediction of earthquake recurrence and/or ground motion attenuation. Amongst these criteria there can remain a degree of subjective judgement, albeit the justifications for selection of the most relevant data sets are provided in the following. An additional criterion is the avoidance of data sets in which prior assumptions about regionalization have been made in the construction. This, for example, excludes the Global Strain Rate Model (Kreemer *et al.* 2014), an important and relevant data set for seismic recurrence modelling, but one in which prior assumptions regarding the location of deforming and non-deforming regions are made in the construction. From this initial appraisal eight global digital data sets are identified, each rendered on to a regular grid with a spacing of  $0.5^\circ$ .

### 3.1 Earthquake catalogues

Earthquake catalogues contain key information for the definition of active regions (Johnston *et al.* 1994; Bird *et al.* 2002; Schulte & Mooney 2005; Kagan *et al.* 2010; Mooney *et al.* 2012). Thus, studying the corresponding pattern of earthquake distributions and the seismic energy release patterns is a primary goal of this study. For this purpose, we make use of the earthquake catalogue developed by Weatherill *et al.* (2016) which contains 562,840 events with homogenized moment magnitudes in the interval  $2.0 \leq M_w < 9.6$  from 1900 to 2014. The catalogue was compiled from various global databases of earthquakes including the ISC Reviewed Bulletin (International Seismological Centre 2011), ISC-GEM (Storchak *et al.* 2013) catalogues and Global Centroid Moment Tensor database (Ekström *et al.* 2012). We applied to this catalogue the smoothing methodology proposed by Kagan & Jackson (1994), using a smoothing kernel width of 100 km to account for the epicentre uncertainty. Fig. 3 shows the result of the smoothing procedure in terms of scalar seismic moment per unit of area and time.

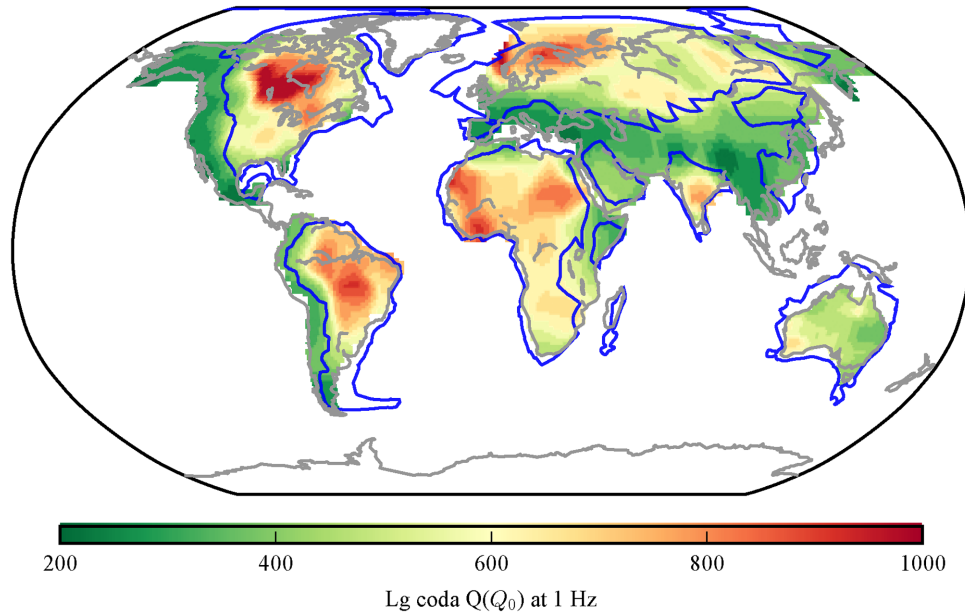
### 3.2 Continent-wide map of 1 Hz Lg coda $Q(Q_0)$

$Q_0$  is a critical property describing the anelastic attenuation of ground motion and has been observed to vary significantly in different tectonic environments. The cause of these variations is posited to be the variation in the density of fluid-filled cracks in the crust (Mitchell *et al.* 1997). Mitchell *et al.* (2008) mapped the  $Q(Q_0)$  for 1 Hz Lg coda waves across Eurasia, and similar results for Africa, Australia, North America and South America have been made available to us by Mitchell (personal communication 2012). The composite map of 1 Hz Lg coda  $Q$  for the five major continents is shown in Fig. 4. It has also been suggested that  $Q_0$  is directly proportional to upper mantle shear wave velocity variations and exhibits low values in crustal regions where seismicity, as well as crustal strain, are high (Mitchell *et al.* 2008). As a result, the  $Q_0$  data set has the potential to provide a good constraint on tectonic regime classification.

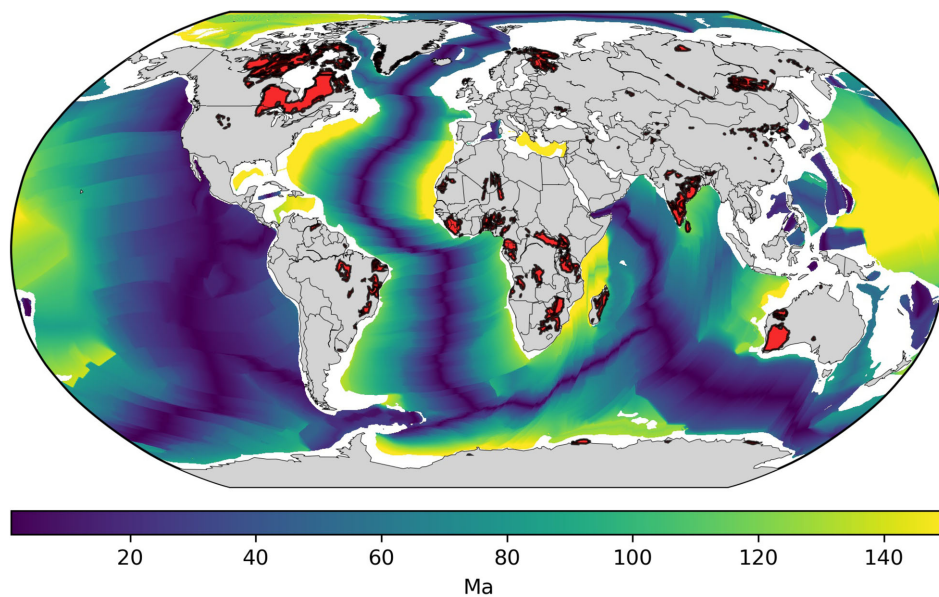
### 3.3 The geological map of the world

The global geological map has been released in 2011 from the Commission for the Geological Map of the World (<https://ccgm.org/en/>, last accessed 2018 February 9). It represents the distribution of the main lithostratigraphic units and the main geological structural features on the Earth surface including the chronostratigraphic units defined for the onshore areas. This geological map classifies the age of the unit according to that of the outcropping rock, which for the current purposes we assume to be an appropriate proxy for the relative age of the lithostratigraphic unit. From this data the Palaeoproterozoic to Archean units are extracted and used as classifier to identify cratonic regions, which are shown in Fig. 5.

The use of geological data in a quantitative framework such as that adopted in due course presents many problems, particularly when seeking global data sets of information. Amongst these are



**Figure 4.** A global data set of 1 Hz Lg coda  $Q(Q_0)$  (Mitchell, personal communication 2012). The stable continental region polygons (in blue) defined by EPRI (Johnston *et al.* 1994) are superposed on the map.



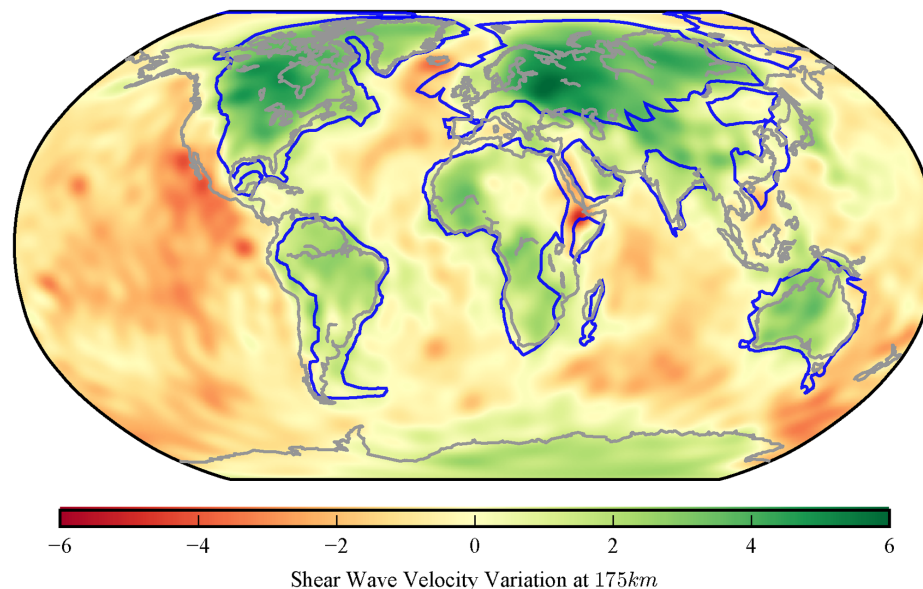
**Figure 5.** Location of oceanic crust and its corresponding age, according to Müller *et al.* (2008), and the location of Palaeoproterozoic to Archean stratigraphic units (onshore red polygons) identified by the Commission for the Geological Map of the World (<https://ccgm.org/en/>)

varying definitions of the geological age of the units and the means by which these are measured, as well as the spatial resolution of the data itself. A trade-off exists between the use of a global data set such as this, in which spatial coverage is achieved at the expense of both the range and accuracy of the reported geological attributes, and a mosaic of local geological data sets, in which further detail may be found but heterogeneity is present in both the definitions and measurement techniques. For the current purpose categorical estimates of age rather than precise quantitative measures are sufficient as a broad classifier of older or younger environments. It is not immediately evident that further refinements of the estimates of geological age provide significant value in predicting regional variation of properties of the continental crust that influence the

characterization of the seismogenic source or ground motion attenuation.

### 3.4 Shear wave velocity variations at 175 km depth

Mooney *et al.* (2012), using a joint analysis of global seismicity and seismic tomography, find that the seismic potential of continental intraplate regions is correlated to the seismic properties of the lithosphere. They use the map of shear-wave velocity variations ( $\delta V_s$ ) from the global shear wave model S40RTS (Ritsema *et al.* 2011) and demonstrate that Archean and Early Proterozoic cratonic lithosphere with higher than average shear-wave velocity variations ( $\delta V_s$ ) at a depth of 175 km have fewer crustal earthquakes



**Figure 6.** A global map of *S*-wave velocity variations at a depth of 175 km developed by Ritsema *et al.* (2011). The stable continental region polygons (in blue) defined by EPRI (Johnston *et al.* 1994) are superposed on the map.

and a lower value of the maximum moment magnitude. Because of the thicker and cooler lithospheric roots beneath the cratons, shear wave velocity variations at 175 km (Fig. 6) exhibits a good correlation with previously assigned stable continental craton regions, which provide a key in mapping the stable continental region in our classification scheme.

### 3.5 Slab 1.0: a 3-D model of global subduction zone geometries

The Slab 1.0 model (Hayes *et al.* 2012) is a collection of geometries defining the subduction interfaces of the twelve most active subduction slabs covering approximately 85 per cent of the active subductions on the globe. We use this subduction interface geometry model into our tectonic regionalization scheme to separate subduction regions from active shallow regions.

### 3.6 Digital model of plate boundaries

A global set of present-day plate boundaries on the Earth are presented in digital form by Bird (2003). From this data set, we extract the subduction related plate boundaries to construct the subduction interfaces for those plate boundaries classified by Bird (2003) as subduction but are absent from Slab 1.0.

### 3.7 Global subduction parameters

In order to define the slab geometry parameters (slab dipping angle, slab azimuth, and the horizontal distance of seismogenic area on slab), we use the online subduction database (<http://submap.gm.univmontp2.fr>, last accessed 2018 February 9) developed by Heuret & Lallemand (2005) to model the other 15 per cent subduction zones which are absent from Slab 1.0 (Hayes *et al.* 2012).

We use the Slab 1.0 model as the major global subduction model since the Slab 1.0 model has been considered as a more updated model and it includes some additional data such as bathymetry,

trench sediment thicknesses and the interpretations of shallow slab interfaces from active source seismic data respect to the Heuret & Lallemand (2005) subduction model.

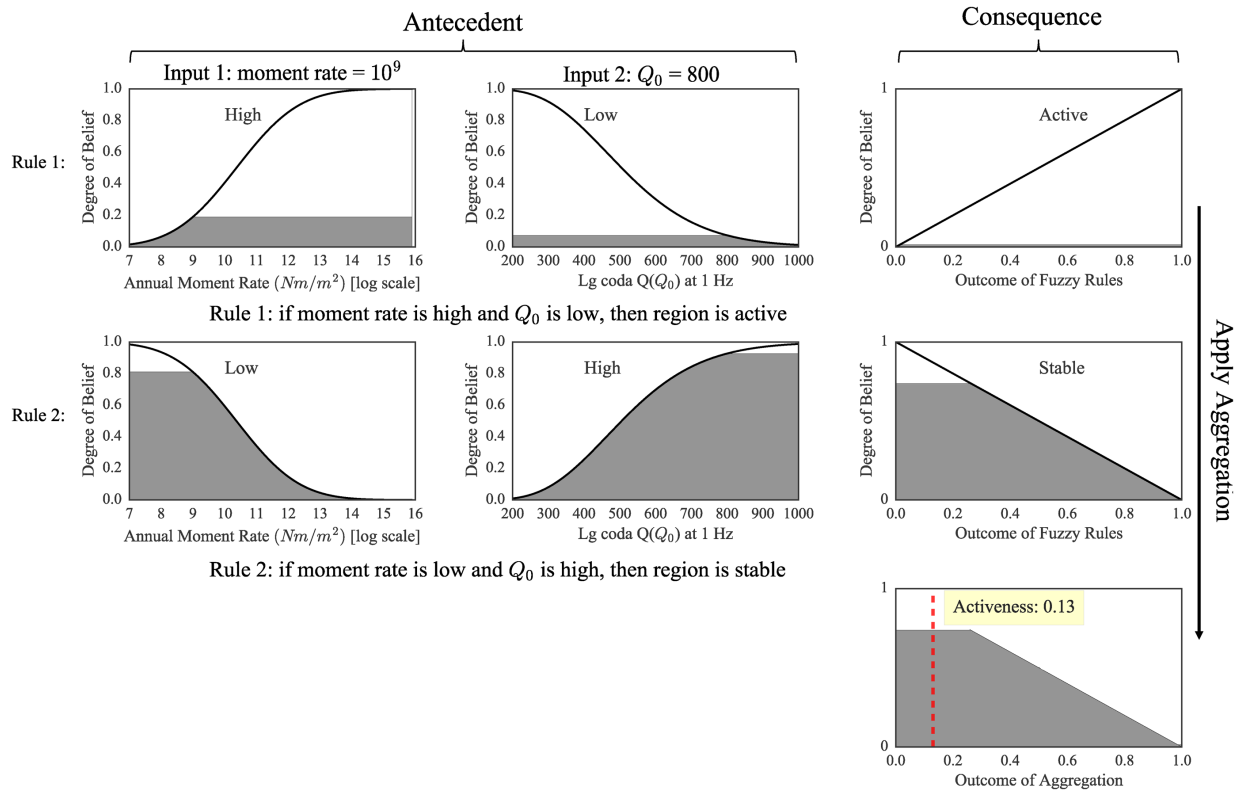
Therefore, with the information of trench coordinates from Bird (2003), and the slab geometry parameters by Heuret & Lallemand (2005), we are able to construct approximate planar 3-D subduction slabs and distinguish the subduction zone from the active shallow region in our tectonic regionalization model.

### 3.8 Oceanic crustal age measurement

For identifying those areas of the Earth surface that are oceanic crust the oceanic crustal age model of Müller *et al.* (2008) is adopted. As discussed subsequently, for application to the assessment of seismic hazard onshore the primary purpose of this classification is to simply distinguish oceanic crust from any other type of environment. As such, this data set is a binary classifier, with a given location being considered oceanic crust if any age estimate is present. For other applications the specific age of oceanic crust in each location may be of interest and further refinements could be considered in order to distinguish, for example, convergent boundaries subducting younger or older crust.

## 4 THE FUZZY LOGIC FRAMEWORK IN TECTONIC REGIONALISATION

Hard or Crisp classifications, that is, classifications without overlap or ambiguity, have often been widely adopted in tectonic regionalization, such as that underlying the Global ShakeMap Atlas (Allen *et al.* 2008; García *et al.* 2012). However, considering the complexity, uncertainty and vagueness in tectonic regionalization, the imposed constraint may be insufficient to correctly classify the large variety of tectonic regimes worldwide. The use of this classification can lead to abrupt changes in tectonic regime over short distances owing to small perturbations of the data samples that lie near a boundary. Furthermore, defining the threshold value of a given quantity as the primary criterion for an assignment to a particular class can be difficult to control and arbitrary. As an alternative, uncertainties, and vagueness can be dealt with a fuzzy classification



**Figure 7.** A Mamdani-type fuzzy inference process for determining activeness of the region. Two inputs are used in this example: seismic moment rate density and quality factor. In this example, the input seismic moment rate density is  $10^9$  ( $\text{Nm km}^{-2}$ ), and for  $Q_0$  is 800.

to be more flexible and tolerant of imprecise data. Through a fuzzy inference system, crisp quantities are instead translated into human linguistic framework, through which an aggregation of individually imprecise decisions can yield a quantitative, reproducible and meaningful output for decision making. The final result is determined via fuzzy logic based deducing mechanism, which is comprised by IF-THEN rules, membership functions and fuzzy logical operations. In short, this process can simulate, in quantitative terms, a set of informed yet imprecise judgements and decisions that would be typical of an ‘expert judgement’ classification. Applications of fuzzy methods in the geoscience field have been proposed in the past decades (e.g. Nordlund 1996; Nikraves & Aminzadeh 2001; Champati ray *et al.* 2007; Grandjean *et al.* 2007; Ansari *et al.* 2015) due to its power on data fusion and as an efficient tool to manipulate the uncertainty related to the interpretations.

In this study, the Mamdani type fuzzy inference process (Mamdani 1974) is applied. It consists of six steps. An example of ‘The activeness of the region’ is introduced to illustrate the steps of fuzzy inference process (Fig. 7).

#### 4.1 Step1: determine a set of fuzzy rules

The fuzzy rules are a collection of linguistic statements that describe how the fuzzy inference system should make a decision regarding classifying an input or controlling an output. Following the fuzzy logic principle and the human thinking process, we define two following fuzzy rules:

- (i) IF seismic moment rate density is ‘High’, THEN the region is ‘Active’.
- (ii) IF seismic moment rate density is ‘Low’, THEN the region is ‘Stable’.

The input variable can be multiple instead of single in the fuzzy rules and be considered simultaneously by connecting with logical operations (e.g. AND/OR). For example, in this problem, instead of considering only the earthquake source, we would like to add information on the crustal wave propagation by adding the quality factor ( $Q_0$ ) as another input variable. Then the fuzzy rules can be set up as:

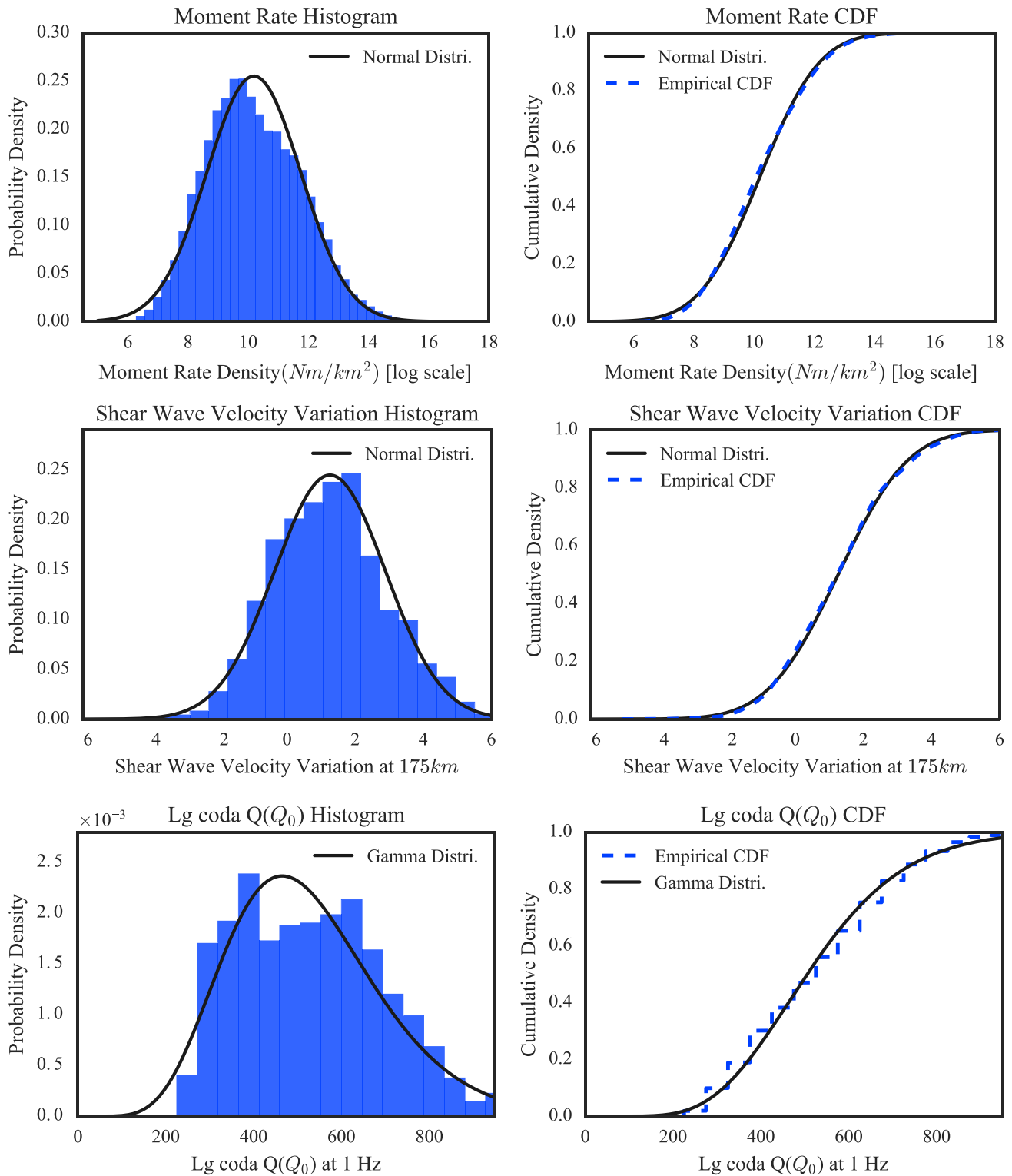
- (i) IF seismic moment rate density is ‘High’, AND quality factor is ‘Low’, THEN the region is ‘Active’.
- (ii) IF seismic moment rate density is ‘Low’, AND quality factor is ‘High’, THEN the region is ‘Stable’.

#### 4.2 Step 2: fuzzify the input variables via membership functions

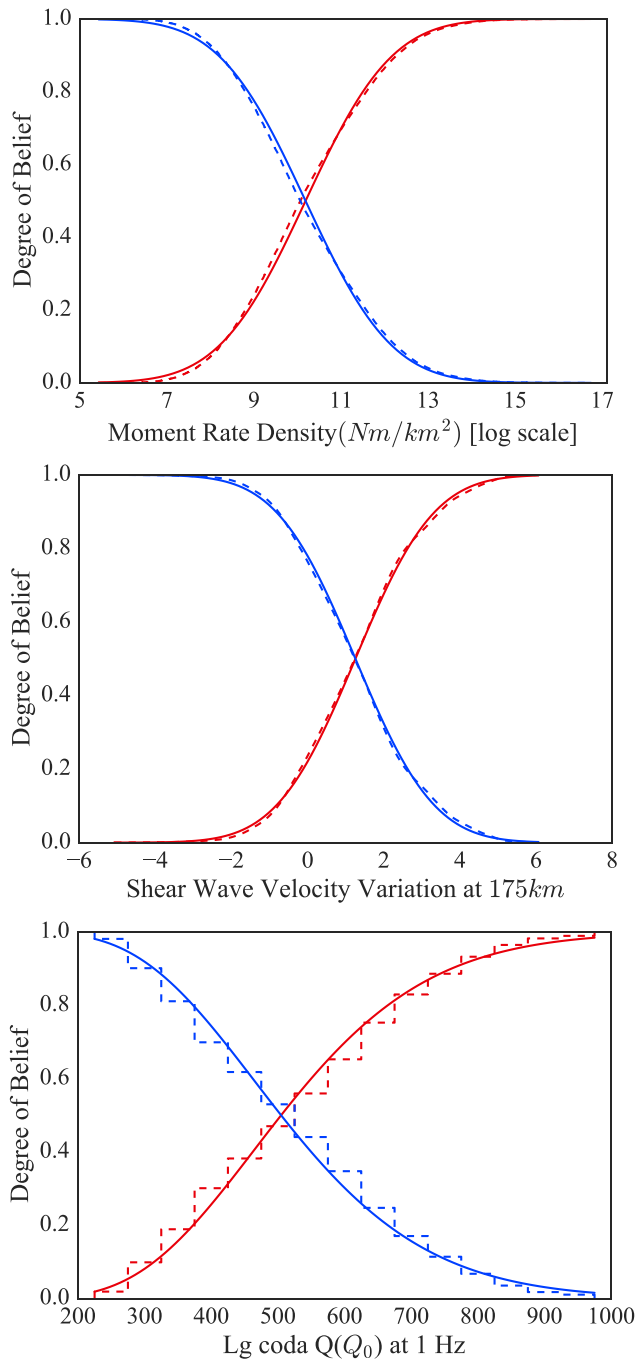
It is necessary to construct membership functions that define what we mean by, for example, high seismic moment rate density and low seismic moment rate density. In fuzzy logic, the truth of any statement becomes a matter of degree. We calculate for the input variables, here the seismic moment rate density and the quality factor, the degree of being ‘Low’, and being ‘High’ via membership functions (a process called *fuzzification*). To design the membership functions of ‘High’ / ‘Low’ seismic moment rate density or ‘High’ / ‘Low’ quality factor, we follow a data-driven approach.

The general approach adopted here uses the histogram plots of the global data sets, and fits them with a continuous distribution. Initially we take the logarithm of seismic moment rate density and generate the histogram plot to observe the global data pattern. In this instance we fit the histogram with a normal distribution (Fig. 8, top left), and take the cumulative distribution function (CDF) of the fitted normal distribution as a membership function of seismic





**Figure 8.** Each panel in the left column shows the histogram of a parameter utilized in this study. The black line also shows the analytical distribution used to describe each empirical distribution. In the right column, we show the corresponding plot in terms of cumulative distributions. Top row: the data of the global seismic moment rate density ( $Nm km^{-2}$ ). The fitted distribution model chosen here is normal distribution with mean ( $\mu$ ) = 10.19, and standard deviation ( $\sigma$ ) = 1.56. Middle row: shear wave velocity variation at 175 km data fitted with normal distribution ( $\mu$  = 1.26, and  $\sigma$  = 1.63). Bottom row: global  $Q_{LG}$  data fitting with Gamma distribution ( $a$  = 8.79, and  $b$  = 59.71).



**Figure 9.** The empirical cumulative curves (dashed lines) and the corresponding membership functions (solid lines) for each linguistic statement of being high (the curves in red) / being low (the curves in blue) for seismic moment rate density, Lg coda  $Q(Q_0)$ , and shear wave velocity variation at 175 km.

moment rate density being ‘High’ (Fig. 8, top left). We then take its complementary function as membership function for defining the extent to which the seismic moment rate density is ‘Low’. The best fitting normal distribution in this example, has a mean value of 10.19, and standard deviation as 1.56. We repeat the process in a similar manner for the quality factor ( $Q_0$ ). The result obtained is shown in Fig. 9.

From these results we obtain a set of data-driven membership functions for the input variables. For example, if the annual seismic

moment rate density at the target site is  $10^9$   $\text{Nm km}^{-2}$ , by using the membership function defined above, the degree of being high seismic moment rate density is 19 per cent, and the degree of being low  $Q_0$  is 7 per cent. (as shown in Fig. 7, top row).

### 4.3 Step 3: combine the fuzzified input variables

In this example, the antecedent of each rule consists of two fuzzy linguistic sets, for example, ‘seismic moment rate density is high’ and ‘quality factor is low’. Here the fuzzy operator is required to combine the two membership values to obtain one numerical value that represents the result of the antecedent for the rule. Fuzzy combinations, ‘AND’, ‘OR’ and sometimes ‘NOT’, can be used. There are many definitions of these fuzzy combinations, but here we introduce only the ones applied in the study. The fuzzy rule, ‘IF seismic moment rate density (A) is ‘High’, AND quality factor (B) is ‘Low’, is written as:

$$\mu_{(A \cap B)} = T(\mu_A(x), \mu_B(x)), \quad (1)$$

where  $\mu_A$  is the membership of A, and  $\mu_B$  the membership of B. T is the result of antecedent for this rule.

There are many ways to compute the fuzzy ‘AND’ combination. Here we apply the technique taking only the product of the two membership values, so in this case, both the membership values are taken into consideration during the process as illustrated in eq. (2):

$$\mu_{(A \cap B)} = T(\mu_A(x), \mu_B(x)) = \mu_A(x) \cdot \mu_B(x). \quad (2)$$

Therefore, when seismic moment rate density is  $10^9$  ( $\text{Nm km}^{-2}$ ), and the  $Q_0$  is 800, the degree of high seismic moment rate density and the degree of low  $Q_0$  is 19 per cent and 7 per cent respectively. Using eq. (2), the result of the antecedent for this rule is 1.33 per cent. Likewise, for the second fuzzy rule: IF seismic moment rate density is ‘Low’, AND the quality factor is ‘High’, the degree of low seismic moment rate density and the degree of high  $Q_0$  are 81 per cent and 93 per cent respectively. The result of antecedent for this rule is 75.33 per cent.

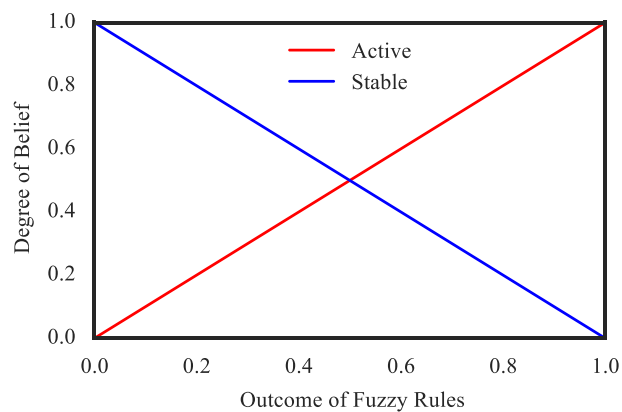
### 4.4 Step 4: Obtain the consequence of the rule

To map the result of fuzzy rules into output linguistic terms (region is ‘Active’/ region is ‘Stable’), output membership functions are required. Here, we simply use a linear function (eq. 3) and its complement function (eq. 4) as membership functions for its region being ‘Active’ and ‘Stable’, respectively, (Fig. 10):

$$f(x) = x \quad (3)$$

$$f(x) = 1 - x. \quad (4)$$

The output variable, degree of being ‘Active’ or of being ‘Stable’, is in the parameter space between 0 and 1. With the membership function of the output variable, we are able to translate the fuzzy result into a single scalar quantity (called defuzzification) indicating the membership for activeness of the region (degree of ‘Active’ or ‘Stable’). For example, from step 3 the results of antecedent for rule 1 and rule 2 are 1.33 per cent and 75.33 per cent respectively. The output membership functions as the result of consequence for each rule are illustrated in Fig. 7.



**Figure 10.** The membership functions for the output linguistic terms. The red curve is the membership function of being in an active region (named ‘Activeness Index’ as well in the study), instead, the blue curve is for being in a stable region.

#### 4.5 Step 5: apply an aggregation method to combine the fuzzy outputs

The final decisions are based on all the rules in a fuzzy inference system, thus the rules must be combined into a single fuzzy set. This process is called aggregation, by which the fuzzy sets that represent the outputs of each rule are combined into an output fuzzy distribution.

The concept of ‘MAXIMUM’, ‘SUM’, and ‘OR’ could each be applicable for the aggregation operation. In this work, the function ‘algebraic sum’ is applied. The aggregation function of ‘algebraic sum’ works as union of the outputs of all the fuzzy rules (Fig. 7, last column). The result of aggregation (fuzzy output distribution) is shown in the bottom right part of Fig. 7.

#### 4.6 Step 6: apply defuzzification from the output distribution

The last step of the fuzzy inference process is *defuzzification*, which provides a single scalar quantity for a potential further classification. As the name implies, *defuzzification* is the opposite operation of fuzzification, where one extracts a precise quantity out of the range of the fuzzy set to the output variable. The defuzzification method we apply is called Mean Maximum (Sugeno 1985) as described in eq. 5). It takes the output distribution and finds its average of maximum (from the plateau) values along the  $x$  axis.  $z^*$  is the average of maximum level along the  $x$  axis,  $a$  and  $b$  is the minimum and the maximum values of the plateau level along the  $x$  axis.

$$z^* = \frac{a + b}{2}. \quad (5)$$

Via the Mean Maximum method, in the example of ‘the activeness of the region’, given seismic moment rate density released is  $10^9$  ( $\text{Nm km}^{-2}$ ), and  $Q_0$  is 800, the fuzzy inference system rates the value of activeness of the region is 0.13. From the same fuzzy inference system, when the seismic moment rate density released is  $10^{12}$  ( $\text{Nm km}^{-2}$ ), and  $Q_0$  is 500, the activeness of the region is 0.68, which is evidently more active than the first case.

The core idea of a soft classification is to replace the two binary logical statements, True and False, by a continuous range between 0 and 1, with all values between representing a transition, which can be modelled quantitatively using functional forms inferred by judgement and/or by data. Where desired, a delineation can be made by using the defuzzification result as the reference input variable for

the classification. According to the application, a user may select the threshold value which fits their own goal for the delineation. Where the fuzzy approach adds significant value in the classification process is the fact that even when a binary classification is a desired output the existence of the memberships provides more information from which the decisions can be understood and, where necessary, challenged. As indicated in Section 1, a binary distinction between ‘active’ and ‘stable’ is not consistent with the physical properties of the Earth and there exists a continuum within these definitions. This continuum is now quantified within the degrees of membership, opening the possibility to both refine the classification to identify intermediate positions and locate those regions on the Earth.

## 5 FUZZY CLASSIFICATION FOR ACTIVENESS OF REGION

The overarching goal of this process is to determine the activeness of a target site in a manner that is appropriate to the context of seismic hazard assessment. This is undertaken using a fuzzy classification scheme. We incorporate data such as smoothed seismic moment rate density,  $Q_0$ , global composite (?) (B.J. Mitchell, personal communication, 2012), and shear wave velocity variation (Mooney *et al.* 2012). In Figs 3, 4 and 6, we show as a matter of reference for each data set the Stable Continent Region polygons delineated by Johnston *et al.* (1994). From the observations shown and the data properties described in Section 3, the chosen data sets correspond well to the tectonic properties implied by these independently derived polygons (here ‘Active’ and ‘Stable’). Where we expect the region to be more active in general, the seismic activity is higher and quality factor  $Q_0$  lower, and with lower S-wave velocity variation observed in the region. Similarly, in the regions we expect to be more stable, the seismic activity is lower, with a higher  $Q_0$  and a lower S-wave velocity variation.

Following the fuzzy inference system described in Section 4, we have set up rules for the further fuzzy classifications. Since the continent-wide map of 1 Hz Lg coda  $Q(Q_0)$  covers only the continental area and the shear wave velocity variation data pattern shows stronger correlation with onshore tectonic features, the classification for continental and oceanic area has been based on different data sets. The continental and oceanic classification will be explained later in Section 7.2.

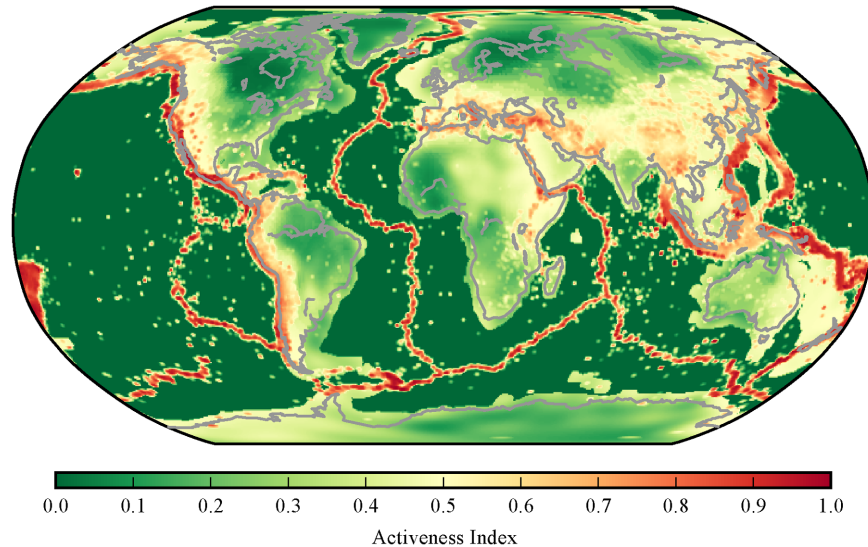
For the continental area, all three data sets are considered in the fuzzy rules:

- (i) IF seismic moment rate density is ‘High’, AND  $Q_0$  is ‘Low’, AND S-wave velocity variation is ‘Low’, THEN the region is ‘Active’.
- (ii) IF seismic moment rate density is ‘Low’, AND  $Q_0$  is ‘High’ AND S-wave velocity variation is ‘High’, THEN the region is ‘Stable’.

For the oceanic area, only seismic moment rate density data are considered:

- (i) IF seismic moment rate density is ‘High’, THEN the region is ‘Active’.
- (ii) IF seismic moment rate density is ‘Low’, THEN the region is ‘Stable’.

Following the same framework as shown in the Section 4, to design the membership functions of each linguistic variable in the fuzzy rules designed in this section, we fit the data with distribution models and take the CDF of the fitted distribution as membership



**Figure 11.** Geographic distribution of the Activeness Index (degree of being active) derived from this study using the fuzzy inference system.

function of being ‘High’. Conversely, we take the survival function as membership function of being ‘Low’ as shown in Figs 9 and 8. The best fitting normal distribution with mean value of 10.19  $\text{Nm km}^{-2}$  and standard deviation of 1.56 for global seismic moment rate density data (in logarithm). For global 1 Hz Lg coda  $Q(Q_0)$ , the best fitting distribution model is a Gamma distribution model with the shape parameter,  $a = 8.79$ , and the scale parameter,  $b = 59.71$ . For shear wave velocity variation at 175 km depth data, the best fitting normal distribution has a mean value of 1.26, and standard deviation of 1.6.

### 5.1 Activeness Index Map from fuzzy inference system

The Earth surface is rendered on to a regular grid with a spacing of  $0.5^\circ$  and this grid is used as a reference for the three input data. The Activeness Index for the centroid of each grid has been calculated using the fuzzy inference system described in Section 4. A global map showing the Activeness Index (inferred as degree of being an active in this study) has then been created and represented in Fig. 11. As expected, the highest values are concentrated along the main subduction zones known globally and are mostly corresponding to the plate boundaries (e.g. spreading ridges). From the map, the gradient change of activeness from more active region to more stable region can be clearly observed.

## 6 FUZZY CLASSIFICATION FOR CRATON REGION

GMPEs for stable continental regions are generally assigned to cratonic regions (low deformation with slow attenuation) and non-cratonic regions (low deformation with fast attenuation) (Akkar *et al.* 2012; Delavaud *et al.* 2012). A craton is defined as a portion of continental crust that has attained and maintained long-term stability, with tectonic activity mostly confined to its margins. Although there is no strict age connotation in this definition, for example, some segments of crust could have attained cratonic stability during the Proterozoic, the term is most commonly applied to stable segments of Archean crust. Long-term stability is thought to be a function, in part, of thicker lithosphere involving a relatively cool but compositionally buoyant keel of Fe-depleted upper mantle

(Bleeker 2003). However, to our knowledge, there is no neat quantitative definition that can be used to distinguish the cratonic region from non-cratonic region in stable continents. On the global scale, it is not trivial to determine if each region has remained undeformed since the Precambrian. Properties such as low attenuation, thick lithospheric mantle root, are mostly distributed in Archen to Palaeoproterozoic crust, however. We use shear wave velocity variation (Mooney *et al.* 2012), 1 Hz Lg coda  $Q(Q_0)$  (Mitchell, personal communication 2012), and worldwide Geologic Stratigraphy map as proxies for these properties.

Using these proxies, we define the following rules:

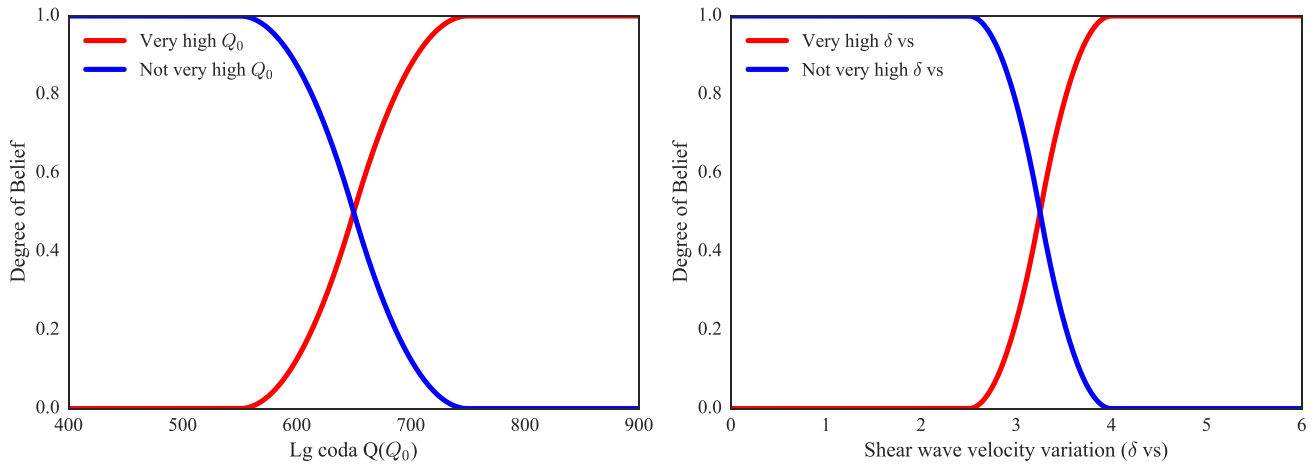
- (i) IF shear wave velocity variation is ‘Very High’, AND  $Q_{LG}$  is ‘Very High’, AND the crustal age is ‘Very Old’, THEN the region is ‘Craton’.
- (ii) IF shear wave velocity variation is ‘Not Very High’, AND  $Q_{LG}$  is ‘Not Very High’, AND the crustal age is ‘Not Very Old’, THEN the region is ‘Non-Craton’.

With these rules established, we then define the membership functions for each linguistic statement. As suggested in Mooney *et al.* (2012), shear wave velocity variation at 175 km has a good correlation with stable continental craton regions. A value of  $\delta V_s$  around 2.5 per cent corresponds to the edges of a craton whereas a  $\delta V_s$  around 3.5 per cent identified the core of the craton. Regions in which  $\delta V_s > 3.5$  per cent do not appear to produce earthquakes of magnitude larger than 6.0.

To define the membership function ‘very high’ shear wave velocity variation ( $\delta V_s$ ), we take the values 2.5 per cent and 4 per cent of  $\delta V_s$  for zero membership and full membership of ‘Very High’ shear wave velocity variation ( $\delta V_s$ ) respectively and then we select an S-shaped function to smooth the in-between values (eq. 6), and a Z-shaped function for the membership function of ‘not very high’  $\delta V_s$  (eq. 7):

Degree of belief, very high  $\delta V_s =$

$$f(x) = \begin{cases} 0, & x < a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a \leq x < \frac{a+b}{2} \\ 1 - 2\left(\frac{x-b}{b-a}\right)^2, & \frac{a+b}{2} \leq x < b \\ 1, & x \geq b \end{cases} \quad (6)$$



**Figure 12.** The membership functions designed for the input variables in the craton classification. Left: the membership functions of being very high  $Q_0$  (red) and not very high  $Q_0$  (blue). Right: the membership functions of being very high  $\delta$ Vs (red) and not very high  $\delta$ Vs (blue)

Degree of belief, not very high  $\delta$ Vs =

$$f(x) = \begin{cases} 1, & x < a \\ 1 - 2\left(\frac{x-b}{b-a}\right)^2, & a \leq x < \frac{a+b}{2} \\ 2\left(\frac{x-a}{b-a}\right)^2, & \frac{a+b}{2} \leq x < b \\ 1, & x \geq b. \end{cases} \quad (7)$$

According to Mitchell *et al.* (2008, 1997) spatial variations of  $Q_0$  show clear correspondence with variations in some upper mantle and crustal properties. It is also expected that higher  $Q_0$  values exist in more stable regions (the low attenuation property in craton region). We observe from the global  $Q_0$  distribution, that the highest  $Q_0$  values usually concentrate in cratons. Using the  $Q_0$  value distribution and the worldwide craton distribution map, we use two values of  $Q_0$ , 550 and 750, which define zero membership and full membership respectively. As in the case of  $\delta$ Vs, we use a S-shaped membership function to define the degree of being very high  $Q_0$ ; the membership function of being very low  $Q_0$  is specified using eqs (6) and (7) (see also Fig. 12).

As cratons are considered to be the oldest parts of continents, in the geologic timescale, the oldest eon (the Archean) has an assigned age of at least 2500 Ma. Most of the cratons are considered to have formed during the Archean (before 2500 Ma) to Palaeoproterozoic (1600 to 2500 Ma), so here we collect the locations of Archean to Palaeoproterozoic crust on the globe. To build the membership function for this categorical data set, we decide that if the geologic unit is Archean, then we assign full membership of being ‘very old in crustal age’. If the geologic unit is Palaeoproterozoic then half-membership of being ‘very old’ is assigned, and if the geologic unit is younger than Palaeoproterozoic then membership of being ‘very old’ is set to zero (eq. 8):

Degree of belief, very old crust =

$$\begin{cases} 1, & \text{for Archean} \\ 0.5, & \text{for Palaeoproterozoic} \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

We apply a Mamdani type of fuzzy inference process again using three input variables ( $Q(Q_0)$ , shear wave velocity variation, and the age of continental crust), and their corresponding membership functions of each statement to obtain ‘The degree of membership to a Cratonic region’ (Craton index) (Fig. 13).

## 7 DIGITAL TECTONIC REGIONALIZATION MODEL

In the active regions, we differentiate between tectonic regimes associated with subduction zones, active shallow crust regions (generally close to plate boundaries where high seismicity occurs in the upper part of the crust) and oceanic regions (earthquakes with mainly oceanic travel paths). In the classification scheme, we differentiate subduction and non-subduction using the slab geometry models (see Section 3). If the target location has been defined as subduction, then an additional separation between interface and within-slab subduction is applied using the location of the investigated site with respect to the location of the slab. For the identification of the oceanic regions, we use the data set describing the oceanic crustal age. If an age is assigned at the target site, then we assume the site belongs to oceanic region.

We subdivide the stable regions according to the crustal material types (i.e. regions of the crust with similar geophysical and geologic characteristics) into continental cratonic regions (low deformation with low attenuation), non-cratonic regions (low deformation but significant attenuation) and oceanic regions (earthquakes with mainly oceanic travel paths). We set the cratonic regions using the results described in Section 6. For a representative delineation, if the degree of cratonic is larger than 0.5, then it is considered as a cratonic region. We consider the cratonic and non-cratonic regions to be mutually exclusive. The determination of oceanic crust in stable region follows the same boundary as in active region.

### 7.1 Identification of subduction zones

In order to define if a target site should be included in a subduction region, we check its surface map position relative to the surface projection of polygons describing the slab surface. If the target site is inside one of these polygons, it is classified as a subduction-related site.

In the subduction-related site, the further discrimination between interface, in-slab and the active shallow zone is performed by taking into account the depth of the target relative to the slab (Fig. 14).

For example, to identify the relative location to the slab at the top shallow layer of the earth (here we define a depth from 0 km to 30 km), instead of planar grids on the earth surface, a cell with volume, with a range of depth is used here for the subclassification.

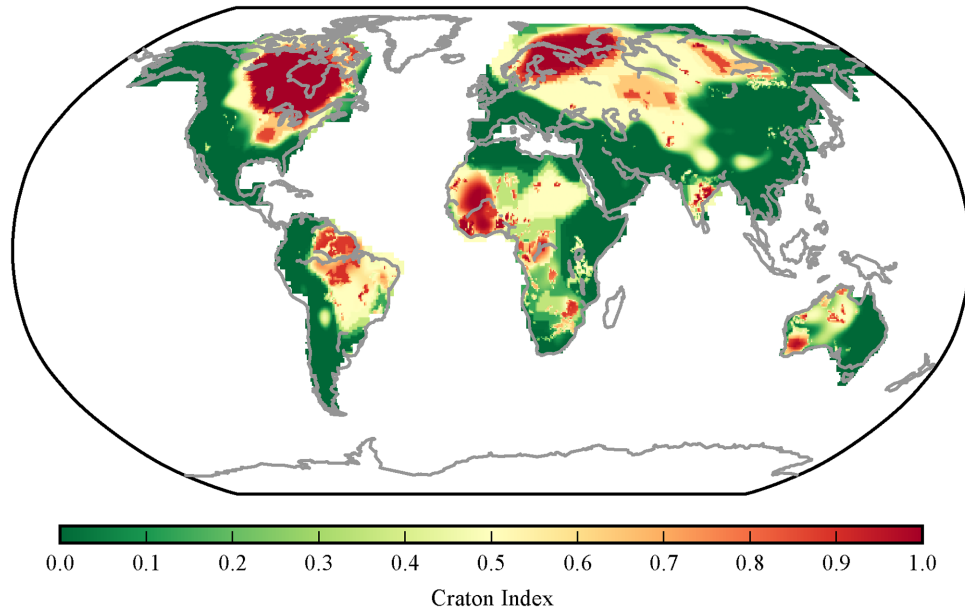


Figure 13. Geographic distribution of the Craton Index (degree of being cratonic) across the globe derived from this study using the fuzzy inference system.

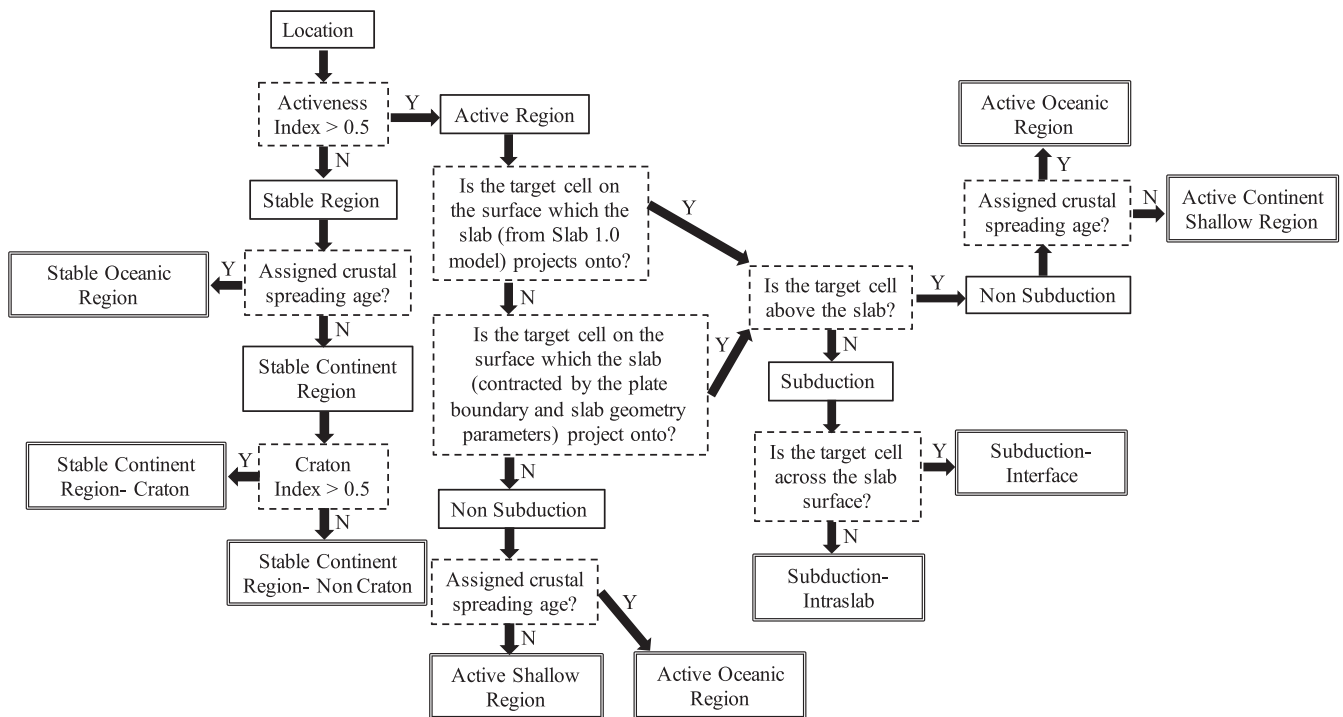


Figure 14. The computational flowchart for the complete tectonic classification.

When the target cell is above the slab, then it is considered as an active shallow region. If the cell is across the slab, then it is classified as Subduction-Interface, or if the cell is below the slab interface, then it is Subduction-Intraslab.

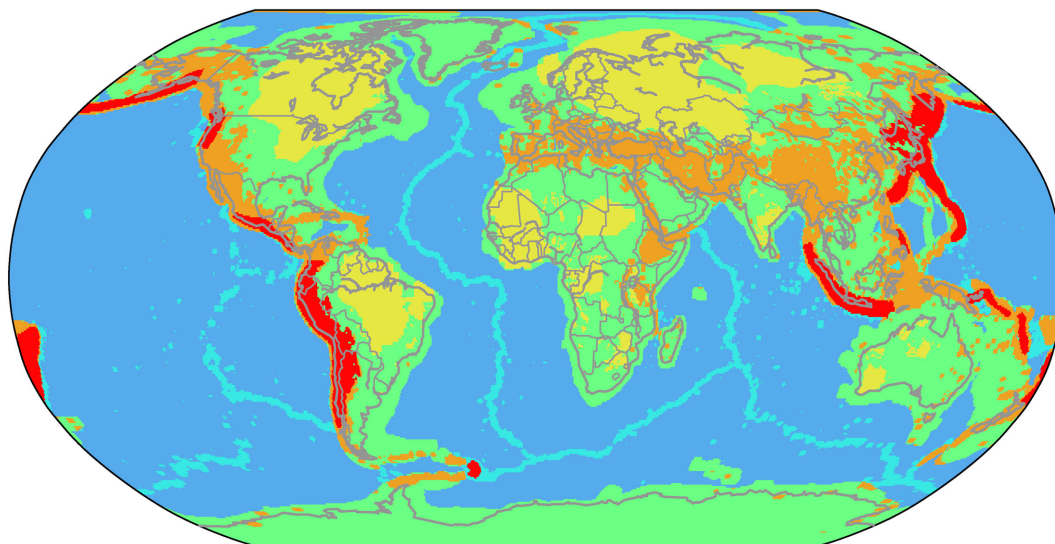
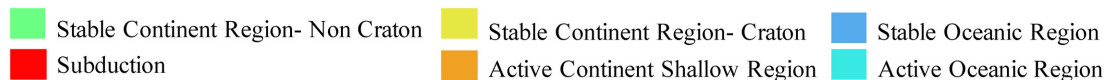
If the target area has a spreading age assigned (no matter what the age is), then it is classified as oceanic region. If there is no spreading age assigned in the target area, then it will be considered as continental region.

### 7.2 Oceanic region and continent region classification

The areas where the travel paths are mainly through oceanic crust exhibit characteristic attenuation properties (Akkar *et al.* 2012). The Müller *et al.* (2008) digital oceanic spreading age measurement has been used as a means of identifying where oceanic crust was found.

### 7.3 An example global tectonic regionalization

We illustrate in Fig. 14 the scheme utilized to provide a global classification of tectonic regimes across the globe. First, for the target site, we check if the activeness (degree of being active) is larger equal than 0.5. If yes, the region will be defined as *Active*



**Figure 15.** The map shows the tectonic regionalization model derived from our study for the top 30 km of the Earth.

Region, and if not, it becomes a *Stable Region*. Depending on the region assigned, a second classification level is applied.

In the case of *Stable Regions*, if the target site has a crustal spreading age assigned, it is classified as oceanic. Otherwise, we check the corresponding degree of being craton (see Fig. 13). If the value is greater than or equal to 0.5, the target site is classified as craton, otherwise, it will be assigned to a non-craton region.

For *Active Region*, we check if the target site belongs to a subduction region. If yes, it will be classified into *Subduction Region*. For *Non Subduction Region*, we check if the target site has been assigned a crustal spreading age. If yes, the region is classified as *Active Oceanic Region*, and if not, the region is classified into *Active Shallow Region*.

By applying for each site of a global regular grid with a spacing of  $0.5^\circ$  the classification scheme just described we obtained a global tectonic regionalization model. In Fig. 15, we show the result of this operation for the shallow layer (0–30 km).

## 8 COMPARISON WITH EXPERT-DRIVEN (GROUND MOTION DATABASES) CLASSIFICATION

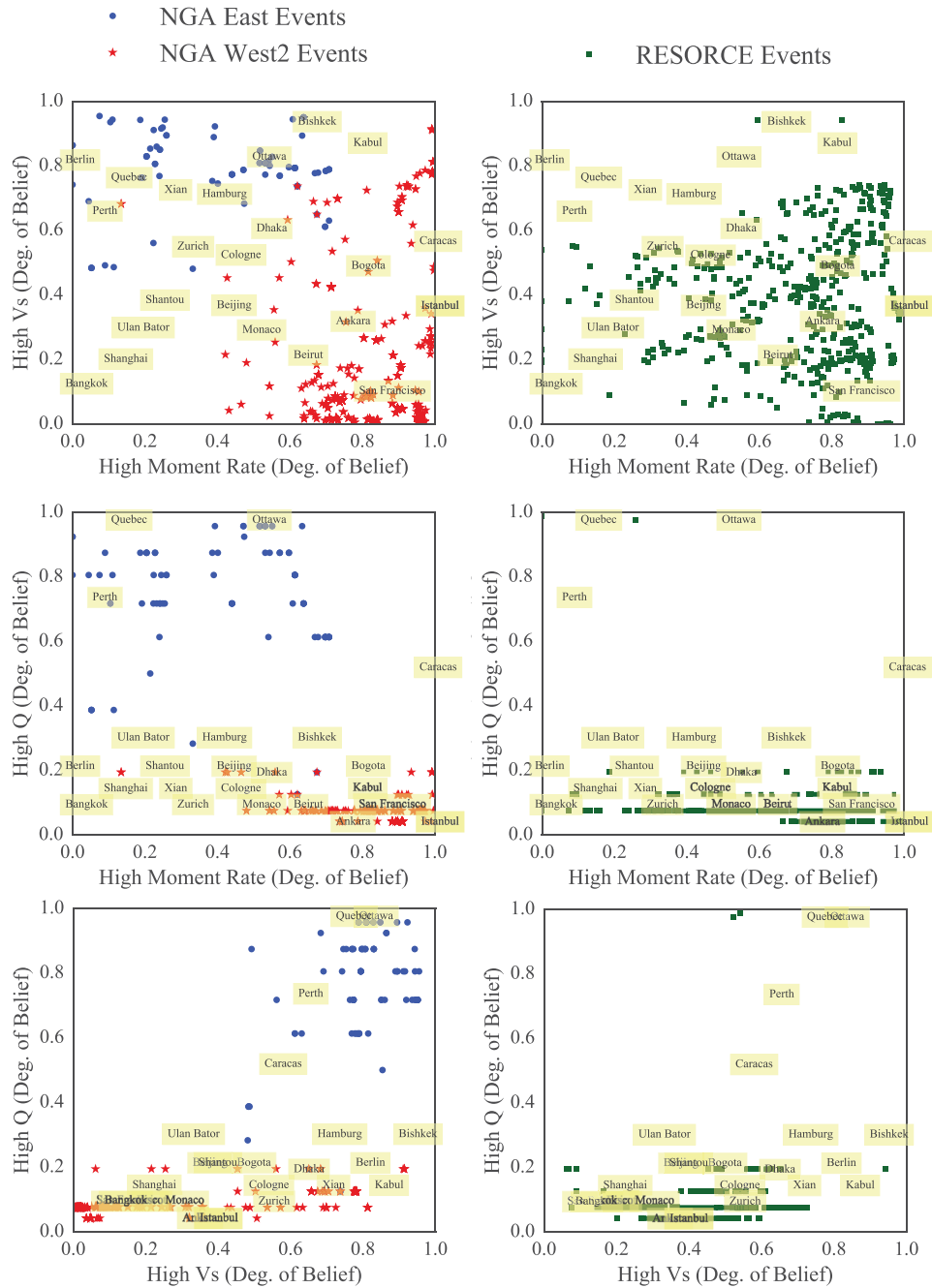
Expert-driven rules have been applied to classify the earthquakes included in the Pacific Earthquake Engineering Research Center NGA-West 2 ground motion database (Ancheta *et al.* 2014) and the European RESORCE database (Akkar *et al.* 2014) for the active regions, and for the NGA-East database (Goulet *et al.* 2014) for the stable regions. Reiterating the seismic hazard assessment context, we make use of these databases in which a prior regionalization is assumed, in order to understand the performance of our data-driven classification result against the expert-driven classification.

From the location of each earthquake, we find the cell in the global regular grid containing its epicentre and consequently we obtain the seismic moment rate density ( $\text{Nm km}^{-2}$ ), Lg coda  $Q$  ( $Q_0$  value, and shear wave velocity variation value to calculate the degree of ‘High’ seismic moment rate density, ‘High’ Lg coda  $Q$  ( $Q_0$ , and ‘High’ shear wave velocity variation according to the

membership functions we have designed (Fig. 9). As shown in Fig. 16, the NGA-West 2 events mostly locate in areas with higher seismic moment, lower  $Q_0$  and lower shear wave velocity variation. NGA-East events on the other hand show low-to-median seismic moment, higher  $Q_0$  and shear wave velocity variation. This observation confirms the ability of the selected parameters to discriminate between active and stable regions defined on the basis of common expert understandings.

The top panel in Fig. 17 illustrates the frequency distribution of the activeness index values obtained for all the earthquakes in the NGA West 2 and NGA East databases. The histograms obtained for the two data sets show a clear separation and distinct domains. This plot demonstrates that in this case, the resolution ability of the computed activeness index is high. In the bottom panel, we show the results of a similar analysis obtained using the earthquakes included in the RESORCE database (Akkar *et al.* 2014), a ground motion database from Europe and the Middle East. This histogram also shows that many earthquakes in the RESORCE database are, in general, closer to the NGA-West event cluster than the NGA-East event cluster although the histogram appears slightly left-skewed due to the low-to-intermediate levels of seismicity and low-to-intermediate levels of crustal attenuation (Fig. 16)

To further illustrate the relevance of the classification in a seismic risk context, it is applied to a selection of major cities from across the globe. The choice of cities reflects the variety of tectonic as well as sociopolitical environments. From the city distribution on the three diagrams (Fig. 16), many cities depart from the general definition of either active region or stable region. Cities such as Paris, London, Cape Town, Bangkok and Ulan Bator, have been located at sites with low seismic moment rate density but also low  $Q(Q_0)$  and low shear wave velocity variation. On the other hand, cities such as Caracas are located at sites with high seismic moment but high  $Q(Q_0)$  and high shear wave velocity variation examples. This evidently suggests the existence of the relevant intermediate environment between the active region and the stable region and echoes the need to adopt a fuzzy framework for the propose.



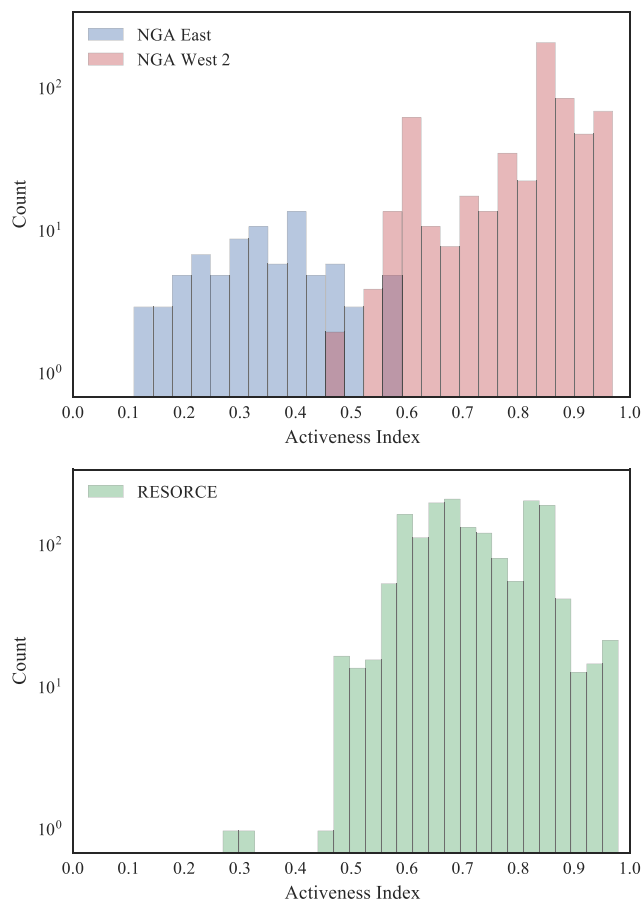
**Figure 16.** Left top: the distribution for degree of belief in high seismic moment rate density and high  $\delta V_s$  for NGA-West2, NGA-East earthquake events. In each plot, as a reference, we also show the corresponding memberships obtained for some major cities. Left middle: the distribution for degree of belief in high seismic moment rate density and high  $Q_0$ . Left bottom: the distribution for degree of belief in high  $\delta V_s$  and high  $Q_0$ . Right top: the distribution for degree of belief in high seismic moment rate density and high  $Q_0$ . Right middle: the distribution for degree of belief in high seismic moment rate density and high  $Q_0$ . Right bottom: the distribution for degree of belief in high  $\delta V_s$  and high  $Q_0$ .

From the distribution of activeness obtained in our study for NGA-East and NGA-West 2 events (Fig. 17, left panel), the fuzzy inference system presented successfully reproduces an existing classification from the two databases. The delineation of two databases is roughly located at an activeness index equals to 0.5 (the median value of activeness index) which has been taken as the threshold value for the delineation for active and stable regions as mentioned in Section 7.

## 9 DISCUSSION

In the scheme of calculating the activeness index, three global data sets (seismic moment rate density,  $Q_0$ , and  $\delta V_s$ ) are used. We demonstrated the capability of the three variables chosen to sufficiently classify the earthquake events occurred in active region (NGA-West 2) and stable region (NGA-East), and to mimic the result of an expert-driven classification. As the whole regionalization scheme is written in a reproducible computational framework,





**Figure 17.** Classification in terms of the Activeness Index of earthquake epicentres included in different databases. Top: NGA-West and NGA-East databases. Bottom: RESORCE database.

it is readily adaptable to plug new data sets into the classification scheme. We do not discount the possibility that other data sets may contribute additional value to this process. Initially, a wider selection of global geophysical databases was appraised, whose properties might be expected to correlate well with certain elements of the earthquake process, but several were discarded. As indicated previously, the global strain rate model, which, although a strong indicator of seismic potential, has a large number of modelling assumptions implicit in its creation, including the prior assignment of deforming and non-deforming cells. The global stress drop database of Allmann & Shearer (2009) was also studied carefully, and whilst stress drop is arguably one of the most important parameters in the earthquake process the spatial sampling was still too sparse to capture the potential stress drop distribution and its spatial variability in a manner that would be meaningful for an analysis such as this. In addition, the global heatflow database (compiled by University of North Dakota) was also consulted, yet the resulting heatflow trends in continental lithosphere remained difficult to interpret in the context of the other data sets, displaying little coherence with other properties of the lithosphere. As such, it was perceived as redundant, with the more pertinent geophysical properties of the crust better captured by the  $Q_0$  and  $\delta V_S$  data sets. These decisions could potentially be revised as the databases themselves develop and new research emerges. They could also be reconsidered if wishing to refine the regionalization at a local or regional scale.

In this study, all data in the classification are equally weighted, inferring the same importance for each data set within the scheme. Various weighting schemes could be adopted if we believe that cer-

tain data are more important for reasons such as its relevance to the problem and/or its reliability and means of acquisition. In the current development, the equal weighing approach gives a satisfactory classification result. An investigation to identify the relevance or the reliability of data to derive a more informative weightings for the classification can be done to explore the possibility to improve the classification result, however, it is beyond the scope of this study.

Our classification of the cratonic regions shows results in agreement with Tang *et al.* (2013), who has defined the craton region at a global scale from a petrological and geochemical perspective. Some discrepancies are nonetheless observed. In Australia, for example, if we compare our results against the ones of Burbidge *et al.* (2012), the cratonic areas appear much narrower in our study than in their model. This might be explained by the use of a broader definition of the cratonic areas (they have included also the non-reactive Proterozoic crust region).

For the subduction classification, we directly apply the digital global subduction geometry models as the primary source of information to inform the classification. Potential further refinements could be possibly made by adding additional information such as bathymetry and Moho depth. No attempt is made in this analysis to distinguish between different types of subduction environments, which may be particularly pertinent in attempting to relate geophysical properties of subductions zones to parameters controlling earthquake recurrence such as maximum magnitude or seismogenic coupling. It is possible to speculate on accessible measures that could further discriminate between different types of subduction zones, including oceanic crustal age or fractal coefficient of roughness in the bathymetry. For a particular location of interest, these could help elucidate those subduction regions that provide the most suitable analogues (in terms of seismogenic productivity or limits on rupture extent) as that under consideration.

For the oceanic regions, we adopt the digital oceanic age model by Müller *et al.* (2008) as a reference data set, from which membership of the oceanic crust categories is simply assigned. We recognize our current approach might not only be affected by high uncertainties, it also neglects the important transition from subduction forearcs to stable oceanic lithosphere.

The capacity to make inferences from data is inseparably connected with uncertainty, which includes the uncertainties in data generation process (e.g. location, magnitude etc. for the seismicity), the uncertainties in the classification algorithm (such as the fuzzy rule induction), or the fuzzy operation selection. Many of these uncertainties are more of an epistemic, and would hopefully be reduced over time as more data are collected. Decreasing the spatial cell resolution may be possible in some regions, though this should be approached with caution. The fuzzy methods incorporate imprecision quite elegantly into the classification process, yet downscaling the spatial resolution may provide a diminishing return as there can exist the potential for the process to over model the apparent spatial variability within a particular data set.

Questions regarding the resolution of the partition, both in terms of spatial scale and of further subdivision of the classes, inevitably lead to the consideration of how to compare, in a quantitative sense, one regionalization against another. Here the context is critical. For selection and application of ground motion models, one can attempt to measure the suitability of a given regionalization using the distribution of ground motion residuals with respect to a single reference model. From an undifferentiated database of ground motion records, it is possible to use the random effects residuals from nonlinear mixed regression of the reference model to determine if a given random effect (i.e. for a given region type) is significant. This concept is illustrated clearly by Stafford (2014). Assessment of the

overall fit, including an appraisal of the parsimony within the regionalization, could be quantified from by comparing the total fits of the regionalized ground motion models. Information theoretic methods may be well suited to this problem in order to penalise over-fitting. Similarly, for seismogenic source modelling a natural test is to use the tectonically analogous regions to develop a smoothed seismicity recurrence model that can be applied within a prospective (or pseudo-prospective) test. In this strategy, it can also be assessed whether further a given regionalization improves the prediction of activity rates for difference magnitudes across a global grid (Bird *et al.* 2015, e.g.). An objective framework for testing how regionalization can improve seismic hazard models is something that should be considered as regionalization itself plays a greater role in both seismogenic source and ground motion modelling.

The outcomes of this work are particularly relevant in the context of the construction of new seismic hazard models. The methodology and the results discussed, are useful tools for the construction of components of seismic hazard models such as the selection or calibration of appropriate GMPEs at regional and national scales (e.g. Delavaud *et al.* 2012). For each site one can obtain a subjective probability of membership of different tectonic classes, which can be mapped directly into analyses of epistemic uncertainty via logic-trees, a widely used tool for quantification of epistemic uncertainties in PSHA. Our revised regionalization model can be used in the GMPEs development process by classifying the seismograms according to the tectonic regime in which they belong.

An interesting circumstance for which the fuzzy methodology may provide new insight is when the earthquake and the receiver (the target site) are located in different tectonic regimes. Currently these cases are seldom addressed, but can be of particular relevance when investigating hazard in regions where high crustal deformation (notionally active) transition rapidly into low deformation (notionally stable). Some possible uses of the degrees of membership returned by the fuzzy approach could be for the assignments of weights for the GMPEs in different tectonic regimes according to the relative length of the earthquake-to-receiver path in each. For example, if the earthquake located in active shallow region and the receiver is in stable continent region, then along the earthquake-to-receiver travel path, we can identify the relative proportion of path which located in active shallow region and that located in stable continent region. This information can be used, in conjunction with the overall degree of membership for the site, as a means of assigning the weights in the logic tree model for different tectonic regions.

## 10 CONCLUSION

Using seismological and tectonic information, a tectonic regionalization model has been obtained using a reproducible computational scheme. The process adopted for delineating tectonic regions is transparent and replicable across the globe. Moreover, the framework is flexible and it can be easily modified in the future when a new, relevant data set becomes available. All the application described herein was demonstrated in a global context, it can be applied also at a local or continental in scale, depending on the availability and resolution of the relevant data set.

A classification-scheme based on fuzzy logic can successfully incorporate concepts that are approximate rather than precise, such as the activeness index (degree of being active) and craton index (degree of being cratonic) derived in the study. This result can be incorporated into logic-tree models, a widely used tool for quantification of epistemic uncertainties in PSHA.

The proposed zonation will offer the possibility of updating and developing new tectonic region-specific proxies. New zonations

tailored for a specific purpose can also be easily implemented following the proposed approach. Our data-driven classification has showed the highly consistency with the expert-driven classification when applied to existing ground motion data sets; however, further testing of new regionalization models should be done in future studies to understand their relative performance and potential impact on seismic hazard assessment.

## ACKNOWLEDGEMENTS

The authors are grateful to the two reviewers, Prof. Peter Bird and Dr. Hilmar Bungum, for providing constructive and insightful comments that helped to improve the quality of the manuscript. The authors acknowledge the support provided to the Global Earthquake Model (GEM) initiative by its public and private sponsors. The authors thank Dr. Roberto Basili, Dr. John Douglas, Dr. Daniel García, Prof. Dr. Corné Kreemer, Prof. Dr. William H.K. Lee, Dr. Damiano Monelli and Dr. Paul Somerville as the members of GEM regionalisation working group who have been supporting initially this idea.

## REFERENCES

- Abrahamson, N.A. & Shedlock, K.M., 1997. Overview (of modern attenuation relationships), *Seismol. Res. Lett.*, **68**(1), 9–23.
- Akkar, S. *et al.*, 2012. Defining a consistent strategy to model ground-motion parameters for the GEM-PEER Global GMPEs Project, in *Proceedings of the 15th World Conference on Earthquake Engineering*, 24–28 September, Paper Number 2743, Lisbon, Portugal.
- Akkar, S. *et al.*, 2014. Reference database for seismic ground-motion in Europe (RESORCE), *Bull. Earthq. Eng.*, **12**(1), 311–339.
- Allen, T.I., Wald, D.J., Hotovec, A.J., Lin, K., Earle, P.S. & Marano, K.D., 2008. *An Atlas of ShakeMaps for selected global earthquakes*, Tech. Rep.
- Allmann, B.P. & Shearer, P.M., 2009. Global stress drop variations for moderate to large earthquakes, *J. geophys. Res.*, **113**(B01310), doi:10.1029/2008JB005821.
- Ancheta, T.D. *et al.*, 2014. Nga-west2 database, *Earthq. Spectra*, **30**(3), 989–1005.
- Ansari, A., Firuzi, E. & Etemadsaeed, L., 2015. Delineation of seismic sources in probabilistic seismic-hazard analysis using fuzzy cluster analysis and Monte Carlo simulation, *Bull. seism. Soc. Am.*, **105**(4), 2174–2191.
- Bird, P., 2003. An updated digital model of plate boundaries, *Geochem. Geophys. Geosyst.*, **4**(3), doi:10.1029/2001GC000252.
- Bird, P., Kagan, Y.Y. & Jackson, D.D., 2002. *Plate Tectonics and Earthquake Potential of Spreading Ridges and Oceanic Transform Faults*, vol. **30**, American Geophysical Union.
- Bird, P., Jackson, D., Kagan, Y., Kreemer, C. & Stein, R., 2015. GEAR1: a global earthquake activity rate model constructed from geodetic strain rates and smoothed seismicity, *Bull. seism. Soc. Am.*, **105**(5), 2538–2554.
- Bleeker, W., 2003. The late Archean record: a puzzle in ca. 35 pieces, *Lithos*, **71**(2), 99–134.
- Boore, D.M., Stewart, J.P., Seyhan, E. & Atkinson, G.M., 2014. NGA-West2 equations for predicting PGA, PGV, and 5 per cent damped PSA for shallow crustal earthquakes, *Earthq. Spectra*, **30**(3), 1057–1085.
- Burbidge, D., Leonard, M., Robinson, D. & Gray, D., 2012. The 2012 Australian earthquake hazard map, *Geoscience Australia, Record*, **71**, 108–116.
- Champati ray, P.K., Dimri, S., Lakhera, R. & Sati, S., 2007. Fuzzy-based method for landslide hazard assessment in active seismic zone of Himalaya, *Landslides*, **4**(2), 101, doi:10.1007/s10346-006-0068-6.
- Coppersmith, K.J. *et al.*, 2012. *Central and eastern United States (CEUS) seismic source characterization (SSC) for nuclear facilities project*, Tech. Rep., Electric Power Research Institute (EPRI).
- Danciu, L., Kale, Ö. & Akkar, S., 2016. The 2014 seismic hazard model of the Middle East: ground motion model, *Bull. Earthq. Eng.*, doi:10.1007/s10518-016-9989-1.

- Delavaud, E. *et al.*, 2012. Toward a ground-motion logic tree for probabilistic seismic hazard assessment in Europe, *J. Seismol.*, **16**(3), 451–473.
- Dreiling, J., Isken, M.P., Mooney, W.D., Chapman, M.C. & Godbee, R.W., 2014. *NGA-East regionalization report: comparison of four crustal regions with Central and Eastern North America using waveform modelling and 5 per cent-damped pseudo-spectral acceleration response*, Tech. Rep., Pacific Earthquake Engineering Research Center, PEER Report 2014/15.
- Dziewonski, A.M., 1971. Upper mantle models from pure-path dispersion data, *J. geophys. Res.*, **76**(11), 2587–2601.
- Ekström, G., Nettles, M. & Dziewoński, A., 2012. The global CMT project 2004–2010: Centroid-moment tensors for 13,017 earthquakes, *Phys. Earth planet. Inter.*, **200**, 1–9.
- Flinn, E.A. & Engdahl, E.R., 1965. A proposed basis for geographical and seismic regionalization, *Rev. Geophys.*, **3**(1), 123–149.
- Flinn, E.A., Engdahl, E.R. & Hill, A.R., 1974. Seismic and geographical regionalization, *Bull. seism. Soc. Am.*, **64**(3–2), 771–992.
- García, D., Wald, D. & Hearn, M., 2012. A global earthquake discrimination scheme to optimize ground-motion prediction equation selection, *Bull. seism. Soc. Am.*, **102**(1), 185–203.
- Goulet, C. *et al.*, 2014. Peer NGA-East database, PEER Report 2014, **17**.
- Grandjean, G., Malet, J.-P., Bitri, A. & Méric, O., 2007. Geophysical data fusion by fuzzy logic for imaging the mechanical behaviour of mudslides, *Bull. soc. géol. France*, **178**(2), 127–136.
- Gudmundsson, Ó. & Sambridge, M., 1998. A regionalized upper mantle (RUM) seismic model, *J. geophys. Res.*, **103**(B4), 7121–7136.
- Hayes, G.P., Wald, D.J. & Johnson, R.L., 2012. Slab 1. 0: A three-dimensional model of global subduction zone geometries, *J. geophys. Res.*, **117**(B1), doi:10.1029/2011JB008524.
- Heeszel, D.S., Wiens, D.A., Nyblade, A.A., Hansen, S.E., Kanao, M., An, M. & Zhao, Y., 2013. Rayleigh wave constraints on the structure and tectonic history of the Gamburtsev Subglacial Mountains, East Antarctica, *J. geophys. Res.*, **118**, 2138–2153.
- Heuret, A. & Lallemand, S., 2005. Plate motions, slab dynamics and back-arc deformation, *Phys. Earth planet. Inter.*, **149**(1), 31–51.
- International Seismological Centre, 2011. *On-line Bulletin*, Int. Seis. Cent., Thatcham, United Kingdom, <http://www.isc.ac.uk>, last accessed 2018 February 9.
- Jacob, K.H., 1972. Global tectonic implications of anomalous seismic P traveltimes from the nuclear explosion Longshot, *J. geophys. Res.*, **77**(14), 2556–2573.
- Johnston, A., Kanter, L., Coppersmith, K. & Cornell, C., 1994. *The earthquakes of stable continental regions. Volume 4, Seismicity data sheets (Part 2)*. Final report, Tech. Rep., Electric Power Research Inst., Palo Alto, CA (United States); Memphis State Univ., TN (United States); Geomatrix Consultants, Inc., San Francisco, CA (United States); Cornell (CA), Portola Valley, CA (United States).
- Jordan, T.H., 1981. Global tectonic regionalization for seismological data analysis, *Bull. seism. Soc. Am.*, **71**(4), 1131–1141.
- Kagan, Y. & Jackson, D., 1994. Long-term probabilistic forecasting of earthquakes, *J. geophys. Res.*, **99**(B7), 13 685–13 700.
- Kagan, Y., Bird, P. & Jackson, D., 2010. Earthquake patterns in diverse tectonic zones of the globe, *Pure appl. Geophys.*, **167**(6–7), 721–741.
- Kagan, Y.Y. & Jackson, D.D., 2012. Whole Earth high-resolution earthquake forecasts, *Geophys. J. Int.*, **190**(1), 677–686.
- Kanamori, H., 1970. Velocity and Q of mantle waves, *Phys. Earth planet. Inter.*, **2**(4), 259–275.
- Khan, A., Boschi, L. & Connolly, J., 2011. Mapping the Earth's thermochemical and anisotropic structure using global surface wave data, *J. geophys. Res.*, **116**(B1), doi:10.1029/2010JB007828.
- Kotha, S.R., Bindi, D. & Cotton, F., 2016. Partially non-ergodic region specific GMPE for Europe and Middle-East, *Bull. Earthq. Eng.*, **14**(4), 1245–1263.
- Kreemer, C., Blewitt, G. & Klein, E.C., 2014. A geodetic plate motion and Global Strain Rate Model, *Geochem. Geophys. Geosyst.*, **15**, 3849–3889.
- Lekic, V. & Romanowicz, B., 2011. Tectonic regionalization without a priori information: a cluster analysis of upper mantle tomography, *Earth planet. Sci. Lett.*, **308**(1), 151–160.
- Mamdani, E.H., 1974. Application of fuzzy algorithms for control of simple dynamic plant, *Proc. Inst. Electr. Eng.*, **121**(12), 1585–1588.
- Mitchell, B.J., 1995. Anelastic structure and evolution of the continental crust and upper mantle from seismic surface wave attenuation, *Rev. Geophys.*, **33**(4), 441–462.
- Mitchell, B.J., Pan, Y., Xie, J. & Cong, L., 1997. Lg coda Q variation across Eurasia and its relation to crustal evolution, *J. geophys. Res.*, **102**(B10), 22 767–22 779.
- Mitchell, B.J., Cong, L. & Ekström, G., 2008. A continent-wide map of 1 Hz Lg coda Q variation across Eurasia and its relation to lithospheric evolution, *J. geophys. Res.*, **113**(B4), doi:10.1029/2007JB005065.
- Mooney, W.D., Ritsema, J. & Hwang, Y.K., 2012. Crustal seismicity and the earthquake catalog maximum moment magnitude ( $m_{\max}$ ) in stable continental regions (SCRs): correlation with the seismic velocity of the lithosphere, *Earth planet. Sci. Lett.*, **357**, 78–83.
- Müller, R.D., Sdrolias, M., Gaina, C. & Roest, W.R., 2008. Age, spreading rates, and spreading asymmetry of the world's ocean crust, *Geochem. Geophys. Geosyst.*, **9**(4), doi:10.1029/2007GC001743.
- Nikravesh, M. & Aminzadeh, F., 2001. Mining and fusion of petroleum data with fuzzy logic and neural network agents, *J. Pet. Sci. Eng.*, **29**(3), 221–238.
- Nordlund, U., 1996. Formalizing geological knowledge—with an example of modeling stratigraphy using fuzzy logic, *J. Sedimentary Res.*, **66**(4), 689–698.
- Okal, E.A., 1977. The effect of intrinsic oceanic upper-mantle heterogeneity on regionalization of long-period Rayleigh-wave phase velocities, *Geophys. J. R. astr. Soc.*, **49**(2), 357–370.
- Petersen, M.D. *et al.*, 2015. The 2014 United States national seismic hazard model, *Earthq. Spectra*, **31**(S1), S1–S30.
- Ritsema, J., van Heijst, H.J. & Woodhouse, J.H., 2011. S40RTS: a degree-40 shear-velocity model for the mantle from new Rayleigh wave dispersion, teleseismic traveltimes and normal-mode splitting function measurements, *Geophys. J. Int.*, **184**, 1223–1236.
- Schulte, S.M. & Mooney, W.D., 2005. An updated global earthquake catalogue for stable continental regions: reassessing the correlation with ancient rifts, *Geophys. J. Int.*, **161**(3), 707–721.
- Somerville, P., Graves, R., Collins, N., Song, S.G., Ni, S. & Cummins, P., 2009. Source and ground motion models for Australian earthquakes, in *Proc. 2009 Annual Conf. Australian Earthquake Engineering Society*, The Australian Earthquake Engineering Society, Newcastle.
- Stafford, P.J., 2014. Crossed and nested mixed-effects approaches for enhanced model development and removal of the ergodic assumption in empirical ground motion models, *Bull. seism. Soc. Am.*, **104**(2), 702–719.
- Stewart, J.P. *et al.*, 2013. *Selection of a global set of ground motion prediction equations: Work undertaken as part of Task 3 of the GEMPEER Global GMPEs project*, Report No. 2013/xx, Pacific Earthquake Engineering Research Center, Berkeley, CA.
- Storchak, D.A., Di Giacomo, D., Bondár, I., Engdahl, E.R., Harris, J., Lee, W.H., Villaseñor, A. & Bormann, P., 2013. Public release of the ISC–GEM Global Instrumental Earthquake Catalogue (1900–2009), *Seismol. Res. Lett.*, **84**(5), 810–815.
- Sugeno, M., 1985. *Industrial Applications of Fuzzy Control*, Elsevier Science Inc.
- Tang, Y.-J., Zhang, H.-F., Ying, J.-F. & Su, B.-X., 2013. Widespread reutilization of cratonic and circum-cratonic lithospheric mantle, *Earth-Sci. Rev.*, **118**, 45–68.
- Toksöz, M.N. & Anderson, D.L., 1966. Phase velocities of long-period surface waves and structure of the upper mantle: 1. Great-Circle Love and Rayleigh wave data, *J. geophys. Res.*, **71**(6), 1649–1658.
- Weatherill, G., Pagani, M. & Garcia, J., 2016. Exploring earthquake databases for the creation of magnitude-homogeneous catalogues: Tools for application on a regional and global scale, *Geophys. J. Int.*, **206**, 1652–1676.
- Woessner, J. *et al.*, 2015. The 2013 European seismic hazard model: key components and results, *Bull. Earthq. Eng.*, **13**(12), 3553–3596.
- Young, J., Presgrave, B., Aichele, H., Wiens, D. & Flinn, E., 1996. The Flinn–Engdahl regionalisation scheme: the 1995 revision, *Phys. Earth planet. Inter.*, **96**(4), 223–297.