Multivariate statistical analysis to investigate the subduction zone parameters favoring the occurrence of giant megathrust earthquakes

S. Brizzi[1], L. Sandri[2], F. Funiciello[1], F. Corbi[1,3], C. Piromallo[4], and A. Heuret[5,3]

[1]Laboratory of Experimental Tectonics, Dipartimento di Scienze, Università degli Studi Roma Tre, Roma, Italia. [2]Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Bologna, Bologna, Italia. [3]Géosciences Montpellier Laboratory, University of Montpellier, Montpellier, France. [4]Istituto Nazionale di Geofisica e Vulcanologia, sezione di Roma, Roma, Italia. [5]Université de Guyane, Campus de Troubiran, Cayenne French Guyana.

Corresponding author: Silvia Brizzi (silvia.brizzi@uniroma3.it)

Abstract

The observed maximum magnitude of subduction megathrust earthquakes is highly variable worldwide. One key question is which conditions, if any, favor the occurrence of giant earthquakes ($M_w \geq 8.5$). Here we carry out a multivariate statistical study in order to investigate the factors affecting the maximum magnitude of subduction megathrust earthquakes. We find that the trench-parallel length of subduction zones and the thickness of trench sediments provide the largest discriminating capability between subduction zones that have experienced giant earthquakes and those having significantly lower maximum magnitude. Monte Carlo simulations show that the observed spatial distribution of great earthquakes cannot be explained by pure chance to a statistically significant level. We suggest that the combination of a long subduction zone with thick trench sediments likely promotes a great lateral rupture propagation, characteristic for almost all giant earthquakes.

Keywords

Giant megathrust earthquakes; maximum magnitude; multivariate statistics; subduction megathrust seismicity; pattern recognition.

1. Introduction

Subduction megathrusts (i.e., large faults between the subducting and overriding plates) produce the Earth's greatest earthquakes, also known as giant earthquakes GEqs (i.e., $M_w \geq 8.5$). Consequently, they account for the majority of seismic energy globally released during the last century *(Pacheco and Sykes*, 1992). As recently demonstrated by the 2004 Sumatra-Andaman (M_w 9.2) and 2011

Tohoku-Oki earthquakes (M_w 9.1), these events are major threats to society and their occurrence provides the motivation to investigate which subduction zones may host such catastrophic events. Where great megathrust earthquakes will occur is not well understood (e.g., *McCaffrey*, 2008). One of the most striking features of subduction megathrust seismicity is indeed the considerable variation in the largest characteristic earthquake observed worldwide (e.g., *Uyeda and Kanamori*, 1979; *Lay and Kanamori*, 1981; *Heuret et al.*, 2012; *Ide*, 2013; *Schellart and Rawlinson*, 2013; *Marzocchi et al.*, 2016). During the last century, some subduction zones - e.g., Alaska, Chile, Japan and Sumatra - have produced $M_w \geq 8.5$ events, while others - e.g., Mariana, New Hebrides and Scotia - have not yet recorded such great earthquakes (Figure 1). This leaves the question open whether any subduction zone can host earthquakes GEqs, given a long-enough observational timespan (*McCaffrey*, 2008), or if specific conditions are needed (e.g., *Ruff and Kanamori*, 1980; *Jarrard*, 1986; *Ruff*, 1989; *Pacheco et al.*, 1993; *Conrad et al.*, 2004; *Heuret et al.*, 2011; *Normile*, 2011; *Marzocchi et al.*, 2016).

Previous works have investigated the potential relationship between the observed maximum magnitude M_{max} and different properties of subduction zones. Over the past decades, the seismic variability of subduction megathrust at the global scale was originally related to the combined effect of plate convergence and age of the subducting plate [*Uyeda and Kanamori*, 1979; *Ruff and Kanamori*, 1980]. It was proposed that giant earthquakes occur at subduction zones that are characterized by rapid subduction of young lithosphere [*Ruff and Kanamori*, 1980]. However, this former idea failed in explaining the occurrence of the 2004 Sumatra-Andaman event, as it violates the relationship both in terms of subducting plate age and subduction rate [*Stein and Okal*, 2005, 2011]. After few years, the 2011 Tohoku-Oki event occurred where the Pacific plate subducting in this region is one of the oldest in the world (about 130 Ma; [*Heuret et al.*, 2011]). Moreover, the relationship is less pronounced if an updated dataset and several historical earthquakes are used [*Stein and Okal*, 2007, 2011; *Heuret et al.*, 2011].

Many other possible links between M_{max} and different geodynamic parameters have also been proposed, including the forearc structure (*Song and Simons*, 2003*; Wells et al.*, 2003), trench migration velocity *(Schellart and Rawlinson*, 2013), upper plate motion *(Peterson and Seno*, 1984*; Schellart and Rawlinson*, 2013*; Scholz and Campos*, 1995) or stress regime (*Heuret et al.*, 2012*; McCaffrey*, 1993), sediment thickness at the trench (e.g., *Ruff*, 1989; *Heuret et al.*, 2012; *Scholl et al.*, 2015) or subducted sediments (*Seno*, 2017), downdip extent of the seismogenic zone (e.g., *Kelleher et al.*, 1974; *Pacheco et al.*, 1993; *Hayes et al.*, 2012; *Schellart and Rawlinson*, 2013; *Corbi et al.*, 2017) or megathrust curvature (*Bletery et al.*, 2016). Besides a few exceptions, *(Jarrard*, 1986*; Ruff and Kanamori*, 1980*)*, these studies are generally based on bivariate linear regression models. Subduction zones, however, are complex dynamic systems where interrelated processes take place and the multi-parameter influence needs to be considered when forecasting the potential M_{max}.

Here we investigate the conditions that possibly controlled the occurrence of GEqs by statistically analyzing the relationships between worldwide subduction zones characteristics and their M_{max} using the database compiled by *Heuret et al*. (2011). In addition to the straightforward linear correlations, we conduct a Pattern Recognition PR analysis (*Sandri et al*, 2004; *Sandri et al.,* 2018) in search of recurrent patterns – combinations of parameters – likely affecting the M_{max} of subduction zones. This approach introduces a new quantitative perspective in assessing the seismic potential of subduction megathrusts, tackling the combined effect that multiple subduction properties may have on the M_{max}.

2. Methods

We use two statistical approaches: the bivariate and PR analyses. Bivariate statistics is performed as a preliminary test on the existence of potential simple cause-effect relationships between subduction zones parameters and megathrust seismicity, with a focus on M_{max}. For this purpose, we calculate Pearson's product-moment R and Spearman's rank ρ correlation coefficients, which allow testing the strength of linear and non-linear (i.e., monotonic) dependence between two variables, respectively. Unlike Pearson's, Spearman's correlation does not require normally-distributed variables and is less susceptible to outliers that can affect the robustness of the analysis. The statistical significance of the correlations is evaluated using p-values.

The PR analysis is performed to investigate whether any combination of variables affects the occurrence of GEqs. The main advantage of this type of analysis is the possibility to obtain information from any possible combination of variables i.e., patterns) that may play a role on the studied process. The basic idea behind PR is to look for quantitative and complex (i.e., related to a multivariate dataset) repetitive patterns that may be common to different objects, each represented by an array of n-features (i.e., qualitative or quantitative parameters characterizing subduction segments; Dataset S1) and belonging to a given class. In our application, this translates into identifying patterns that may discriminate between subduction segments (the objects) with $M_{max} < 8.5$ (class 1) and with $M_{max} \geq 8.5$ (class 2). The features characterizing our objects, preliminary and independently compiled by *Heuret et al.* (2011, 2012), are described in section 3.

From a technical point of view, PR methods are used to classify objects, based on an array of characterizing features. The analysis generally consists of three main steps: *i)* the learning phase, *ii)* the voting phase and *iii)* the control experiments. During the learning phase, a set of known and classified objects is used to identify all the possible patterns characterizing each class. The identified patterns are used in the voting phase to classify new objects, whose class is unknown to the algorithm. Finally, results stability is evaluated with control experiments by repeating both the learning and voting phases with different values of the algorithm input parameters. Due to the limited amount of available data, we performed only the learning phase aiming at identifying any recurrent pattern that discriminate segments that have experienced GEqs from those that have not.

The analysis is based on two different PR non-parametric algorithms, i.e., the Binary Decision Tree BDT (*Mulargia et al.,* 1992*; Rounds,* 1980) and Fisher discriminant analysis FIS (e.g., *Duda and Hart,* 1973). Both algorithms have been previously used on synthetic data with known patterns to test their ability of recognizing recurrent schemes and extracting relevant features also on small data sets, with non-normal and discrete/categorical data (e.g.*, Sandri and Marzocchi,* 2004*).*

BDT (Figure 2a) builds up a decisional tree where the progressive branching gives all the possible patterns. The subset of the most relevant features in discriminating the two classes is automatically provided by means of the non-parametric Kolmogorov-Smirnov two-sample statistics (e.g., *Hollander and Wolfe,* 1999). BDT computes the empirical cumulative distribution function ECDF for each feature of both classes and looks for the "root" of the pattern, i.e., the first-order feature. This is the feature for which the significance level of the statistical difference between the ECDF of the two classes is lower than *i*) a significance level α (fixed a-priori; $\alpha = 0.01$), representing the risk we accept for a wrong attribution at each step, and *ii*) the significance level of the statistical difference calculated for any other feature. Based on the root of the pattern and its threshold value, each object (subduction segment) is assigned to one of the two classes. As long as it possible to find a feature for which the CDFs in the two classes are statistically different at a significance level lower than α, the algorithm provides progressively higher-order features.

FIS (Figure 2b) is based on the projection of the data along the direction that maximize the ratio of "between-class" variance to "within-class" variance in order to reduce data variation in the same class and increase the separation between the classes. This direction is the linear combination of features (i.e., pattern) affecting the most the M_{max} of subduction segments. The algorithm is here applied through a combinatorial approach, i.e., we tested all the possible combinations of the N features considered, taken in groups of k features at a time (k = 1, ... , N). For every possible k value, we selected the optimal pattern, which is the one leading to the lowest classification error (i.e., the number of subduction segments incorrectly classified out of the total number of subduction segments). Theoretically, the classification error should decrease as k increases, up to an optimal value beyond which the algorithm performance either remains stable (i.e., adding new features does not improve the classification) or deteriorates because of the noise introduced by irrelevant features. Therefore, among all the optimal patterns, we selected the one with the lower classification error and consisting of the smallest number of features. Using very different PR approaches allows checking whether the results are dependent on the type of algorithm used. Although the risk of possible overfit can be excluded only by applying the pattern found on independent data (i.e., voting phase), the stability of the results, which is also indirectly checked by running multiple PR tests with different combinations of input features (Table 1) and using different M_{max} datasets (Table 2), provides indirect evidence that this risk is reduced.

3. The database

To investigate the conditions favoring the occurrence of GEqs, we used the global database of subduction zones and interplate seismicity (*Heuret et al.*, 2011), which covers a wide range of seismological, geometric, kinematic and physical subduction characteristics (Dataset S1 and Table 2 for data and variables notation, respectively) of 62 worldwide subduction segments (*Heuret et al.*, 2011*)*. The segmentation procedure involved two different steps: i) definition of 505 trench-normal transects (2° wide and 1° spaced in the trench-parallel direction) and geometric properties of the interface (e.g., dip, width and strike), and ii) grouping of transects into 62 segments, so that the seismogenic zone characteristics of one segment can be considered homogeneous (*Heuret et al.*, 2011). The seismogenic zone of each of the 505 transects was mapped by selecting the shallow (depth ≤ 70 km) thrust-fault type earthquakes from both the Centennial ($M_w \geq 7$; 1900-1976) and Harvard CMT ($5.5 \leq M_w \leq 7$; 1976-2007) catalogs, using some specific features such as location, depth, focal mechanism and orientation of the fault plane (*Heuret et al.*, 2011; *McCaffrey*, 1994). For the Centennial catalog, all the earthquakes located between the volcanic arc and 50 km before the trench on the subducting plate were used. For the Harvard CMT catalog, the selection of the events included the earthquakes with at least one nodal plane consistent with the local geometry and orientation of the megathrust. The mapping of the seismogenic zone was further improved by using, for each of the identified thrust earthquakes, the location given in the EHB catalog.

The segmentation criteria used to merge different transects into segments are as follows in order of importance: *i)* the rupture area inferred for $M_w \geq 8$ earthquakes is included in a single segment; *ii)* transects with homogeneous interplate seismicity were grouped in a single segment (e.g., N-Kermadec was differentiated from S-Kermadec because of a higher number of events with higher magnitude); iii) transects with homogeneous seismogenic zone geometry (e.g., dip, downdip width) were grouped in a single segment (e.g., New Britain was differentiated from Bougainville because of its flatter geometry and narrower seismogenic zone width). Although ruling out potential unconscious biases related to the segmentation model and the selection of megathrust earthquakes is impossible, we believe that performing our analyses on a database that was independently and previously compiled helps minimizing this risk. More details about megathrust earthquake selection and segmentation methodology can be found in the auxiliary information of Heuret et al. (2011). The M_{max} of each segment was extracted from the ISC-GEM Global Instrumental Earthquake Catalogue (*Storchak et al.*, 2013*)*, which includes recently improved and homogeneous data of large global earthquakes (1900-2007; $M_w \geq 5.5$). To ensure continuity with the previous work of *Heuret et al.* (2011), we also used the Centennial-Harvard CMT catalogs (1900-1975, $M_w \geq 7$ and 1976-2007, $M_w \geq 5.5$, respectively). Therefore, our statistical analyses are performed on the following M_{max} datasets: i) M_{max} ISC-GEM from 1900 to 2007 ($M_{max\ GEM1900}$; Figure 1) and ii) M_{max} Centennial + CMT ($M_{max\ Cent+CMT}$; Figure S1) from 1900 to 2007. In addition, to account for the potential inaccuracy of earthquake data before the World-Wide Network of Standard Seismographs came into existence (*Oliver and Murphy*, 1971), we subsampled the ISC-GEM catalog from 1960 to obtain a M_{max} dataset ($M_{max\ GEM1960}$; Figure S2) which is likely more homogeneous in terms of uncertainties in the M_{max} estimates. The M_{max} of N-Chile and Japan segments, which have experienced a GEq after 2007, are updated considering the M_w of 2010 Maule (M_w 8.8) and 2011 Tohoku-Oki (M_w 9.1) events. For clarity, the M_{max} of a segment is the maximum M_w of that segment according to the dataset used. We choose to not include pre-1900 earthquakes (e.g., 1700 Cascadia, 1833 Sumatra and 1868 Peru) to avoid introducing potential biases, also considering that an accurate estimate of the magnitude of historical earthquake is difficult.

The investigated parameters, depicting the geometric, kinematic, and physical characteristics of global subduction zones, are listed in Table 2 and illustrated in Figure 3. The geometric parameters mostly describe the geometry of the seismogenic zone, including the horizontal and vertical coordinates of the updip and downdip limits (x_{min}, x_{max} and z_{min} and z_{max}), the dip and curvature radius of the slab at the trench (θ and R), the mean arc-trench distance ($d_{arc\text{-}trench}$, a proxy of megathrust dip) and the slab downdip length ($W_{intraslab}$). We also considered the entire trench-parallel extent of a given subduction zone (L_{trench}), which represents its largest available potential rupture length. This is calculated as the sum of the trench-parallel lengths of all the segments showing spatial lateral continuity, regardless of the potential presence of discontinuities such as aseismic ridges, change in dip or trench curvature. Where this continuity cannot be univocally defined (e.g., Japan vs. Izu-Bonin-Mariana or Philippines) we grouped subduction segments belonging to the same subducting plate. The kinematic parameters include absolute (i.e., upper plate V_{upn}, trench V_{tn} and subducting plate V_{spn}) and relative (convergence V_{cn} and subduction V_{sn}) plate velocities. For the statistical analyses, the absolute plate motion is described using the hotspot reference frame HS3 (*Gripp and Gordon*, 2002), while for the relative velocities we used the trench-normal component. The physical parameters include the subducting plate age at the trench (A), the average sediment thickness at the trench (T_{sed}; *Heuret et al.*, 2012), the type of margin in terms of accretion vs erosion (AvsE), the upper plate nature (UPN) and the upper plate strain (UPS). Several database parameters are not fully independent, as indicated by the correlation coefficients derived from least square linear regression analysis (figure S3). High correlation coefficients are observed between θ and R (R = 0.69), θ and $d_{arc\text{-}trench}$ (R = 0.68), R and $d_{arc\text{-}trench}$ (R = 0.83), V_c and V_s (R = 0.62), V_{up} and V_t (R = 0.82), V_{up} and V_{sp} (R = 0.75), V_t and V_{sp} (R = 0.65), and T_{sed} and AvsE (R = 0.62). Scatter plots of the other significant correlations ($R \geq 0.5$ and p-value ≤ 0.05) reveals the lack of a clear trend, suggesting that low p-values are related mostly to the presence of outliers. Because of this interdependence, which may introduce spurious statistical relationships, the PR analysis is performed using only a subset of the database parameters (Table 1), which is still meant to cover the geometric, kinematic and physical subduction characteristics of convergent margins while excluding redundant variables. Among θ, R and $d_{arc\text{-}trench}$, we choose to use $d_{arc\text{-}trench}$ as a proxy of the megathrust dip because of its higher estimate accuracy. For the kinematic parameters, we decided to test different combinations to avoid excluding a priori any potential useful information. To consider the possible effect of sediments along the plate interface (*Ruff*, 1989; *Heuret et al.*, 2102, *Scholl et al.*, 2015), we used T_{sed} (instead of AvsE) because it is a continuous variable. Despite the potential biases, we assumed that trench sediment thickness is representative of the amount of material subducted at seismogenic zone depths.

Some of the database parameters (e.g., T_{sed}, θ, $W_{intraslab}$) are also not defined for all the 62 subduction segments, because of the scarce availability of data (e.g., seismic reflection profiles to reasonably constrain T_{sed} (*Heuret et al.*, 2012), and number of interplate earthquakes used to map the geometric characteristics of the seismogenic zone (*Heuret et al.*, 2011)). This implies that we could use 38 or 40 subduction segments (over 62) for the PR analysis depending on the M_{max} dataset (38 for $M_{max\ GEM1960}$ and 40 for both $M_{max\ GEM1900}$ and $M_{max\ Cent+CMT}$), because all the objects with a missing value for at least one parameter were discarded. Finally, since the database parameters are measured at different scales, the PR analysis was performed using standardized values to ensure an equal contribution of each feature to the identified patterns.

4. Results

4.1. Bivariate analysis

Pearson's and Spearman's correlations (Figure 4) between seismological and subduction zones parameters are generally very weak (mean |R| = 0.21 ± 0.04 and mean |ρ| = 0.22 ± 0.05). Significant (i.e., |R| ≥ 0.5 and p-values ≤ 0.05) positive Pearson's correlations (Figure 4a) are observed between the number of earthquake N_{eq} and V_{sn} (|R| = 0.58), and between the seismicity rate τ and V_{sn} (| R| = 0.64). N_{eq} is also positively correlated with the subducting plate velocity V_{spn} (|R| = 0.51). These outcomes are coherent with previous work by *Heuret et al.* (2011).

Spearman's analysis (Figure 4b) shows the same results, though in most cases |ρ| is higher than |R| (i.e., |ρ| = 0.73 and |ρ| = 0.75 for correlations between N_{eq} and V_{sn}, and τ and V_{sn}, respectively).

Figure 5 shows scatter plots between the 19 subduction zone parameters and the M_{max} of the segments. The M_{max} is in the range 6.4-9.6, 5.7-9.6 and 5.1-9.6 for $M_{max\ GEM1900}$, $M_{max\ Cent+CMT}$ and $M_{max\ GEM1960}$ datasets, respectively. For most the parameters, we observe a considerable scatter of the data and lack of correlation, with |R| ranging from 0.008 to 0.544, 0.003 to 0.396, 0.032 to 0.557 for $M_{max\ GEM1900}$, $M_{max\ Cent+CMT}$ and $M_{max\ GEM1960}$ datasets, respectively. The highest correlation is observed with L_{trench}, with |R| = 0.55, |R| = 0.56 and |R| = 0.40 for $M_{max\ GEM1900}$, $M_{max\ Cent+CMT}$ and $M_{max\ GEM1960}$ datasets (Figure 5), respectively. In particular, an increase in L_{trench} is related to an increase in Mmax. Spearman's coefficients (Figure 4b) also confirm the positive relationship between M_{max} and L_{trench}, with |ρ| = 0.54, |ρ| = 0.50 and |ρ| = 0.56 for $M_{max\ GEM1900}$, $M_{max\ Cent+CMT}$ and $M_{max\ GEM1960}$, respectively. The second and third statistically significant correlations (p-values ≤ 0.05) are observed between $M_{max\ GEM1900}$, $W_{intraslab}$ and V_c, with |R| = 0.42 and 0.41 respectively. These correlations are confirmed also for the other two M_{max} datasets (|R| = 0.45 and |R| = 0.29 for $M_{max\ Cent+CMT}$; |R| = 0.46 and |R| = 0.26 for $M_{max\ GEM1960}$). Despite the consistent trend lines, a $M_{max\ GEM1900}$ ≥ 8.5 is observed for a wide (ca. 75% and 85% of the total) range of $W_{intraslab}$ and V_c values, suggesting that specific conditions are not needed. Parameters showing a clearer distinction between the observed range of GEqs and the total range are V_{sn}, θ and A (i.e., 36% , 38% and 43% of the total range, respectively). However, scatter plots suggest that outliers possibly drive the observed trend.

4.2.PR analysis

BDT results show a very simple pattern, consisting only of the first-order feature independently of the combination of input features and of the M_{max} datasets used. The algorithm identifies L_{trench} as the only feature that is able to discriminate between the two classes. Subduction segments with L_{trench} > 3900 km are classified as belonging to class 2, suggesting they have the propensity for hosting GEqs. The statistical significance of this outcome is given by the statistical significance level $\alpha = 0.01$ imposed a-priori. Thus, L_{trench} (and the corresponding threshold value) splits the data into two subsets, whose empirical cumulative distribution function ECDF are statistically different at 1% significant level. For all the combinations of input tested, the classification error is 15% when considering $M_{max\ GEM1900}$. This means that 6 over 40 subduction segments (i.e., Java, Mariana, Izu-Bonin, N-Kurili, Colombia, N-Peru) are characterized by high L_{trench}, but have not experienced a GEq. The classification error increases up to 20% and 21% if $M_{max\ Cent+CMT}$ and $M_{max\ GEM1960}$ are used to classify the segments, respectively. This means that 8 over 40 or 38 subduction segments have high L_{trench} but have not experienced a GEq. No further branching of the classification tree is observed, probably because of the low number of objects belonging to either one of the classes that prevents BDT from finding significant differences.

FIS results are mostly consistent among the different combinations of input and M_{max} datasets used. The patterns derived with $M_{max\ GEM1900}$ (Figure 6a) include L_{trench} as first-order feature, as its standardized coefficient (absolute value) is the highest. T_{sed} plays a significant role as well, since it is identified as second-order feature. Assuming the other pattern features fixed, the sign of the L_{trench} and T_{sed} standardized coefficients indicates a positive relationship with M_{max}. These results are also confirmed when considering the $M_{max\ Cent+CMT}$ and $M_{max\ GEM1960}$ datasets (Figure 6b-c), as the combination of L_{trench} and T_{sed} still provides the largest discriminating capability.

The choice of the M_{max} dataset affects the higher-order features appearing in the patterns. For both $M_{max\ GEM1900}$ and $M_{max\ Cent+CMT}$, the higher-order feature is $d_{arc\text{-}trench}$ (Figure 6a-b). When considering the $M_{max\ GEM1960}$ dataset (Figure 6c) instead, the higher-order feature is A. Assuming again the other pattern features fixed, the sign of the standardized coefficients highlights a negative correlation of M_{max} with $d_{arc\text{-}trench}$ and A.

The classification errors is ca. 17% for $M_{max\ GEM1900}$ and $M_{max\ Cent+CMT}$ (i.e., 7 over 40 subduction segments misclassified) and 16% for $M_{max\ GEM1960}$ (i.e., 6 over 38 segments misclassified).

Considering only the two most important features of the pattern to classify subduction segments, the equation describing the plane that maximize the ratio of the dispersion between the classes to the dispersion within the classes is:

$$1.70 * L_{trench} = -\ 0.43 * T_{sed} + 0.47$$

This means that a subduction segment characterized by a given L_{trench} and T_{sed} is classified as belonging to the GEqs class if $1.70 * L_{trench} + 0.43 * T_{sed} - 0.47 > 0$. Therefore, subduction segments need to have long trench-parallel extent and relatively high sediment thickness to be classified as "potentially generating GEqs". The thresholds defining the GEqs class (in terms of L_{trench} and T_{sed}) have uncertainties related to the relatively low number of subduction segments used for the classification. However, our findings highlight that GEqs have occurred preferentially along subduction segments with high L_{trench} (> 3900 km) and T_{sed} (≥ 1 km).

5. Discussion

5.1. do L_{trench} and T_{sed} influence M_{max} by pure chance?

PR analysis highlighted the primary role of L_{trench} and T_{sed} on tuning the M_{max} of megathrusts. Long subduction zones (i.e., high L_{trench}) have already been associated with the largest earthquakes (*Schellart and Rawlinson*, 2013). This may not be very surprising since, as already discussed by some authors (e.g., *Schellart and Rawlinson*, 2013), long subduction zones (i.e., those with L_{trench} >3000 km) represent the vast majority (≈ 75%) of the worldwide extent of subduction zones However, in our segmentation (blindly borrowed from *Heuret et al.* 2011), they cover 49% of the global extent of subduction zones. If we include only the 4 longest subduction zones identified by our PR analysis (L_{trench} > 3900 km), this fraction decreases to 43%. We thus statistically tested whether GEqs since 1900 (Table 3) occurred on long subduction zones just because they account for such a portion of subductions, by means of a Bernoulli trial scheme in which a "success" was the occurrence of a GEq on a long subduction zone. In this view, a "success" had a probability of 49% or 43% (depending on the L_{trench} threshold used, see above). Our empirical evidence is that we have 14 "success" out of 14 trials, and the theoretical probability of such occurrence is well below 1% (i.e., 5 10^{-5} and 8 10^{-6}, respectively). Assuming we may not expect GEqs at short subduction zones (e.g., Calabria; L_{trench} ≪ 1000 km) unless unrealistically high coseismic slip, we repeated the test by considering only subduction zones with L_{trench} ≥ 1000 km. only to subduction zones with L_{trench} ≥ 1000 km. In this case, the probability of "success" increased to 62% and 54% for the two threshold of L_{trench}, respectively. Accordingly the probabilities of 14 success out of 14 trials are 1.2 10^{-3} and 2.0 10^{-4}, still much lower than 1%. In other words, we could reject (at 1% significance level) the null hypothesis that all GEqs have occurred at long subduction zones by chance, just because of the fraction of the total subduction zones of the Earth they cover.

Looking at the spatial distribution of $M_{max\ GEM1900}$, we observe that the 12 segments with M_{max} ≥ 8.5 belong to the 4 longest subduction zones (i.e., Aleutians-Alaska, NW-Pacific, Indian, South America; Figure 1; Dataset S1). We tested whether this evidence could be explained by pure chance by assigning, through 10^6 Monte Carlo simulations, 14 GEqs randomly per unit length of subduction zone to *n* subduction segments (where *n* ranged from 1 to 14, as we allowed for repetitions). Then we counted how many times we observe M_{max} ≥ 8.5 on at least 12 segments belonging to the 4 longest subduction zones: this happened only 37 times in 10^6 simulations, i.e., a p-value of roughly 4 10^{-5}. The same test was repeated considering only subduction zones with L_{trench}

≥ 1000 km, and the corresponding p-value was 4 10^{-4}. For both tests, we could again reject the null hypothesis at 1% significance level. Therefore, accounting for the length of subduction zones, even when considering only the longer ones, does not explain why GEqs are observed only at the longest.
T_{sed} has been proposed as an important controlling factor for the genesis of GEqs as well. The topographic relief of the subducting plate may be smoothened by the presence of abundant subducting sediments. Such thick sediment layer along the plate interface is supposed to provide homogeneous strength, which may promote the rupture to propagate over longer trench-parallel distances (e.g., *Ruff*, 1989; *Heuret et al.*, 2012; *Scholl et al.*, 2015). Thick sediments may also act as a barrier for the fluid flow toward the megathrust creating a stronger interface (*Seno*, 2017).
Looking at the spatial distribution of $M_{max\ GEM1900}$, we observe that among the segments belonging to the 4 longest subduction zones (as defined by PR results), relatively high T_{sed} appears to be a preferred condition for GEqs occurrence. The empirical cumulative distribution functions of T_{sed} for the 4 longest subduction zones (Figure 7) show that all the segments (11) hosting events with $M_{max\ GEM1900} \geq 8.5$ have T_{sed} higher than the 30^{th} percentile of the respective subduction zone. Moreover, for the majority of these segments (9 out of 11), T_{sed} is also higher than the median of the respective subduction zone. Accordingly, all the GEqs (13; Table 3) took place along segments where T_{sed} is higher than the 30^{th} percentile of the respective subduction zone, most of them (11 out of 13) being located where T_{sed} is higher than the respective median.
Aiming to test whether these 4 observations could be explained by pure chance (Supplementary Text S1), we performed 10^5 Monte Carlo simulations, with the null hypothesis of GEqs occurring randomly on the 4 longest subduction zones regardless their T_{sed}. In each simulation, we assigned 13 GEqs randomly to the subduction segments belonging to the 4 longest subduction zones. Note that the number of GEqs now considered is lower with respect to the previous simulations as T_{sed} of one of the segments is unknown. We designed two tests, in which we constrained: *i)* the observed number of GEqs of each subduction zone (e.g., 3 GEqs for the South American subduction zone; #1; Figure 7; Table 4) and *ii)* the observed total number for all the subduction zones (#2; Table 4). Then, we counted how many times our simulated M_{max} matched the 4 observations described above. For both tests, we could reject the null hypothesis of each observation at 1% significance level (Table 4). In other words, given a long subduction zone, the occurrence of GEqs on relatively thick sediment segments seems very unlikely to be related to pure chance.
Repeating the Monte Carlo simulations presented in this section with $M_{max\ GEM1960}$ dataset leads to the same results.

5.2. What favors the occurrence of GEqs?

The bivariate statistical analysis clearly highlighted that none of the analyzed subduction parameters can univocally account for the great M_{max} diversity observed worldwide. This is because subduction zones are very complex systems and the megathrust seismicity is the result of the joint effect of various parameters. Limitations for understanding the occurrence of GEqs are also related to the observational record, which is very short compared to the recurrence time of GEqs (*McCaffrey*, 2008). Having only a century's worth of detailed earthquake history implies that we have not yet observed a complete seismic cycle with reasonable spatio-temporal resolution. Therefore, subduction megathrusts may not have different capabilities of producing GEqs and it is likely that any segment could host a M_w 9 event given hundreds to thousands of years (*McCaffrey*, 2008). However, recent analysis of the frequency-magnitude distribution of subduction interplate earthquakes over the last half-century showed that the energy release is variable among global subduction zones (*Marzocchi et al.*, 2016). Our results suggest that on a shorter timescale, at least as long as that covered by the available seismic catalogs, there may be favorable conditions for GEqs occurrence. Some subduction segments appear more prone to host GEqs than others do, and the different seismic behavior may be linked to different seismicity rates (*Heuret et al.*, 2011).

Probabilities calculated by means of Bernoulli trial scheme and Monte Carlo simulations highlighted that GEqs preferentially occurred at long subduction zones not by pure chance but rather because they likely allow for a longer (along-strike) rupture. Additionally, within the same subduction zone, GEqs are promoted at sediment-rich segments. Locked megathrust sub-segments, characterized by homogeneous strength conditions, have been associated to a smooth interface likely due to the presence of a thick sediment layer (e.g., *Kostoglodov*, 1988; *Ruff*, 1989; *Tichelaar and Ruff*, 1991; *Cloos and Shreve*, 1996; *Wang and Bilek*, 2011 ,2014; *Heuret et al.*, 2012; *Tan et al.*, 2012; *Kopp*, 2013; *Scholl et al.*, 2015).

The combination of a long subduction with a smooth or smoothened interface therefore enhances the conditions for large trench-parallel extent of the rupture and, in turn, higher earthquake magnitudes. It should be noted, though, that there may be also other factors influencing the occurrence of GEqs. Indeed, the Cascadia subduction zone is relatively short (L_{trench} = 1152 km), but is supposed to have experienced a M_w 9 earthquake in 1700 (*Satake et al.*, 2003). Another important exception is the 2011 Tohoku-Oki earthquake, which occurred instead at an erosive and poorly sedimented margin, characterized by subducting horst-graben structures. Interestingly, the main rupture area of this event features a rather low-relief subducting lithosphere, suggesting that high slip can be achieved even without the presence of thick-sediment at the trench (e.g., *Wang and Bilek*, 2014*; Scholl et al.*, 2015). In contrast, subduction of rugged seafloor has been suggested to lead to fault creeping and low magnitude earthquakes, because rupture areas may be geometrically constrained (e.g., *Gao and Wang*, 2014; *Wang and Bilek*, 2011). Furthermore, a subducting seamount may cause the development of a fracture network in which distributed deformation (*Wang and Bilek*, 2011) and even rupture termination (e.g., *Mochizuki et al.*, 2008) tend to occur, as it has been argued for the 2011 Tohoku-Oki earthquake (*Wang and Bilek*, 2014).

The patterns highlighted by the PR analysis include also $d_{arc\text{-}trench}$ or A (depending on the considered M_{max} dataset) as higher-order features playing a minor role on M_{max}. Our results suggest that GEqs are favored in areas characterized by a small $d_{arc\text{-}trench}$ (i.e., < 250 km, corresponding to megathrusts dipping from 13° to 35°) and young (i.e., A < 60 Myr) subducting plate. Since a low $d_{arc\text{-}trench}$ implies relatively steep dip, this result seems at odd with those studies suggesting a positive correlation between the megathrust dip and the M_{max}, based on geometrical considerations. In fact, shallow dipping angles determine an increased downdip extent of the seismogenic zone, hence allowing ruptures to extend over wider fault areas with respect to steeper subductions (*Corbi et al*., 2017; *Kelleher et al*., 1974; *Lay et al*., 1982; *Muldashev and Sobolev*, 2017; *Schellart and Rawlinson*, 2013; *Uyeda and Kanamori*, 1979). However, this issue is debated since no significant direct correlation between M_{max} and the dip angle or the downdip extent of the seismogenic zone has been found so far (*Heuret et al.*, 2011; *Pacheco et al.*, 1993).

Studies over the last decades have also suggested that the plate convergence rate and the age of subducting plate affect the M_{max} of subduction zones. Young and buoyant plates, subducting fast, attain relatively flat morphologies which imply a wider interface area potentially leading to great M_w events (e.g., *Uyeda and Kanamori*, 1979; *Ruff and Kanamori*, 1980). This model seemed very reasonable, until the 2004 Sumatra-Andaman earthquake occurred unexpectedly where a relatively young – 70 Myr old – plate is subducting at 15-25 mm/yr (*Stein and Okal*, 2011, 2005).

The minor influence of these two features, suggested by PR results, is very likely not related to the simple geometrical effect of widening the potential downdip rupture width, especially considering that the downdip extent of the seismogenic zone is at least ≈ 6 times smaller than the trench-parallel one. Rather, there may be a relationship with the state of stress of the subduction interface, which obviously is an important control on earthquake occurrence. For instance, *Bletery et al*. (2016) found a stronger correlation between earthquake size and the curvature of the megathrust, compared the dip angle: flat (low-curvature) megathrusts have homogeneous strength conditions over large areas and, therefore, are more likely to favor the occurrence of GEqs (*Bletery et al.*, 2016). Recently, it has been shown that the age of the subducting plate correlates positively with b-values (i.e., the slope of the earthquake size distribution) of global subduction zones, thus implying that large earthquakes occur preferentially in subduction zone with younger slabs (*Nishikawa and Ide*, 2014). According to the authors, the buoyancy of the slab would thus influence the stress state of the subduction megathrust. However, the analysis has been restricted to a small subset of parameters, i.e., subducting plate age and plate motion. Including also geometric and physical characteristics of subduction zones to estimate the b-value worldwide will possibly provide new insights on the conditions favoring the occurrence of GEqs.

6. Conclusions

The statistical analyses presented in this paper highlight the major role of L_{trench} and T_{sed}, the parameters concurring to enhance long ruptures in the trench-parallel direction. The Monte Carlo tests showed that the short-term spatial distribution of GEqs does not appear to be random. Rather, these great events are more likely to be observed along segments belonging to the longest subduction zones and characterized by a relatively high sediment supply. Recent GEqs (except the anomalous case of 2011 Tohoku-Oki earthquake) demonstrate that great magnitudes result from a rupture spanning laterally for several hundreds of kilometers (e.g., *Subarya et al.*, 2006; *Moreno et al.*, 2009), as a result of the joint failure of neighboring sub-segments of the megathrust (e.g., *Kaneko et al.*, 2010). It should not be forgotten that faults, especially the largest ones, are not simply interfaces of frictional contact but areas of structural complexity (e.g., *Wang*, 2010). Among the factors controlling the seismogenic behavior of subduction megathrust, mechanical and physical properties of the plate interface are of first-order importance (e.g., *Wang and Bilek*, 2011, 2014; *Moreno et al.*, 2012; *Kopp*, 2013). Despite the efforts, the key question of where GEqs are more likely to occur is far from being answered. However, constraining the trench-parallel distribution of large seismogenic patches and understanding how these relate to the excess of sediments supply or to the presence of topographical features (e.g., *Scholz and Small*, 1997; *Robinson*, 2006; *Morgan et al.*, 2008; *Müller and Landgrebe*, 2012; *Basset and Watts*, 2015) will greatly improve our understanding of the conditions limiting earthquake size.

Acknowledgments

We thank D. Scholl, an anonymous reviewer and the Editor K. Wang for helping improving manuscript. FC received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 658034 (AspSync). All codes and data used in this work are available upon request to the corresponding author.

References

Basset, D., Watts, A., 2015. Gravity anomalies, crustal structure, and seismicity at subduction zones: 1. Seafloor roughness and subduction relief. Geochemistry Geophys. Geosystems 16, 1541–1576. doi:10.1002/2014GC005684

Bletery, Q., Thomas, A.M., Rempel, A.W., Karlstrom, L., Sladen, A., De Barros, L., 2016. Mega-earthquakes rupture flat megathrusts. Science 80- 354, 1027–1031. doi:10.1126/science.aag0482

Cloos, M., Shreve, R.L., 1996. Shear-zone thickness and the seismicity of Chilean- and Marianas-type subduction zones. Geology 24, 107–110. doi:10.1130/0091-7613(1996)024<0107:SZTATS>2.3.CO;2

Conrad, C.P., Bilek, S., Lithgow-Bertelloni, C., 2004. Great earthquakes and slab pull: Interaction between seismic coupling and plate-slab coupling. Earth Planet. Sci. Lett. 218, 109–122. doi:10.1016/S0012-821X(03)00643-5

Corbi, F., Herrendörfer, R., Funiciello, F., van Dinther, Y., 2017. Controls of seismogenic zone width and subduction velocity on interplate seismicity: Insights from analog and numerical models. Geophys. Res. Lett. 44, 6082–6091. doi:10.1002/2016GL072415

Duda, R.O., Hart, P.E., 1973. Pattern classification and scene analysis. Wiley, New York.

Gao, X., Wang, K., 2014. Strength of stick-slip and creeping subduction megathrusts from heat flow observations. Science 80345, 1038–1041. doi:10.1126/science.1255487

Gripp, A.E., Gordon, R.G., 2002. Young tracks of hotspots and current plate velocities. Geophys. J. Int. 150, 321–361. doi:10.1046/j.1365-246X.2002.01627.x

Hayes, G.P., Wald, D.J., Johnson, R.L., 2012. Slab1.0: A three-dimensional model of global subduction zone geometries. J. Geophys. Res. Solid Earth 117, 1–15. doi:10.1029/2011JB008524

Heuret, A., Conrad, C.P., Funiciello, F., Lallemand, S., Sandri, L., 2012. Relation between subduction megathrust earthquakes, trench sediment thickness and upper plate strain. Geophys. Res. Lett. 39, 1–6. doi:10.1029/2011GL050712

Heuret, A., Lallemand, S., Funiciello, F., Piromallo, C., Faccenna, C., 2011. Physical characteristics of subduction interface type seismogenic zones revisited. Geochemistry, Geophys. Geosystems 12, 1–26. doi:10.1029/2010GC003230

Hollander, M., Wolfe, D.A., 1999. Nonparametric statistical methods, John Wiley and Sons Perry, P. and S. Wolff. Wiley, New York.

Ito, G., Martel, S.J., 2002. Focusing of magma in the upper mantle through dike interaction. J. Geophys. Res. Solid Earth 107, 2223. doi:10.1029/2001JB000251

Jarrard, R.D., 1986. Relations among subduction parameters. Rev. Geophys. 24, 217–284. doi:10.1029/RG024i002p00217

Kaneko, Y., Avouac, J.-P., Lapusta, N., 2010. Towards inferring earthquake patterns from geodetic observations of interseismic coupling. Nat. Geosci. 3, 363–369. doi:10.1038/ngeo843

Kelleher, J., Savino, J., Rowlett, H., McCann, W., 1974. Why and where great thrust earthquakes occur along island arcs. J. Geophys. Res. 79, 4889–4899. doi:10.1029/JB079i032p04889

Kopp, H., 2013. Invited review paper: The control of subduction zone structural complexity and geometry on margin segmentation and seismicity. Tectonophysics 589, 1–16. doi:10.1016/j.tecto.2012.12.037

Kostoglodov, V., 1988. Sediment Subduction - a Probable Key for Seismicity and Tectonics at Active Plate Boundaries. Geophys. J. 94, 65–72.

Lay, T., Kanamori, H., Ruff, L., 1982. The asperity model and the nature of large subduction zone earthquakes. Earthq.

Predict. Res. 1, 3–71.

Marzocchi, W., Sandri, L., Heuret, A., Funiciello, F., 2016. Where giant earthquakes may come. J. Geophys. Res. Solid Earth 121, 7322–7336. doi:10.1002/2016JB013054

McCaffrey, R., 2008. Global frequency of magnitude 9 earthquakes. Geology 36, 263–266. doi:10.1130/G24402A.1

McCaffrey, R., 1993. On the Role of the Upper Plate in Great Subduction Zone Earthquake. J. Geophys. Res. 98, 11953–11966.

Mochizuki, K., Yamada, T., Shinohara, M., Yamanaka, Y., Kanazawa, T., 2008. Weak Interplate Coupling by Seamounts and Repeating M 7 Earthquakes. Science 321, 1194–1197. doi:10.1126/science.1160250

Moreno, M., Melnick, D., Rosenau, M., Baez, J., Klotz, J., Oncken, O., Tassara, A., Chen, J., Bataille, K., Bevis, M., Socquet, A., Bolte, J., Vigny, C., Brooks, B., Ryder, I., Grund, V., Smalley, B., Carrizo, D., Bartsch, M., Hase, H., 2012. Toward understanding tectonic control on the M w 8.8 2010 Maule Chile earthquake. Earth Planet. Sci. Lett. 321–322, 152–165. doi:10.1016/j.epsl.2012.01.006

Moreno, M.S., Bolte, J., Klotz, J., Melnick, D., 2009. Impact of megathrust geometry on inversion of coseismic slip from geodetic data: Application to the 1960 Chile earthquake. Geophys. Res. Lett. 36, 1–5. doi:10.1029/2009GL039276

Morgan, E.C., McAdoo, B.G., Baise, L.G., 2008. Quantifying geomorphology associated with large subduction zone earthquakes. Basin Res. 20, 531–542. doi:10.1111/j.1365-2117.2008.00368.x

Mulargia, F., Marzocchi, W., Gasperini, P., 1992. Statistical identification of physical patterns which accompany eruptive activity on Mount Etna, Sicily. J. Volcanol. Geotherm. Res. 53, 289–296. doi:10.1016/0377-0273(92)90087-T

Muldashev, I., Sobolev, S., 2017. Estimation of Maximum Magnitudes of Subduction Earthquakes, in: EGU General Assembly Conference Abstracts. p. 11466.

Müller, R.D., Landgrebe, T.C.W., 2012. The link between great earthquakes and the subduction of oceanic fracture zones. Solid Earth 3, 447–465. doi:10.5194/se-3-447-2012

Normile, D., 2011. Devastating Earthquake Defied Expectations. Science 331, 1375–1376.

Pacheco, J.F., Sykes, L.R., 1992. Seismic moment catalog of large shallow earthquakes, 1900 to 1989. Bull. Seismol. Soc. Am. 82, 1306–1349. doi:10.1130/0091-7613(2001)029<0347:TDFMEE>2.0.CO;2

Pacheco, J.F., Sykes, L.R., Scholz, C.H., 1993. Nature of seismic coupling along simple plate boundaries of the subduction type. J. Geophys. Res. 98, 14133. doi:10.1029/93JB00349

Peterson, E.T., Seno, T., 1984. Factors Affecting Seismic Moment Release Rates in Subduction Zones. J. Geophys. Res. 89, 10233–10248.

Robinson, D.P., 2006. Earthquake Rupture Stalled by a Subducting Fracture Zone. Science 312, 1203–1205. doi:10.1126/science.1125771

Rounds, E.M., 1980. A combined nonparametric approach to feature selection and binary decision tree design. Pattern Recognit. 12, 313–317. doi:10.1016/0031-3203(80)90029-1

Ruff, L., Kanamori, H., 1980. Seismicity and the subduction process. Phys. Earth Planet. Inter. 23, 240–252. doi:10.1016/0031-9201(80)90117-X

Ruff, L.J., 1989. Do trench sediments affect great earthquake occurrence in subduction zones? Pure Appl. Geophys. PAGEOPH 129, 263–282. doi:10.1007/BF00874629

Sandri, L., Marzocchi, W., 2004. Testing the performance of some nonparametric pattern recognition algorithms in realistic cases. Pattern Recognit. 37, 447–461. doi:10.1016/j.patcog.2003.08.009

Sandri, L., Marzocchi, W., Zaccarelli, L., 2004. A new perspective in identifying the precursory patterns of eruptions. Bull. Volcanol. 66, 263–275. doi:10.1007/s00445-003-0309-7

Sandri, L., Acocella, V., Newhall, C., 2017. Searching for patterns in caldera unrest. Geochem. Geophys. Geosyst., 18, 2748–2768, doi:10.1002/ 2017GC006870Schellart, W.P., Rawlinson, N., 2013. Global correlations between maximum magnitudes of subduction zone interface thrust earthquakes and physical parameters of subduction zones. Phys. Earth Planet. Inter. 225, 41–67. doi:10.1016/j.pepi.2013.10.001

Scholl, D.W., Kirby, S.H., von Huene, R., Ryan, H., Wells, R.E., Geist, E.L., 2015. Great (≥Mw8.0) megathrust earthquakes and the subduction of excess sediment and bathymetrically smooth seafl oor. Geosphere 11, 236–265. doi:10.1130/GES01079.1

Scholz, C.H., Campos, J., 1995. On the mechanism of seismic decoupling and back arc spreading at subduction zones. J. Geophys. Res. 100, 22103. doi:10.1029/95JB01869

Scholz, C.H., Small, C., 1997. The effect of seamount subduction on seismic coupling. Geology 25, 487–490. doi:10.1130/0091-7613(1997)025<0487:TEOSSO>2.3.CO;2

Seno, T., 2017. Subducted sediment thickness and Mw 9 earthquakes. J. Geophys. Res. Solid Earth 122, 470–491. doi:10.1002/2016JB013048

Song, T.-R.A., Simons, M., 2003. Large Trench-Parallel Gravity Variations Predict Seismogenic Behavior in Subduction Zones. Science 301, 630–633. doi:10.1126/science.1085557

Stein, S., Okal, E.A., 2011. The size of the 2011 Tohoku earthquake need not have been a surprise. Eos, Trans. Am. Geophys. Union 92, 227–228. doi:10.1029/2011EO270005

Stein, S., Okal, E.A., 2005. Speed and size of the Sumatra earthquake. Nature 434, 581–582. doi:doi:10.1038/434581a

Stein, S., Okal, E. a., 2007. Ultralong period seismic study of the December 2004 Indian Ocean earthquake and implications for regional tectonics and the subduction process. Bull. Seismol. Soc. Am. 97, S279–S295. doi:10.1785/0120050617

Storchak, D.A., Di Giacomo, D., Bondár, I., Engdahl, E.R., Harris, J., Lee, W.H.K., Villaseñor, A., Bormann, P., 2013. Public release of the ISC-GEM global instrumental earthquake catalog (1900-2009). Seismol. Res. Lett. 84, 810–815. doi:10.1785/0220130034

Subarya, C., Chlieh, M., Prawirodirdjo, L., Avouac, J.-P., Bock, Y., Sieh, K., Meltzner, A.J., Natawidjaja, D.H., McCaffrey, R., 2006. Plate-boundary deformation associated with the great Sumatra-Andaman earthquake. Nature 440, 46–51. doi:10.1038/nature04522

Tan, E., Lavier, L.L., Van Avendonk, H.J.A., Heuret, A., 2012. The role of frictional strength on plate coupling at the subduction interface. Geochemistry, Geophys. Geosystems 13. doi:10.1029/2012GC004214

Tichelaar, B.W., Ruff, L.J., 1991. Seismic coupling along the Chilean Subduction Zone. J. Geophys. Res. 96, 11997. doi:10.1029/91JB00200

Uyeda, S., Kanamori, H., 1979. Back-arc opening and the mode of subduction. J. Geophys. Res. 84, 1049. doi:10.1029/JB084iB03p01049

Wang, K., 2010. Finding fault in fault zones. Science 329, 152–153. doi:10.1126/science.1192223

Wang, K., Bilek, S.L., 2014. Invited review paper: Fault creep caused by subduction of rough seafloor relief. Tectonophysics 610, 1–24. doi:10.1016/j.tecto.2013.11.024

Wang, K., Bilek, S.L., 2011. Do subducting seamounts generate or stop large earthquakes? Geology 39, 819–822. doi:10.1130/G31856.1

Wells, R.E., Blakely, R.J., Sugiyama, Y., Scholl, D.W., Dinterman, P.A., 2003. Basin-centered asperities in great subduction zone earthquakes: A link between slip, subsidence, and subduction erosion? J. Geophys. Res. Solid Earth 108. doi:10.1029/2002JB002072

Table 1. Combinations of features used for each PR test

Input	Geometric			Physical			Kinematic				
1	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{sn}	V_{cn}	V_{upn}	V_{tn}	V_{spn}
2	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{upn}	V_{tn}	V_{spn}		
3	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{sn}	V_{cn}			
4	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{sn}	V_{upn}			
5	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{sn}	V_{tn}			
6	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{sn}	V_{spn}			
7	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{c}	V_{upn}			
8	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{c}	V_{tn}			
9	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{c}	V_{spn}			
10	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{sn}				
11	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{cn}				
12	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{upn}				
13	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{tn}				
14	$d_{arc\text{-}trench}$	L_{trench}	$W_{intraslab}$	A	T_{sed}	UPS	V_{spn}				

Table 2. Parameters notation

Parameter	Explanation	Units	Category
N_{eq}	number of earthquakes	-	seismological
τ	seismicity rate	number of events per century and per 10^3 km of trench	
CSM	Cumulated Seismic Moment	N m	
M_{mrr}	equivalent representative magnitude sensu *Ruff and Kanamori* (1980)	-	
$M_{max\ GEM1900}$	Maximum M_w from ISC-GEM catalog during 1900 – 2007 period	-	
$M_{max\ Cent+CMT}$	Maximum M_w from Centennial + CMT catalogs during 1900 – 2007 period	-	
$M_{max\ GEM1960}$	Maximum M_w from ISC-GEM catalog during 1960 – 2007 period	-	
z_{min}	depth of the updip limit of the seismogenic zone	km	geometric
z_{max}	depth of the downdip limit of the seismogenic zone	km	
x_{min}	distance from the trench of the updip limit of the seismogenic zone	km	
x_{max}	distance from the trench of the downdip limit of the seismogenic zone	km	
L_{trench}	trench-parallel length of subduction zone	km	
$W_{intraslab}$	Downdip length of the slab	km	
$d_{arc\text{-}t}$	mean arc-trench distance	km	
R	curvature radius at the trench	km	
θ	dip of the megathrust	°	
A	age of the subducting plate at the trench	Ma	physical
T_{sed}	sediment thickness at the trench	km	
UPN	Upper Plate Nature *1 = continental; 2 = oceanic*		
AvsE	Accretionary vs erosive margin *1 = accretionary; 2 = erosive*		
UPS	Upper Plate Strain *1 = Extensional; 2 = Neutral; 3 = Compressive* (Heuret et al., 2011)		
V_{sn}	trench-normal subduction velocity ($V_{sn} = V_{tn} + V_{spn}$)	mm/yr	kinematic
V_{cn}	trench-normal convergence velocity ($V_{cn} = V_{snn} + V_{tn}$)	mm/yr	
V_{upn}	trench-normal upper plate velocity; trenchward motion is positive	mm/yr	
V_{tn}	trench-normal trench velocity; migration towards subducting plate (rollback) is positive	mm/yr	
V_{spn}	trench normal component of subducting plate velocity; trenchward motion is positive	mm/yr	

Table 3. List of the great earthquakes considered for the Monte Carlo simulations, from ISC-GEM 1900 dataset.

Name	Subduction segment	Subduction zone	Mw	Date
Andaman	Ad	Indonesia	9.0	December 26, 2004
Sumatra	Sm		8.6	March 28, 2005
Sumatra	Sm		8.6	September 12, 2007
Timor*	Tm		8.5	February 2, 1938
S-Kuril	S-Ku	North-West Pacific	8.5	October 13, 1963
Japan	Jp		9.1	March 11, 2011
Kamchatka	Km		8.9	November 4,1952
Ws-Aleutians	Ws-At	Aleutians-Alaska	8.7	February 2, 1965
C-Aleutians	C-At		8.6	March 9, 1957
E-Aleutians	E-At		8.6	April 1, 1946
E-Alaska	E-Ak		9.3	March 28, 1964
N-Chile	N-Ch	South America	8.8	February 27, 2010
S-Chile	S-Ch		9.6	May 22, 1960
S-Chile	S-Ch		8.6	May 22, 1960

* T_{sed} unknown

Table 4. p-values of Monte Carlo simulations

Test	Observation	p-value
#1	a	$1.6e^{-4}$
	b	0
	c	$2.8e^{-3}$
	d	$7.0e^{-5}$
#2	a	$3.0e^{-5}$
	b	0
	c	$5.7e^{-2}$
	d	$7.0e^{-5}$

Table

Click here to download Table: Brizzi_et_al_2017 -revised_Dataset S1.xlsx

Name	Subduction segment	Subduction zone	N_{eq}
Calabria*	Cb		0
W-Aegean	W-Ae	Mediterranean	17
E-Aegean*	E-Ae		1
Makran*	Mk	Makran	2
Andaman	Ad		67
Sumatra	Sm	Indian	121
Java	Jv		39
Timor*	Tm		17
Seram*	Se	Seram	21
Wetar*	We	Wetar	7
Flores*	F	Flores	9
Halmahera*	H	Halmahera	8
Sangihe*	S	Sangihe	49
Sulawesi*	Sw	Sulawesi	37
Sulu*	Su	Sulu	3
Cotobato*	C	Cotobato	14
Manila	Mn	Manila	27
Philippines	Ph	Philippines	121
S-Ryukyu*	S-Ry		42
N-Ryukyu	N-Ry	Nankai-Ryukyu	35
Nankai	Na		6
Palau*	Pl	Palau	0
Yap*	Yp	Yap	4
Marianas	Mr		46
Izu-Bonin	Iz		54
Japan	Jp	North-West Pacific	215
S-Kuril	S-Ku		141
N-Kuril	N-Ku		84
Kamchatka	Ka		102
W-Aleutians*	W-At		5
Ws-Aleutians	Ws-At		93
C-Aleutians	C-At	Aleutians-Alaska	125
E-Aleutians	E-At		73
W-Alaska	W-Ak		31
E-Alaska	E-Ak		15
Cascades*	Cs	Cascades	0
Mexico	Me		62
Costa Rica	Cr	Central America	119
Cocos	Co		24
Colombia	Cl		19
N-Peru	N-Pe		25
S-Peru*	S-Pe	South America	35
N-Chile	N-Ch		161
S-Chile	S-Ch		8
Patagonia*	Pt		3
Antilles	An	Antilles	16
Muertos*	Mu	Muertos	3
Venezuela*	Ve	Venezuela	0
Panama*	Pn	Panama	0

Figure (with caption below and on the same page)

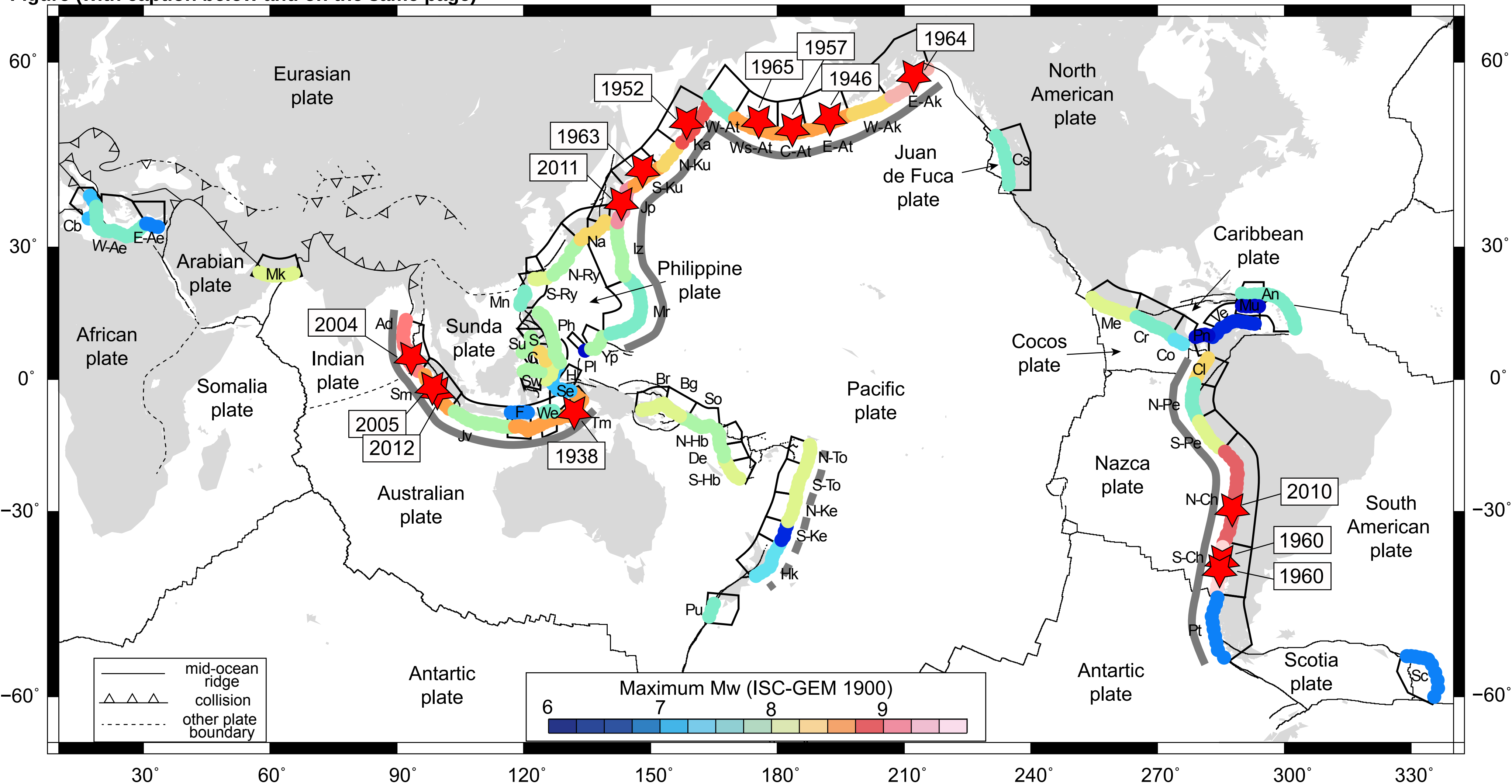

Figure 1. Observed maximum M_w of subduction megathrust earthquakes according to ISC-GEM 1900 dataset. The black boxes represent subduction segments as defined by Heuret et al. (2011). The thick grey lines mark subduction zones with long trench-parallel extension (L_{trench} > 3900 km). Red stars show the location of recent giant earthquakes (i.e., $M_w \geq 8.5$). Subduction segment are labelled by abbreviations; full names are listed in Dataset S1.

Figure (with caption below and on the same page)

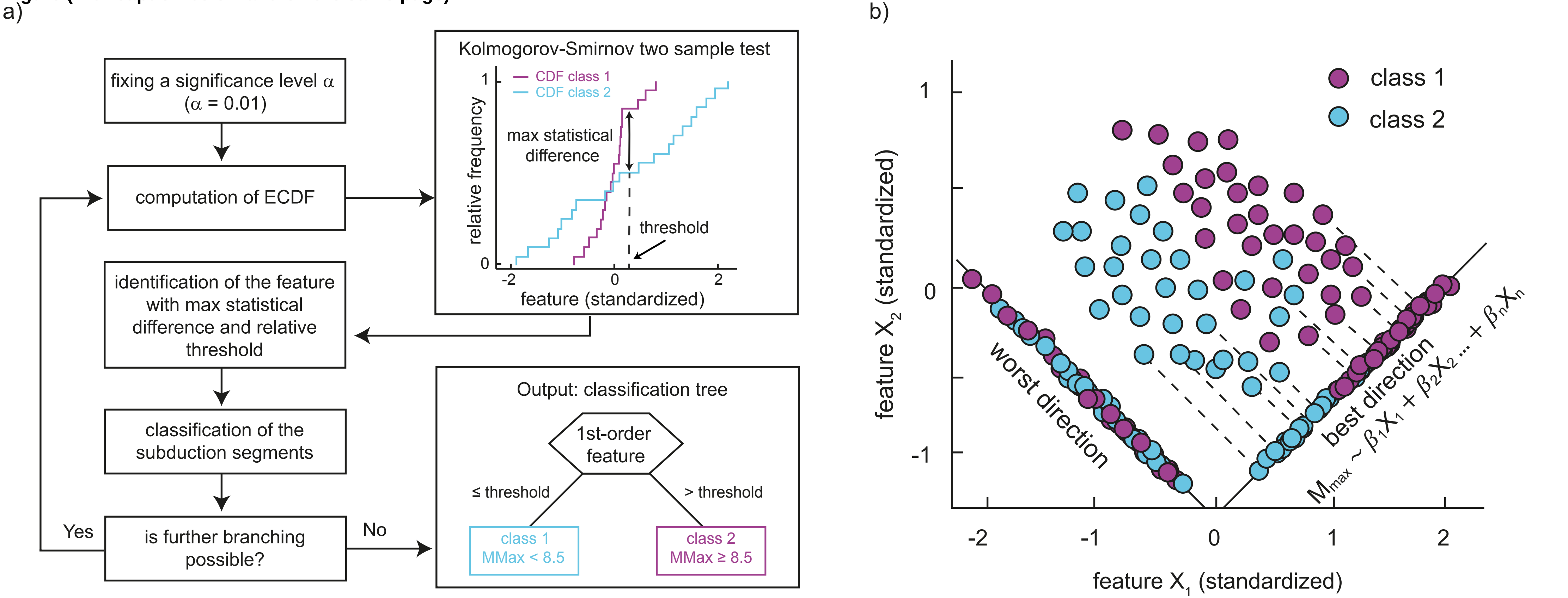

Figure 2. a) Flow chart describing the procedure of BDT alghoritm. The algorithm is based on the Kolmogorov-Smirnov two sample test. Each parameter included in the classification tree is able to split the data in two subset, whose empirical cumulative distribution functions ECDF are statistically different at 1% significance level. **b)** Schematic representation of the projection perfomerd by FIS algorithm to classify objects belonging to two different classes in a n = 2 feature space. The direction along which objects (subduction segments) are projected (i.e., best direction) maximizes the ratio of the dispersion between the classes to the dispersion within the classes. This direction - or pattern - is a linear combination of parameters affecting the $M_{max.}$

Figure (with caption below and on the same page)

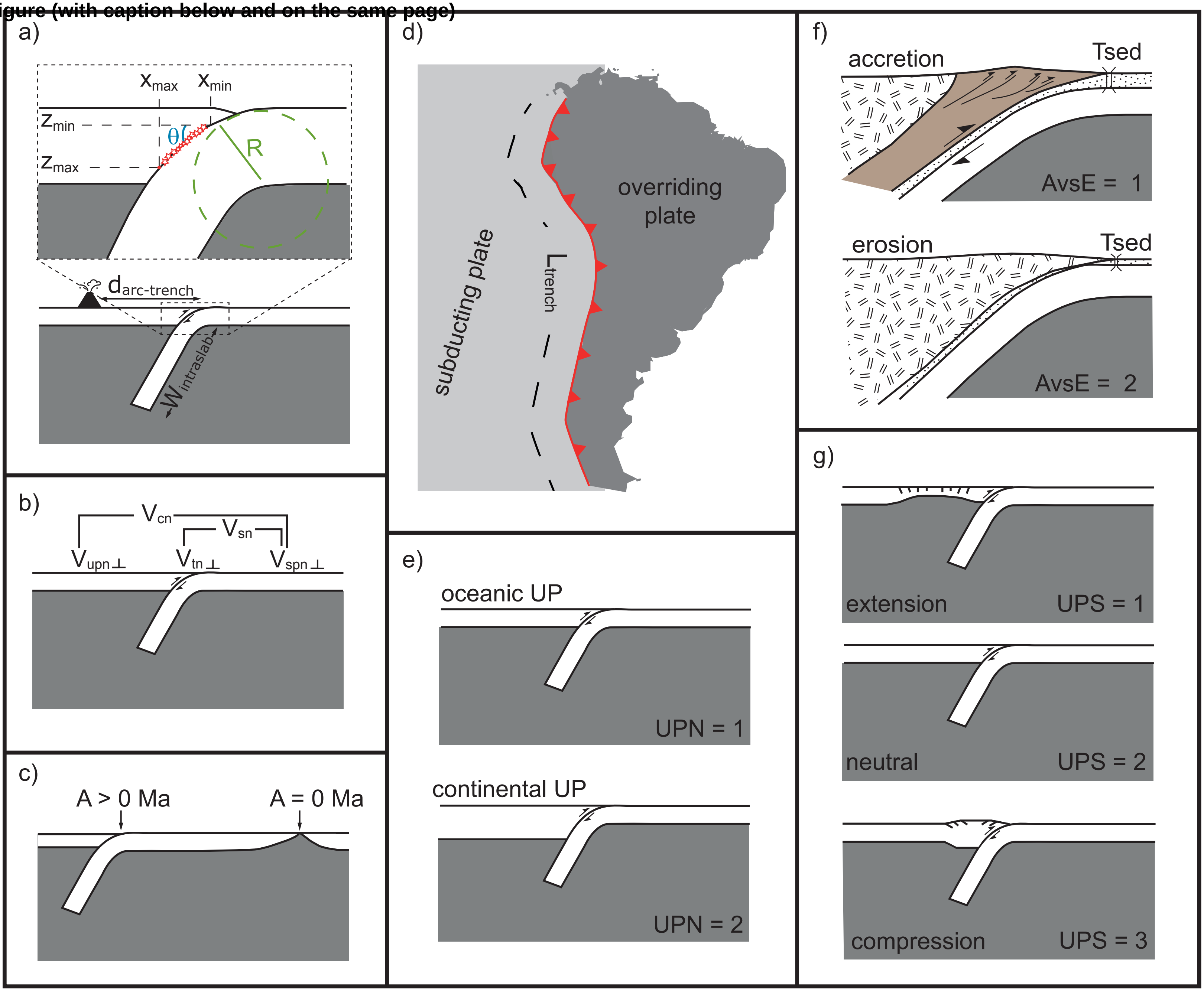

Figure 3. Schematic diagram showing the 19 subduction zone parameters that are investigated in relation to their potential effect on the maximum magnitude M_{max} of megathrust earthquakes. **a)** horizontal (x_{min}, x_{max}) and vertical (z_{min}, z_{max}) coordinates of the updip and downdip limits of the seismogenic zone, dip of the megathrust (θ), curvature radius of the slab at the trench (R), mean arc-trench distance ($d_{arc\text{-}trench}$), slab downdip length ($W_{intraslab}$); **b)** subduction (V_{sn}) and convergence (V_{cn}) velocities, upper plate velocity (V_{upn}), trench velocity (V_{tn}), and subducting plate velocity (V_{spn}); **c)** age of the subducting plate at the trench (A); **d)** trench-parallel extent of the subduction zone (L_{trench}); **e)** upper plate nature (UPN), f) sediment thickness at the trench (T_{sed}), and accretionary vs erosive margin (AvsE); upper plate strain (UPS).

Figure (with caption below and on the same page)

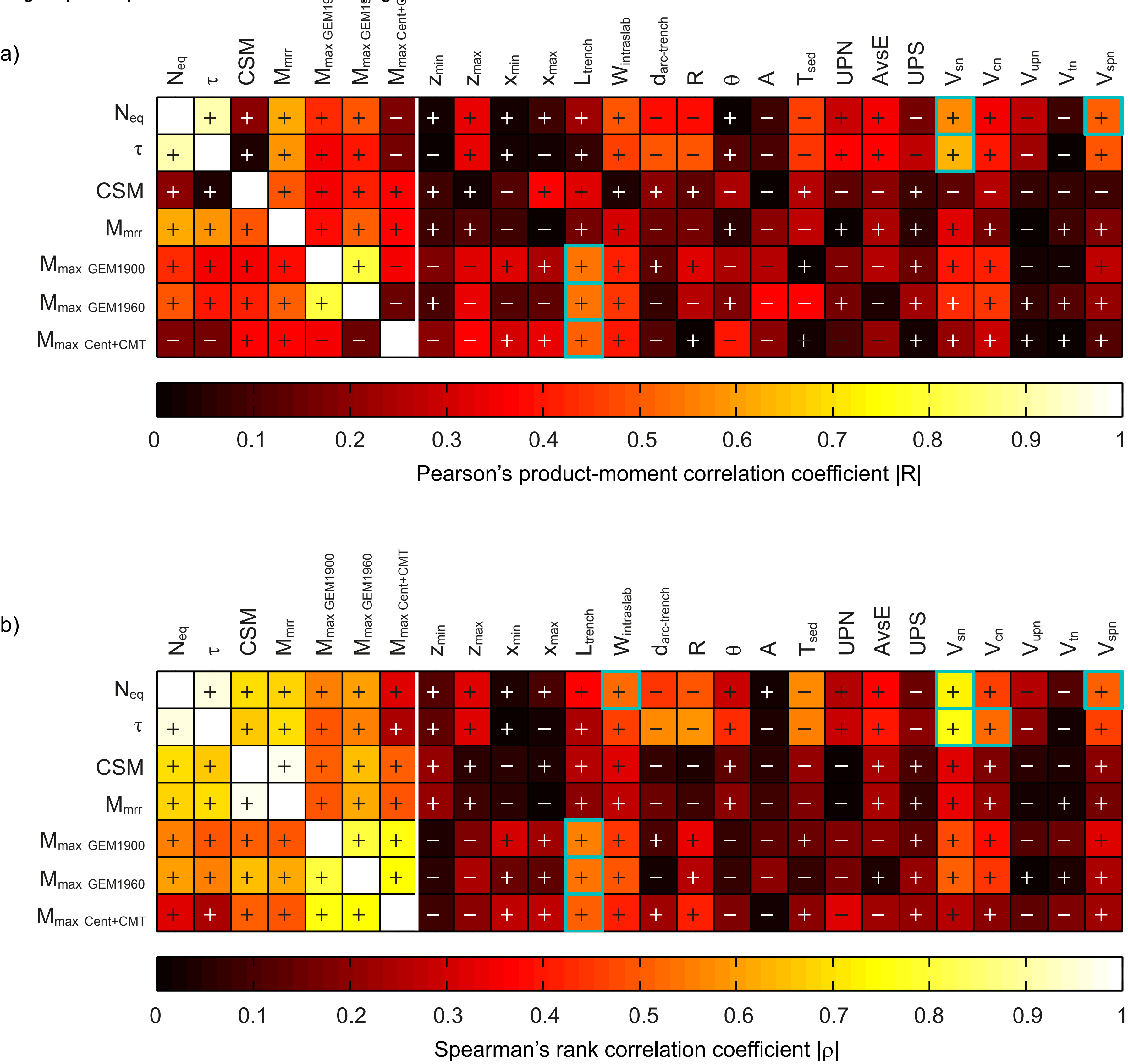

Figure 4. Bivariate correlations. **a)** Pearson's product-moment R and **b)** Spearman's rank ρ correlation coefficients between seismological and subduction segments parameters. Symbols identify the sign (positive or negative) of the correlation coefficient and the color (black or white) refers to the p-value of the correlation (≤ 0.05 or > 0.05, respectively). Cyan rectangles highlight the most significant correlations ($|R|$ or $|\rho| \geq 0.5$ and p-value ≤ 0.05). Seismological and subduction parameters are defined in Table 2.

Figure (with caption below and on the same page)

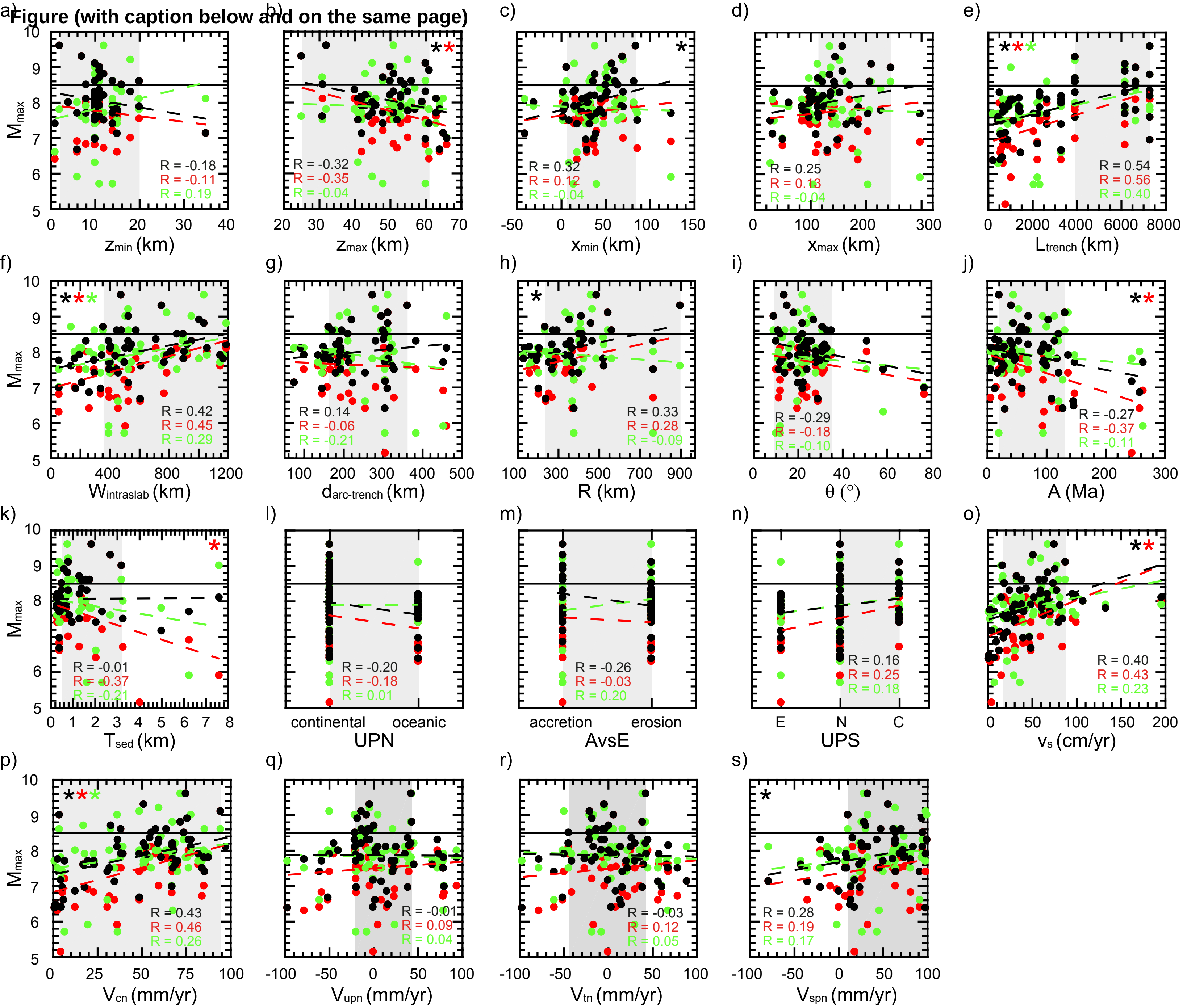

Figure 5. Scatter plots showing the dependence between the maximum M_w M_{max} and 19 subduction zone parameters (see also Figure 3). **a)** depth of the updip limit of the seismogenic zone; **b)** depth of the downdip limit of the seismogenic zone; **c)** distance from the trench of the updip limit of the seismogenic zone; **d)** distance from the trench of the downdip limit of the seismogenic zone; **e)** trench-parallel extent of the subduction zone; **f)** downdip length of the slab; **g)** mean arc-trench distance; **h)** curvature radius of the slab at the trench; **i)** dip of the megathrust; **j)** age of the subducting plate at the trench; **k)** sediment thickness at the trench; **l)** upper plate nature; **m)** accretionary vs erosive margin; **n)** upper plate strain; **o)** subduction velocity; **p)** convergence velocity; **q)** upper plate velocity; **r)** trench velocity; **s)** subducting plate velocity. Colors refer to the M_{max} dataset used for the analysis: $M_{max\ GEM\ 1900}$ (black), $M_{max\ Cent+CMT}$ (red), $M_{max\ GEM1960}$ (green). The dashed lines represent the best least-square regression fit, with correlation coefficients R reported at the bottom of each panel. The asterisks highlight correlations with p-value ≤ 0.05. The light grey area indicates the observed range of a given parameter for earthquakes with $M_w \geq 8.5$ according to $M_{max\ GEM1900}$. The continous black line highlights $M_w \geq 8.5$.

Figure (with caption below and on the same page)

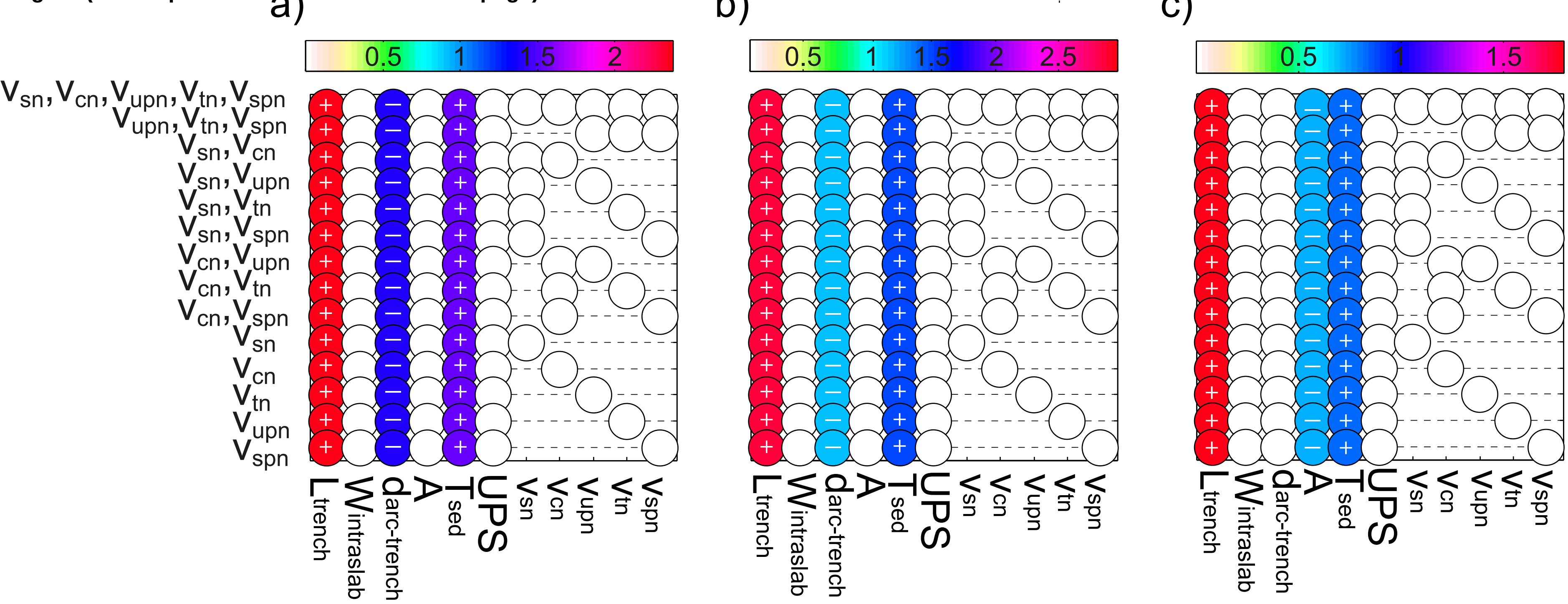

Figure 6. FIS classification patterns derived from **a)** $M_{max\ GEM1900}$, **b)** $M_{max\ Cent+CMT}$ and **c)** $M_{max\ GEM1960}$ datasets. Each row of the plot refers to one PR test; the combination of kinematic features used as input for the corresponding test is listed in the left-side labels (see Table 1 for all the input features included in each PR test). The bottom-side labels show all the features that may potentially contribute to the pattern. The absolute value of the coefficients of the features included in the patterns is displayed according to the color bar. Symbols inside the circles refer to the sign (positive or negative) of the coefficients.

Figure (with caption below and on the same page)

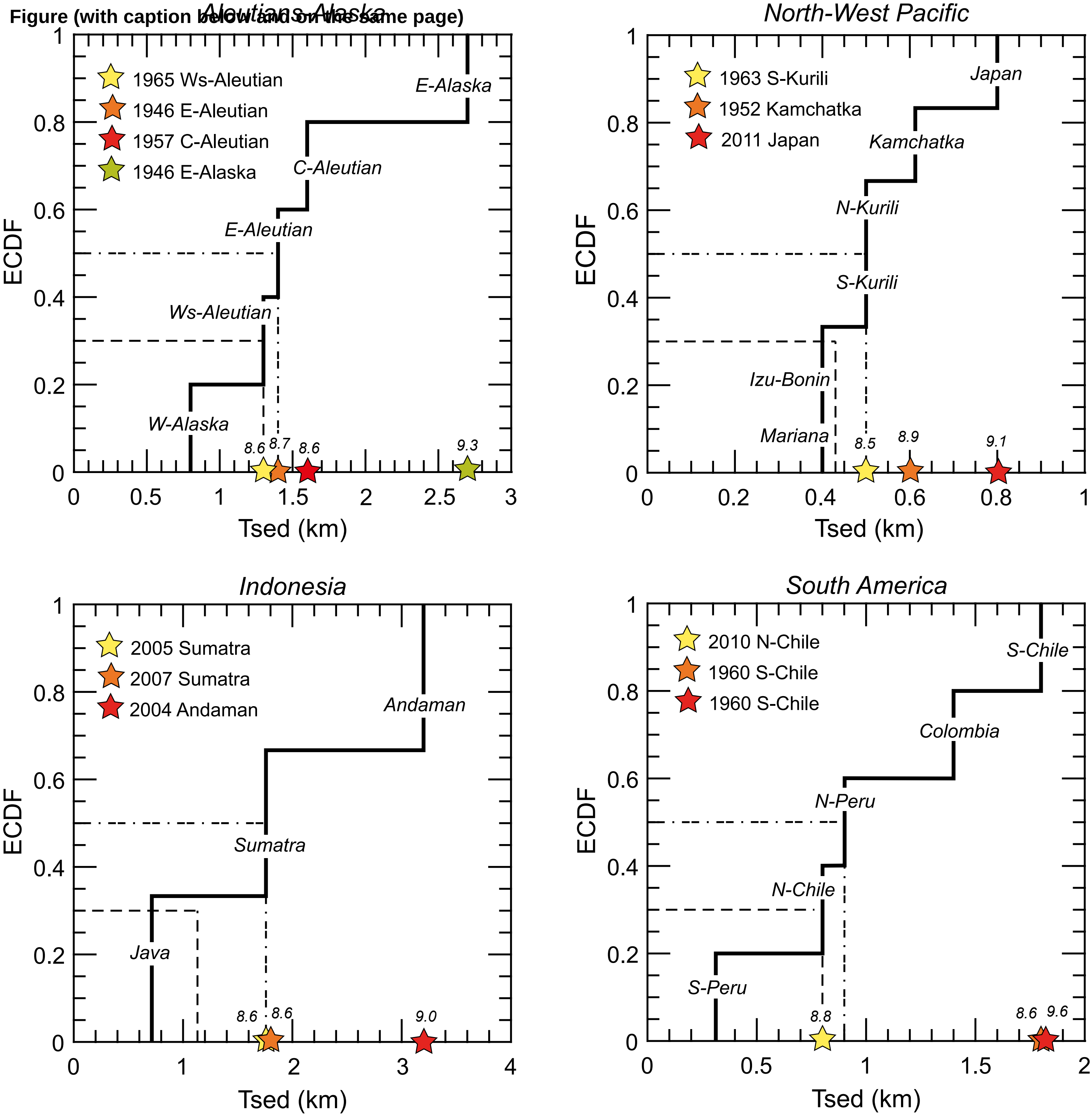

Figure 7. Empirical Cumulative Distribution Function ECDF of the sediment thickness at the trench T_{sed} for the 4 longest subduction zones: Aleutians-Alaska, North-West Pacific, Indian and South America. The stars mark T_{sed} values of the subduction segments that have experienced a giant-earthquake according to the ISC-GEM 1900 dataset (see also Table 3). The numbers above the stars refer to the M_w of the events. Giant-earthquakes occuring on the same segment have been slightly shifted for a clearer graphical representation. The dashed and dashed-dotted lines highlight the 30th percentile and the median of the ECDF, respectively.