POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Advanced Sensitivity Analysis Methods for Seismic-Induced Tsunami NaTech Risk Assessment

TESI DI LAUREA MAGISTRALE IN SAFETY AND PREVENTION ENGINEERING IN THE PROCESS INDUSTRY – INGEGNERIA DELLA PREVENZIONE E DELLA SICUREZZA NELL'INDUSTRIA DI PROCESSO

Author: **Nicola Gallo**

Student ID:	922299
Advisor:	Enrico Zio
Co-advisor:	Francesco Di Maio, Jacopo Selva, Matteo Taroni
Academic Year:	2020-21

Abstract

NaTech events (Natural Hazard Triggering Technological Disasters) are industrial accidents triggered by natural hazards which may lead to potentially tremendous impact on the environment and the population. Seismic-induced tsunami NaTech risk assessment entails the seismic sources to be characterised and modelled to provide the Peak Ground Acceleration (PGA) information as input to the seismic-induced tsunamis models and simulation needed for a Seismic Probabilistic Tsunami Hazard Analysis (SPTHA). In this thesis, we propose two Sensitivity Analysis (SA) methods to deal with the computational issues related with:

1. The identification of the model parameters most affecting the PGA;
2. The identification most relevant features of the seismic model, for deciding a priori the seismic scenarios to be simulated.

With respect to the first issue, we propose a novel Bootstrapped Modularised Global Sensitivity Analysis (BMGSA) method. The method is tested on a benchmark case study. The results are compared with a standard variance-based Global SA method. The strength of the proposed method is that its application only requires input-output data and not the direct accessibility to the code.

With respect to the second issue, we propose a wrapper-based heuristic approach to select the set of most relevant features of the seismic model, for deciding a priori the seismic scenarios to be simulated. The proposed approach is based a Multi-Objective Differential Evolution Algorithm (MODEA) and is developed with reference to a case study whose objective of the analysis is calculating the annual rate of a threshold exceedance of the height of tsunami waves caused by subduction earthquakes that might be generated on a section of the Hellenic Arc and propagated to a target site on the eastern coast of Sicily (Siracusa). The comparison between the mean values of annual rate of exceedance of the tsunami wave height estimated considering only the selected scenarios and the full set of scenarios shows that the proposed approach allows a reduction of 95% of the number of scenarios with half of the features to be considered, and with no appreciable loss of accuracy.

Keywords: Probabilistic Seismic Hazard Assessment (PSHA), Modularised Global Sensitivity Analysis (MGSA), Bootstrapped Modularised Global Sensitivity Analysis (MGSA), Seismic Probabilistic Tsunami Hazard Analysis (SPTHA); Scenario selection; Feature selection; Wrapper approach; Multi-Objective Differential Evolution Algorithm (MODEA).

Abstract in lingua italiana

Gli eventi NaTech (*Natural Hazard Triggering Technological Disasters*) sono incidenti industriali causati da calamità naturali con possibili impatti disastrosi su ambiente e popolazione. La valutazione del rischio di NaTech dovuti a tsunami sismogenerati implica la caratterizzazione e la modellazione delle sorgenti sismiche per calcolare *Peak Ground Acceleration* (PGA), l'input di modelli e simulazioni di tsunami, necessaria per una *Seismic Probabilistic Tsunami Hazard Analysis* (SPTHA). In questa tesi proponiamo due metodi di *Sensitivity Analysis* (SA) per affrontare due problemi computazionali:

1. Individuazione dei parametri del modello che influenzano maggiormente la PGA;
2. Individuazione delle *feature* del modello sismico più rilevanti, per decidere a priori gli scenari sismici da simulare.

Riguardo il primo problema, proponiamo un nuovo metodo di *Bootstrapped Modularised Global Sensitivity Analysis* (BMGSA), testandolo su un caso di studio di riferimento e confrontando i risultati con un metodo standard di Global SA *variance-based*. Il punto di forza del metodo proposto è che la sua applicazione necessita solo di dati di input-output e non dell'accesso diretto al codice.

Riguardo il secondo problema, proponiamo un approccio euristico *wrapper-based*, per selezionare l'insieme delle caratteristiche più rilevanti del modello sismico, per decidere a priori gli scenari sismici da simulare. L'approccio proposto si basa su un Algoritmo di Evoluzione Differenziale Multi-Obiettivo (MODEA) ed è sviluppato con riferimento ad un caso studio il cui obiettivo è il calcolo del tasso annuale di superamento di una altezza soglia delle onde di tsunami causate da terremoti di subduzione che potrebbero essere generate su un tratto dell'Arco Ellenico e propagate ad un sito target sulla costa orientale della Sicilia (Siracusa). Il confronto tra i valori medi del tasso annuale di superamento dell'altezza dell'onda di tsunami stimata considerando solo gli scenari selezionati e l'insieme completo degli scenari mostra che l'approccio proposto permette una riduzione del 95% del numero di scenari con la metà delle *feature* da considerare, e senza apprezzabili perdite di precisione.

Parole chiave: Probabilistic Seismic Hazard Assessment (PSHA); Modularised Global Sensitivity Analysis (MGSA); Bootstrapped Modularised Global Sensitivity Analysis (MGSA); Seismic Probabilistic Tsunami Hazard Analysis (SPTHA); Scenario selection; Feature selection; Wrapper approach; Multi-Objective Differential Evolution Algorithm (MODEA).

Contents

Introduction

NaTech events (Natural Hazard Triggering Technological Disasters) are industrial accidents triggered by natural hazards (i.e., hurricanes, floods, earthquakes, tsunamis, etc.) which may lead to losses of hazardous materials with potentially tremendous impact on the environment and the surrounding population [1], [2]. Industrial facilities located in coastal areas are exposed to tsunami NaTech and the associated potential flooding resulting in damage or collapse of buildings, tanks or other equipment, possibly causing the release of contaminants [3]. The Niigata (1964) and Tohoku (2011) earthquakes and tsunamis, for example, resulted in oil spread from an oil refinery plant [4] and radioactive release of material from a nuclear power plant [5], respectively. This emphasises the hazard posed by earthquakes and subsequent tsunamis which may trigger accidents (e.g., Tohoku earthquake and tsunami) [6]–[8]. Tsunami NaTech are, indeed, typically triggered by earthquakes occurring offshore or in the proximity of the coastline in active subduction zones, resulting in a sudden deformation of the seafloor that perturbates and displaces the entire water column above it, i.e., generating a tsunami [9]. To manage tsunami threat, tsunami hazard and risks methodologies have been developed through time to quantify the tsunami hazard and the potential consequent risks [9], [10].

Early on, “worst credible”/ “worst case” scenarios approaches have been adopted [11]–[14]. These approaches consist in the postulation of conservative scenarios, which are then simulated by high-resolution codes to verify the response of the system and its safety barriers. For example, in [3] a worst case analysis of tsunamis impacting an oil refinery is reported, whereas in [15] an application to a nuclear power plant is described. However, these approaches have proven to be limited in modelling seismic sources as well as tsunamis [16], due to the large uncertainty, both epistemic and aleatory, given by the scarcity of tsunami observations [17].

In this thesis, we focus on an approach called Seismic Probabilistic Tsunami Hazard Analysis (SPTHA) [18]. To overcome “worst credible”/ “worst case” scenarios analyses, SPTHA is aimed at estimating, for a certain location, the annual rate of exceedance of a seismic-induced tsunami wave with respect to a predefined threshold. The analysis relies on computationally demanding numerical simulations of seismic-induced tsunami wave generation and propagation.

SPTHA entails performing:

i) Seismic sources characterisation and modelling,

ii) Seismic-induced tsunamis modelling and simulation.

Probabilistic Seismic Hazard Analysis (PSHA) is aimed at characterising and modelling seismic sources by assessing, at a given target location and for a given exposure time window ΔT, the probability that a given intensity measure (IM) of the ground motion, typically the Peak Ground Acceleration (PGA), exceeds a threshold value γ [19]. The output of the PSHA are hazard curves, defined by quantifying the mean annual rates of exceedance of a set of intensity measures (IM) values. Considering PGA as the IM and assuming a Poisson process, as the model of earthquake occurrence, with parameter λ_H denoting the mean annual rate of exceedance of the γ-th PGA level, the probability of interest is calculated:

$$P(PGA > \gamma, \Delta T) = 1 - \exp[-\lambda_H(PGA > \gamma)\Delta T] \tag{1}$$

Since the propagation of the earthquake wave in the soil is typically evaluated by empirical relationships, called Ground Motion Prediction Equations (GMPEs), λ_H is quantified by means of the total probability theorem as [20]:

$$\lambda_H(PGA > \gamma) = \lambda \int_{m_{min}}^{m_{max}} \int_0^r P(PGA > \gamma | m, r) f_m(m) f_r(r) dmdr \tag{2}$$

where λ is the mean annual rate of earthquake occurrence at a given source location (i.e., the number of occurrence of earthquakes with intensity of PGA above a given threshold per year); the distribution $f_m(m)$ describes the probability distribution of different earthquake magnitudes, typically assumed to follow a truncated Gutenberg-Richter distribution within the interval of values $[m_{min}; m_{max}]$ and slope parameter b [20]; $f_r(r)$ describes the probability distribution of the source-to-target distance r, assuming a spatial distribution for earthquakes [19]. These input distributions are typically determined from historical, instrumental, and geological observations [19], [21], but large epistemic uncertainty exists, so that many alternative parametrisations of the model are possible [22], [23].

Tsunami hazard is classically assessed by simulation of either one "worst credible" or few representative scenarios [18], [24]. This can be an effective approach when (i) the effects of frequent, small magnitude earthquakes are expected to be negligible compared to those less frequent large magnitude earthquakes, and (ii) the analysis is conducted in a relatively simple geophysical context where tsunami hazard is dominated by the large magnitude earthquakes occurring in subduction zones, whose geometries are reasonably well constrained [18]. On the other hand, when tsunamis are generated in complex and fragmented tectonic environments (e.g., the Caribbean Sea and the Mediterranean Sea) or when relatively short return periods need to be considered, the tsunami hazard might be severely biased [18]. To explicitly account for the whole spectrum of seismic triggering events and their related uncertainty, a probabilistic analysis of a large set of potential tsunamis can be performed (Seismic Probabilistic Tsunami Hazard Analysis, SPTHA) [9], [24]. Specifically, SPTHA aims to estimate the probability that the height ψ of an earthquake-induced tsunami wave exceeds a threshold $\tilde{\psi}$, within in an exposure time ΔT, at a location of coordinates $\bar{a}$ [9]. Each tsunami is assumed to be generated by a seismic scenario σ_x belonging to the space of possible seismic scenarios Σ ($\sigma_{\bar{x}} \in \Sigma$), characterized by parameters $\bar{x}$ and occurring with annual frequency $\lambda(\sigma_{\bar{x}})$ considering a Poisson process for the wave exceedance event occurrence in time, the probability of exceedance P_e can be written as:

$$P_e = Pr(\psi_{\bar{a}} \geq \tilde{\psi}; \Delta T) \approx 1 - exp(-\Lambda(\psi_{\bar{a}} \geq \tilde{\psi})\, \Delta T) \tag{3}$$

where $\Lambda(\psi_{\bar{a}} \geq \tilde{\psi})$ is the annual rate of occurrence of a tsunami of intensity $\psi_{\bar{a}} \geq \tilde{\psi}$ at location $\bar{a}$. This rate is calculated by integrating, over the space Σ, the annual frequency $\lambda(\sigma_{\bar{x}})$ of occurrence of the seismic scenario $\sigma_{\bar{x}}$ times the probability $Pr(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}})$ that the tsunami wave generated by the scenario exceeds $\tilde{\psi}$:

$$\Lambda(\psi_{\bar{a}} \geq \tilde{\psi}) = \int_{\Sigma} \lambda(\sigma_{\bar{x}}) Pr(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}}) d\sigma_{\bar{x}} \tag{4}$$

Considering, without loss of generality and for the sake of simplicity, a set of Q discretized seismic scenarios $\sigma_{\bar{x}_q}$ ($q = 1, \dots, Q$) with $\lambda\left(\sigma_{\bar{x}_q}\right)$ and $Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_q}\right)$, Eq. (4) can be approximated as:

$$\Lambda(\psi_{\bar{a}} \geq \tilde{\psi}) \approx \sum_{q=1}^{Q} \lambda\left(\sigma_{\bar{x}_q}\right) Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_q}\right) \tag{5}$$

To account for epistemic uncertainty, M alternative formulations of $\lambda\left(\sigma_{\bar{x}_q}\right)$ and $Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_q}\right)$ can be considered, producing M alternative quantifications of both factors in Eq. (5). The mean hazard rate can, then, be evaluated as:

$$\Lambda(\psi_{\bar{a}} \geq \tilde{\psi}) \approx \frac{1}{M} \sum_{m=1}^{M} \sum_{q=1}^{Q} \lambda\left(\sigma_{\bar{x}_q}\right)_m Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_q}\right)_m \tag{6}$$

where $\lambda\left(\sigma_{\bar{x}_q}\right)_m$ is the generic entry of the matrix $\overline{\overline{\lambda(\sigma_{\bar{x}})}}$:

$$\overline{\overline{\lambda(\sigma_{\bar{x}})}} = \begin{pmatrix} \lambda(\sigma_{\bar{x}_1})_1 & \cdots & \lambda(\sigma_{\bar{x}_1})_m & \cdots & \lambda(\sigma_{\bar{x}_1})_M \\ \vdots & & \vdots & & \vdots \\ \lambda\left(\sigma_{\bar{x}_q}\right)_1 & \cdots & \lambda\left(\sigma_{\bar{x}_q}\right)_m & \cdots & \lambda\left(\sigma_{\bar{x}_q}\right)_M \\ \vdots & & \vdots & & \vdots \\ \lambda\left(\sigma_{\bar{x}_Q}\right)_1 & \cdots & \lambda\left(\sigma_{\bar{x}_Q}\right)_m & \cdots & \lambda\left(\sigma_{\bar{x}_Q}\right)_M \end{pmatrix} \tag{7}$$

and $Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_q}\right)_m$ is the generic entry of the matrix $\overline{\overline{Pr(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}})}}$:

$$\begin{aligned} &\overline{\overline{Pr(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}})}} \\ &= \begin{pmatrix} Pr(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_1})_1 & \cdots & Pr(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_1})_m & \cdots & Pr(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_1})_M \\ \vdots & & \vdots & & \vdots \\ Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_q}\right)_1 & \cdots & Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_q}\right)_m & \cdots & Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_q}\right)_M \\ \vdots & & \vdots & & \vdots \\ Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_Q}\right)_1 & \cdots & Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_Q}\right)_m & \cdots & Pr\left(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}_Q}\right)_M \end{pmatrix} \end{aligned} \tag{8}$$

Note that the calculation of the entries of $\overline{\overline{Pr(\psi_{\bar{a}} \geq \tilde{\psi} | \sigma_{\bar{x}})}}$ may result computationally burdensome, for example when using highly non-linear tsunami simulation models.

In the particular case of a local SPTHA for the estimation of inundation hazard curves for a small target site, e.g., a refinery, high-resolution inundation simulations are needed. This requires either large High Performance Computing (HPC) resources [25], [26] or a reduction of the number of simulations by, for example a two-stage filtering procedure [18], [26], [27], or training a metamodel, for example an Adaptive

Kriging that mimics the behaviour of the computationally demanding tsunami inundation simulator, e.g., [28], [29].

On the other hand, when the interest is in a regional SPTHA for the estimation of inundation hazard curves for large areas, such as countries or continents, the computation is usually performed by using simplified relationships between the water elevation at the shoreline and the maximum inundation height [30], [31].

In this thesis, we propose two novel Sensitivity Analysis (SA) methods to address the aforementioned computational issues related with SPTHA, namely:

1. A Bootstrapped Modularised method of Global Sensitivity Analysis for Probabilistic Seismic Hazard Assessment to identify the model parameters most affecting the PGA (that ultimately affects the tsunami wave height);
2. A heuristic features selection approach for scenario analysis of a Regional Seismic Probabilistic Tsunami Hazard Assessment to identify the features of the seismic model worthy to be fed to a seismic-induced tsunami simulation code.

SA can aid the understanding of how the uncertainty in the model is apportioned among the model input parameters uncertainties [32], [33]. In other words, through SA, one can identify the most sensitive parameters and better focus the uncertainty analysis without losing accuracy. Different SA techniques have been proposed in literature, which can be sorted into three main categories: local, regional, and global [34]. Local and regional analyses limit inputs variations to a subset of their overall ranges. Local methods evaluate at low computational costs the effects on the system response of small perturbations in the model input variables around fixed values [34]. Then, local SA provides information on the sensitivity of the model output to the input variability at some fixed points. Regional analyses, on the contrary, focus on calculating the sensitivity of the model output to the variability of the inputs varying in given ranges of the inputs; yet, they do not give complete account to the uncertainty of the model inputs, in terms of their distributions [35], [36]. Global Sensitivity Analysis (GSA) methods, instead, explore the whole distribution range of the model inputs and the effects of their mutual combination on the model output, but they do so at larger computational costs than local and regional methods [35], [36]. GSA methods can be regression-based [37], variance-based [36], [38], distribution-based [39], [40] and expected value of information (EVI)-based [41].

Among the GSA methods, variance decomposition based on Sobol indices is most widely used [42].

Sobol indices measure that contribution the input variables provide, individually or in groups, to the variability of the model output [35], [43]. They are usually computed via a double-loop Monte Carlo Simulation (MCS), with computational cost equal to $C = v \cdot n_1 \cdot n_2$, where v is the number of input variables, n_1 the sample size for estimating the inner loop and n_2 the sample size for the outer loop [44], [45]. When the model runs are time consuming, the computational cost is high and strategies have been proposed to reduce it, including: reduced-order models calibrated on input-output data obtained by few runs of the original model, e.g., Bayesian approaches [46], kriging [47], and polynomial chaos expansion [48], [49]; sampling schemes tailored to efficiently characterise the sensitivity of the model inputs, e.g., Fourier Amplitude Sensitivity Test (FAST) [50], [51] and Effective Algorithm for computing global Sensitivity Indices (EASI) [52]; finally, data-driven approaches allow exploiting available datasets to calculate sensitivity measures [42], [53], which is quite of interest for many practical applications in which an input-output dataset is available and models to perform a GSA using MCS-based methods cannot be run [42], [54].

The novel Modularised GSA (MGSA) method, developed to identify the input variables which the output of a seismic model is most sensitive to, assuming that only an input-output dataset is given and with no need of repeating hazard computations, is sketched in the flowchart of Figure 1 and described in "A Bootstrapped Modularised method of Global Sensitivity Analysis for Probabilistic Seismic Hazard Assessment". It consists of, first, applying a Bootstrap technique to the available input-output dataset to artificially increase the amount of data available [55], [56]. Then, for each d-th dataset, a sensitivity index is calculated for each input variable. Without loss of generality, in this work we propose to calculate the first-order Sobol index, which measures the input variables individual contributions to the variability of the model output [35], [43]: in practice, for each input variable, the d-th dataset is modularised (i.e., partitioned) into sub-sets that are used to calculate the variance of the model output Y and the first-order Sobol index [57]. Finally, the D independent rankings of the input variables, obtained based on their first-order Sobol indices values, are ensembled to provide an aggregated ranking of the input variables which the output is sensitive to. Typical ensemble strategies are Bottom-Up

(BU) and All-Out (AO) strategies: the former computes a ranking order of the input variables out of each d-th dataset and combines the D alternative rankings a posteriori to generate a final aggregated ranking order [32]; on the contrary, the latter merges a priori the information from the D datasets by averaging the D Sobol indices for each input parameter and, then, provides the final ranking [32].

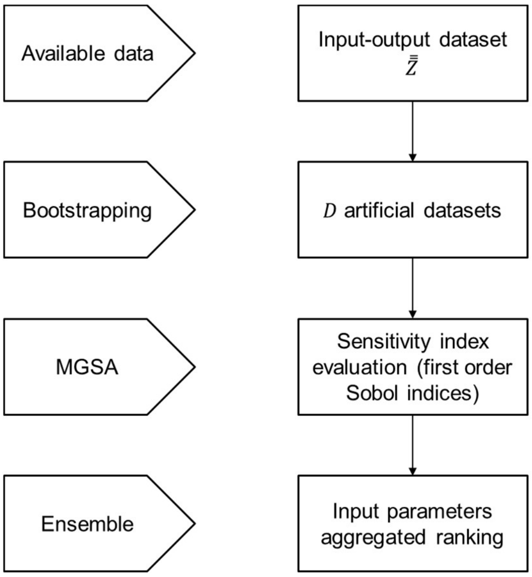

Figure 1: Flowchart of the proposed method.

The proposed method is tested on a hypothetical case study with a point seismic source and a nearby target point, where the hazard intensity corresponding to 10% probability of being exceeded in 50y is to be calculated. The results are compared to those obtained by a standard variance-based GSA method [33], which is the state-of-practice approach when the simulation model is available.

This thesis considers regional SPTHA and illustrates a novel approach for identifying the relevant features of the seismic scenarios and the selection of a limited number of them needed for performing the annual rate estimation with sufficient accuracy. Specifically, a Multi-Objective Differential Evolution Algorithm (MODEA) is used to select the features [58].

The proposed approach, described in the chapter "A heuristic feature selection approach for scenario analysis of a Regional Seismic Tsunami Hazard Assessment", is developed with reference to a case study whose objective of the analysis is calculating the annual rate of exceedance of a threshold $\tilde{\psi} = 1m$ of tsunami wave height, resulting from subduction earthquakes in a section of the Hellenic Arc. The target site $\bar{a}$ for the propagation of the wave is at Siracusa, on the eastern coast of Sicily. The case study considers the crustal seismicity generated in the Kefalonia-Lefkada region, thus developing outside the subduction interface of the Hellenic Arc [59], [60]; this is one of the regions considered in the tsunami hazard model recently released for the NEAM region [60], [61]. The source area comprises a total of $Q_{tot} = 23272$ seismic scenarios and $M = 1000$ alternative models for the calculation of $\Lambda(\psi_{\bar{a}} \geq 1m)$.

A comparison is provided between the value of the mean annual rate of exceedance estimated considering only the selected scenarios SPTHA and the full set of scenarios SPTHA. The outcome of the comparison shows that the proposed approach allows a significant reduction of the number of scenarios needed without affecting the accuracy of the estimate.

The structure of the thesis is as follows: in Chapter 1 "A Bootstrapped Modularised method of Global Sensitivity Analysis for Probabilistic Seismic Hazard Assessment" is presented; in Chapter 2 "A heuristic feature selection approach for scenario analysis of a Regional Seismic Tsunami Hazard Assessment" is presented; finally in Chapter 3 conclusions are drawn.

1. A Bootstrapped Modularised method of Global Sensitivity Analysis for Probabilistic Seismic Hazard Assessment

With regards to the possibility of earthquake occurrence at a given location, Probabilistic Seismic Hazard Assessment (PSHA) evaluates the probability of exceedance of a given earthquake intensity measure like the Peak Ground Acceleration (PGA), at a target point for a given exposure time. The stochasticity of the occurrence of seismic events is modelled by stochastic processes and the propagation of the earthquake wave in the soil is typically evaluated by empirical relationships called Ground Motion Prediction Equations (GMPEs). The large epistemic uncertainty affecting PSHA is quantified by defining alternative model settings and/or model parametrisations. In this work, we propose a novel Bootstrapped Modularised Global Sensitivity Analysis (BMGSA) method for identifying the model parameters most important for the epistemic uncertainty in PSHA. The method consists in:

1. Generating alternative artificial datasets by bootstrapping an available input-output dataset;
2. For each alternative bootstrapped dataset, calculating a sensitivity index with the modularised method;
3. Aggregating the individual rankings obtained from each alternative bootstrapped dataset, with Bottom-Up/All-Out strategies.

The proposed method is tested on a benchmark case study. The results are compared with a standard variance-based Global Sensitivity Analysis (GSA) method of literature. The novelty and strength of the proposed BMGSA method is that its application only requires input-output data and not the direct accessibility to the PSHA code.

1.1 The novel Bootstrapped Modularised GSA

Let us consider a model g whose output value $Y \in \mathbb{R}$ depends on the values of uncertain input parameters $\bar{X} = (X_1, X_2, \dots, X_N)$:

$$Y = g(\bar{X}) \tag{9}$$

Let us also assume that an input-output data set $\bar{\bar{Z}}$ is given (i.e., the analyst may not dispose the simulation code):

$$\bar{\bar{Z}} = \begin{pmatrix} \bar{Z}^1 \\ \vdots \\ \bar{Z}^S \end{pmatrix} = \begin{pmatrix} x_1^1 & \cdots & x_N^1 & y^1 \\ \vdots & \ddots & \vdots & \vdots \\ x_1^S & \cdots & x_N^S & y^S \end{pmatrix} \tag{10}$$

where $\bar{Z}^s = [\bar{x}^s, y^s], \forall\, s = 1, \dots, S$ is the s-th input-output pattern of $\bar{Z} = (X_1, X_2, \dots, X_N, Y)$

The proposed methodology consists in:

1. Generating D alternative bootstrapped artificial datasets from the available input-output dataset $\bar{\bar{Z}}$ [62];
2. From each d-th alternative dataset and for each input variable, calculating a sensitivity index (here the first-order Sobol index) with the modularised method [57];
3. Aggregating the D individual rankings (one for each alternative dataset) with Bottom-Up/All-Out strategies [32].

1.1.1 Generation of the Bootstrapped datasets

Bootstrap is a computer-based method usually employed to assess the accuracy of statistical estimates with minimum assumptions [56]. The main benefit is avoiding additional computational burden (for example, when simulation codes are computationally demanding or not available, as in the current case) by relying only on the available data [56], [63], which makes it particularly fit for the purpose of this work. The basic idea is to generate a number D of artificial datasets $\bar{\bar{Z}}_d$ by random sampling, with replacement from $\bar{\bar{Z}}$, the input-output patterns. The generic bootstrapped $\bar{\bar{Z}}_d$, thus, consists in a $S \times (N+1)$ matrix (see Figure 2).

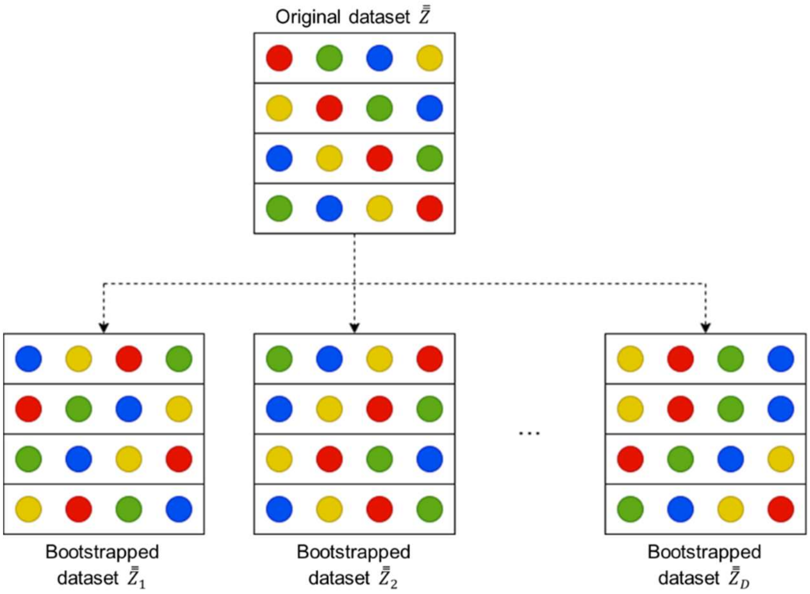
Original dataset $\bar{\bar{Z}}$
...
Bootstrapped dataset $\bar{\bar{Z}}_1$
Bootstrapped dataset $\bar{\bar{Z}}_2$
Bootstrapped dataset $\bar{\bar{Z}}_D$

Figure 2: Bootstrap replicates of an original dataset.

1.1.2 The modularised method to calculate the Sobol index

The Sobol index is the result of the application of a variance-based method that apportions the output variance into the single (or groups of) variables variances [64], [65]. No hypotheses are made on the structure of the model g from which the data have been generated. The variance $Var[Y]$ of the output Y can be, indeed, decomposed as follows [35], [36]:

$$Var[y] = Var_{X_n}\left[\mathbb{E}_{X_{\sim n}}(y|x_n)\right] + \mathbb{E}_{X_n}\left[Var_{X_{\sim n}}(y|x_n)\right] \tag{11}$$

where:

- $Var_{X_n}\left[\mathbb{E}_{X_{\sim n}}(y|x_n)\right]$ is the variance of Y caused by X_n without considering its interactions with other input variables (i.e., $X_{\sim n}$)
- $\mathbb{E}_{X_n}\left[Var_{X_{\sim n}}(y|x_n)\right]$ is the variability of Y depending on all variables but on n (i.e., $X_{\sim n}$);
- $\mathbb{E}(\cdot)$ is the expectation operator;

The first-order Sobol index for the n-th generic input variable is defined as [35], [36]:

$$S_n = \frac{Var_{X_n}\left[\mathbb{E}_{X_{\sim n}}(y|x_n)\right]}{Var[y]}, \forall\, n = 1, \dots, N \tag{12}$$

The larger S_n, the more X_n contributes to the variance of Y [64]. As mentioned in the Introduction, the computation of the Sobol indices usually requires a double-loop MCS, which can be computationally burdensome. Since, in practice, the numerator is solved by a double-loop MCS [36]:

- The inner loop computes $\mathbb{E}_{X_{\sim n}}(y|x_n)$ using n_1 random samples of $\bar{X}_{\sim n}$ with fixed $\bar{X}_n$;
- The outer loop computes $Var_{X_n}\left[\mathbb{E}_{X_{\sim n}}(y|x_n)\right]$ by iterating the inner loop n_2 times, with different values of X_n;

In total, for each S_n, the number of model evaluations is $C_n = n_1 \cdot n_2$, that is unaffordable if each evaluation is time-consuming (notice that, in many practical applications, each loop must be of order greater than 1000 [34], [45]). To address this issue and avoid calling the simulation code, we propose a modularised approach, that partitions the $\bar{\bar{Z}}_d$ into subsets and proceeds as follows [44], [45].

Step 1: construct the reduced matrix

For each n-th input variable, append the n-th input column of $\bar{\bar{Z}}_d$ to the output column; then, shuffle the rows in ascending order to obtain $\bar{\bar{Z}}_d^*$, where $x_n^{1*} \leq x_n^{2*} \leq \dots \leq x_n^{S*}$:

$$\bar{\bar{Z}}_d^* = \begin{pmatrix} x_n^{1*} & y^{1*} \\ \vdots & \vdots \\ x_n^{S*} & y^{S*} \end{pmatrix} \tag{13}$$

Step 2: partition the reduced matrix in subsets

Partition the support of X_n in $k = 1, \dots, K$ mutually exclusive subsets $\bar{\bar{Z}}_d^{*k}$, such that $\cup_{k=1}^{K} X_n^k = X_n \wedge X_n^k \cap X_n^{l \neq k} = \emptyset$. Operatively, divide the resulting matrix $\bar{\bar{Z}}_d^*$ into $k = 1, \dots, K$ submatrices $\bar{\bar{Z}}_d^{*k}$ of J rows, each retaining the order of the Step 1:

$$\bar{\bar{Z}}_d^{*k} = \begin{pmatrix} x_n^{1*k} & y^{1*k} \\ \vdots & \vdots \\ x_n^{J*k} & y^{J*k} \end{pmatrix} \tag{14}$$

Note that $J \cdot K = S$, where S is the total input-output pattern size and $K = int(\sqrt{S})$ [44], [53]. Notice that the large K improves the accuracy of the estimation of $Var_{X_n}[\cdot]$, while worsening the accuracy of the estimation of $\mathbb{E}_{X_{\sim n}}(y|x_n)$, and vice versa [44], [66].

Step 3: estimation of the Sobol index

The Sobol index $S_{n,d}$ is, finally, calculated as:

$$S_{n,d} = \frac{Var_{X_n}[\mathbb{E}_{X_{\sim n}}(y|x_n)]}{Var[y]} \approx \frac{\frac{1}{K}\sum_{k=1}^{K}(\bar{y}_k - \bar{y})^2}{\frac{1}{S}\sum_{s=1}^{S}(y_s - \bar{y})^2} \tag{15}$$

where: $\bar{y} = \frac{1}{S}\sum_{s=1}^{S} y^s$ and $\bar{y}_k = \frac{1}{J}\sum_{j=1}^{J} y^{j*k}$.

The calculation of $S_{n,d}$, as a result of the modularisation, depends solely on X_n and Y, and can be performed even if the input values of $\bar{X}_{\sim n}$ are not available [44], [45].

1.1.3 Ensemble of the alternative rankings

Each input variable X_n has been, thus, assigned a $S_{n,d}$. Input variables can, accordingly, be ranked from the most important (largest $S_{n,d}$) to the least contributor to the variance (smallest $S_{n,d}$). Two strategies are explored for ensembling the D available alternative rankings: the BU and AO strategies [32].

1.1.3.1 Bottom-Up strategy

Each d-th bootstrapped dataset $\bar{\bar{Z}}_d$ is treated separately from the others to generate its input ranking $\bar{R}_{BU,d}$. Then, the set of input rankings obtained from all the D datasets is processed a posteriori to give the final aggregated ranking order $\bar{R}_{BU}$ (Figure 3) [32]. For each $\bar{\bar{Z}}_d$ the $S_{n,d}$ are computed and, by sorting in ascending order, the corresponding ranking $\bar{R}_{BU,d}$ is obtained. The final ranking order $\bar{R}_{BU}$ is obtained applying the Borda method, that consists in computing the Borda Count (BC) for each input variable [32]. Denoting by $p_{n,d}$ the n-th variable order inside the d-th ranking, the BC for the input variable X_n is given by [32]:

$$BC_n = \sum_{d=1}^{D} p_{n,d}. \tag{16}$$

A small value of BC_n means that the n-th input variable is among the most important (top ranked) input variables [32].

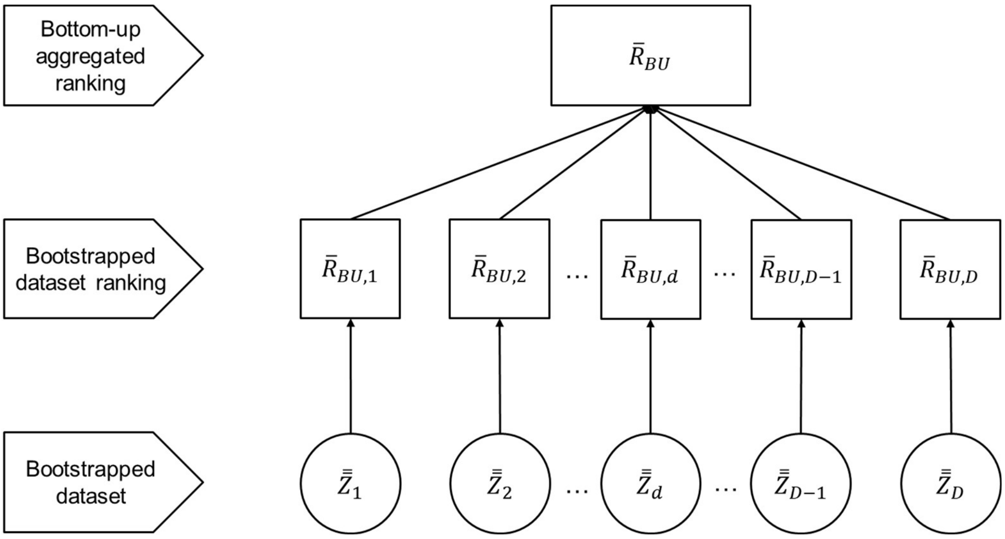

Figure 3: The proposed bottom-up aggregation strategy.

1.1.3.2 All-Out strategy

The AO strategy a priori merges the information coming from each dataset $\bar{\bar{Z}}_d$ (see Figure 4). For each input variable X_n, the expected value $\mathbb{E}(S_n)$ of each Sobol index S_n is computed over the D datasets [32]:

$$\mathbb{E}(S_n) = \frac{1}{D}\sum_{d=1}^{D} S_{n,d} \tag{17}$$

Sorting $\mathbb{E}(S_n)$ in ascending order provides the AO aggregated ranking $\bar{R}_{AO}$ (the larger the value of $\mathbb{E}(S_n)$, the more important the n-th input variable) [32].

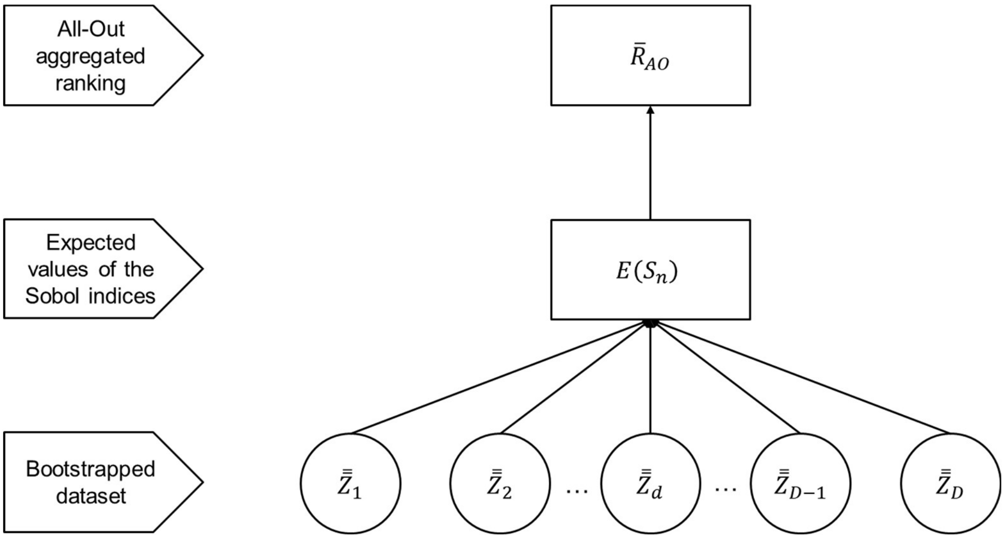

Figure 4: The proposed all-out aggregation strategy.

1.2 Case study

The proposed GSA methodology has been tested on a benchmark case study regarding the PSHA of one hypothetical seismic point source and one target point located in its proximity. The source model consists of a point source with a given mean annual rate and generating seismicity with magnitudes following a truncated Gutenberg-Richter distribution [67]. The source model is, then, coupled with a standard Ground Motion Prediction Equation (GMPE) for the propagation of the earthquake waves in the soil [68]. A reference target point is selected in the near field, at distance of approximately 10 km.

The epistemic uncertainty of the PSHA is evaluated with respect to six input parameters, accounting for a total of 16384 alternative computational settings, resulting in $\bar{\bar{Z}} = [16384 \times 7]$, as in Eq. (18) below. The purpose is to quantify the impact of these parameters on the epistemic uncertainty of the IM value corresponding to an exceedance probability of 10% in 50 years, with a mean return period of 475 years. The input parameters are $\bar{X} = (\sigma_{GMPE}, \lambda, m_{max}, m_{min}, b, r)$ where: σ_{GMPE} is the standard deviation of the GMPE, λ is the mean annual rate of seismic activity at the source location (i.e., the number of earthquakes per year of intensity magnitude m a minimum magnitude m_{min}), m_{min} and m_{max} are the minimum and the maximum magnitude parameters of the truncated Gutenberg-Richter distribution, whose slope is b [20], [67], and r is the source-to-target distance [20].

The latter parameter is here added to the other five input to emulate the dependence of GMPEs on source characteristics like earthquake depth, size or geometry. The output variable is the IM PGA, i.e., the reference peak ground acceleration at the target location that has annual rate of exceedance λ_H assumed to be equal to 1/475y.

$$\bar{\bar{Z}} = \begin{pmatrix} \sigma_{GMPE_1} & \lambda_1 & m_{max_1} & m_{min_1} & b_1 & r_1 & PGA_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{GMPE_s} & \lambda_s & m_{max_s} & m_{min_s} & b_s & r_s & PGA_s \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{GMPE_S} & \lambda_S & m_{max_S} & m_{min_S} & b_S & r_S & PGA_S \end{pmatrix} \tag{18}$$

The epistemic distributions of all the six input variables $\bar{X}$ are reported in Table 1. Notably, the source-target distance is set around 10 km, thus in the very near field, in which the dependence of distance on the source characteristics (geometry, depth, and dimension) is more pronounced. Consequently, a quite large variance is set for r. The other parameters are inspired from the ones adopted in the areal sources of the PSHA study that is enforced by law in Italy, MPS04 [69], [70]. In particular, the parameters of the source model ($\lambda, m_{min}, m_{max}$ and b) are inspired by source zone 920 (Val di Chiana-Ciociaria) of MPS04, with a reduced value m_{min} (from 4.76 to 4.5) and λ (as we are considering a point source). The central value of σ_{GMPE} is instead taken from [68]. Variance value representing the epistemic uncertainty on the parameters, are set based on expert judgement.

Input variable	Units	Type of distribution	Mean value	Standard
σ_{GMPE}	g_0 (m/s^2)	Normal	0.3446	0.0490
λ	yr^{-1}	Normal	0.0600	0.0021
m_{max}	-	Normal	5.6791	0.2430
m_{min}	-	Normal	4.5005	0.1000
b	-	Normal	1.9597	0.0580
r	km	Normal	10.0142	2.9639

Table 1: Model input variables and output, with their associated distributions.

The results of the proposed method are compared to those obtained by a standard variance-based GSA method [33], which is the state-of-practice approach when the simulation model is available.

1.3 Results

The BMGSA methodology described in Section 2 has been applied to the case study presented in Section 3. The original dataset $\bar{\bar{Z}}$ has been replicated by bootstrap to generate $D = 1000$ datasets. Each replicate matrix $\bar{\bar{Z}}_d$ is comprised of $S = 16384$ rows (the input-output patterns) and 7 columns (6 input variables, 1 output). The partition size chosen to test the proposed methodology is $K = int(\sqrt{S}) = 128$.

The results of the assessment carried out with the standard GSA on the case study of Section 3 are taken as a benchmark, i.e., as correct ranking. This ranking is reported in Table 2, along with the values of the Sobol indices.

Notably, the main drivers of the epistemic uncertainty on the reference PGA are m_{min} and σ_{GMPE}. While the dependence on σ_{GMPE} is a well-established result [71], the dependence on m_{min} is not that straightforward, and it is probably due to the selection of a target point in the very near field (about 10 km) and the use of a very large value for the slope b-value (about 2). The combined effect of these two parameters is to produce a very large number of events with a magnitude very close to m_{min}, resulting in a critical dependence of the reference PGA on the selected minimum magnitude value.

Rank	1	2	3	4	5	6
Input variable	m_{min}	σ_{GMPE}	λ	r	m_{max}	b
$\boldsymbol{S_n}$	0.5896	0.3811	0.0470	0.0320	0.0264	0.0221

Table 2: Input variables ranking obtained with the standard GSA [2].

1.3.1 Results of the Bottom-Up strategy

The application of the ensemble BU strategy produces the final ranking obtained by the Borda method, shown in Table 3. The major limitation of the BU strategy is that the result is lumped in a ranking table that is not transparent with respect to the actual Sobol indices that generate that ranking and, finally, the analyst is not provided with any confidence measure on the resulting rank: in other words, it cannot be quantitatively assessed how much the generic n-th input X_n contributes to the variance of Y.

Rank	1	2	3	4	5	6
Input variable	m_{min}	σ_{GMPE}	λ	r	m_{max}	b

Table 3: Input variables ranking obtained through MGSA, D=1000 bootstrapped datasets, BU strategy.

1.3.2 Results of the All-Out strategy

The ranking and the expected values of the Sobol indices obtained with the ensemble strategy AO are reported in Table 4. As stated for the BU strategy, one can observe that:

1. The variables σ_{GMPE} and m_{min} are the first two (by far) more relevant inputs (Figure 5);
2. The other input variables bring a negligible contribution to the variability of the output.

Rank	1	2	3	4	5	6
Input variable	m_{min}	σ_{GMPE}	λ	m_{max}	r	b
$\mathbb{E}(\boldsymbol{S_n})$	0.5760	0.3693	0.0508	0.0359	0.0336	0.0249

Table 4: Input variables ranking obtained through MGSA, D=1000 bootstrapped datasets, AO strategy.

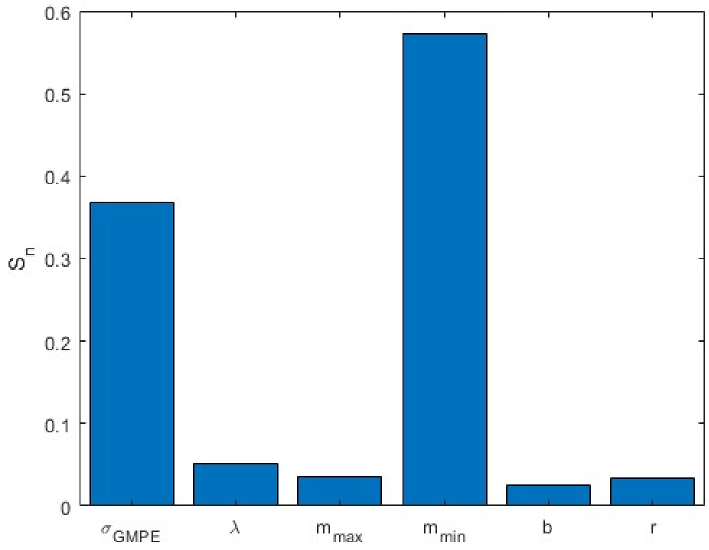

Figure 5: Sobol indices obtained with the AO strategy.

1.3.3 Comparison with the benchmark results

The results of the proposed methodology have been then compared with the ranking results of the standard GSA, reported in Table 5.

Rank	1	2	3	4	5	6
Standard GSA	m_{min}	σ_{GMPE}	λ	r	b	m_{max}
BU (BMGSA)	m_{min}	σ_{GMPE}	λ	m_{max}	r	b
AO (BMGSA)	m_{min}	σ_{GMPE}	λ	m_{max}	r	b
No bootstrap (MGSA)	m_{min}	σ_{GMPE}	r	m_{max}	λ	b

Table 5: Input variables rankings (sample size S=16384).

Both ensemble strategies and the standard GSA identify the σ_{GMPE} and m_{min} as the most important variables, whereas the sensitivity indices of the other input variables are negligible (Table 5, Figure 6). The disagreement regarding the ranking for the positions 4-6 may be due to hidden dependences and/or correlations between the input variables, as well as to the quantity of data upon which the rankings are drawn.

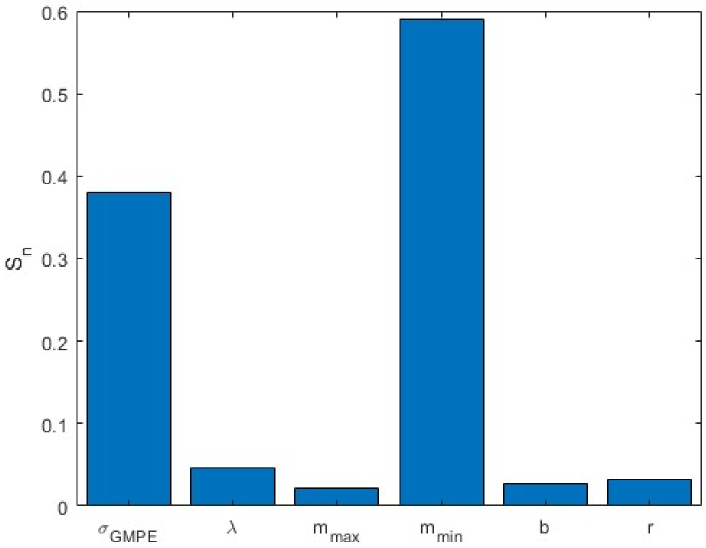

Figure 6: Sobol indices obtained with the standard GSA [2].

To highlight the important role played by the bootstrapping (Section 1.1.1) in obtaining such results, we show (Figure 7-Figure 11) the results that would have been obtained with a given input-output dataset $\bar{\bar{Z}}$ of decreasing size ($S = 16384, 8192, 4096, 2048, 1024$), employing the more transparent AO ensemble strategy (green squares in the Figure 7-Figure 11). These results are compared with i) the benchmark values (Standard GSA, blue diamonds in the Figure 7-Figure 11) and ii) the results obtained with the MGSA without bootstrap (magenta circles in the Figure 7-Figure 11). The relative rankings are reported in Table 6-Table 9.

When $\bar{\bar{Z}} = [16384 \times 7]$, as shown in Figure 7 and Table 5, the Standard GSA (blue diamonds in Figure 7), the BMGSA (green squares in Figure 7) and the MGSA (magenta circles in Figure 7) agree on the identification of m_{min} and σ_{GMPE} as the most important variables, whereas for the third most important variable only Standard GSA and BMGSA agree on λ. Then, the approaches provide different rankings for lower ranking positions.

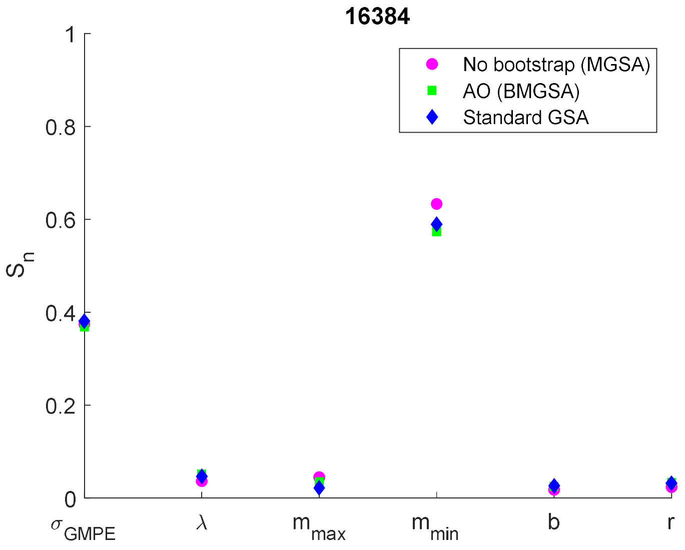

Figure 7: Sobol indices estimates at sample size S=16384.

When $\bar{\bar{Z}} = [8192 \times 7]$, as shown in Figure 8 and Table 6, the Standard GSA (blue diamonds in Figure 8), the BMGSA (green squares in Figure 8) and the MGSA (magenta circles in Figure 8) agree on the identification of m_{min} and σ_{GMPE} as the most important variables, whereas for the third most important variable only Standard GSA and BMGSA agree on λ. Then, the approaches provide different rankings for lower ranking positions.

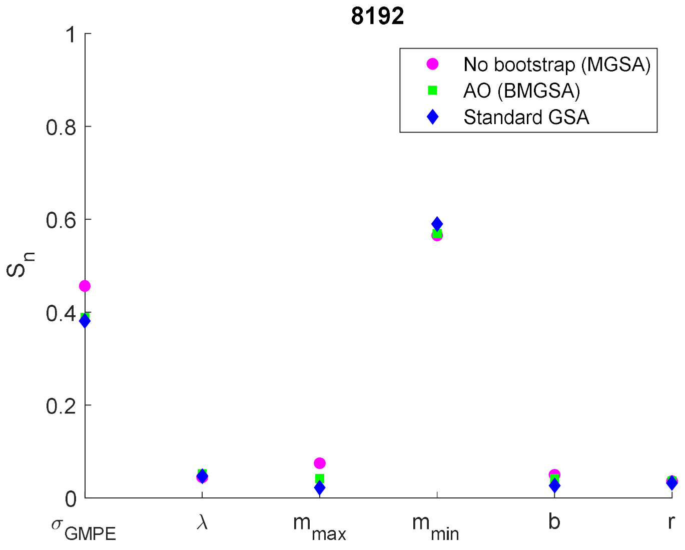

Figure 8: Sobol indices estimates at sample size S=8192.

Rank	1	2	3	4	5	6
Standard GSA	m_{min}	σ_{GMPE}	λ	r	b	m_{max}
AO (BMGSA)	m_{min}	σ_{GMPE}	λ	b	m_{max}	r
No bootstrap (MGSA)	m_{min}	σ_{GMPE}	m_{max}	b	λ	r

Table 6: Input variables rankings (sample size S=8192).

When $\bar{\bar{Z}} = [4096 \times 7]$, as shown in Figure 9 and Table 7, the Standard GSA (blue diamonds in the Figure) and the BMGSA (green squares in Figure 9) agree on the identification of m_{min} and σ_{GMPE} as the most important variables, as well as on third (λ) and fourth (r) most important variables. Then, the approaches provide different rankings for lower ranking positions. The MGSA (magenta circles in Figure 9) instead yields a completely different ranking (except for position 4). Notice that, when the dimension of $\bar{\bar{Z}}$ decreases, even if the most important variables are correctly identified, a less accurate estimation of the Sobol indices is provided and the differences between the GSA (blue diamonds in Figure 9), the BMGSA (green circles in Figure 9) and MGSA (magenta circles in Figure 9) increase.

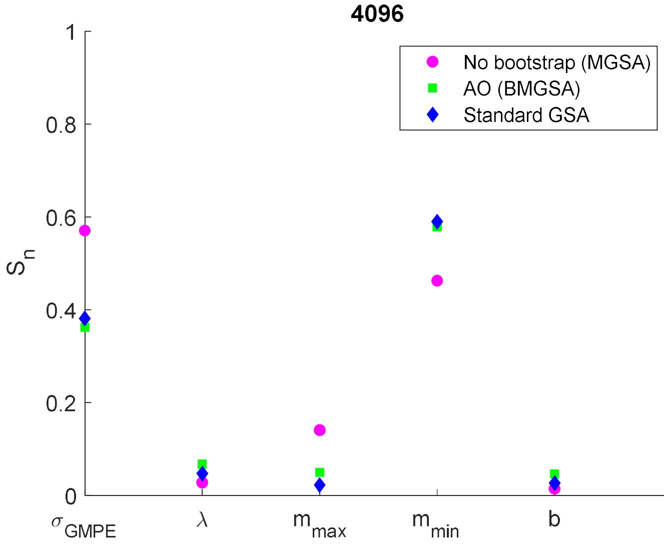

Figure 9: Sobol indices estimates at sample size S=4096.

Rank	1	2	3	4	5	6
Standard GSA	m_{min}	σ_{GMPE}	λ	r	b	m_{max}
AO (BMGSA)	m_{min}	σ_{GMPE}	λ	r	m_{max}	b
No bootstrap (MGSA)	σ_{GMPE}	m_{min}	m_{max}	r	λ	b

Table 7: Input variables rankings (sample size S=4096).

When $\bar{\bar{Z}} = [2048 \times 7]$, as shown in Figure 10 and Table 8, the Standard GSA (blue diamonds in Figure 10) and the BMGSA (green squares in Figure 10) agree on the identification of m_{min} and σ_{GMPE} as the most important variables, as well as on third (λ) and fourth (r) most important variables. Then, the approaches provide different rankings for lower ranking positions. The MGSA (magenta circles in Figure 10) instead yields a completely different ranking (except for position 5). Notice that, when the dimension of $\bar{\bar{Z}}$ decreases, even if the most important variables are correctly identified, a less accurate estimation of the Sobol indices is provided and the differences between the GSA (blue diamonds in Figure 10), the BMGSA (green circles in Figure 10) and MGSA (magenta circles in Figure 10) further increase with respect to the case with $\bar{\bar{Z}} = [4096 \times 7]$.

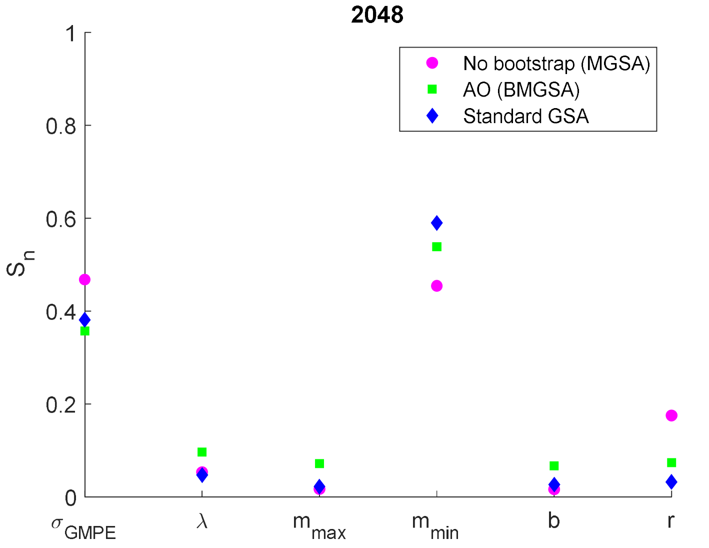

Figure 10: Sobol indices estimates at sample size S=2048.

Rank	1	2	3	4	5	6
Standard GSA	m_{min}	σ_{GMPE}	λ	r	b	m_{max}
AO (BMGSA)	m_{min}	σ_{GMPE}	λ	r	m_{max}	b
No bootstrap (MGSA)	σ_{GMPE}	m_{min}	r	λ	m_{max}	b

Table 8: Input variables rankings (sample size S=2048).

When $\bar{\bar{Z}} = [1024 \times 7]$, as shown in Figure 11 and Table 9, the Standard GSA (blue diamonds), the BMGSA (green squares in Figure 11) and the MGSA (magenta circles in Figure 11) agree on the identification of m_{min} as the most important variable, whereas for the second (σ_{GMPE}) and third (λ) most important variables only Standard GSA and BMGSA agree. Then, the approaches provide different rankings for lower ranking positions. Nevertheless, as Figure 11 clearly shows, the numerical values of the Sobol indices obtained with the proposed BMGSA may not be considered satisfactory. Furthermore, Figure 12-Figure 16 show that, when the dimension of $\bar{\bar{Z}}$ decreases, the distributions of $S_{n,d}$ become wider (i.e., bootstrap replicates are subject to noise and, as a result, the Sobol indices estimate are not precise).

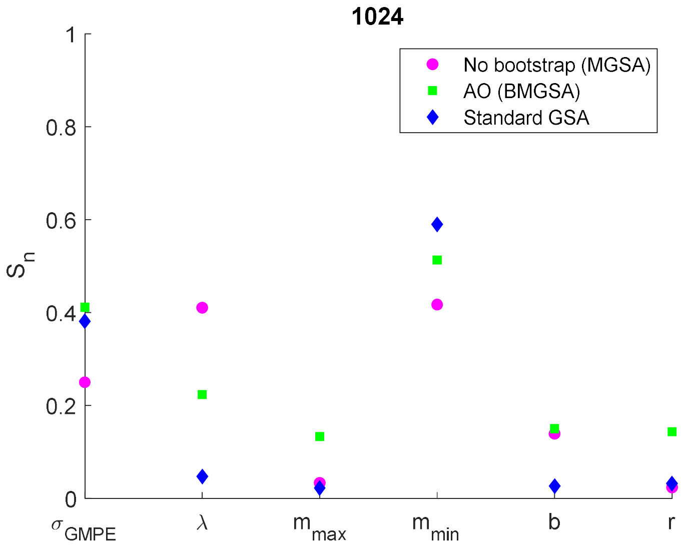

Figure 11: Sobol indices estimates at sample size S=1024.

Rank	1	2	3	4	5	6
Standard GSA	m_{min}	σ_{GMPE}	λ	r	b	m_{max}
AO (BMGSA)	m_{min}	σ_{GMPE}	λ	b	r	m_{max}
No bootstrap (MGSA)	m_{min}	λ	σ_{GMPE}	b	m_{max}	r

Table 9: Input variables rankings (sample size S=1024).

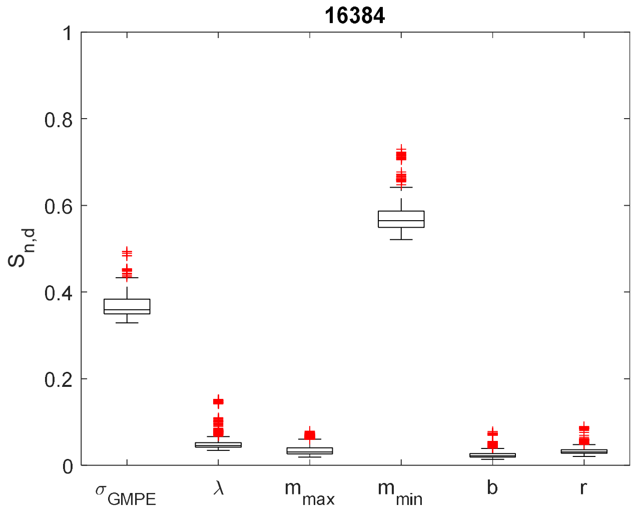

Figure 12: $S_{n,d}$ distributions at sample size S=16384

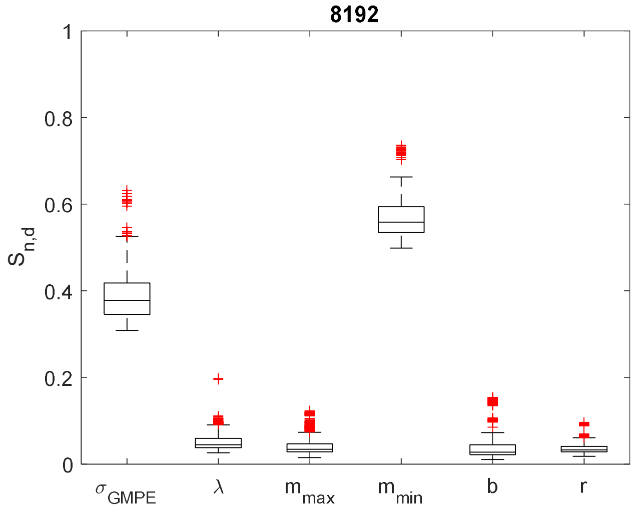

Figure 13: $S_{n,d}$ distributions at sample size S=8192

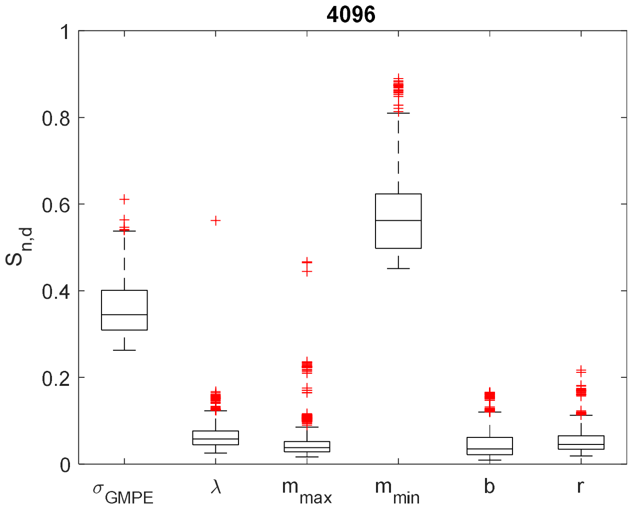

Figure 14: $S_{n,d}$ distributions at sample size S=4096

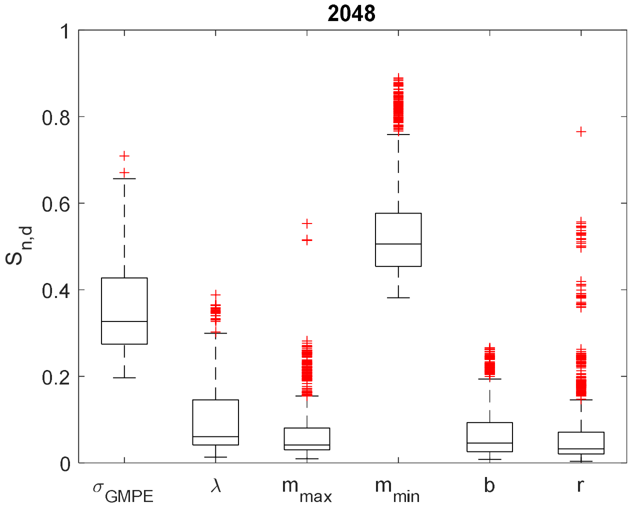

Figure 15: $S_{n,d}$ distributions at sample size S=2048

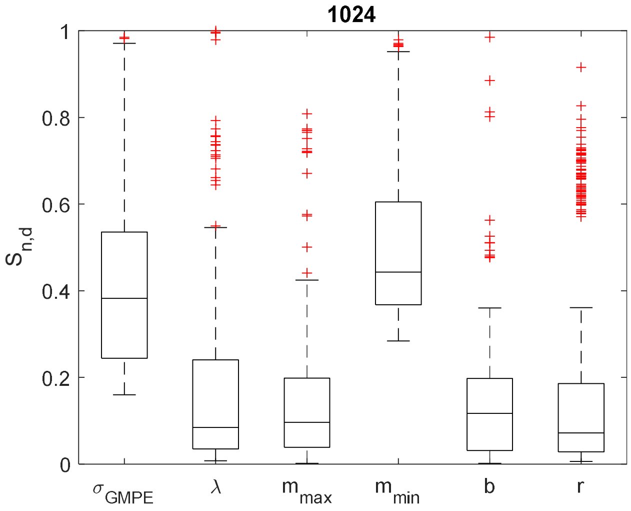

Figure 16: $S_{n,d}$ distributions at sample size S=1024

As general conclusion, we can state that bootstrapping allows relying on a very small dataset. Indeed, a sample size of $S = 2048$ (Figure 10) allows correctly identifying the most important input variables, while $S = 4096$ (Figure 9) yields already a very satisfactory estimate of the Sobol indices values (compared with the GSA estimates). Thus, as a general recommendation, we may conclude that a ratio of 4:1 of $S{:}\,D$ (dataset size vs number of bootstrap replicates) is enough to guarantee satisfactory results, without resorting further to demanding computations.

For the case study at hand, we can conclude that, m_{min}, σ_{GMPE}, and λ have been identified as the input variables which most influence the reference PGA, whereas r, b, and m_{max} influence is negligible (for whatever dataset size $S = 16384, 8192, 4096, 2048, 1024$). The analyst, once identified the input parameters which most influence the epistemic uncertainty on reference PGA, may decide to further investigate the choice made regarding such inputs and proceed with the uncertainty analysis.

We underline that, obviously, the numerical results obtained are relative to the specific case and cannot be generalized to other PSHA case studies. In particular, while the strong impact of σ_{GMPE}, and λ on hazard quantifications is well known (e.g., [33]), the reasons behind the importance of m_{min} must be further investigated. In Figure 17, we show the impact of small events (i.e., with a magnitude near Mw 4.5) in our case study. The relative short source to site distance (10 km) and the low

PGA level (0.07g) make very important the contribution of small events: the probability of exceedance of such PGA level for a magnitude Mw 4.5 is larger than 0.2. In the case of larger distances (e.g., 40 km) of larger PGA levels (e.g., 0.20 g) the impact of small events noticeably decreases.

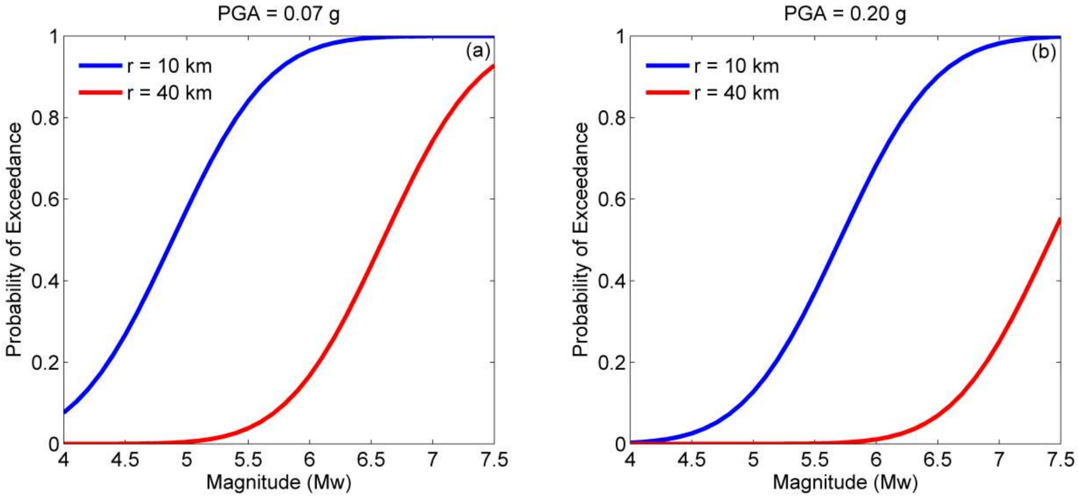

Figure 17: Probability of exceedance of a specific PGA level, 0.07g for panel (a), 0.20g for panel (b), as a function of the magnitude, using the GMPE adopted in this study, for two different distances (10 km for the blue curve, 40 km for the red curve)

The large probabilities shown by the blue curve in panel (a) of Figure 17 clearly explain the unexpected large importance of the m_{min} as input variable in PSHA computation (see [71] or [72] for a deeper investigation of the effect of small magnitude events in PSHA). This highlights that the proposed method represents also an important sanity check for any hazard quantification, as PSHA results are assumed independent from the selection of m_{min}. In this case, indeed, we show that the tails of the GMPE for small magnitudes are sufficiently populated to strongly impact the hazard quantification also at a relatively high mean return period (475 years in this case), at least in the near field of the source areas.

Notably, this effect, as pointed out in other studies [71], [72], could be due to the extension of the validity of the GMPEs to small magnitude events that, in some cases, can even lead to a bias in the hazard estimation.

2. A heuristic feature selection approach for scenario analysis of a Regional Seismic Tsunami Hazard Assessment

Seismic Probabilistic Tsunami Hazard Analysis (SPTHA) is aimed at estimating the annual rate of exceedance of an earthquake-induced tsunami wave of a certain location with reference to a predefined height threshold. The analysis relies on computationally demanding numerical simulations of seismic-induced tsunami wave generation and propagation. A large number of scenarios needs to be simulated to account for the aleatory and epistemic uncertainties. However, the exceedance of tsunami wave threshold height is a rare event so that most of the simulated scenarios bring little statistical contribution to the estimation of the annual rate yet increasing the computational burden. To efficiently address this issue, we propose a wrapper-based heuristic approach to select the set of most relevant features of the seismic model, for deciding a priori the seismic scenarios to be simulated. The proposed approach is based a Multi-Objective Differential Evolution Algorithm (MODEA) and is developed with reference to a case study whose objective of the analysis is calculating the annual rate of a threshold exceedance of the height of tsunami waves caused by subduction earthquakes that might be generated on a section of the Hellenic Arc and propagated to a target site on the eastern coast of Sicily (Siracusa). The comparison between the mean values of annual rate of exceedance of the tsunami wave height estimated considering only the selected scenarios and the full set of scenarios shows that the proposed approach allows a reduction of 95% of the number of scenarios with half of the features to be considered, and with no appreciable loss of accuracy.

2.1 Case study

We consider the regional SPTHA for the target site $\bar{a}$ on the eastern coast of Sicily (Siracusa, red cross in Figure 18), exposed to tsunamis triggered by crustal earthquakes occurring outside the subduction interface of the Hellenic Arc in the Kefalonia-Lefkada region [60]. Earthquakes are assumed to be generated at specific epicentral locations H_i (i=1,…, 42, blue points in Figure 18) with different magnitudes, depths, and faulting mechanisms. Without loss of generality, the following assumptions are made:

i. The threshold is of $\tilde{\psi} = 1m$ at 50m from the coastline.

ii. One epicentral location (star 14 in Figure 18) is considered, since a large number $Q = 721$ of seismic scenarios $\sigma_{\bar{x}}$ is available, making, $\Lambda(\psi_{\bar{a}} \geq 1m|H_{14})$ equal to:

$$\Lambda(\psi_{\bar{a}} \geq 1m|H_{14}) \approx \frac{1}{M}\sum_{m=1}^{M}\sum_{q=1}^{Q}\lambda\left(\sigma_{\bar{x}_q}|H_{14}\right)_m Pr\left(\psi_{\bar{a}} \geq 1m|\sigma_{\bar{x}_q}, H_{14}\right)_m \tag{19}$$

(Herein after, for the sake of readability, H_{14} will be omitted).

iii. Each $\sigma_{\bar{x}}$ is characterised by the set of parameters $\bar{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$ [59], [60]**,** whose support and values are listed in **Errore. L'origine riferimento non è stata trovata.**. These parameters (see Figure 19 for a schematic representation) are:

1. x_1 Magnitude
2. x_2 Depth (top of the fault)
3. x_3 Strike (of the focal mechanism)
4. x_4 Dip (of the focal mechanism)
5. x_5 Rake (of the focal mechanism)
6. x_6 Area (of the fault), i.e., the product of its width by its length
7. x_7 Length (of the fault)
8. x_8 Slip (of the fault)

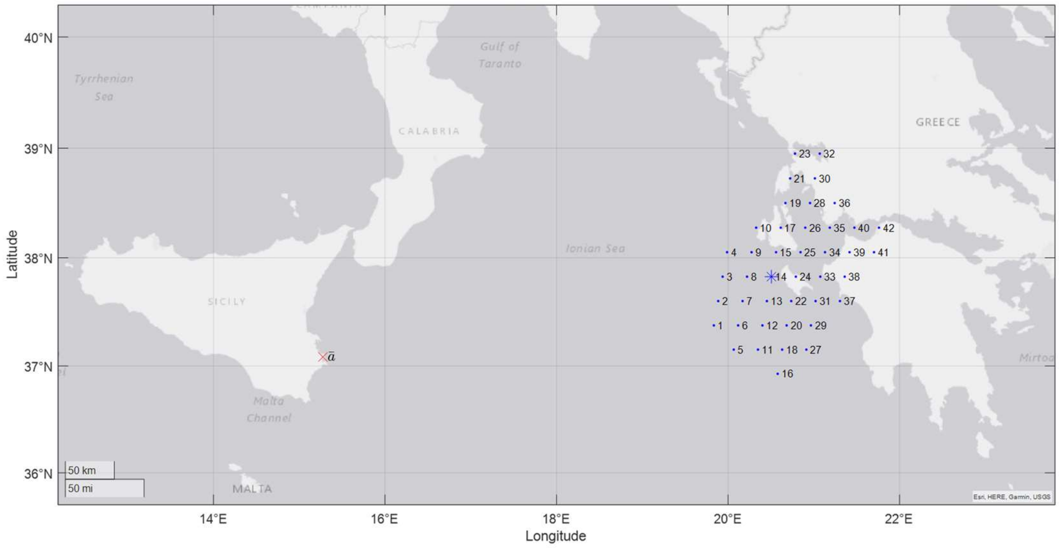

Figure 18: Seismic zones H_i (i=1,…,42) of the Hellenic Arc (blue dots), case study source (blue asterisk), target site ā (red cross), (Siracusa).

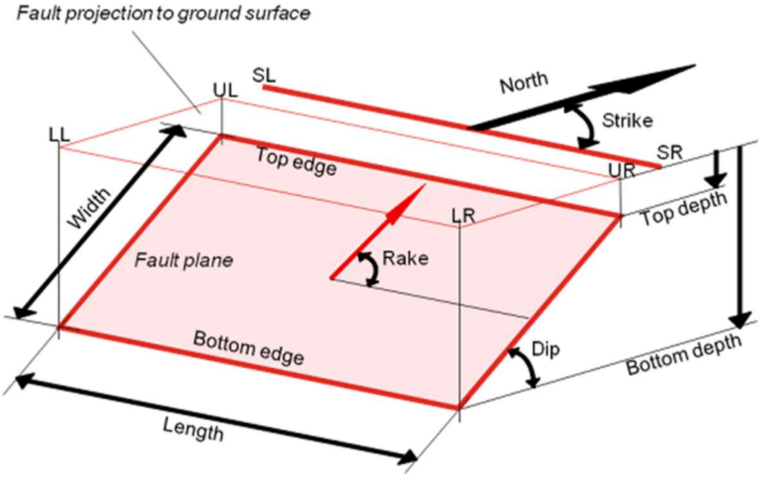

Figure 19: Schematic representation of an earthquake and its parameters [73].

Parameter	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Support	[6.5000-8.0933]	[1.00-17.86]	[22.5-337.5]	[10-90]	[0-270]	[318.95-12648.92]	[22.68-665.73]	[0.67-4.21]
Values	6.5000	1.00	22.5	10	0	318.95	22.68	0.67
	6.8012	5.97	67.5	30	90	558.32	34.39	0.95
	7.0737	7.56	112.5	50	180	638.11	37.88	1.09
	7.3203	9.43	157.5	70	270	1194.98	50.10	1.29
	7.5435	10.94	202.5	90		1205.54	63.64	1.30
	7.7453	1158	247.5			2108.29	70.44	1.71
	7.9280	14.00	292.5			2133.31	95.87	1.73
	8.0933	14.12	337.5			3566.59	112.28	2.21
		16.65				3524.55	126.69	2.24
		17.86				5608.92	163.06	2.79
						5676.15	187.72	2.82
						8541.96	204.89	3.44
						8644.77	298.74	3.48
						12497.92	454.99	4.16
						12648.92	665.73	4.21

Table 10: Parameters of the seismic scenarios.

2.2 Methodology

To alleviate the computational burden of the SPTHA, the procedure sketched in Figure 20 is developed. Firstly, an optimisation problem is solved to identify the optimal set of seismic scenarios that contribute most to $\Lambda(\psi_{\bar{a}} \geq 1m)$ of Eq. (19). Then, their features values are identified. The optimisation is performed by a wrapper-based heuristic approach: based on a Multi-Objective Differential Evolution Algorithm (MODEA) wherein the DE engine [58], [74] iteratively searches for candidate sets of scenarios, among the original dataset of $Q = 721$ scenarios, whose performance is evaluated with respect to a given cost function. Once the optimal set of scenarios is identified, their common features are retrieved by statistical analysis.

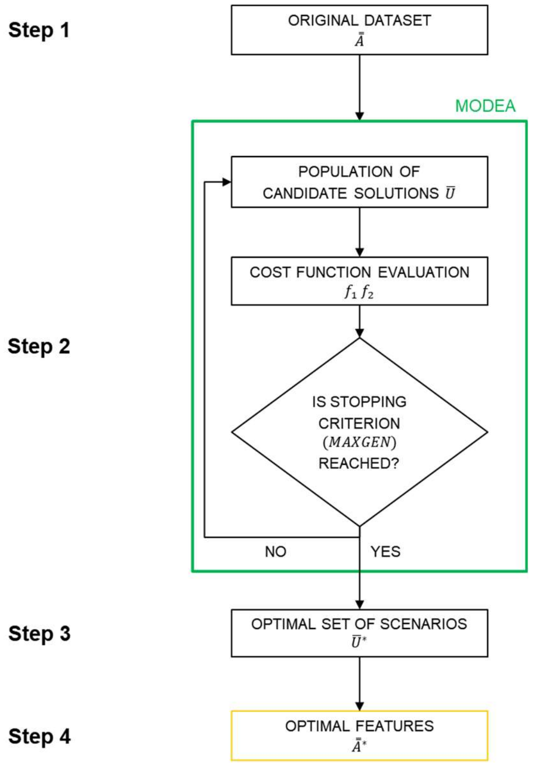
Step 1
ORIGINAL DATASET $\bar{\bar{A}}$
MODEA
POPULATION OF CANDIDATE SOLUTIONS $\bar{U}$
COST FUNCTION EVALUATION $f_1\ f_2$
Step 2
IS STOPPING CRITERION $(MAXGEN)$ REACHED?
NO
YES
Step 3
OPTIMAL SET OF SCENARIOS $\bar{U}^*$
Step 4
OPTIMAL FEATURES $\bar{\bar{A}}^*$

Figure 20: Wrapper approach for optimal set of scenarios selection based on MODEA

The procedure is explained in detail here below.

Step 1: Consider the original dataset

The original dataset $\bar{\bar{A}} = [Q \times 9]$ is:

$$\bar{\bar{\mathrm{A}}} = \begin{pmatrix} x_{1,1} & \cdots & x_{8,1} & \Lambda\left(\psi_{\bar{a}} \geq 1m | \sigma_{\bar{x}_1}\right) \\ \vdots & \vdots & \vdots & \vdots \\ x_{1,q} & \cdots & x_{8,q} & \Lambda\left(\psi_{\bar{a}} \geq 1m | \sigma_{\bar{x}_q}\right) \\ \vdots & \vdots & \vdots & \vdots \\ x_{1,Q} & \cdots & x_{8,Q} & \Lambda\left(\psi_{\bar{a}} \geq 1m | \sigma_{\bar{x}_Q}\right) \end{pmatrix} \tag{20}$$

where $x_{1,q}$ is the value of the parameter x_1 in the q-th scenario, $x_{2,q}$ is the value of the parameter x_2 in the q-th scenario, etc., and $\Lambda\left(\psi_{\bar{a}} \geq 1m|\sigma_{\bar{x}_q}\right) = \frac{1}{M}\sum_{m=1}^{M}\lambda\left(\sigma_{\bar{x}_q}\right)_m Pr\left(\psi_{\bar{a}} \geq 1m|\sigma_{\bar{x}_q}\right)_m$ is the annual rate of exceedance of the q-th scenario.

Step 2: Apply MODEA to identify the most relevant scenarios

The MODEA searches the global minimum of a set of objective (cost) functions $F = \{f(\cdot)\}$, of one (or more) decision vector(s) $\bar{U}$ (typically a string of binary digits) [75], [76]. In the case of interest for this work, $\bar{U}$ indicates whether the q-th seismic scenario is considered in the candidate solution (q-th bit equal to 1) or not (q-th bit equal to 0).

The MODEA search is performed by initially randomly sampling the bits of the NP vectors that compose the initial population strings [75]. Then, iteratively, the population is enriched by the solution $\bar{U}$ that best fits the objective functions, through a selection process driven by a set of parameters, i.e., the scaling factor F and the crossover probability CR [58]. For a thorough description of the process based on DE and its controlling parameters, the interested reader may refer to the Appendix A or to [58].

The two objective functions considered are:

1. Minimisation of Q (i.e., the number of scenarios $\sigma_{\bar{x}_q}$ considered in the solution):

$$f_1 = \sum_{q=1}^{Q} U_q \tag{21}$$

2. Minimisation of the squared error SE between the annual rate of exceedance $\Lambda(\psi_{\bar{a}} \geq 1m)$ and the annual rate of exceedance calculated considering exclusively the $Q^* = min\left(\sum_{q=1}^{Q} U_q\right)$ selected scenarios $\Lambda^*(\psi_{\bar{a}} \geq 1m)$:

$$f_2 = \left(\Lambda(\psi_{\bar{a}} \geq 1m) - \Lambda^*(\psi_{\bar{a}} \geq 1m)\right)^2 \tag{22}$$

where $\Lambda^*(\psi_{\bar{a}} \geq 1m)$ is calculated as:

$$\Lambda^*(\psi_{\bar{a}} \geq 1m) = \sum_{q=1}^{Q} \lambda\left(\sigma_{\bar{x}_q}\right) Pr\left(\psi_{\bar{a}} \geq 1m|\sigma_{\bar{x}_q}\right) U_q \tag{23}$$

The search procedure ends when the stopping criterion (e.g., the maximum number of generations $MAXGEN$) is reached.

Step 3: Optimal set of scenarios

The optimal solution vector $\bar{U}^*$ (i.e., the optimal set of scenarios) that optimizes the multi-objective function of Eqs. (21) and (22) is selected from the Pareto optimal front [75], as the solution with the minimum number Q^* of entries equal to 1 (i.e., the scenarios considered in the candidate solution).

Step 4: Optimal features identification

To identify the most relevant features to be considered for the SPTHA, we first calculate the optimal features matrix $\bar{\bar{A}}^* = [Q^* \times 9]$, as the Hadamard product of the original dataset $\bar{\bar{A}}$ with $\bar{U}^*$ (with $(Q - Q^*)$ null vector rows):

$$\bar{\bar{A}}^* = \bar{\bar{A}} \circ \bar{U}^* \tag{24}$$

$$\bar{\bar{A}}^* = \begin{pmatrix} x_{1,1} & \cdots & x_{8,1} & \Lambda(\psi_{\bar{a}} \geq 1m|\sigma_{\bar{x}_1}) \\ \vdots & \vdots & \vdots & \vdots \\ x_{1,q^*} & \cdots & x_{8,q^*} & \Lambda\left(\psi_{\bar{a}} \geq 1m|\sigma_{\bar{x}_{q^*}}\right) \\ \vdots & \vdots & \vdots & \vdots \\ x_{1,Q^*} & \cdots & x_{8,Q^*} & \Lambda\left(\psi_{\bar{a}} \geq 1m|\sigma_{\bar{x}_{Q^*}}\right) \end{pmatrix} \tag{25}$$

Then, the matrix $\bar{\bar{A}}^*$ is columnwise compared with the original dataset $\bar{\bar{A}}$ to assess their commonality (i.e., the optimal features subset).

2.3 Results

The approach described in 2.2 has been applied to the case study presented in Section 2.1. The search for optimal scenarios among the $Q = 721$ of the original dataset is performed by a MODEA (DE/rand/1/bin strategy, see Appendix A for further details), with objective functions f_1 and f_2 (respectively Eq. (21) and Eq. (22)), where f_2 is calculated referring to the benchmark value of the annual rate of exceedance $\Lambda(\psi_{\bar{a}} \geq 1m) = 3.3193 \cdot 10^{-12} yr^{-1}$ calculated from the full set of scenarios. In practice, each candidate solution $\overline{U}$ is a binary string of $Q = 721$ bits. The population size NP, the scaling factor F, the crossover probability CR and the generation bound $MAXGEN$, have been expertly set equal to 20, 0.5, 0.9 and 10000, respectively: specifically, NP has been set equal to $10 \cdot (\#of\ objectives = 2) = 20$ in line with [58]; CR has been set equal to 0.9 for a fast convergence [58]; F has been set equal to 0.5 in line with [58], [77]; the stopping criterion $MAXGEN = 10000$ has been set following a trial-and-error procedure [74].

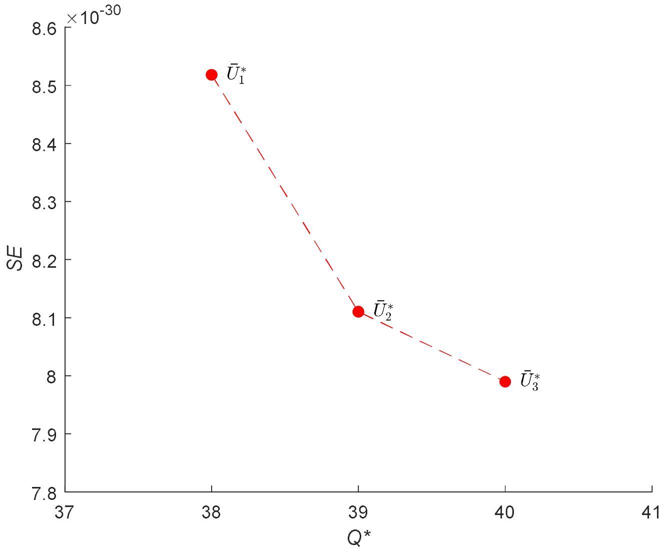

Figure 21: Pareto optimal front after MAXGEN iterations

When the stopping criterion is reached, the Pareto front shown in Figure 21 is obtained:

1. $\overline{U}_1^*$ yields $Q^* = 38$ scenarios with a $SE = 8.5^{-30} yr^{-2}$ and a percentage error of 0.085%
2. $\overline{U}_2^*$ yields $Q^* = 39$ scenarios with a $SE = 8.1^{-30} yr^{-2}$ and a percentage error of 0.066%
3. $\overline{U}_3^*$ yields $Q^* = 40$ scenarios with a $SE = 8.0^{-30} yr^{-2}$ and a percentage error of 0.063%

In this work, the solution $\overline{U}_1^*$ is preferred because it yields the minimum number of $Q^* = 38$ scenarios (i.e., a 95% reduction with respect to Q) with a reasonably small

$SE = 8.5^{-30} years^{-2}$ (i.e., a percentage error of 0.085%) in the estimation of $\Lambda(\psi_{\bar{a}} \geq 1m|H_{14})$.

In Table 11, all the features and Q^* scenarios selected by the MODEA are listed, without discarding low-frequency scenarios [28]. All Q^* selected scenarios contribute to $\Lambda(\psi_{\bar{a}} \geq 1m|H_{14})$ with a relatively large probability of threshold exceedance $\Lambda\left(\psi_{\bar{a}} \geq 1m|H_{14}, \sigma_{\bar{x}_q}\right)$. On the contrary, most of the $Q = 721$ seismic scenarios in the original dataset have a $\Lambda\left(\psi_{\bar{a}} \geq 1m, \sigma_{\bar{x}_q}\right) < 10^{-20}$, i.e., bring a negligible contribution to the estimation of $\Lambda(\psi_{\bar{a}} \geq 1m|H_{14})$ but increase the computational burden. Furthermore, regarding the features selected to characterise the $Q^* = 38$ scenarios, these are reduced with respect to those that characterise the Q scenarios as shown in Table 12 and Figure 21-Figure 29.

In what follows, a geophysical interpretation of the results obtained is provided.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	$\Lambda\left(\psi_{\bar{a}} \geq 1m|H_{14}, \sigma_{\bar{x}_q}\right)$
6.5000	1.00	337.5	50	90	318.95	22.68	0.67	3.98E-17
6.8012	1.00	157.5	50	90	638.11	34.39	0.95	4.17E-16
6.8012	1.00	157.5	70	90	638.11	34.39	0.95	2.79E-16
6.8012	1.00	157.5	90	270	638.11	34.39	0.95	3.28E-16
6.8012	1.00	157.5	90	90	638.11	34.39	0.95	1.35E-16
6.8012	1.00	337.5	70	90	638.11	34.39	0.95	4.16E-17
6.8012	1.00	337.5	50	270	638.11	34.39	0.95	2.12E-17
6.8012	1.00	337.5	50	90	638.11	34.39	0.95	4.17E-16
6.8012	1.00	337.5	30	90	638.11	34.39	0.95	4.04E-13
6.8012	7.56	337.5	50	90	638.11	34.39	0.95	2.95E-16
6.8012	7.56	337.5	30	90	638.11	34.39	0.95	4.38E-13
6.8012	14.12	337.5	30	90	638.11	34.39	0.95	8.72E-15

7.0737	1.00	22.5	50	90	1194.98	50.10	1.30	2.41E-17
7.0737	1.00	157.5	50	270	1194.98	50.10	1.30	1.33E-16
7.0737	1.00	157.5	50	90	1194.98	50.10	1.30	1.50E-15
7.0737	1.00	157.5	70	90	1194.98	50.10	1.30	1.59E-15
7.0737	1.00	157.5	90	270	1194.98	50.10	1.30	1.11E-15
7.0737	1.00	337.5	70	90	1194.98	50.10	1.30	1.64E-16
7.0737	1.00	337.5	50	270	1194.98	50.10	1.30	4.02E-16
7.0737	1.00	337.5	50	90	1194.98	50.10	1.30	2.05E-15
7.0737	1.00	337.5	30	270	1194.98	50.10	1.30	3.26E-17
7.0737	1.00	337.5	30	90	1194.98	50.10	1.30	9.94E-13
7.0737	9.43	157.5	70	90	1194.98	50.10	1.30	2.32E-17
7.0737	9.43	337.5	30	90	1194.98	50.10	1.30	7.13E-13
7.0737	9.43	337.5	10	90	1194.98	50.10	1.30	4.25E-17
7.3203	1.00	157.5	50	90	2108.29	70.44	1.73	3.30E-16
7.3203	1.00	157.5	70	90	2108.29	70.44	1.73	1.05E-16
7.3203	1.00	157.5	90	270	2108.29	70.44	1.73	1.94E-16
7.3203	1.00	337.5	30	90	2108.29	70.44	1.73	3.45E-13
7.3203	11.58	157.5	50	90	2108.29	70.44	1.73	5.44E-17
7.3203	11.58	337.5	50	90	2108.29	70.44	1.73	7.61E-17
7.3203	11.58	337.5	30	90	2108.29	70.44	1.73	1.48E-13
7.5435	1.00	157.5	50	90	3524.55	95.87	2.24	1.96E-16
7.5435	1.00	157.5	70	90	3524.55	95.87	2.24	7.34E-17
7.5435	1.00	337.5	50	90	3524.55	95.87	2.24	1.78E-16
7.5435	1.00	337.5	30	90	3524.55	95.87	2.24	2.51E-13

7.7453	1.00	337.5	30	90	5608.92	126.69	2.82	3.83E-15

Table 11: Features and $\Lambda\left(\psi_{\bar{a}} \geq 1\mathrm{m}|\mathrm{H}_{14}, \sigma_{\bar{x}_q}\right)$ of the $\mathrm{Q}^{*} = 38$ selected scenarios

Parameter	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Values	6.5000	1.00	22.5	10	90	318.95	22.68	0.67
	6.8012	7.56	157.5	30	270	638.11	34.39	0.95
	7.0737	9.43	337.5	50		1194.98	50.10	1.30
	7.3203	1158		70		2108.29	70.44	1.73
	7.5435	14.12		90		2133.31	95.87	2.24
	7.7453					3524.55	126.69	2.82
						5608.92		

Table 12: Support values of the selected scenarios

2.3.1 Magnitude

The DE search engine has not selected, because negligible (red in Figure 22), those scenarios characterised by large magnitudes ($x_1 = 7.7453, 8.0933$): in such cases, the annual rates are negligible and do not bring any significant contribution to the hazard curve estimation (see Eq. (19)), even if a relatively large threshold of $\tilde{\psi} = 1m$ at 50m is assumed.

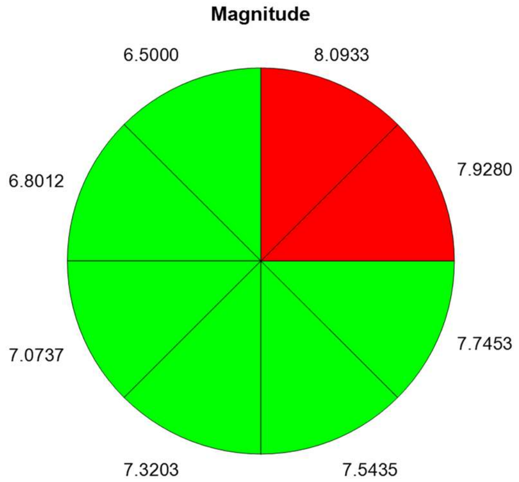

Figure 22: Values of the magnitude (green) and non-selected ones (red) in the selected scenarios

2.3.2 Depth

The DE search engine has identified as relevant mainly those scenarios characterised by a depth value of $1km$, along with a few scenarios characterised by depth values of $7.56km$, $9.43km$, $11.58km$, and $14.12km$ calculated in line with NEAMTHM18 documentation [60], [61] (green in Figure 23). This result is justified by the dependence of the depth on the magnitude: a depth equal to $1km$ is considered for all magnitudes whereas larger depths, instead, are modelled for smaller magnitudes only, that have been found as important (see Figure 23).

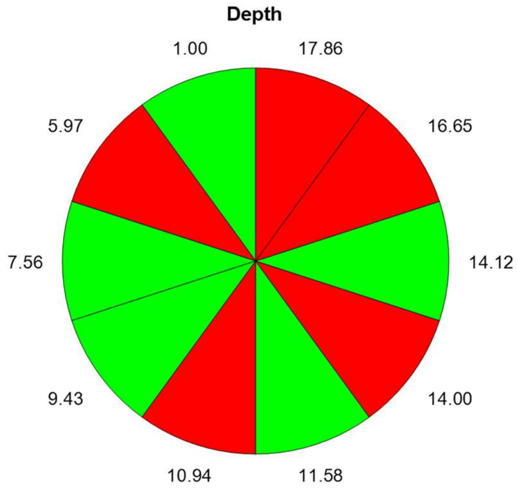

Figure 23: Values of the depth (green) and non-selected ones (red) in the selected scenarios

2.3.3 Strike

The DE search engine has identified as relevant (green in Figure 24) those scenarios characterised by strike angle values of 22.5°, 157.5°, 337.5°, i.e., directions approximately perpendicular to the source-to-site tsunami propagation path (see Figure 18).

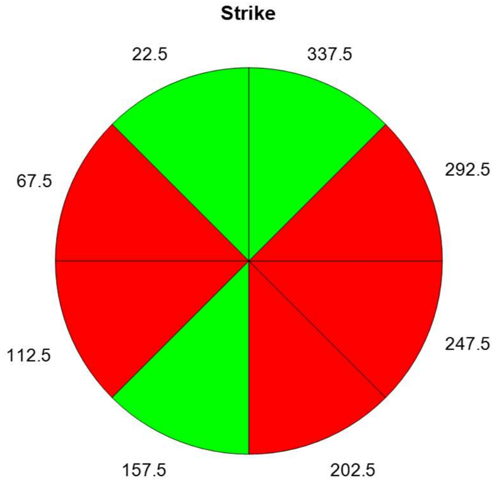

Figure 24: Values of the strike (green) and non-selected ones (red) in the selected scenarios

2.3.4 Dip

The DE search engine has identified as relevant (green in Figure 25) all dip angles in the selected scenarios: thus, dip is not a distinguishing characteristic of the scenarios.

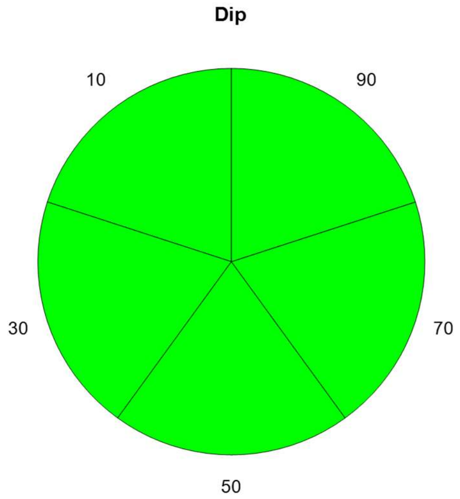

Figure 25: Values of the dip (green) and non-selected ones (red) in the selected scenarios

2.3.5 Rake

Only scenarios with rake values of 90° and 270° have been selected (green in Figure 26). This result is expected, as only dip-slip earthquakes can generate a significant deformation of the sea bottom, thus generating higher tsunami waves.

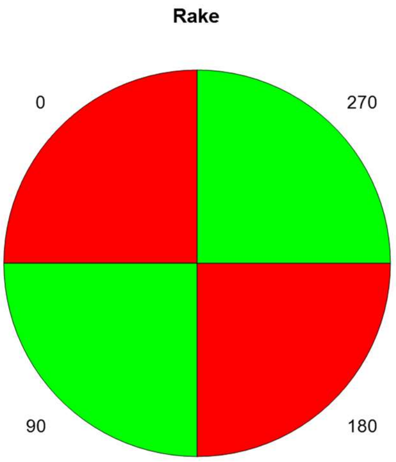

Figure 26: Values of the rake (green) and non-selected ones (red) in the selected scenarios

2.3.6 Area

Area values are computed relying on empirical scaling relationships from magnitude (e.g. $\log Area = A + B \times Magnitude$), using different relationships for dip-slip ($x_5 = 90°, 270°$) and strike-slip ($x_5 = 0°, 180°$) earthquakes [73]. Only the scenarios with area values of $318.5km^2$, $638.11km^2$, $1194.98km^2$, $2108.29km^2$, $3524.55km^2$, $5608.92km^2$ (green in Figure 27) have been selected, i.e., those scenarios corresponding to small magnitude dip-slip earthquakes. In other words, larger area values, corresponding to larger magnitudes, have not been coherently selected as well as smaller area values, corresponding to smaller magnitudes and strike-slip earthquakes.

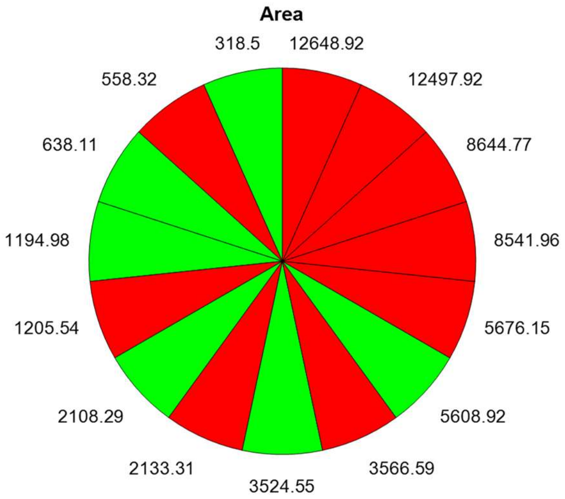

Figure 27: Values of the area (green) and non-selected ones (red) in the selected scenarios

2.3.7 Length

Length values are computed relying on empirical scaling relationships from magnitude (e.g. $\log Length = A + B \times Magnitude$), using different relationships for dip-slip ($x_5 = 90°, 270°$) and strike-slip ($x_5 = 0°, 180°$) earthquakes [73]. Only the scenarios with length values of $22.68km$, $34.39km$, $50.10km$, $70.44km$, $95.87km$, $126.69km$ (green in Figure 28), i.e., those scenarios corresponding to small magnitude dip-slip earthquakes. In other words, larger length values, corresponding to larger magnitudes, have not been coherently selected as well as smaller length values, corresponding to smaller magnitudes and strike-slip earthquakes.

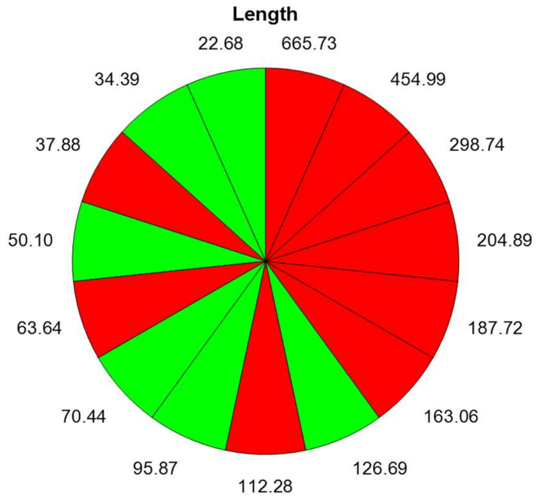

Figure 28: Values of the length (green) and non-selected ones (red) in the selected scenarios

2.3.8 Slip

Slip values are computed relying on empirical scaling relationships from magnitude (e.g. $Slip \propto Magnitude/Area$), using different relationships for dip-slip ($x_5 = 90°, 270°$) and strike-slip ($x_5 = 0°, 180°$) earthquakes [73]. Only the scenarios with slip values of 0.67, 0.95, 1.30, 1.73, 2.24, 2.82 (green in Figure 29) have been selected, i.e., those scenarios corresponding to small magnitude dip-slip earthquakes. In other words, larger slip values, corresponding to larger magnitudes, have been coherently selected as well as smaller slip values, corresponding to smaller magnitudes and strike-slip earthquakes.

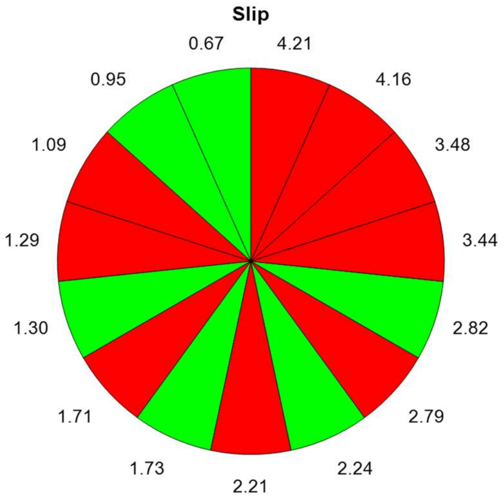

Figure 29: Values of the slip (green) and non-selected ones (red) in the selected scenarios

As a result of the MODE selection, the analyst may simulate the scenarios characterised by:

- Magnitude $x_1 \in (6.5000, 6.8012, 7.0737, 7.3203, 7.5435, 7.7453)$
- Depth $x_2 \in (1, 7.56, 9.43, 11.58, 14.12)$;
- Strike $x_3 \in (22.5, 157.5, 337.5)$;
- Dip $x_4 \in (10, 30, 50, 70, 90)$;
- Rake $x_5 \in (90, 270)$;
- Area $x_6 \in (318.5, 638.11, 1194.98, 2108.29, 3524.55, 5608.92)$;
- Length $x_7 \in (22.68, 34.39, 50.10, 70.44, 95.87, 126.69)$;
- Slip $x_8 \in (0.67, 0.95, 1.30, 1.73, 2.24, 2.82)$.

These results are expected, based on the tsunamigenic capability of earthquakes (see [78] and references therein). They depend both on the particular case study analysed and on the specific tsunami threshold of $\psi_{\bar{a}} \geq 1m$ chosen. Larger tsunami intensities, e.g., $\psi_{\bar{a}} \geq 10m$, would have involved different (probably larger) magnitudes. On the other hand, the results for the Strike, Dip, and Rake angles are probably more general, and they are possibly still valid for larger tsunami intensities.

3. Conclusions

In this thesis, we have proposed two SA methods to deal with the computational issues of the SPTHA related with:

i) The identification of the model parameters most affecting the PGA;

ii) The identification of the features of the seismic model worthy to be fed to the seismic-induced tsunami simulation code.

In Chapter 1 "A Bootstrapped Modularised method of Global Sensitivity Analysis for Probabilistic Seismic Hazard Assessment", we have proposed a novel Bootstrapped Modularised Sensitivity Analysis (BMGSA) method based on bootstrapping, MGSA and ensemble strategies to identify the input parameters which the output of a PSHA model is most sensitive to, assuming that only an input-output dataset is given whereas the model is not available. The novelty and strength of the proposed BMGSA method is that to be applied it only needs data and not the source simulation code.

The capability of the proposed method is tested on a benchmark case study. The results have been compared with a standard variance-based GSA method of literature, showing that the proposed method and the standard GSA agree on the identification of the three by-far most important input variables. Furthermore, the BMGSA has proved to be reliable even when applied to very small datasets.

The application of the developed technique to PSHA demonstrates its capability of scoring correctly the importance of existing epistemic uncertainty factor, needing only the input and the output data. This allows applying the technique to any hazard model in which epistemic uncertainty is to be evaluated. Its systematic application to hazard studies to detect the most influential parameters, would allow hazard practitioners to both improve the sanity checks during the assessment and to focus future research toward the reduction of epistemic uncertainty by further characterisation of the important factors.

The results of our applications, for example, highlight the importance of small magnitudes near to the seismic source areas, showing the importance of the definition of the minimum magnitude and the potential impact of the tail of the uncertainty distributions on GMPE on seismic hazard evaluation.

In Chapter 2 "A heuristic feature selection approach for scenario analysis of a Regional Seismic Tsunami Hazard Assessment", a novel approach for reducing the number of seismic scenarios to be considered for SPTHA has been presented. The approach is a wrapper-based feature selection heuristic approach based on MODEA. It selects the relevant features of the seismic scenarios to be simulated.

The proposed approach has been applied to a case study with reference to the estimation of the annual rate of exceedance of a height threshold $\tilde{\psi} = 1m$ of tsunami waves caused by crustal earthquakes that might be generated on the Kefalonia-Lefkada region in North-western Greece and propagated to a target site $\bar{a}$ on the eastern coast of Sicily (Italy).

The proposed approach is shown to be able to significantly reduce the number of features describing the seismic source variability and, thus, the number of scenarios to be considered in the analysis without affecting the accuracy of the estimate of the annual rate of exceedance. A geophysical interpretation of the results has been provided.

Further research work will be devoted to the comparison of the proposed approach to other existing methods that may be applied with similar goals, e.g., a standard disaggregation procedure [25], [59], [79].

Bibliography

[1] V. C. Moreno, F. Ricci, R. Sorichetti, A. Misuri, and V. Cozzani, "Analysis of Past Accidents Triggered by Natural Events in the Chemical and Process Industry," *Chem. Eng. Trans.*, vol. 74, pp. 1405–1410, May 2019.

[2] A. Mesa-Gomez, J. Casal, and F. MuNoz, "Risk analysis in Natech events: State of the art," *J. Loss Prev. Process Ind.*, vol. 64, p. 104071, 2020.

[3] A. M. Cruz, G. Franchello, and E. Krausmann, "Assessment of Tsunami Risk to an Oil Refinery in Southern Italy," *JRC Sci. Tech. Reports*, p. 58, 2009.

[4] Y. Iwabuchi, S. Koshimura, and F. Imamura, "Study on Oil Spread Caused by the 1964 Niigata Earthquake Tsunami," *J. Disaster Res.*, vol. 1, no. 1, pp. 157–168, 2006.

[5] A. Sato and Y. Lyamzina, "Diversity of Concerns in Recovery after a Nuclear Accident: A Perspective from Fukushima," 2018.

[6] F. Lin and H. Li, "Safety analysis of nuclear containment vessels subjected to strong earthquakes and subsequent tsunamis," *Nucl. Eng. Technol.*, vol. 49, no. 5, pp. 1079–1089, 2017.

[7] H. Park, D. T. Cox, M. S. Alam, and A. R. Barbosa, "Probabilistic seismic and tsunami hazard analysis conditioned on a megathrust rupture of the cascadia subduction zone," *Front. Built Environ.*, vol. 3, no. June, pp. 1–19, 2017.

[8] R. De Risi and K. Goda, "Probabilistic earthquake–Tsunami multi-hazard analysis: Application to the tohoku region, Japan," *Front. Built Environ.*, vol. 2, no. October, pp. 1–19, 2016.

[9] A. Grezio *et al.*, "Probabilistic Tsunami Hazard Analysis: Multiple Sources and Global Applications," *Reviews of Geophysics*, vol. 55, no. 4. pp. 1158–1198, 2017.

[10] J. Behrens *et al.*, "Probabilistic Tsunami Hazard and Risk Analysis: A Review of Research Gaps," vol. 9, p. 1, 2021.

[11] S. Tinti and A. Armigliato, "The use of scenarios to evaluate the tsunami

impact in southern Italy," *Mar. Geol.*, vol. 199, no. 3–4, pp. 221–243, Sep. 2003.

[12] R. Tonini, A. Armigliato, G. Pagnoni, F. Zaniboni, and S. Tinti, "Tsunami hazard for the city of Catania, eastern Sicily, Italy, assessed by means of Worst-case Credible Tsunami Scenario Analysis (WCTSA)," *Nat. Hazards Earth Syst. Sci.*, vol. 11, no. 5, pp. 1217–1232, 2011.

[13] S. Tinti *et al.*, *Handbook of Tsunami Hazard and Damage Scenarios*. 2011.

[14] I. El-Hussain, R. Omira, Z. Al-Habsi, M. A. Baptista, A. Deif, and A. M. E. Mohamed, "Probabilistic and deterministic estimates of near-field tsunami hazards in northeast Oman," *Geosci. Lett.*, vol. 5, p. 30, 2018.

[15] R. Prasad, *Tsunami hazard assessment at nuclear power plant sites in the United States of America*. 2012.

[16] E. L. Geist and T. Parsons, "Probabilistic analysis of tsunami hazards," *Nat. Hazards*, vol. 37, no. 3, pp. 277–314, Mar. 2006.

[17] E. L. Geist and T. Parsons, "Undersampling power-law size distributions: effect on the assessment of extreme natural hazards," vol. 72, pp. 565–595, 2014.

[18] S. Lorito, J. Selva, R. Basili, F. Romano, M. M. Tiberti, and A. Piatanesi, "Probabilistic hazard for seismically induced tsunamis: Accuracy and feasibility of inundation maps," *Geophys. J. Int.*, vol. 200, no. 1, pp. 574–588, Jan. 2015.

[19] R. K. Mcguire, "Probabilistic seismic hazard analysis: Early history," *Earthq. Eng. Struct. Dyn. Earthq. Engng Struct. Dyn*, vol. 37, pp. 329–338, 2008.

[20] J. Selva and L. Sandri, "Probabilistic seismic hazard assessment: Combining Cornell-like approaches and data at sites through Bayesian inference," *Bull. Seismol. Soc. Am.*, vol. 103, no. 3, pp. 1709–1722, Jun. 2013.

[21] R. K. McGuire and W. J. Arabasz, "12. An Introduction to Probabilistic Seismic Hazard Analysis," in *Geotechnical and Environmental Geophysics*, 1990, pp. 333–354.

[22] M. Kowsari, N. Eftekhari, A. Kijko, E. Yousefi Dadras, H. Ghazi, and E. Shabani, "Quantifying Seismicity Parameter Uncertainties and Their Effects on Probabilistic Seismic Hazard Analysis: A Case Study of Iran," *Pure Appl. Geophys.*, vol. 176, no. 4, pp. 1487–1502, Apr. 2019.

[23] L. Hofer and M. A. Zanini, "The role of uncertainty of model parameters in PSHA," in *COMPDYN Proceedings*, 2019, vol. 3, pp. 5527–5534.

[24] T. Rikitake and I. Aida, "Tsunami hazard probability in Japan," *Bull. Seismol. Soc. Am.*, vol. 78, no. 3, pp. 1268–1278, Jun. 1988.

[25] S. J. Gibbons *et al.*, "Probabilistic Tsunami Hazard Analysis: High Performance Computing for Massive Scale Inundation Simulations," *Front. Earth Sci*, vol. 8, p. 1, 2020.

[26] F. I. González *et al.*, "Probabilistic tsunami hazard assessment at Seaside, Oregon, for near- and far-field seismic sources," *J. Geophys. Res*, vol. 114, p. 11023, 2009.

[27] M. Volpe, S. Lorito, J. Selva, R. Tonini, F. Romano, and B. Brizuela, "From regional to local SPTHA: Efficient computation of probabilistic tsunami inundation maps addressing near-field sources," *Nat. Hazards Earth Syst. Sci.*, vol. 19, no. 3, pp. 455–469, Mar. 2019.

[28] F. Di Maio, M. Belotti, M. Volpe, J. Selva, and E. Zio, "Seismic Probabilistic Tsunami Hazard Assessment: a novel approach based on Parallel density scanned Adaptive Kriging," 2021.

[29] V. Bacchi *et al.*, "Using Meta-Models for Tsunami Hazard Analysis: An Example of Application for the French Atlantic Coast ," *Frontiers in Earth Science* , vol. 8. p. 41, 2020.

[30] A. Gailler, H. Hébert, F. Schindelé, and D. Reymond, "Coastal Amplification Laws for the French Tsunami Warning Center: Numerical Modeling and Fast Estimate of Tsunami Wave Heights Along the French Riviera," *Pure Appl. Geophys.*, vol. 175, no. 4, pp. 1429–1444, 2018.

[31] S. Glimsdal *et al.*, "A New Approximate Method for Quantifying Tsunami Maximum Inundation Height Probability," *Pure Appl. Geophys.*, vol. 176, no. 7, pp. 3227–3246, 2019.

[32] F. Di Maio, A. Bandini, E. Zio, S. C. Alberola, F. Sanchez-Saez, and S. Martorell, "Bootstrapped-ensemble-based Sensitivity Analysis of a trace thermal-hydraulic model based on a limited number of PWR large break loca simulations," *Reliab. Eng. Syst. Saf.*, vol. 153, pp. 122–134, Sep. 2016.

[33] C. Molkenthin, F. Scherbaum, A. Griewank, H. Leovey, S. Kucherenko, and F. Cotton, "Derivative-based global sensitivity analysis: Upper bounding of sensitivities in seismic-hazard assessment using automatic differentiation," *Bull. Seismol. Soc. Am.*, vol. 107, no. 2, pp. 984–1004, Apr. 2017.

[34] E. Borgonovo and E. Plischke, "Sensitivity analysis: A review of recent

advances," *European Journal of Operational Research*, vol. 248, no. 3. Elsevier B.V., pp. 869–887, 01-Feb-2016.

[35] E. Zio, "Computational Methods for Reliability and Risk Analysis," 2009.

[36] A. Saltelli *et al.*, *Global Sensitivity Analysis. The Primer*. 2008.

[37] A. Saltelli and J. Marivoet, "Non-parametric statistics in sensitivity analysis for model output: A comparison of selected techniques," *Reliab. Eng. Syst. Saf.*, vol. 28, no. 2, pp. 229–253, Jan. 1990.

[38] A. Saltelli and I. M. Sobol', "Sensitivity analysis for nonlinear mathematical models: numerical experience (in Russian)," *Math. Model. Comput. Exp.*, vol. 7, no. 11, pp. 16–28, 1995.

[39] E. Borgonovo, "A new uncertainty importance measure," *Reliab. Eng. Syst. Saf.*, vol. 92, no. 6, pp. 771–784, Jun. 2007.

[40] Q. Liu and T. Homma, "A new computational method of a moment-independent uncertainty importance measure," *Reliab. Eng. Syst. Saf.*, vol. 94, no. 7, pp. 1205–1211, Jul. 2009.

[41] J. E. Oakley, A. Brennan, P. Tappenden, and J. Chilcott, "Simulation sample sizes for Monte Carlo partial EVPI calculations," *J. Health Econ.*, vol. 29, no. 3, pp. 468–477, May 2010.

[42] Z. Hu and S. Mahadevan, "Probability models for data-Driven global sensitivity analysis," *Reliab. Eng. Syst. Saf.*, vol. 187, pp. 40–57, Jul. 2019.

[43] E. Plischke, "An adaptive correlation ratio method using the cumulative sum of the reordered output," in *Reliability Engineering and System Safety*, 2012, vol. 107, pp. 149–156.

[44] C. Li and S. Mahadevan, "An efficient modularized sample-based method to estimate the first-order Sobol index," *Reliab. Eng. Syst. Saf.*, vol. 153, pp. 110–121, Sep. 2016.

[45] E. Plischke, E. Borgonovo, and C. L. Smith, "Global sensitivity measures from given data," *Eur. J. Oper. Res.*, vol. 226, no. 3, pp. 536–550, May 2013.

[46] J. E. Oakley and A. O'Hagan, "Probabilistic sensitivity analysis of complex models: A Bayesian approach," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 66, no. 3, pp. 751–769, 2004.

[47] J. P. C. Kleijnen, "Kriging metamodeling in simulation: A review," *European Journal of Operational Research*, vol. 192, no. 3. pp. 707–716, 01-Feb-2009.

[48] B. Sudret, "Global sensitivity analysis using polynomial chaos expansions," *Reliability Engineering and System Safety*, vol. 93, no. 7. Elsevier, pp. 964–979, 01-Jul-2008.

[49] Y. Caniou and B. Sudret, "Distribution-based global sensitivity analysis using polynomial chaos expansions," in *Procedia - Social and Behavioral Sciences*, 2010, vol. 2, no. 6, pp. 7625–7626.

[50] S. Tarantola, D. Gatelli, and T. A. Mara, "Random balance designs for the estimation of first order global sensitivity indices," *Reliab. Eng. Syst. Saf.*, vol. 91, no. 6, pp. 717–727, Jun. 2006.

[51] C. Xu and G. Gertner, "Extending a global sensitivity analysis technique to models with correlated parameters," *Comput. Stat. Data Anal.*, vol. 51, no. 12, pp. 5579–5590, Aug. 2007.

[52] E. Plischke, "An effective algorithm for computing global sensitivity indices (EASI)," *Reliab. Eng. Syst. Saf.*, vol. 95, no. 4, pp. 354–360, Apr. 2010.

[53] M. Strong and J. E. Oakley, "An efficient method for computing single-parameter partial expected value of perfect information," *Med. Decis. Mak.*, vol. 33, no. 6, pp. 755–766, 2013.

[54] E. Borgonovo, X. Lu, E. Plischke, O. Rakovec, and M. C. Hill, "Making the most out of a hydrological model data set: Sensitivity analyses to open the model black-box," *Water Resour. Res.*, vol. 53, no. 9, pp. 7933–7950, Sep. 2017.

[55] S. M. Hoseyni, F. Di Maio, M. Vagnoli, E. Zio, and M. Pourgol-Mohammad, "A Bayesian ensemble of sensitivity measures for severe accident modeling," *Nucl. Eng. Des.*, vol. 295, pp. 182–191, Dec. 2015.

[56] J. P. C. Kleijnen and D. Deflandre, "Validation of regression metamodels in simulation: Bootstrap approach," *Eur. J. Oper. Res.*, vol. 170, no. 1, pp. 120–131, Apr. 2006.

[57] E. Plischke, E. Borgonovo, and C. L. Smith, "Global sensitivity measures from given data," *Eur. J. Oper. Res.*, vol. 226, no. 3, pp. 536–550, May 2013.

[58] R. Storn and K. Price, "Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces," *J. Glob. Optim.*, vol. 11, no. 4, pp. 341–359, 1997.

[59] J. Selva *et al.*, "Quantification of source uncertainties in Seismic Probabilistic Tsunami Hazard Analysis (SPTHA)," *Geophys. J. Int.*, vol. 205, no. 3, pp. 1780–1803, Jun. 2016.

[60] R. Basili *et al.*, “The Making of the NEAM Tsunami Hazard Model 2018 (NEAMTHM18),” *Front. Earth Sci.*, vol. 8, p. 753, 2021.

[61] R. Basili *et al.*, “NEAM Tsunami Hazard Model 2018 (NEAMTHM18): online data of the Probabilistic Tsunami Hazard Model for the NEAM Region from the TSUMAPS-NEAM project.” Istituto Nazionale di Geofisica e Vulcanologia (INGV), 2018.

[62] P. Baraldi, G. Gola, E. Zio, D. Roverso, and M. Hoffmann, “A randomized model ensemble approach for reconstructing signals from faulty sensors,” *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9211–9224, Aug. 2011.

[63] P. Baraldi, E. Zio, G. Gola, D. Roverso, and M. Hoffmann, “Two novel procedures for aggregating randomized model ensemble outcomes for robust signal reconstruction in nuclear power plants monitoring systems,” *Ann. Nucl. Energy*, vol. 38, no. 2–3, pp. 212–220, Feb. 2011.

[64] T. Homma and A. Saltelli, “Importance measures in global sensitivity analysis of nonlinear models,” *Reliab. Eng. Syst. Saf.*, vol. 52, no. 1, pp. 1–17, 1996.

[65] F. Maio, G. Nicola, Y. Yu, E. Zio, and F. Di Maio, “Sensitivity Analysis and Failure Damage Domain Identification of the Passive Containment Cooling System of an AP1000 Nuclear Reactor,” 2014.

[66] M. Strong, J. E. Oakley, and J. Chilcott, “Managing structural uncertainty in health economic decision models: A discrepancy approach,” *J. R. Stat. Soc. Ser. C Appl. Stat.*, vol. 61, no. 1, pp. 25–45, Jan. 2012.

[67] Y. Y. Kagan, “Seismic moment distribution revisited: I. Statistical results,” 2002.

[68] C. Cauzzi and E. Faccioli, “Broadband (0.05 to 20 s) prediction of displacement response spectra based on worldwide digital records,” 2008.

[69] Gruppo di Lavoro MPS, “Redazione della mappa di pericolosità sismica prevista dall’Ordinanza PCM del 20 marzo 2003. Rapporto Finale,” 2003.

[70] M. Stucchi, C. Meletti, V. Montaldo, H. Crowley, G. M. Calvi, and E. Boschi, “Seismic hazard assessment (2003-2009) for the Italian building code,” *Bull. Seismol. Soc. Am.*, vol. 101, no. 4, pp. 1885–1911, Aug. 2011.

[71] J. J. Bommer and N. A. Abrahamson, “Why do modern probabilistic seismic-hazard analyses often lead to increased hazard estimates?,” *Bulletin of the Seismological Society of America*, vol. 96, no. 6. pp. 1967–1977, Dec-2006.

[72] R. M. W. Musson, “Ground motion and probabilistic hazard,” *Bull. Earthq.*

Eng., vol. 7, no. 3, pp. 575–589, Aug. 2009.

[73] R. Basili *et al.*, "The Database of Individual Seismogenic Sources (DISS), version 3: Summarizing 20 years of research on Italy's earthquake geology," *Tectonophysics*, vol. 453, no. 1–4, pp. 20–43, 2008.

[74] G. Reynoso-Meza, "Multi-Objective Optimization Differential Evolution Algorithm." MATLAB Central File Exchange, 2021.

[75] E. Zio, P. Baraldi, and N. Pedroni, "Selecting features for nuclear transients classification by means of genetic algorithms," *IEEE Trans. Nucl. Sci.*, vol. 53, no. 3, pp. 1479–1493, 2006.

[76] E. Zio, P. Baraldi, and G. Gola, "Ensemble feature selection for diagnosing multiple faults in rotating machinery," *Proc. Inst. Mech. Eng. Part O J. Risk Reliab.*, vol. 221, no. 1, pp. 29–41, 2007.

[77] A.-A. Ahmed, "Feature subset selection using ant colony optimization," *Int. J. Comput.*, vol. 2, no. 1, pp. 53–58, 2005.

[78] A. Grezio, L. Sandri, W. Marzocchi, A. Argnani, P. Gasparini, and J. Selva, "Probabilistic tsunami hazard assessment for Messina Strait Area (Sicily, Italy)," *Nat. Hazards*, vol. 64, no. 1, pp. 329–358, Sep. 2012.

[79] P. Bazzurro and C. A. Cornell, "Disaggregation of seismic hazard," *Bull. Seismol. Soc. Am.*, vol. 89, no. 2, pp. 501–520, 1999.

A. Appendix A

Differential Evolution (DE) is a parallel direct search method which utilizes NP D-dimensional parameter vectors $x_{i,G}, i = 1,2, \dots, NP$ as a population for each generation G. NP does not change during the minimisation process. The initial vector population is chosen randomly and should cover the entire parameter space. DE generates new vectors by adding the weighted difference between two population vectors to a third vector in an operation called "mutation". The mutated vector's parameters are then mixed with the elements of another predetermined vector, the target vector, to yield the so-called trial vector, in an operation referred to as "crossover". If the trial vector yields a lower cost function value than the target vector, the trial vector replaces the target vector in the following generation. This last operation is called selection. Each population vector has to serve once as the target vector so that NP competitions take place in one generation. DE's basic strategy can be described as follows.

Mutation

For each target vector $x_{i,G}, i = 1,2, \dots, NP$, a mutant vector is generated according to:

$$v_{i,G+1} = x_{r_1,G} + F \cdot \left(x_{r_2,G} - x_{r_3,G}\right) \tag{A.1}$$

with random indexes $r_1, r_2, r_3 \in \{1,2, \dots, NP\}$, integer, mutually different and $F > 0$. The randomly chosen integers r_1, r_2, r_3 are also chosen to be different from the running index i, so that NP must be greater or equal to four to allow for this condition. F is a real and constant factor $\in [0, 2]$ which controls the amplification of the differential variation $\left(x_{r_2,G} - x_{r_3,G}\right)$.

Crossover

In order to increase the diversity of the perturbed parameter vectors, crossover is introduced. To this end, the trial vector:

$$u_{i,G+1} = \left(u_{1i,G+1}, u_{2i,G+1}, \dots, u_{Di,G+1}\right) \tag{A.2}$$

is formed, where

$$u_{ji,G+1} = \begin{cases} v_{ji,G+1} \; if \; (randb(j) \leq CR) \; or \; j = rnbr(i) \\ x_{ji,G} \; if \; (randb(j) > CR) \; and \; j \neq rnbr(i) \end{cases} \\ j = 1,2, \dots, D \tag{A.3}$$

In (A.3), $randb(j)$ is the j-th evaluation of a uniform random number generator with outcome ∈ [0; 1]. CR is the crossover constant ∈ [0; 1] and has to be determined by the user. $rnbr(i)$ is a randomly chosen index $\in 1, 2, \dots, D$ which ensures that $u_{i,G+1}$ gets at least one parameter from $v_{i,G+1}$.

Selection

To decide whether or not it should become a member of generation $G + 1$, the trial vector $u_{i,G+1}$ is compared to the target vector $x_{i,G}$ using the greedy criterion. If vector $u_{i,G+1}$ yields a smaller cost function value than $x_{i,G}$, then $x_{i,G+}$ is set to $u_{i,G+1}$; otherwise, the old value $x_{i,G}$ is retained.

The above scheme is not the only variant of DE which has proven to be useful. In order to classify the different variants, the notation: DE/x/y/z is introduced where: x specifies the vector to be mutated which currently can be “rand” (a randomly chosen population vector) or “best” (the vector of lowest cost from the current population); y is the number of difference vectors used; z denotes the crossover scheme. The current variant is “bin” (Crossover due to independent binomial experiments). Using this notation, the basic DE-strategy described can be written as DE/rand/1/bin.

This whole section is extracted from [58].

List of Figures

List of Tables

List of Symbols

ΔT	Exposure time window
γ	Threshold value of the given intensity measure
λ_H	Mean annual rate of exceedance of the γ-th PGA level
λ	Mean annual rate of earthquake occurrence at a given source location
$f_m(m)$	Probability distribution of different earthquake magnitudes
m_{min}	Minimum magnitude of the probability distribution $f_m(m)$
m_{max}	Maximum magnitude of the probability distribution $f_m(m)$
b	Slope parameter of the probability distribution $f_m(m)$
$f_r(r)$	Probability distribution of the source-to-target distance
r	Source-to-target distance
C	Computational cost of the double-loop Monte Carlo Simulation
v	Number of input variables
n_1	Sample size for estimating the inner loop of the double-loop MCS
n_2	Sample size for estimating the outer loop of the double-loop MCS
d	Index of the generic bootstrapped dataset
D	Total number of the bootstrapped dataset
g	Model

Y	Output of the model g
$\bar{X}$	Input parameters of the PSHA
$\bar{\bar{Z}}$	Input-output dataset of the PSHA
s	Index of the input-output pattern of $\bar{\bar{Z}}$
S	Total number of the input-output pattern of $\bar{\bar{Z}}$, i.e., number of rows of $\bar{\bar{Z}}$
N	Number of input parameters, i.e.
n	Index of the input parameter
$Var[\cdot]$	Variance
$\mathbb{E}[\cdot]$	Expectation operator
S_n	Sobol index of the n-th parameter
$\bar{\bar{Z}}_d^*$	Reduced matrix
K	Number of mutually exclusive subset of the reduced matrix
k	Index of the mutually exclusive subsets of the reduced matrix
$\bar{\bar{Z}}_d^{*k}$	Generic mutually exclusive subset of the reduced matrix of the generic bootstrapped dataset
J	Number of rows of the $\bar{\bar{Z}}_d^{*k}$
$S_{n,d}$	Sobol index of the n-th parameter of the generic d-th bootstrapped dataset
$\bar{R}_{BU,d}$	Input ranking (bottom-up strategy) of the generic bootstrapped dataset
$\bar{R}_{BU}$	Final aggregated input ranking (bottom-up strategy)
BC_n	Borda count for the n-th input variable
$p_{n,d}$	n-th input variable order inside the d-th ranking
$\bar{R}_{AO}$	Final aggregated input ranking (all-out strategy)
σ_{GMPE}	Standard deviation of the GMPE
$\bar{a}$	Target site coordinates of the SPTHA
$\psi_{\bar{a}}$	Tsunami intensity in $\bar{a}$
ΔT	Exposure time
$\bar{x}$	Seismic scenario parameters vector of the SPTHA

$\sigma_{\bar{x}}$	Seismic scenario of parameters $\bar{x}$
$\varSigma$	Space of possible seismic scenarios
$\lambda(\sigma_{\bar{x}})$	Annual frequency of the seismic scenario $\sigma_{\bar{x}}$
Pr	Probability
P_e	Probability of exceedance
$\tilde{\psi}$	Tsunami intensity threshold
$\varLambda(\psi_{\bar{a}} \geq \tilde{\psi})$	Annual frequency of occurrence of a tsunami of intensity $\psi_{\bar{a}} \geq \tilde{\psi}$ at location $\bar{a}$
q	Generic simulated seismic scenario
Q	Total number of seismic scenarios to be simulated
Q^*	Optimised (i.e., minimum) number of seismic scenarios to be simulated
H_i	Generic seismic zone
H_n	Total number of seismic zones
$\varTheta$	Set of alternative ET models
ϑ	Generic alternative ET model
M	Number of alternative models for the calculation of $\Lambda(\psi_{\bar{a}} \geq \tilde{\psi})$
x_1	Magnitude of an earthquake
x_2	Depth of the fault
x_3	Strike of the focal mechanism
x_4	Dip of the focal mechanism
x_5	Rake of the focal mechanism
x_6	Area of the fault
x_7	Length of the fault
x_8	Slip of the fault
F	Set of the objective functions of the MODEA
f	Objective function of the MODEA
$\bar{U}$	Decision variable vector

$\bar{U}^*$	Solution vector
$\bar{\bar{A}}$	Features matrix
$\bar{\bar{A}}^*$	Optimal features matrix
SE	Squared error
NP	Population size of the DE
F	Scaling factor of the DE
CR	Crossover probability of the DE
$MAXGEN$	Maximum number of DE generations
D	Number of objectives of the MODEA

List of Acronyms

PSHA	Probabilistic Seismic Hazard Assessment
PGA	Peak Ground Acceleration
GMPE	Ground Motion Prediction Equation
SA	Sensitivity Analysis
GSA	Global Sensitivity Analysis
MGSA	Modularised Global Sensitivity Analysis
BMGSA	Bootstrapped Modularised Global Sensitivity Analysis
IM	Intensity Measure
MCS	Monte Carlo Simulation
BU	Bottom-Up
AO	All-Out
BC	Borda Count
DE	Differential Evolution
GA	Genetic Algorithm
MODEA	Multi Objective Differential Evolution Algorithm
SPTHA	Seismic Probabilistic Tsunami Hazard

Acknowledgements

This study was developed within the research project "Assessment of Cascading Events triggered by the Interaction of Natural Hazards and Technological Scenarios involving the release of Hazardous Substances" funded by MIUR - Italian Ministry for Scientific Research under the PRIN 2017 program (grant 2017CEYPS8).