Soil Dynamics and Earthquake Engineering 148 (2021) 106734

Contents lists available at ScienceDirect

Soil Dynamics and Earthquake Engineering

journal homepage: www.elsevier.com/locate/soildyn

Efficient selection of up to three-component ground excitation sets for earthquake engineering applications

Konstantinos T. Tsalouchidis [a], Lukas Moschen [a,b], Christoph Adam [a,*]

[a] University of Innsbruck, Unit of Applied Mechanics, Innsbruck, Austria
[b] VCE Vienna Consulting Engineers ZT GmbH, Unit of Earthquake Engineering and Structural Dynamics, Vienna, Austria

ARTICLE INFO

Keywords:
Ground motions
Efficient
Selection
Genetic algorithm
Multi-component
Multi-objective optimization

ABSTRACT

Ground motion record selection methodologies are commonly developed to ensure that the input excitation used in response history analyses embodies essential conditions such as spectral compatibility, hazard and intensity measure consistency, seismological and site-specific criteria, always performing in a computationally efficient manner. A methodology utilizing genetic algorithms is revisited here, expanded to select multi-component ground motions and satisfying the typically required selection objectives of earthquake engineering applications, while ensuring increased efficiency. Multi-objective optimization is performed, claimed to be superior in delivering robust results that account for spectral compatibility in first and second order statistics (mean and standard deviation) in a wide range of spectral values, as well as satisfying seismological and site-specific criteria. A unique contribution is the ability to include probability distribution targets in specific ordinates of the spectrum, on top of the mean and standard deviation, allowing for more refined ground motion sets that can be used to reduce the number of records required in response history analyses. Additionally, a novel benchmarking process to assess the efficiency of ground motion record methodologies is introduced here, in terms of providing sets that are globally-optimal solutions to the optimization problem. Through this benchmarking algorithm, the proposed methodology appears to be impeccable in extracting the best possible ground motion sets.

1. Introduction

1.1. Motivation and state of the art

In many earthquake engineering applications, it is typically required to determine the seismic demands of structures through a series of nonlinear response history analyses i.e. by integrating step-by-step the equations of motion in the time domain. This type of analysis is widely used as a reference in research, as it is the most accurate approach [1]. The required input excitation is the time-history of ground accelerations, i.e. ground motion (GM). The most common choice is to use records of previous earthquake events, appropriately selected and (if necessary) scaled to match specific spectral targets (e.g. Ref. [2]). Other choices are simulated or artificial seismograms [3], or seismograms obtained through physics-based simulations of the seismic source, the propagation of the seismic waves and the site effects. GM selection defines the input excitation that connects the seismic hazard of a given region to the response of the analyzed structure. Since GM uncertainty contributes significantly to the uncertainty of structural analysis output, this process can have considerable impact on conclusions regarding structural safety [1,4]. Therefore, numerous research efforts have been made to establish hazard consistent selection and to eliminate pitfalls in defining the input excitation with respect to seismic hazard.

GM selection approaches are commonly classified into two main categories: “scenario-based selection” and “target-based selection”. According to the first one, the selection of records is primarily based on seismological and site parameters corresponding to earthquake scenarios of the seismicity of interest, while then trying to meet spectral targets (or exceed them, as required in most seismic codes) [5]. In the second case, these criteria of causal nature are significantly relaxed and the GM selection is performed based on the spectral target, since spectral values are the main parameters affecting the seismic demand. The uniform hazard spectrum (UHS) is a commonly used target spectrum that has been shown to be a conservative choice because it implies that the large amplitudes of spectral values will occur at all periods [6]. To address this, Ref. [7] introduced the conditional mean spectrum (CMS), which provides the expected (i.e. mean) spectral values, conditioned on the occurrence of a target spectral acceleration value at a specific period

* Corresponding author.
E-mail address: christoph.adam@uibk.ac.at (C. Adam).

https://doi.org/10.1016/j.soildyn.2021.106734
Received 9 December 2020; Received in revised form 12 February 2021; Accepted 18 March 2021
Available online 12 June 2021

of interest. To further expand this rationale, Ref. [8] introduced the generalized conditional intensity measure (GCIM), which considers more GM properties (e.g. significant duration and Arias intensity) as conditional values. An important element in GM selection is the record-to-record variability in the resulting GM set. Therefore, Ref. [9] has developed a semi-automated procedure capable of selecting and scaling GM to fit a target acceleration spectrum while controlling variability in the set. To account for the response spectrum variance in addition to the mean, the conditional spectrum (CS) was introduced by Ref. [10]. In this work, conditional GM selection is achieved by an algorithm that uses Monte Carlo simulations to generate response spectra and then chooses recorded GM to match them, while a "greedy optimization" technique is then employed to further improve the match. In Ref. [11], the algorithm of Ref. [10] was revisited to improve its utility and speed, where the efficiency of the proposed procedure was assessed through its computational expense (i.e., the run time). In that work, the problem of GM selection is approached from an algorithmic point of view rather than focusing on hazard consistency or the appropriate use of GM features i.e. intensity measures (IM), which gather the main research attention around this topic.

Addressing inconsistencies and pitfalls in selecting records and computing target spectra has also been in the center of attention. For example, in Ref. [12], attention is drawn to the difference between spectral accelerations as obtained from any horizontal component of GM, typically used in structural analysis, and the geometric mean of spectral accelerations from the two horizontal components, typically used in hazard analysis. Linking the hazard to the structural response under such inconsistency can yield significantly altered results. Analogically, since probabilistic seismic hazard analysis (PSHA) estimates the seismic hazard by incorporating multiple earthquake scenarios and GM prediction models, Ref. [13] has introduced the "exact CS calculation" which, unlike typical CS calculations, considers multiple causal earthquakes to produce more accurate target spectra. A comparison and thorough review of the exact CS and the GCIM method can be found in Ref. [14], while state-of-the-art reviews of GM selection approaches are presented in Refs. [1,15,16].

Apart from hazard consistency in GM selection, a significant amount of research has focused on practical schemes. Using the Adaptive Harmony Search (a meta-heuristic optimization algorithm), Ref. [17] proposed a GM selection framework that facilitates both code-based and CMS-based selection. In Ref. [18], a genetic algorithm (GA) was used to select GM sets that are compatible with a predefined response spectrum, supported by a graphical user interface (GUI). To support structure-specific GM selection for seismic design and assessment, Ref. [19] introduced the "ISSARS" software environment integrated with a finite element analysis package, while the same authors in Ref. [20] aimed to support reliable design and assessment by presenting a decision support process for structure-specific selection of code-compatible GM. Structure-specific and hazard-consistent selection for risk-targeted ground motion selection was also addressed in Ref. [21], where a web-based system was developed for the selection of characteristic ground motions. To accommodate code-based selection of real records for practitioners, "REXEL", an open-source software tool that is accompanied with a GUI, was introduced in Ref. [22].

The vast majority of applications incorporating GM excitation involve 2D structural models that are horizontally excited with one component of the GM. Accordingly, research on hazard consistency and the appropriate use of IM to ensure efficiency and sufficiency is mostly focused on these types of applications (definitions regarding efficient and sufficient IM can be found in Ref. [23]). The shortcoming of first-mode spectral acceleration is addressed in Ref. [24], which is often used as a conditioning IM in CS, but is only useful when GM selection is required for specific single-mode dominated structures. In that work, the log-average spectral acceleration over a period range is suggested instead, which seems to be a superior scalar IM particularly suitable for multiple demand parameters and also regional risk assessments of

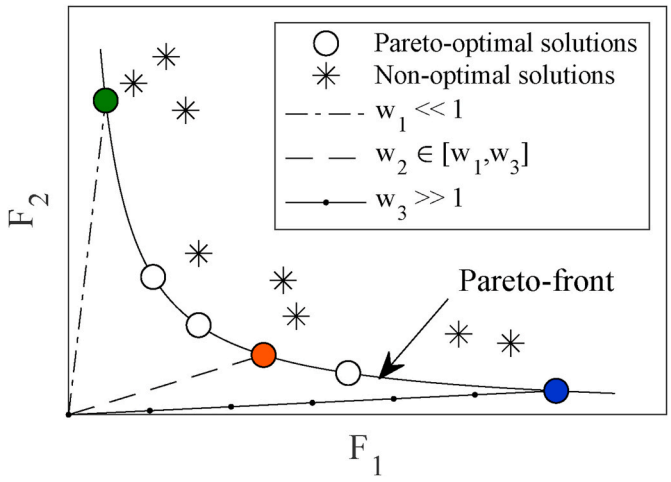

Fig. 1. Schematic representation of two-objective optimization (F_1 , F_2) showing pareto-optimal and non-optimal solutions, the pareto-front and the influence of choosing one of three different weight factors $w = w_1$, w_2 and w_3, respectively, when one-objective optimization is performed instead.

building portfolios, also supported by other studies (e.g. Refs. [25–30]). Focusing on one-dimensional excitation is useful because usually one of the horizontal axes of the structure is critical and therefore the analyses related to this axis are most important. However, this is not the case for torsion-sensitive structures, such as irregular buildings. In such cases, bidirectional horizontal excitation of a 3D structural model is a necessity, along with the appropriate use of IM that refer to both axes. Several approaches have been proposed to address the key elements of bidirectional selection, including the appropriate use of IM, e.g. Refs. [31–35]. One step further is to incorporate the vertical component as well, resulting in three dimensional excitation. This is essential in applications where the evaluation of non-structural components is intended, as they are usually more sensitive to accelerations than displacements, and it has been shown that the vertical acceleration component is necessary in the definition of vertical acceleration demands [36,37]. In addition, for some more complex structures, such as buildings with cantilevers and bridges, the vertical component of the excitation may also be critical to the structural safety. In Ref. [38], a methodology for selecting, scaling and orienting three components of GM for intensity-based assessments is presented, in which each record is selected to match the target spectra as closely as possible. A discussion of multicomponent hazard identification is presented in that work, emphasizing the need to clarify appropriate multicomponent target spectra.

The previous methods aim to select a single GM set, which usually has a mean spectrum matching a mean target spectrum, while some of them also consider the variance spectrum matching a target. To achieve spectral match in the first and second order statistics (i.e., mean and variance), it is common to use a single objective function that is the weighted sum of the two individual objectives (e.g. Refs. [10,11,24,26,31,38–40]). In contrast, several selection strategies praise the use of multi-objective optimization, which (normally) results in more than one optimum solution forming the pareto-front, i.e., an ensemble of GM sets referred as pareto-optimal, where no other set can be found to improve at least one objective while not reducing the other objectives (Fig. 1). The benefit of pursuing a multi-objective optimization is that it sidesteps the use of weight factors that give relative importance between the objectives, whose values are somehow arbitrary, without physical meaning, and therefore difficult to choose.

In Fig. 1, a schematic representation of a two-objective optimization problem is presented, where F_1 and F_2 represent the scores of the two objectives (i.e. the fitness or objective functions) that the optimal solutions minimize. In the same figure, it is evident that selecting a positive

value for a weight factor w (e.g. $w = w_1, w_2, w_3$) to perform a single-objective optimization as the weighted sum of the two objectives F_1 and F_2 (i.e. $F = F_1 + wF_2$) affects the resulting solution. Choosing, for example, $w = w_1$ leads to a single-objective optimization that returns the green GM set as a solution,[1] which could have significantly worse performance in F_2 than a different GM set that satisfies both objectives well, e.g., for $w = w_2$ resulting in the red GM set. Even if F_1 and F_2 are fairly well minimized for all $w = w_1, w_2, w_3$, different properties might exist in the resulting GM sets. That is, blue, green and red GM sets in Fig. 1 might differ significantly in their seismological and site properties, their number of records or the scale factors used. By extending the rationale of using weight factors to three-component selection, the problem of appropriately selecting their values becomes even greater because a weight for each additional objective must be selected.

The starting point of the proposed approach is the work in Ref. [41], where a GA was used to perform the multi-objective selection of GM sets that have a median and standard deviation spectrum that match predefined targets in a period range. Multi-objective selection using GA was also suggested by Ref. [42], where more objectives were introduced such as "earthquake scenario" parameters (e.g., magnitude, source-to-cite distance, soil conditions). In that work, a thorough review of selection schemes is presented and the need for guidance regarding the use of metaheuristic optimization algorithms (such as GA) is emphasized, since they become more and more popular in GM selection applications. One more benefit in using GA for the process of GM selection, as pointed out in Ref. [43], is the elimination of Monte Carlo simulations along with the increased efficiency. The notion of efficiency in the optimization process is frequently used to describe that the resulting GM sets have a good match with the targets and that the optimization process is performed quickly. It is well understood that GM selection is an optimization problem that aims to find the optimal solution(s) that satisfy the objectives without exploring the entire solution space, as this is impractical, if not unfeasible. The number of solutions are the possible combinations of records given by the binomial coefficient, which leads to huge solution spaces, even for small databases with a few hundred records [41–43].

1.2. Objective of the proposed work

In the present work, the concept of multi-objective selection is extended for the first time in up to three-component ground excitation sets, having spectral targets as objectives and using GA to perform the optimization process. In this way, given the desired spectral targets in up to three orthogonal directions, the goal of the GM selection optimization process is to satisfy the targets by extracting from a suitable database of GM records, subsets with a desirable number of records. Typically, these spectral targets refer to first and second order statistics, i.e. specific mean and standard deviation spectra in a range of periods. A unique contribution of the proposed framework is the ability to specify the whole probability distribution function as a target, on top of the mean and standard deviation targets, at specific periods of interest in the spectrum. This is accomplished through a concept inspired by the stratified sampling technique, and its impact in GM selection is demonstrated through an application. In another application that demonstrates the efficiency of the proposed methodology, a novel procedure is introduced to assess the quality of the resulting GM sets in reaching (or at least approaching) the globally optimal solutions of the optimization problem. This is of paramount importance and is investigated for the first time in a GM selection scheme. It clearly shows the performance of the optimization process as opposed to the less informative measure of quality in the spectral match (also shown here), since the later is largely dependent on the targets as well as the database of GM used. In terms of computational cost, the run-time of the proposed methodology is in the range of minutes to a few hours, which is typically not critical since GM selection is not a repetitive process in most applications. Therefore, computation times are only approximately indicated when necessary to promote understanding rather than benchmark the proposed procedure. The overview of the whole framework, the implementation of the various selection options and relevant applications are presented in the following sections.

[1] If the applied optimization scheme is able to return the optimal solution, otherwise a non-optimal solution is returned.

2. A comprehensive framework for selecting ground motion record sets through genetic algorithms

The broad family of evolutionary algorithms (EA) are methods to solve optimization problems inspired by the biological evolution and its key mechanism, the natural selection, where the heritable genes change in the population over the progression of generations. These changes are achieved in nature through some biological operations such as selection, crossover and mutation and the corresponding (genetic) operators in the current EA optimization scheme have an important role and therefore will be described in detail below. In the problem formulation, the population is a number of candidate solutions (i.e. GM record sets) that iteratively change their characteristics (often referred to as genes) through selection, crossover and mutation. Each iteration corresponds to a new generation of the population. As the generation progresses, the candidate solutions in the population should optimally satisfy the objective functions and form a near-global Pareto front.

The algorithm used in this work is a subclass of EA, an elitist genetic algorithm variant of NSGA-II [44]. It is part of the Matlab Global Optimization Toolbox and all procedures are implemented in Matlab specifically for the purpose of the proposed GM record selection approach, so that the configuration of the optimization problem is tailored to facilitate its needs. Three phases can be distinguished, similarly to Ref. [41]: pre-processing, processing and post-processing. In pre-processing, the parameters and input data of the selection are defined. Processing follows, where the proposed optimization scheme converges to the solution. The outcome is a list of GM sets with detailed information about the properties of each GM record utilized. To assess these GM sets and choose the interesting one(s), the post-processing phase is a tool to display, rank and evaluate them. The three phases are described in detail below.

2.1. Pre-processing: defining selection criteria and targets

In the pre–processing phase the input criteria regarding the selection process are defined. The required information for the selection process is:

- directions of the GM components defined in up to three orthogonal axis (i.e. x, y and z)
- target spectra, which are the objectives that the selection process is designed to meet, two for each direction of interest: a mean and a standard deviation target spectrum
- GM screening criteria, reducing the initial size of the GM database, mainly based on seismological and site parameters in order to predominantly account for the earthquake scenarios that are expected to contribute to the seismicity of interest
- properties of the resulting GM record sets regarding their scaling and the number of records in the sets, in terms of admissible ranges

The above input choices for the selection process are explained in greater detail here to better clarify them and demonstrate the versatility of the proposed selection approach.

2.1.1. Single- and multi-component multi-objective selection

Firstly, the desirable dimensions of the GM record sets are specified in terms of the considered components, which can be up to three

Table 1
Possible record selection schemes performed through the proposed methodology.

Notation	Record Selection Description
One-component target selection	
H1	One arbitrary horizontal component is selected to match one horizontal target spectrum
Hgm1	The geometric mean of two orthogonal horizontal components is calculated and matched with the target spectrum of the geometric mean
V1	The vertical component is selected to match the vertical target spectrum
Two-component target selection	
H2	Two orthogonal horizontal components are selected to match two horizontal target spectra
HV2	One arbitrary horizontal and the vertical components are matched with a horizontal and a vertical target spectrum
HgmV2	Similar to HV2, with the difference that the horizontal component is the geometric mean of two orthogonal horizontal ones
Three-component target selection	
HV3	Two orthogonal horizontal and the vertical components are matched with three target spectra

(namely x, y and z components). In the case of three-component selection, the outcome of the selection process will be used as excitation of 3D structural models in the three orthogonal directions, thus utilizing each component of the GM (two horizontal and the vertical). Separately from the dimensionality of the record sets, the targets of the selection process are also defined here in terms of the above-mentioned desirable directions. It is noted that the target spectra might be provided for just one of these directions, hence the remaining components will be utilized regardless of their spectral properties. Alternatively, target spectra in more directions of interest can be provided, driving the selection to satisfy one-, two- or three-component targets described in more detail in Table 1. In this table, seven options are presented in total, referred to as $'H1'$, $'Hgm1'$, $'H2'$, $'V1'$, $'HV2'$, $'HgmV2'$ and $'HV3'$, where $'H'$ denotes horizontal, $'V'$ vertical, $'gm'$ geometric mean of two orthogonal horizontal components, and the number indicates the dimensionality of the targets (1, 2 or 3).

2.1.1.1. Single-component target selection. As previously explained, most earthquake engineering applications concern 2D structural models and therefore the majority of the record selection approaches focus on one horizontal component of the excitation to match target spectra that are consistent with the hazard at the site of interest. Typically used in real applications, this component of the selected records is oriented along the critical building axis to perform nonlinear response history analyses, as it is expected that the accumulation of damage and collapse primarily originates from this direction for a specific level of shaking. This type of selection is referred here as $'H1'$ since one target in the horizontal direction is employed. A common issue in record selection observed in Ref. [12] is the inconsistency between selecting arbitrary horizontal components of GMs to match targets that have been calculated with respect to the geometric mean. To address this, under the choice $'Hgm1'$, the geometric mean of the two horizontal components is calculated and then matched to the target spectrum. Finally, if the vertical excitation is of interest, namely $'V1'$ (similar to 'H1' but in the vertical direction), the vertical component of the GMs is matched to the target.

2.1.1.2. Two-component target selection. Maintaining hazard consistency in both horizontal components when performing nonlinear dynamic analyses in 3D models is of great importance [31]. Therefore, two orthogonal horizontal spectra can be defined by $'H2'$ as targets for the selection process and the resulting record sets will be (at least) two-dimensional. Two-dimensional record sets may also be of interest with respect to one arbitrary horizontal component and the vertical component, denoted here as $'HV2'$. If the geometric mean of the horizontal components is preferred instead of the one arbitrary, the $'HgmV2'$ selection option is pursued.

2.1.1.3. Three-component target selection. Matching spectral targets in all three orthogonal directions can be pursued by the option $'HV3'$, where the proposed procedure performs selection of three-component GM that fit first and second order spectral targets in each direction (i.e., six objectives in total).

2.1.2. Spectral targets and pre-selection criteria

The spectral targets must be defined for each direction of interest. The targets are provided here as mean and standard deviation spectra of the natural logarithm of spectral acceleration together with the period range in which they are matched. Additionally to these first and second order statistics in a period range, the probability distribution function at specific spectral ordinates can be requested as a target (e.g. a target of lognormal probability distribution for the spectral accelerations at T = 1s). This is a unique element in the proposed approach, not supported by other selection strategies. This is achieved through a concept similar to latin hypercube sampling (LHS), schematically shown for one variable in Fig. 2 (a), where the probability space is divided into equal sub-spaces and one random number is generated inside each corresponding interval in the variable domain. Here, since we seek to select appropriate records rather than generating numbers, the selection process is forced to favor the record sets that have spectral acceleration values, at specific periods, falling into appropriate 'strata' (i.e. sub-spaces in the spectral acceleration domain). These sub-spaces are obtained by the same process shown in Fig. 2 (a), e.g. at T = 1s in Fig. 2 (b), thus resulting in the required probability distribution efficiently. More details regarding the implementation of this requirement is provided in section 2.2.

In addition to selecting GMs based on their spectral shape as defined by the targets, additional preliminary criteria based on seismological, site, and other parameters are typically used in the literature to exclude GMs with very different characteristics from the target hazard. Quantitative recommendations on causal parameter bounds are provided in Ref. [45] for PSHA-based GM selection. Here, the screening criteria are the fault rupture mechanism, the magnitude (M_w), the source-to-cite

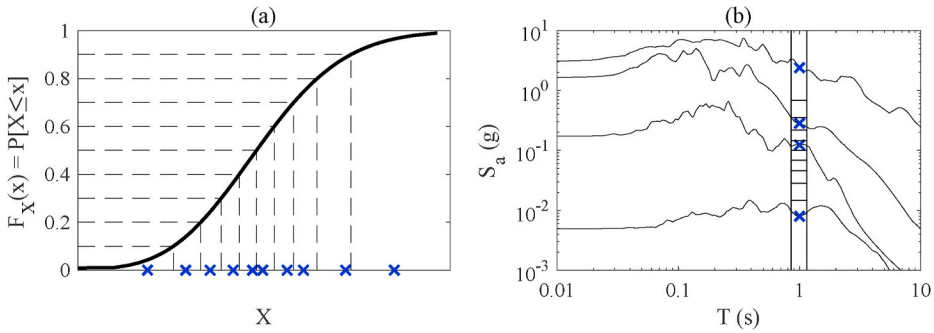

Fig. 2. (a) The LHS concept for one variable; (b) the proposed scheme for GM selection, equivalent to LHS.

distance (R_{JB}) (as defined by Joyner and Boore), the soil conditions at the site expressed through the average shear wave velocity at the top 30 m ($V_{s,30}$), as well as scaling limits to exclude records that would require excessive scale factors to reach the targets in the period ranges of interest. All these criteria reduce the initial database of GM records (*RecPool*) to a subset (*BoundedRecPool*) of more appropriate ones to the application at stake.

2.1.3. Scaling of ground motions

The proposed approach is one of the scaling methods in which the scale factors of the records in the GM set are appropriately chosen so that its statistical properties match with the targets while circumventing the introduction of bias due to excessive scaling [4].

Controlling the scaling properties of GM sets are of particular interest and setting bounds is largely supported in the research community (e.g. Refs. [31,38]), while a scale factor as close to unity (unscaled record) is preferred e.g. Refs. [42,46]. The bounds of the scale factors are input variables in this framework, changeable depending on the application. A reasonable value of 5 for an upper bound of scaling factors is considered in this work, as was proposed by Ref. [41]. The lower bound is set to 0.5, although it is not an important parameter since scaling is normally used because of the lack of recorded GMs at high intensities, therefore increasing rather than decreasing the intensity of records is typically sought. Another parameter of the scaling factors is the flexibility in their usage. Constant scaling among the records in a specific set can be considered beneficial because the relative intensities between unscaled motions in the scaled set are preserved [41]. However, this constraint limits the ability to meet specific targets, especially when multiple components are of interest, which means that equal scaling would be applied to all records and components. For this reason, in Ref. [33], following the rationale in Ref. [47], it is argued that different scaling factors between components could be allowed. In Ref. [31], supporting these views on the topic, a differentiation of 5% between the scale factors of the two horizontal components of the GMs is permitted.

Considering the above, the proposed method employs a constant scaling factor for all records in a record set, namely $SF_{constant}$, for which appropriate bounds are provided as input choices (with $SF_{constant} \in [0.5-5]$ being the default). To allow for some flexibility between the different components in x, y and z directions, a differentiation of 5% between each component can be chosen, resulting in the relation

$$0.95 \le \frac{SF_i}{SF_{constant}} \le 1.05,\ i = x,y,z \tag{1}$$

Consequently, the scale factors SF_x, SF_y and SF_z, which correspond to scale factors in x, y and z directions, differ less than 5% between each other, constant for the entire set. It is noted that if the above is not considered important in a specific record selection application, complete independence of the scale factors between records is also supported by the proposed procedure, offering even more flexibility.

2.2. Processing: extracting pareto-optimal solutions

After the parameters of the selection are defined and the initial database of GM is reduced based on the preliminary criteria, the selection of the sets is initiated. Here, the formulation of the optimization problem and its basic solution procedures are explained.

2.2.1. Optimization problem formulation

As previously mentioned, for each direction of interest $i = x,y,z$ there are two targets in the period range $T \in [T_{start}, T_{end}]$, the mean $\widehat{m}^{(tgt)}_{lnSa_i}(T)$ and the standard deviation $\widehat{\sigma}^{(tgt)}_{lnSa_i}(T)$ spectrum targets of the logarithm of spectral accelerations. By the end of the optimization problem, the $k = 1,\ldots,K$ solutions (i.e. GM sets) should have mean $\widehat{m}^{(set-k)}_{lnSa_i}(T)$ and standard deviation $\widehat{\sigma}^{(set-k)}_{lnSa_i}(T)$ spectra such that the objective functions

$$F^{(set-k)}_{m,i} = \sqrt{\sum_{T=T_{start}}^{T_{end}} \left(\widehat{m}^{(set-k)}_{lnSa_i}(T) - \widehat{m}^{(tgt)}_{lnSa_i}(T)\right)^2} \tag{2}$$

$$F^{(set-k)}_{\sigma,i} = \sqrt{\sum_{T=T_{start}}^{T_{end}} \left(\widehat{\sigma}^{(set-k)}_{lnSa_i}(T) - \widehat{\sigma}^{(tgt)}_{lnSa_i}(T)\right)^2} \tag{3}$$

are minimized forming a near global pareto-front (similar to Fig. 1). Only for the case where a distribution function (e.g., lognormal distribution) at a specific period (or periods) is an additional target, the process described in section 2.1.2 is performed and the sub-spaces at the spectral acceleration domain are obtained. The n_k records of set k are forced to fall into the n_k sub-spaces by minimizing the number of sub-spaces not occupied with a spectral acceleration value (N_{Empty}). For the case of multiple periods of interest, N_{Empty} is the sum of empty sub-spaces in all periods. The additional objective function is given by

$$F_{LHS,i} = N_{empty} \tag{4}$$

To obtain the mean and standard deviation spectra of the logarithm of spectral accelerations for a GM set k containing n_k records, used in Equations (2) and (3), one has to compute them for each period of interest $T \in [T_{start}, T_{end}]$ by the following equations,

$$\widehat{m}^{(set-k)}_{lnSa_i}(T) = \frac{1}{n_k}\sum_{j=1}^{n_k} ln\left(SF^{(set-k)}(j)Sa_i(j,T)\right) \tag{5}$$

$$\widehat{\sigma}^{(set-k)}_{lnSa_i}(T) = \sqrt{\frac{1}{n_k-1}\sum_{j=1}^{n_k}\left(ln\left(SF^{(set-k)}(j)Sa_i(j,T)\right) - \widehat{m}^{(set-k)}_{lnSa_i}(T)\right)^2} \tag{6}$$

Note that the scale factors $SF^{(set-k)}(j)$ in these equations are written for the general case where they are independent between the records $j = 1,\ldots,n_k$, whereas for the case of constant scaling between the records in set k, it is $SF^{(set-k)}(j) = SF^{(set-k)} = SF_{constant}$, see Equation (1). The term $Sa_i(j,T)$ refers to the spectral acceleration of the component i, record j at period T.

Therefore, the optimization problem comes down to finding the proper scale factors as well as an index to identify which records take part in each set, similar to related research on the topic [41,42]. In this work however, to increase the speed and efficiency of the optimization scheme, we minimize the number of variables by including the above information in the same variable. For the GM set k, there are N variables $SF = [SF_{rec1}, SF_{rec2}, \ldots, SF_{recN}]$, taking the values $SF_{rec} \in \{0, [SF_{min}, SF_{max}]\}$ for each of the N records available in the database. Zero denotes that the record is not included in the set and the nonzero variables are bounded by the limits of the scale factors. Moreover, their total number (n_k) is bounded by the allowed range of records in a set. Finally, implementing the above in vectorized form through Matlab allows for great speed in the computations of Equations (2)–(6) at each step of the optimization until convergence is reached. The steps and procedures of the optimization process are explained in section 2.2.2.

2.2.2. Convergence to solution: Creation, selection, crossover, mutation and reproduction

To converge to a near-global (or global) pareto-optimal solution, a series of five processes take place, as subsequently described.

2.2.2.1. Creation. The first step is to create an initial population of candidate solutions also referred to as individuals (i.e. GM record sets), each of them represented by a vector of N variables $[SF_{rec1}, SF_{rec2}, \ldots, SF_{recN}]$ corresponding to each one of the N records of the database. The number of the population size is crucial for the behavior of the optimization. A large population size increases the chances of converging in a near-global pareto-optimal solution by exploring the solution space more thoroughly, but also increases the computation time. Since N is

typically in the range of hundreds, a population of $PopSize = 200$ individuals is suggested and implemented as the default choice in the built-in procedures in Matlab, following the work in Ref. [44]. Due to the measures taken in the proposed procedure to efficiently formulate the optimization problem, this optimization option is adjustable and increased in our scheme, since a population size of $PopSize = 1000$ converges in several minutes.[2]

To create the initial population, a random number n_k in the range of desirable records is generated for each of the individuals, and n_k out of N randomly chosen variables are assigned with nonzero values in the range of the admissible scale factors. In this way, the initial population that satisfies all the constraints of the optimization scheme is randomly selected in the form of the matrix

$$SF_{iteration-0} = \begin{bmatrix} SF_{rec1}^{(k=1)} & SF_{rec2}^{(k=1)} & \cdots & SF_{recN}^{(k=1)} \\ SF_{rec1}^{(k=2)} & SF_{rec2}^{(k=2)} & \cdots & SF_{recN}^{(k=2)} \\ \vdots & \vdots & \ddots & \vdots \\ SF_{rec1}^{(k=PopSize)} & SF_{rec2}^{(k=PopSize)} & \cdots & SF_{recN}^{(k=PopSize)} \end{bmatrix} \quad (7)$$

where each row corresponds to an individual. If constant scaling is chosen, the nonzero values in each row are equal. After the initial population $SF_{iteration-0}$ is created, the iterative process begins, where selection, crossover, mutation and reproduction choices will control its behavior.

2.2.2.2. Selection. In each iteration, a new population is formed, starting with the first $SF_{iteration-0}$ to the last, and each individual in the population is assigned a fitness score based on its ability to minimize the objective (or fitness) functions, e.g. Equations (2) and (3). Through selection, some individuals are chosen from the population to become the parents for the next generation, which means that these individuals are used to generate the children through the operations of crossover and mutation (described in the following paragraphs). In this work, the tournament selection method is used [48], where the individuals with the highest fitness scores have a greater chance to be selected as parents.

2.2.2.3. Crossover. Through the process of crossover, characteristics from two parents are combined to create the crossover children, with the idea that combining the characteristics of two well-performing individuals will potentially result in an even better individual. There are multiple ways to perform crossover. In this work, the operation needs to generate crossover children that respect the constraints of the problem, which are the bounds of the scale factors, constant scaling (if required) and a constrained number of nonzero variables. For this reason, a record that is not included in either parent (i.e., both parents have a common zero value) will not be included in the child as well, and the records common in both parents will be included in the child with their scale factor being the average of the parents, i.e. $SF^{(child)} = (SF^{(k=parent1)} + SF^{(k=parent2)})/2$. The remaining part of the non-similar scale factors of the parents (one zero value and one non-zero value) will be chosen randomly to ensure that the child has appropriate number of non-zero scale factors, following all the necessary constraints.

2.2.2.4. Mutation. A typical problem with the use of genetic algorithms is premature convergence to a local optimum. This is caused by the lack of diversity in the population and disproportionate exploitation (i.e., using existing solutions to obtain new, refined ones with better fitness scores) versus exploration (i.e., searching for new solutions in different regions of the solution space). Mutation, which can be done in various ways, can bring diversity to the population, as it randomly alters a small amount of the genes of the parents.

For each parent that takes part in mutation, a uniform multiplication of the scale factors can be found that minimizes Equation (2) (or the sum of them in the case of multi-component optimization) and is performed replacing their original value. This does not affect the score of Equation (3). Moreover, the parents chosen for mutation, are changed by randomly swapping nonzero with zero variables in their vector. Depending on the fitness score of the parents, the number of variables swapped varies from one (for the best performing ones) to three (for the worst performing ones). These randomly changed vectors of the parents are the vectors of the mutation children.

2.2.2.5. Reproduction. The population to be created in the next generation is a combination of high performing individuals from the previous generation, referred to as elitist selection, and the children of the parents obtained trough crossover and mutation. The reproduction options control the contribution of elitist selection, crossover children and mutation children that will take part in the next generation. Elitist selection ensures that the quality of the solution does not decrease from one generation to the next. Through crossover, the algorithm converges to a solution and therefore it normally accounts for the greater part of the next generation. Finally, through mutation divergence is pursued, further exploring the solution space and avoiding local optima and typically makes up a smaller portion of the next generation. Here, 5% of the population size is assigned to elite individuals (e.g., 50 individuals for a population size of 1000), while 80% of the remaining is assigned to crossover and 20% to mutation children (e.g., 760 and 190 individuals for a population size of 1000).

2.3. Post-processing: choosing between the optimal solutions

The last part of the selection process is to assess the solutions, as they are expected to be more than one and form a near-global Pareto front. Three metrics are introduced according to which the solutions are ranked, namely $R_{spectral}$, R_{causal} and R_{dist}, which are explained below.

2.3.1. Spectral match ($R_{spectral}$)

At this point, the concept of weight factors is used to rank the different solutions according to how well they match the spectral targets. By changing and adjusting the weights, one quickly obtains a different rank of the resulting solutions after the entire optimization scheme is completed, with the best solution having the lowest $R_{spectral}$ value with respect to this metric. This is considered superior here to the alternative of using weight factors to merge the objective functions, which leads to single-objective optimization, since this would require rerunning the optimization process for each combination of weight factors, resulting in tedious computations. In the general case of three-component selection, where $i = x, y, z$, each solution has an $R_{spectral}$ value obtained by

$$R_{spectral} = \sum_i \left(w_{m,i} F_{m,i} + w_{\sigma,i} F_{\sigma,i} \right) \quad (8)$$

2.3.1.1. Earthquake scenario (R_{causal}). A rank of the GM sets (in ascending order) by criteria related to the earthquake scenarios and site conditions is obtained through R_{causal}. In this way, stricter criteria than those introduced in section 2.1.2 can be explored as ranking objective. For example, based on the PSHA results and disaggregation of the hazard at the site of interest, the values of M_w, R_{JB} and $V_{s,30}$ of the most contributing scenario(s) can be chosen, resulting in $BoundedRecPool^*$, a subset of the database $BoundedRecPool$ used for selection. Therefore, the value of R_{causal} for each set based on the number of records Rec included in $BoundedRecPool^*$ is obtained by

$$R_{causal} = 1 - \frac{|Rec \in BoundedRecPool^*|}{|Rec|} \quad (9)$$

[2] on a personal computer with 3 GHz Intel Xeon W CPU processor.

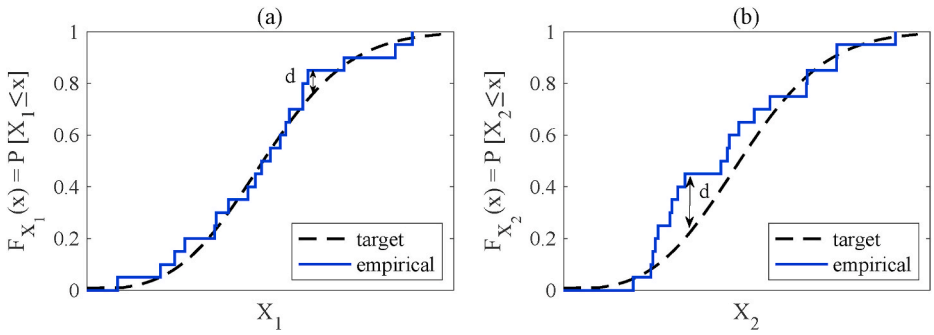

Fig. 3. Schematic representation of the comparison of the empirical distribution functions of (a) X_1, (b) X_2, with a common target distribution by the metric value d.

2.3.1.2. Target distribution match (R_{dist}). An important element of the proposed selection process is the ability to specify a target distribution of spectral accelerations (instead of only mean and standard deviation targets), at specific periods in the spectrum. To assess and to rank the resulting GM sets based on this objective, the metric value d from the Kolmogorov-Smirnov goodness of fit test is used [49], which is the maximum absolute value of the difference between the empirical (i.e. the one obtained from the sets) and the target cumulative distribution function. As an example, Fig. 3 shows the metric value d for two different empirical distributions X_1 and X_2, where the comparison to the same target distribution appears to be better in X_1 compared to X_2, even though X_1 and X_2 have the same means and standard deviations. In Fig. 3, it is clear that a lower d value corresponds to a better fit to the target distribution and is therefore assigned as R_{dist} for each GM set, or in case of multiple periods of interest, the sum of d values in all of them is assigned, resulting in a rank of the solutions in ascending order.

3. Applications

In this section, several applications of GM selection using the proposed procedure are presented to demonstrate its main features and contributions. Since multi-dimensional target selection is one of the main extensions of the current work, the ultimate case of three-component target selection is presented along with the most common, one-component horizontal target selection. It is considered that the other record selection options supported through this work, as presented in Table 1, are essentially extensions or simplifications of the one- or three-component target selection respectively, and therefore, performing applications with respect to them does not lead to a deeper understanding of the proposed approach or its efficiency.

As mentioned earlier, efficiency in GM selection usually refers to the ability of the procedure to select GM sets fast and with good agreement with respect to the target spectra. The latter can be qualitatively assessed by the values of the objective functions obtained from Equations (2) and (3), or most commonly, quantitatively assessed schematically by the spectra. In this work, the applications that demonstrate successful spectral matching are presented in section 3.2. Reduced computation time is always a requirement in engineering applications. Since in the current procedure the time required for GM selection is in the range of minutes (up to a couple of hours if the application requires it), in the next applications the computation time is indicated only if it provides understanding and not to evaluate the process in terms of computation speed.

A different notion of efficient GM selection is explicitly investigated in section 3.1, where the proposed methodology is tested for its ability to extract GM set(s) that form a near-global (in contrast to a local) optimum solution to the optimization problem. A "global" optimum means that the resulting GM set(s) is the optimal solution among all possible solutions. Not converging to global optima is a typical concern when optimization techniques are used to solve problems. It is proposed here that this application is used as a benchmark process for the various GM selection optimization approaches to address this issue.

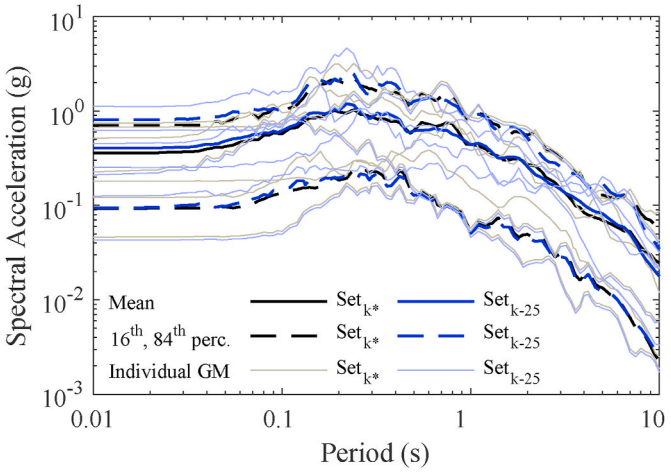

Fig. 4. The globally optimal solution Set_{k^*} with $n_{k^*} = 7$ GM records represented by the individual spectra (gray lines), their mean spectrum (black solid line), and 16th and 84th percentile spectra (black dashed lines), together with the results of the Algorithm 1 (corresponding blue lines) of the pareto-optimal solution Set_{k-25}, obtained after 25 iterations. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.1. Extracting the globally optimal solution

To examine whether a GM selection strategy provides the globally optimal solution, the entire solution space must be explored (i.e., test the outcome of the objective functions for all possible combinations of records and keep the best ones), which is computationally very expensive. In this work, an alternative strategy is introduced explained by the algorithm presented above (subsequently referred as "Algorithm 1") to showcase whether the proposed methodology approaches the global optimum.

The process starts by providing a pool of GM records, referred to as *RecPool*, along with the admissible ranges of the causal parameters $M_w^{(range)}$, $R_{JB}^{(range)}$, $V_{s,30}^{(range)}$ that form the pre-selection criteria for screening the pool of GM records (as would be the case in any typical GM selection strategy), and the number of records (n_{k^*}) in the GM set k^* along with a constant scale factor (SF_{k^*}). After screening the *RecPool*, to create the *BoundedRecPool*, n_{k^*} GM records are selected randomly and scaled with SF_{k^*}, forming the set k^*. The mean and standard deviation spectra of this set are computed by Equations (5) and (6), and they are assigned as targets for the optimization process. The process is then performed with the objective of satisfying Equations (2) and (3), where the solutions, i.

e., GM sets $k = 1, \ldots, K$, that all have n_{k^*} records, are assessed in approaching the already known globally optimal solution, i.e., GM set k^*. The ratio of the number of common records in each of the $k = 1, \ldots, K$ sets and k^* over n_{k^*} is obtained and the maximum value, denoted *GlobalSolutionRatio*, is the output of the process, where *GlobalSolutionRatio* $= 1$ means that the global solution is found. It should be noted that this process is identical regardless of the number of components utilized from the GM records.

Algorithm 1. Globally optimal solution benchmark

Algorithm 1 Globally optimal solution benchmark

Input: $RecPool, n_{k^*}, SF_{k^*}, M_w^{(range)}, R_{JB}^{(range)}, V_{s,30}^{(range)}$
Output: $GlobalSolutionRatio$
1: $BoundedRecPool \leftarrow \{RecPool : M_w \in M_w^{(range)} \text{ and } R_{JB} \in R_{JB}^{(range)} \text{ and } V_{s,30} \in V_{s,30}^{(range)}\}$
2: $Set_{k^*} \leftarrow$ Select n_{k^*} records from $BoundedRecPool$ randomly
3: $SSet_{k^*} \leftarrow Set_{k^*} \times SF_{k^*}$
4: Calculate $\hat{m}_{lnSai}^{(set-SSet_{k^*})}(T)$ and $\hat{\sigma}_{lnSai}^{(set-SSet_{k^*})}(T)$ from Equations 5, 6
5: $\hat{m}_{lnSai}^{(tgt)}(T) \leftarrow \hat{m}_{lnSai}^{(set-SSet_{k^*})}(T)$
6: $\hat{\sigma}_{lnSai}^{(tgt)}(T) \leftarrow \hat{\sigma}_{lnSai}^{(set-SSet_{k^*})}(T)$
7: $Set_k \leftarrow$ Run the optimization procedure to obtain the $k = 1, \ldots, K$ GM sets, that minimize Equations 2, 3 ▷ $|Set_k| = |Set_{k^*}|$
8: **for** $k = 1, \ldots, K$ **do**
//Compute the cardinality of the intersection of Set_k and Set_{k^*}
9: $InterSet_k = Set_k \cap Set_{k^*} = \{Rec : Rec \in Set_k \text{ and } Rec \in Set_{k^*}\}$
10: $C(k) = |InterSet_k|$
11: **end for**
12: $GlobalSolutionRatio \leftarrow max(C)/n_{k^*}$
13: **return** $GlobalSolutionRatio$

The initial pool of records *RecPool* comes from the NGA-West2 database [50], which contains 21,336 (mostly) three-component GM records from 599 events covering a wide range of magnitudes, source-to-site distance and soil conditions. To screen this database, considering less diverge (i.e., more realistic to be used in a GM selection process) earthquake scenario parameters, a choice of magnitudes with $M_w^{(range)} \in [5.5, 7.9]$, source-to-site distances in *km* with $R_{JB}^{(range)} \in [0, 100]$ and shear wave velocity in m/s with $V_{s,30}^{(range)} \in [180, 360]$ is chosen, resulting in *BoundedRecPool* containing 682 GM records.

With these pre-selection criteria and $SF_{k^*} = 2.5$, the Algorithm 1 is tested for GM sets of $n_{k^*} = 7$ records and repeated 10 times, because its random nature might play an important role in the outcome. The population size is increased to 50,000 for this application to explore the solution space more thoroughly.[3] The genetic algorithm is terminated after 100 iterations, resulting in a total calculation of 5 million individuals. After the completion of each one of the 10 repetitions, the exact global optimum solution Set_{k^*} is found (i.e *GlobalSolutionRatio* = 7/7), as well as the exact value of the scale factor $SF_{k^*} = 2.5$. For each of the 10 repetitions of the algorithm, approximately 30 min of computation time is required in order to compute all 100 iterations.[4]

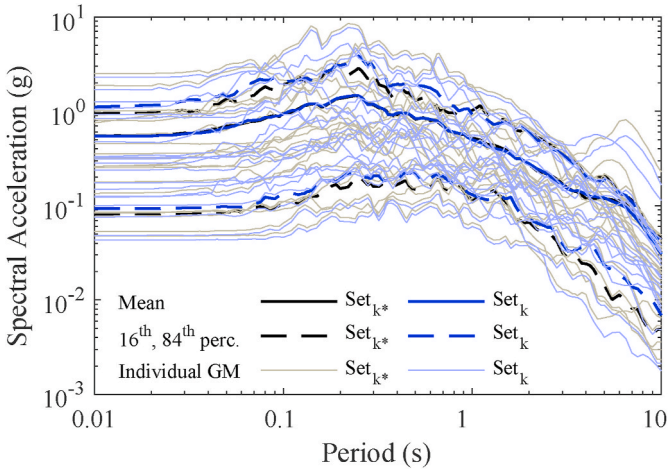

Fig. 5. The globally optimal solution Set_{k^*} with $n_{k^*} = 20$ GM records represented by the individual spectra (gray lines), their mean spectrum (black solid line) and 16th and 84th percentile spectra (black dashed lines), together with the results of the Algorithm 1 (corresponding blue lines) of the pareto-optimal solution Set_k (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

In Fig. 4 one run of the algorithm is demonstrated through the individual spectra of the records in Set_{k^*} (gray lines), their mean[5] spectrum (black solid line) and their 16th and 84th percentile spectra (black dashed lines). Note that this is the globally optimal solution and is obtained when the algorithm is stopped after 100 iterations. The early progress of the algorithm is also shown (by the corresponding blue lines), via the pareto-optimal solution Set_{k-25}, which is obtained after only 25 iterations. It is evident that a near-global optimal solution is reached after this small number of iterations, since *GlobalSolutionRatio* $= 3/7$ and the spectral match is of high quality. For a better understanding of the convergence process, note that after 25 iterations, all 7 records of Set_{k^*} are present in the different pareto-optimal solutions, and hence the globally optimal solution Set_{k^*} is returned within the following 75 iterations due to the effect of the crossover operation.

It is noteworthy that examining a total of 5×10^6 solutions during this optimization scheme is only a fraction of the total solution space, which is defined by the binomial coefficient $\binom{682}{7} \approx 1.3 \times 10^{16}$. In other words, if these 5×10^6 solutions were randomly selected to find

[3] A discussion on possible changes of the population size for the optimal performance of the algorithm is presented in Appendix A.

[4] on a personal computer with 3 GHz Intel Xeon W CPU processor.

[5] It refers to the geometric mean. Since the arithmetic mean of the natural logarithm of spectral accelerations is associated with the target in Equation (2), it is equivalent to show and compare the geometric mean of spectral accelerations in the non-logarithmic domain.

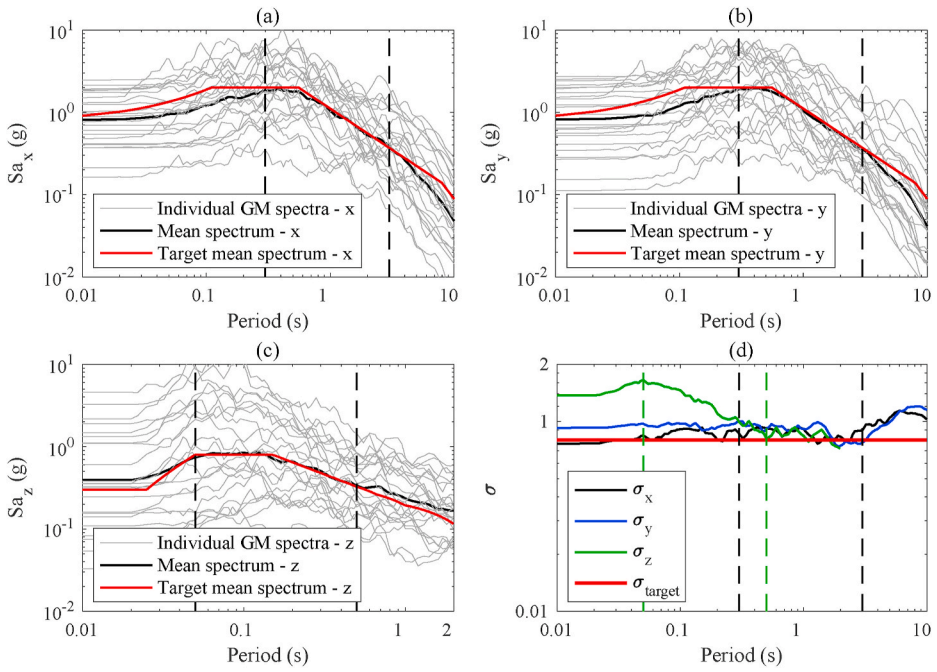

Fig. 6. Spectra of the individual GMs of the pareto-optimal solution Set_k (gray), their mean spectra (black) and the target mean spectra (red) in (a) x, (b) y, and (c) z directions; (d) target standard deviation spectra (red) of the natural logarithm of spectral accelerations and the corresponding spectra of the pareto-optimal solution Set_k in x, y and z directions. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

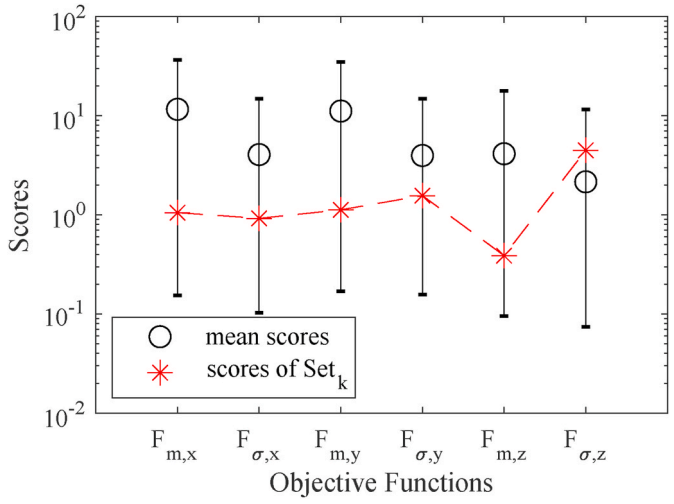

Fig. 7. Minimum, maximum and mean scores of the objective functions for all pareto-optimal solutions and those of Set_k.

the globally optimal solution, the probability of obtaining it would be approximately 3.8×10^{-10}. Consequently, the efficiency of the proposed algorithm appears to be impeccable and the consistency in its performance, regardless of the random nature of the benchmark process, demonstrates its robustness.

To aggravate the optimization problem of returning the globally optimum solution, $n_{k^*} = 20$ records is also tested here. Increasing the number of records in the GM set results in a smoother (i.e. less unique) shape in the target mean and standard deviation spectra, which means that multiple different solutions can converge to the globally optimum one. Moreover, the volume of the solution space increases significantly to $\binom{682}{20} \approx 1.5 \times 10^{38}$, i.e., approximately 10^{22} times larger than in the case of $n_{k^*} = 7$ records. The same procedure is performed without constraining the iterations, resulting in pareto-optimal solutions that show good agreement with the target spectra and *GlobalSolutionRatio* = 10/20, here presented by one solution Set_k in Fig. 5, similar to Fig. 4. Even though the exact globally optimal solution Set_{k^*} is not returned in this case, it is considered that a near-global one is reached since the spectral match is of high quality and a high percentage is achieved in *GlobalSolutionRatio*. Details of the GM record sets appearing in Figs. 4 and 5 are provided in Appendix A, in Table A.3 and A.2, respectively.

3.2. Extracting GM sets that meet spectral multi-objective criteria

3.2.1. Meeting targets in first and second order statistics of the response spectra

The application presented in the predecessor of the proposed procedure [41] is revisited here, extracting a three- instead of one-component GM set for Century City, California. In that work [41], the design spectrum according to ASCE (2010) was employed as the target mean spectrum while the target standard deviation spectrum of the natural logarithm of spectral accelerations was constant with $\sigma = 0.8$ and the maximum value of scale factors was 5, similarly to other research endeavors and guidelines on the topic.

Here, the same targets are utilized for both horizontal directions, a rather difficult multi-component target objective since the two horizontal components of each record are usually very similar in spectral shape but with some small differences in spectral amplitude. For this reason, similar to Ref. [31], a differentiation of 5% between the scale factors of the components of each record is allowed, while the same scale factor bound of 5 is preserved. To adopt a compatible choice for the 5% damped vertical design acceleration spectrum, the suggestion of the U.S. National Earthquake Hazards Reduction Program (NEHRP) as proposed in Ref. [51] is followed, which utilizes a modified version of the methodology developed by Campbell and Bozorgnia [52]. Details about this methodology and the exact calculations leading to the adopted vertical design acceleration spectrum can be found in Ref. [36]. The NGA-West2 database is utilized [50], with the following earthquake scenario and site parameters as screening criteria: magnitudes with $M_w^{(range)} \in [5.5, 7.9]$, source-to-site distances in *km* with $R_{JB}^{(range)} \in [0, 100]$ and shear

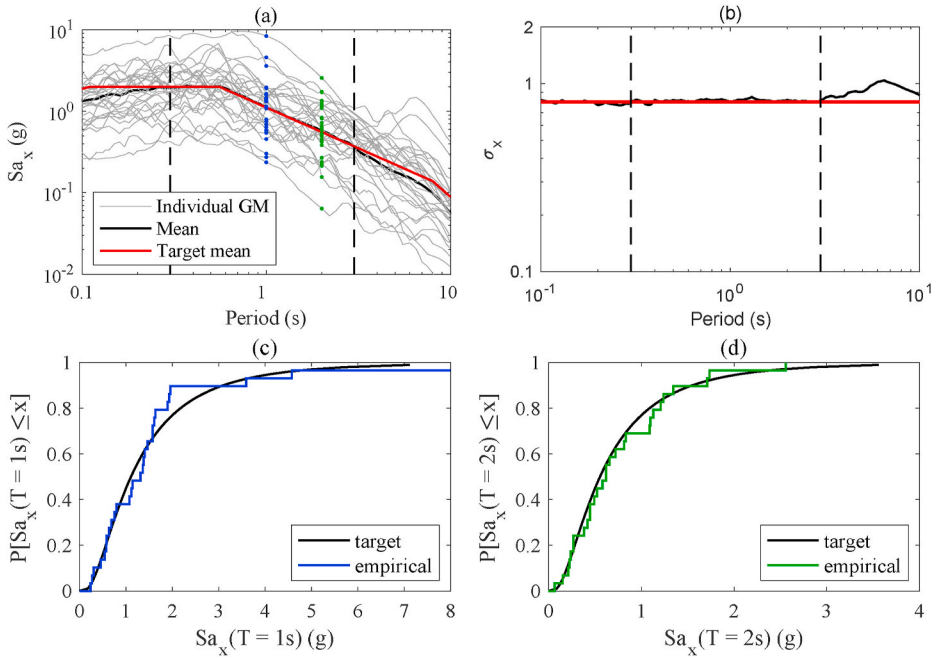

Fig. 8. Results for one-component GM selection when F_{LHS} is not applied, represented by (a) spectra of individual GMs, their mean and target mean; (b) their standard deviation spectra and the target; empirical and target cumulative distribution of spectral accelerations at (c) $T = 1s$; and (d).$T = 2s$

wave velocity in m/s with $V_{s,30}^{(range)} \in [180,\ 360]$, resulting in 682 three-component GM records. A wide range of periods in the spectrum is selected from 0.3s to 3s for both horizontal components and 0.05s–0.5s for the vertical component, reflecting realistic ranges of interest for multi-story building frame structures. Finally, it is chosen that a range of 20–30 GM should be included in the resulting sets.

The results of the optimization process are presented by a pareto-optimal solution Set_k, where in Fig. 6 (a), (b) and (c) show respectively the x, y, and z components of the individual GM spectra (gray lines), their mean spectrum (black lines) and the mean target spectrum (red lines). In Fig. 6 (d), the standard deviation spectra of the natural logarithm of the spectral accelerations for x (black), y (blue), and z (green) components are shown along with their common target (red). The period ranges of interest for the horizontal and vertical components are enclosed by corresponding vertical dashed lines in Fig. 6 (a)–(d). Overall, except for the standard deviation of the vertical component ($F_{\sigma,z}$), the quality of the match in the desired period ranges appears to be excellent. This is also portrayed in Fig. 7, where the scores of the six objective functions are shown for Set_k along with the minimum, maximum and the mean of scores for all pareto-optimal solutions. Compared to the scores of the other objective functions of Set_k, $F_{\sigma,z}$ appears worse.

The GM set Set_k is the pareto-optimal solution obtained by the post-processing metric $R_{spectral}$, presented in section 2.3 when Equation (8) is calculated for $w_{m,i} = w_{\sigma,i} = 1$, $i = x, y, z$. This choice of weights seems reasonable, since it assigns equal importance to the six objective functions when they are summed. However, adding absolute values of the first and second order statistics of the spectral accelerations, from the three orthogonal directions, is rather arbitrary and has no physical meaning. The resulting spectra of a different GM set obtained for a different selection of weights are presented in Appendix A, through Figs. A.10 and A.11 similar to Figs. 6 and 7. This example demonstrates that a given set of weights can provide a GM set that minimizes some objective functions to a great extent, but is insufficient overall. It supports the superiority of multi-objective selection, since it is not hindered by this choice prior to the optimization process, and therefore multiple alternatives of weight factors can be explored posteriorly, without any computational cost.

3.2.2. *Meeting distribution function targets at periods of the response spectra*

In addition to good fitting of the target mean and standard deviation spectra, the ability to match probability distribution targets is also investigated, through the effect of spectral accelerations falling into appropriate bins, as explained in section 2.1.2. To this end, the application presented in section 3.2.1 is revisited, this time performing one-component (instead of three) horizontal GM selection. The targets, earthquake scenario and site parameters, as well as the other options (e.g., the number of records in the GM sets being in the range of 20–30) remain the same here. Considering the lognormal distribution as the target, the effect of the additional objective F_{LHS} from Equation (4) is examined using the resulting cumulative distributions of spectral accelerations at two specific periods in the spectrum range, at $T = 1s$ and $T = 2s$. The GM selection process is repeated three times, assigning F_{LHS} in different ways explained hereafter. After the optimization is complete, the distribution-matching post-processing metric R_{dist} presented in section 2.3 is employed to rank the pareto-optimal solutions. This is applied consistently in all three applications to avoid bias[6] and the best performing GM set from each application is compared, denoted Set_{k1}, Set_{k2} and Set_{k3}, respectively.

The first application, represented in Fig. 8 by Set_{k1}, has no distribution function requirements, i.e., the objective function F_{LHS} is not applied. Essentially, this is a one-component simplification of the application in section 3.2.1. Fig. 8 (a) shows the individual GM spectra of Set_{k1}, their mean and the target mean, along with the spectral accelerations of interest at $T = 1s$ and $T = 2s$ highlighted with blue and green markers, respectively. Fig. 8 (b) shows the standard deviation spectra with their target. Fig. 8 (c) and (d) show the empirical cumulative distributions of the spectral accelerations at $T = 1s$ and $T = 2s$, respectively, along with the corresponding lognormal distribution function with the same mean and standard deviation. In the second example, F_{LHS} is applied for $T = 1s$ and in the third example, F_{LHS} is applied simultaneously at $T = 1s$ and $T = 2s$. The results are depicted in

[6] R_{dist} is applied for both $T = 1s$ and $T = 2s$ simultaneously in all three applications.

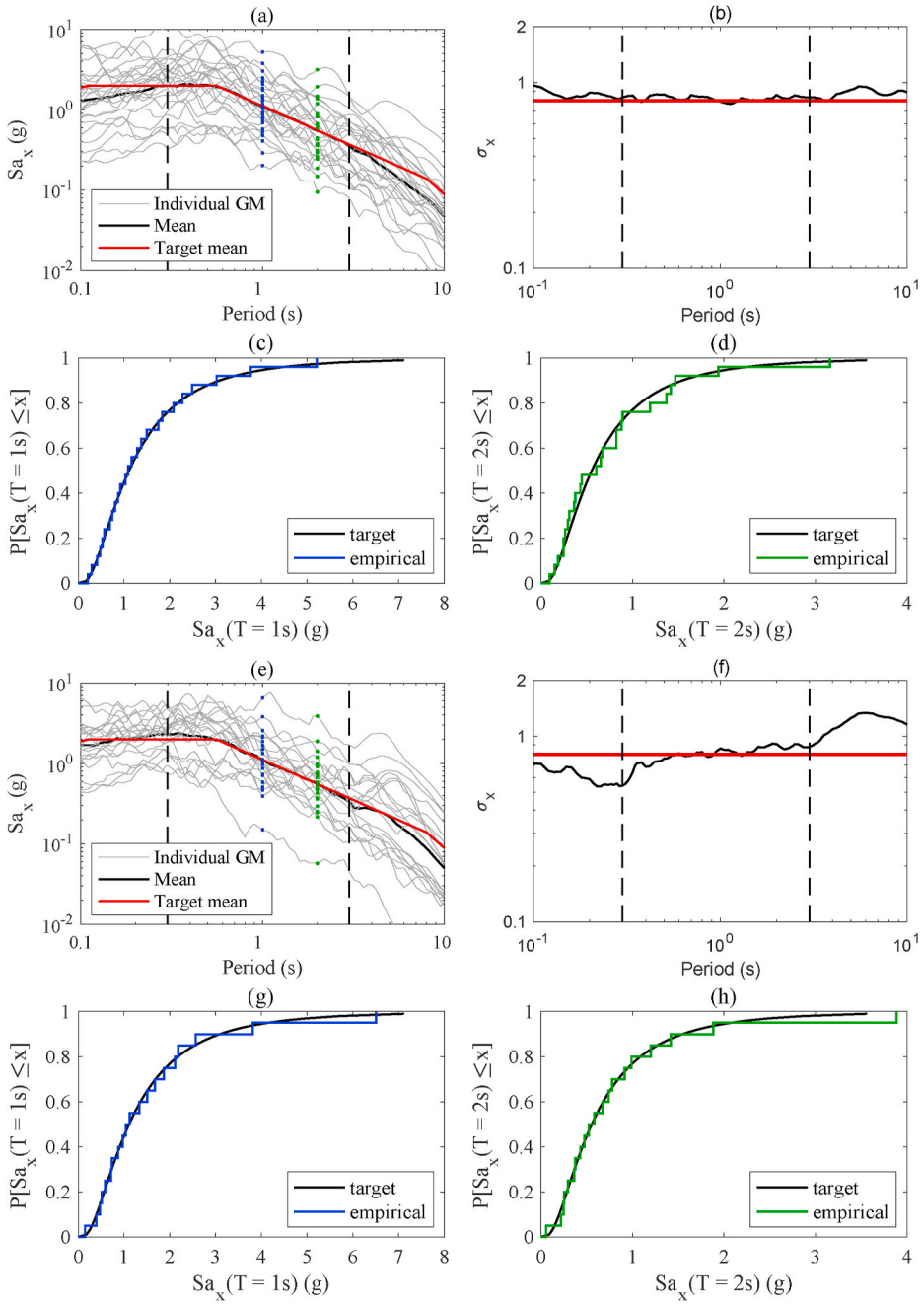

Fig. 9. Results for one-component GM selection, similar to Fig. 8, when F_{LHS} is applied on $T = 1s$ (a)–(d); and when F_{LHS} is applied on both $T = 1s$ and $T = 2s$ (e)–(h).

Fig. 9 (a) - (d) (Set_{k2}) and Fig. 9 (e)–(h) (Set_{k3}), similarly to Fig. 8 (a)–(d). There are 25 records in Set_{k1} as well as Set_{k2} and 20 records in Set_{k3}.

The quality in matching the mean and standard deviation targets of Set_{k1}, Set_{k2} and Set_{k3} (Fig. 8 (a), (b) compared to the results of Fig. 9 (a), (b), (e), (f)) leads to the conclusion that the additional requirement of matching distribution functions does not obstruct the main objective of meeting first and second order statistics in a range of periods. Even though the quality of the match of Set_{k1} is slightly better compared to Set_{k2} and Set_{k3}, it seems that all of them meet these targets almost perfectly.

The improved match between empirical and target cumulative distributions based on F_{LHS} becomes apparent when comparing Fig. 9 (c), (d), (g), (h) with Fig. 8 (c) and (d). Moreover, it is interesting to compare Set_{k2} with Set_{k3}, i.e., in the former case F_{LHS} is applied at $T = 1s$ and in the latter case F_{LHS} is applied at $T = 1s$ and $T = 2s$. It is observed that the match of the distribution at $T = 2s$ is better for Set_{k3}, without a compromise on the quality of the distribution at $T = 1s$. Therefore, the application of F_{LHS} in more than one periods appears to be advantageous. To quantitatively assess the effect of F_{LHS}, it is found that the number of empty bins N_{Empty} in $T = 1s$ and $T = 2s$ cumulatively is 12 out of $2 \times 25 = 50$ for Set_{k1} (having 25 records) and only 1 out of $2 \times 20 = 40$ for Set_{k3} (having 20 records), showing the large influence it can have. It is well known that an increased number of records in a GM set causes the probability distribution of the spectral accelerations to resemble the lognormal [4] and this is arguably the case in all three applications here. However, these results show that although there are fewer records in Set_{k3} compared to Set_{k1}, their distribution fits the lognormal distribution better. This is a promising effect of stratified sampling, which is approximated here by an LHS equivalent scheme, resulting in a "favorable" distribution of spectral accelerations. It appears that it can

be used to reduce the number of records in the GM sets while preserving the same quality of the distributions of spectral accelerations. Fewer records in a GM set means fewer response history computations and is generally desired.

4. Summary and conclusions

A robust methodology for selecting and scaling GM records has been proposed that is versatile for meeting targets in first and second order statistics of the acceleration response spectrum in a wide range of periods, as well as meeting distribution function targets of spectral accelerations at specific ordinates of the spectrum. The latter is a unique contribution achieved through a novel strategy in GM selection that mimics the LHS scheme to yield GM sets that not only fit the desired distribution function targets, but also very efficiently yield better results on a reduced number of records. Efficiency is also studied here in terms of the ability of the proposed optimization scheme to provide optimal results. To this end, a significant contribution in this work is a benchmarking algorithm introduced to assess GM selection optimization schemes. Through this benchmarking process, it is shown that the proposed methodology is able to deliver results consisting of the global (or near global) optimum solutions, which means that no other solution can be found that matches better at least one target without reducing the quality of the others. The current methodology utilizes evolutionary GA to satisfy the above targets by performing multi-objective optimization, which is claimed here to provide robust results. Moreover, the above selection capabilities are applicable in up to three-component GM, making the proposed approach ideal for numerous GM selection applications. The methodology has been scrutinously implemented in Matlab and is available from the authors. When developing the software, emphasis was placed on guiding the user through the selection phases, ensuring ease of use and allowing complete flexibility in calibrating all of the parameters involved in the selection process.

CRediT authorship contribution statement

Konstantinos T. Tsalouchidis: Conceptualization, Methodology, Investigation, Software, Visualization, Validation, Formal analysis, Data curation, Writing – original draft. **Lukas Moschen:** Conceptualization, Methodology, Supervision. **Christoph Adam:** Conceptualization, Methodology, Supervision, Funding acquisition, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The funding from the Austrian Science Fund (FWF) for the research project No. I 3414-N30 is gratefully acknowledged.

Appendix A. Supplementary data and results for the applications

Appendix A.1. Extracting the globally optimal solution

The GM records of Set_{k^*} and Set_k that appear in Figure 5 are presented here in Table A.2 and the ones of Set_{k^*} and Set_{k-25} that appear in Figure 4 are presented in Table A.3. The common records in the sets appear at the top and are highlighted in bold, and for each record the earthquake name and the unique sequence number in the NGA-West2 database [50] is provided. It is noteworthy that both in Table A.2 and A.3, in addition to the common GM records in the sets, the non-common ones frequently come from the same earthquake event, share common features, and thus effectively substitute each other.

Note that in this application where the goal is to extract the globally optimal solutions, the population size is increased to 50000 (from the default value being 1000 in the proposed method). This ensures that the solution space is explored thoroughly, while keeping the computation time relatively low. Normally, the potential user of the method would not need to adjust this value, unless a similar need to thoroughly explore the solution space is considered necessary. For example, if the solution space is significantly larger than the ones in the applications presented, then this could be a good practise to ensure better quality in the expense of speed in computations. Note that the solution spaces presented here are already large, representing realistic applications where a couple hundreds of records are available and a couple tens of records are required in the GM sets.

Table A.2
GM records of Set_{k^*} and Set_k appearing in Figure 5 with the earthquake names and unique sequence numbers of the NGA-West2 database [50].

	Set_{k^*}		Set_k	
no.	earthquake name	NGA no.	earthquake name	NGA no.
1	**Chi-Chi, Taiwan-04**	**2729**	**Chi-Chi, Taiwan-04**	**2729**
2	**Chi-Chi, Taiwan-04**	**2733**	**Chi-Chi, Taiwan-04**	**2733**
3	**Chi-Chi, Taiwan-03**	**2592**	**Chi-Chi, Taiwan-03**	**2592**
4	**Chuetsu-oki**	**4895**	**Chuetsu-oki**	**4895**
5	**Duzce, Turkey**	**1602**	**Duzce, Turkey**	**1602**
6	**Landers**	**862**	**Landers**	**862**
7	**Chi-Chi, Taiwan**	**1598**	**Chi-Chi, Taiwan**	**1598**
8	**Chi-Chi, Taiwan**	**1264**	**Chi-Chi, Taiwan**	**1264**
9	**Imperial Valley-06**	**174**	**Imperial Valley-06**	**174**
10	**Iwate**	**5781**	**Iwate**	**5781**
11	Chi-Chi, Taiwan	1539	Chi-Chi, Taiwan	1224
12	Chi-Chi, Taiwan-03	2651	Chi-Chi, Taiwan	1180
13	Darfield, New Zealand	6975	Darfield, New Zealand	6927
14	Iwate	5782	Iwate	5812
15	Tottori, Japan	3963	Tottori, Japan	3937
16	Chi-Chi, Taiwan-05	3215	Imperial Valley-06	175
17	Chi-Chi, Taiwan-03	2648	Chi-Chi, Taiwan-03	2501

(*continued on next page*)

Table A.2 (*continued*)

	Set_{k^*}		Set_k	
no.	earthquake name	NGA no.	earthquake name	NGA no.
18	Kocaeli, Turkey	1155	El Mayor-Cucapah	5985
19	Whittier Narrows-01	638	Chuetsu-oki	4997
20	Chi-Chi (aftershock 4), Taiwan	3860	Chi-Chi, Taiwan-06	3494

Appendix A.2. Meeting targets in first and second order statistics

An additional pareto-optimal solution to the one presented as a result of the application in 3.2.1 is shown here to demonstrate that the selection of weights is not straightforward and should be repeated to ensure that the final choice from the pareto-optimal solutions satisfies the requirements of the problem in the best way. In Figure 6, the solution Set_k is obtained when all weights are assigned the same value $w_{m,i} = w_{\sigma,i} = 1$, $i = x,y,z$, and the scores of the objective functions are shown in Figure 7. Here, in Figure A.10 and A.11, the pareto-optimal solution is obtained when more importance is given to the standard deviation spectra with $w_{m,i} = 0.05$ and $w_{\sigma,i} = 1, i = x,y,z$. It can be seen that the resulting solution strongly favors the standard deviation match in all components and does not provide acceptable results for the mean targets.

Table A.3
GM records of Set_{k^*} and Set_{k-25} appearing in Figure 4 with the earthquake names and unique sequence numbers of the NGA-West2 database [50].

	Set_{k^*}		Set_{k-25}	
no.	earthquake name	NGA no.	earthquake name	NGA no.
1	**Chi-Chi, Taiwan-04**	**2732**	**Chi-Chi, Taiwan-04**	**2732**
2	**Chi-Chi, Taiwan-05**	**2945**	**Chi-Chi, Taiwan-05**	**2945**
3	**Loma Prieta**	**778**	**Loma Prieta**	**778**
4	Chi-Chi, Taiwan	1320	Chi-Chi, Taiwan	1598
5	Kobe, Japan	1118	Niigata, Japan	4207
6	El Mayor-Cucapah	5832	Morgan Hill	457
7	Chi-Chi, Taiwan-05	2939	Taiwan SMART1(45)	3679

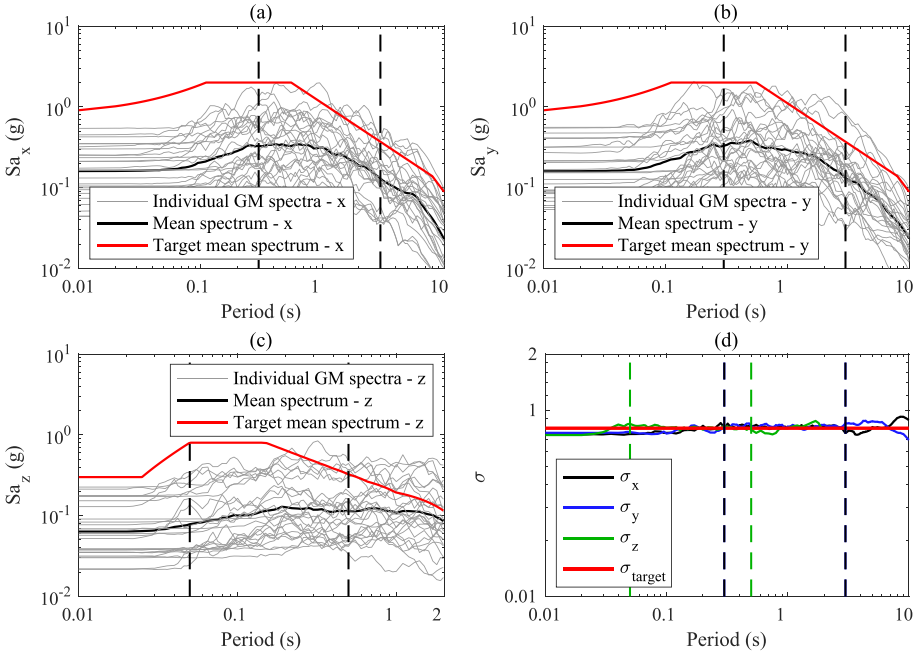

Figure A.10. Spectra of the individual GMs of the pareto-optimal solution Set_k (grey), their mean spectra (black) and the target mean spectra (red) in (a) x, (b) y, and (c) z directions; (d) target standard deviation spectra (red) of the natural logarithm of spectral accelerations and the corresponding spectra of the pareto-optimal solution Set_k in x, y and z directions.

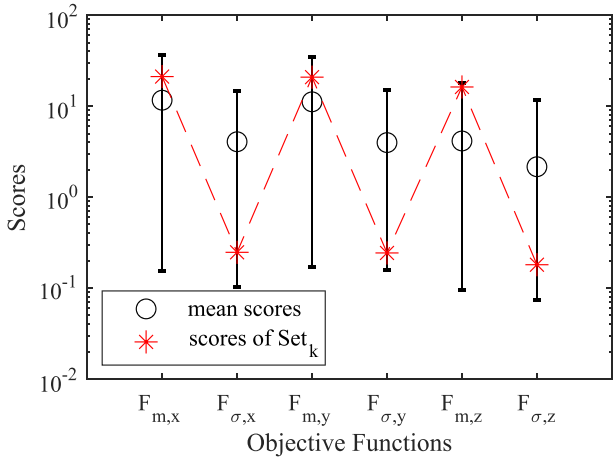

Figure A.11. Scores of the objective functions for all pareto-optimal solutions, the mean scores of all solutions and those of Set_k.

References

[1] Sextos AG. Selection of ground motions for response history analysis. In: Encyclopedia of earthquake engineering. Berlin, Heidelberg: Springer; 2014. p. 1–10.

[2] Haselton C, Whittaker A, Hortacsu A, Baker J, Bray J, Grant D. Selecting and scaling earthquake ground motions for performing response-history analyses. In: Proceedings of the 15th world conference on earthquake engineering, earthquake engineering research institute; 2012. p. 4207–17.

[3] Giaralis A, Spanos P. Wavelet-based response spectrum compatible synthesis of accelerograms—eurocode application (EC8). Soil Dynam Earthq Eng 2009;29(1):219–35.

[4] Shome N, Cornell CA, Bazzurro P, Carballo JE. Earthquakes, records, and nonlinear responses. Earthq Spectra 1998;14(3):469–500.

[5] Stewart JP, Chiou S-J, Bray JD, Graves RW, Somerville PG, Abrahamson NA. Ground motion evaluation procedures for performance-based design. Soil Dynam Earthq Eng 2002;22(9):765–72.

[6] Baker JW, Allin Cornell C. Spectral shape, epsilon and record selection. Earthq Eng Struct Dynam 2006;35(9):1077–95.

[7] Baker JW. Conditional mean spectrum: tool for ground-motion selection. J Struct Eng 2011;137(3):322–31.

[8] Bradley BA. A ground motion selection algorithm based on the generalized conditional intensity measure approach. Soil Dynam Earthq Eng 2012;40:48–61.

[9] Kottke A, Rathje EM. A semi-automated procedure for selecting and scaling recorded earthquake motions for dynamic analysis. Earthq Spectra 2008;24(4):911–32.

[10] Jayaram N, Lin T, Baker JW. A computationally efficient ground-motion selection algorithm for matching a target response spectrum mean and variance. Earthq Spectra 2011;27(3):797–815.

[11] Baker JW, Lee C. An improved algorithm for selecting ground motions to match a conditional spectrum. J Earthq Eng 2018;22(4):708–23.

[12] Baker JW, Cornell CA. Which spectral acceleration are you using? Earthq Spectra 2006;22(2):293–312.

[13] Lin T, Harmsen S, Baker J, Luco N. Conditional spectrum computation incorporating multiple causal earthquakes and ground-motion prediction models. Bull Seismol Soc Am 2013;103:1103–16.

[14] Kwong NS, Chopra AK. Evaluation of the exact conditional spectrum and generalized conditional intensity measure methods for ground motion selection. Earthq Eng Struct Dynam 2016;45(5):757–77.

[15] C. B. Haselton, J. W. Baker, Y. Bozorgnia, C. Goulet, E. Kalkan, N. Luco, T. Shantz, N. Shome, J. Stewart, P. Tothong, et al, Evaluation of ground motion selection and modification methods: predicting median interstory drift response of buildings, PEER report 1.

[16] Katsanos EI, Sextos AG, Manolis GD. Selection of earthquake ground motion records: a state-of-the-art review from a structural engineering perspective. Soil Dynam Earthq Eng 2010;30(4):157–69.

[17] Macedo L, Castro J. SelEQ: an advanced ground motion record selection and scaling framework. Adv Eng Software 2017;114:32–47.

[18] Naeim F, Alimoradi A, Pezeshk S. Selection and scaling of ground motion time histories for structural design using genetic algorithms. Earthq Spectra 2004;20(2):413–26.

[19] Katsanos EI, Sextos AG. ISSARS: an integrated software environment for structure-specific earthquake ground motion selection. Adv Eng Software 2013;58:70–85.

[20] Katsanos E, Sextos A. Structure-specific selection of earthquake ground motions for the reliable design and assessment of structures. Bull Earthq Eng 2018;16(2):583–611.

[21] Klinc R, Šebenik Ž, Dolšek M, Brozovič M, Dolenc M. A web-based system for the selection of characteristic ground motions. Adv Eng Software 2019;135:102688.

[22] Iervolino I, Galasso C, Cosenza E. REXEL: computer aided record selection for code-based seismic structural analysis. Bull Earthq Eng 2009;8:339–62.

[23] Luco N, Cornell CA. Structure-specific scalar intensity measures for near-source and ordinary earthquake ground motions. Earthq Spectra 2007;23(2):357–92.

[24] Kohrangi M, Bazzurro P, Vamvatsikos D, Spillatura A. Conditional spectrum-based ground motion record selection using average spectral acceleration. Earthq Eng Struct Dynam 2017;46(10):1667–85.

[25] De Biasio M, Grange S, Dufour F, Allain F, Petre-Lazar I. Intensity measures for probabilistic assessment of non-structural components acceleration demand. Earthq Eng Struct Dynam 2015;44(13):2261–80.

[26] Kohrangi M, Vamvatsikos D, Bazzurro P. Site dependence and record selection schemes for building fragility and regional loss assessment. Earthq Eng Struct Dynam 2017;46(10):1625–43.

[27] Adam C, Kampenhuber D, Ibarra LF, Tsantaki S. Optimal spectral acceleration-based intensity measure for seismic collapse assessment of p-delta vulnerable frame structures. J Earthq Eng 2017;21(7):1189–95.

[28] Adam C, Kampenhuber D, Ibarra LF. Optimal intensity measure based on spectral acceleration for P-delta vulnerable deteriorating frame structures in the collapse limit state. Bull Earthq Eng 2017;15(10):4349–73.

[29] Tsantaki S, Adam C, Ibarra LF. Intensity measures that reduce collapse capacity dispersion of p-delta vulnerable simple systems. Bull Earthq Eng 2017;15(3):1085–109.

[30] Eads L, Miranda E, Lignos DG. Average spectral acceleration as an intensity measure for collapse risk assessment. Earthq Eng Struct Dynam 2015;44(12):2057–73.

[31] Kohrangi M, Bazzurro P, Vamvatsikos D. Conditional spectrum bidirectional record selection for risk assessment of 3D structures using scalar and vector IMs. Earthq Eng Struct Dynam 2019;48(9):1066–82.

[32] Kohrangi M, Vamvatsikos D, Bazzurro P. Implications of intensity measure selection for seismic loss assessment of 3-D buildings. Earthq Spectra 2016;32(4):2167–89.

[33] Beyer K, Bommer JJ. Selection and scaling of real accelerograms for Bi-directional loading: a review of current practice and code provisions. J Earthq Eng 2007;11(sup1):13–45.

[34] Liang X, Mosalam KM. Ground motion selection and modification evaluation for highway bridges subjected to bi-directional horizontal excitation. Soil Dynam Earthq Eng 2020;130:105994.

[35] Sextos AG, Katsanos EI, Manolis GD. EC8-based earthquake record selection procedure evaluation: Validation study based on observed damage of an irregular R/C building. Soil Dynam Earthq Eng 2011;31(4):583–97.

[36] Moschen L, Medina RA, Adam C. Vertical acceleration demands on column lines of steel moment-resisting frames. Earthq Eng Struct Dynam 2016;45(12):2039–60.

[37] Gremer N, Adam C, Medina RA, Moschen L. Vertical peak floor accelerations of elastic moment-resisting steel frames. Bull Earthq Eng 2019;17(6):3233–54.

[38] Kwong NS, Chopra AK. Selecting, scaling, and orienting three components of ground motions for intensity-based assessments at far-field sites. Earthq Spectra 2020;36(3):1013–37.

[39] Ha SJ, Han SW. An efficient method for selecting and scaling ground motions matching target response spectrum mean and variance. Earthq Eng Struct Dynam 2016;45(8):1381–7.

[40] Ha SJ, Han SW. A method for selecting ground motions that considers target response spectrum mean and variance as well as correlation structure. J Earthq Eng 2016;20(8):1263–77.

[41] Moschen L, Medina RA, Adam C. A ground motion record selection approach based on multiobjective optimization. J Earthq Eng 2019;23(4):669–87.

[42] Mergos PE, Sextos AG. Selection of earthquake ground motions for multiple objectives using genetic algorithms. Eng Struct 2019;187:414–27.
[43] Georgioudakis M, Fragiadakis M. Selection and scaling of ground motions using multicriteria optimization. J Struct Eng 2020;146(11):04020241.
[44] Deb K. Multi-objective optimization using evolutionary algorithms, vol. 16; 2001. John Wiley & Sons.
[45] Tarbali K, Bradley BA. The effect of causal parameter bounds in PSHA-based ground motion selection. Earthq Eng Struct Dynam 2016;45(9):1515–35.
[46] Luco N, Bazzurro P. Does amplitude scaling of ground motion records result in biased nonlinear structural drift responses? Earthq Eng Struct Dynam 2007;36(13):1813–35.
[47] Bommer JJ, Acevedo AB. The use of real earthquake accelerograms as input to dynamic analysis. J Earthq Eng 2004;8(spec01):43–91.
[48] Miller BL, Goldberg DE, et al. Genetic algorithms, tournament selection, and the effects of noise. Complex Syst 1995;9(3):193–212.
[49] Massey Jr FJ. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc 1951;46(253):68–78.
[50] Ancheta TD, Darragh RB, Stewart JP, Seyhan E, Silva WJ, Chiou BS-J, Wooddell KE, Graves RW, Kottke AR, Boore DM, et al. NGA-West2 database. Earthq Spectra 2014;30(3):989–1005.
[51] Building Seismic Safety Council, NEHRP recommended seismic provisions for new buildings and other structures (FEMA P-750), Washington, DC: Federal Emergency Management Agency.
[52] Campbell KW, Bozorgnia Y. Updated near-source ground-motion (attenuation) relations for the horizontal and vertical components of peak ground acceleration and acceleration response spectra. Bull Seismol Soc Am 2003;93(1):314–31.