

EFFICIENT COMPUTATION OF ACCURATE SEISMIC FRAGILITY FUNCTIONS  
THROUGH STRATEGIC STATISTICAL SELECTION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Francisco Peña

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2019

Purdue University

West Lafayette, Indiana

**THE PURDUE UNIVERSITY GRADUATE SCHOOL**  
**STATEMENT OF APPROVAL**

Dr. Shirley Dyke, Co-Chair

Lyles School of Civil Engineering

Dr. Ilias Billionis, Co-Chair

School of Mechanical Engineering

Dr. Ayhan Irfanoglu

Lyles School of Civil Engineering

Dr. Julio Ramirez

Lyles School of Civil Engineering

**Approved by:**

Dr. Dulcy Abraham

Head of the Burke Graduate Program

*To my wife, Natalia, who has been an unconditional source of support and encouragement.*

*I am truly thankful for all her love, patience, and support.*

*To the memory of my father-in-law, William, who helped shape me into the person I am  
today.*

## ACKNOWLEDGMENTS

My sincere gratitude and appreciation goes to my advisors, Dr. Shirley Dyke and Dr. Ilias Bilonis. Their continuous support, guidance, and encouragement made of this Ph.D. an enriching and rewarding journey. I am grateful for all they did for my entire family within these years.

I also want to thank the rest of the committee members for this dissertation, Dr. Ayhan Irfanoglu and Dr. Julio Ramirez, for their valuable input to this dissertation, advice, and challenging questions that helped shape this research.

Besides the committee members, I would like to thank Dr. George Mavroeidis and Yanan Cao, from the University of Notre Dame, for providing the earthquake database implemented for this analysis. I want to acknowledge my fellow graduate students, current and former members of the Intelligent Infrastructure Systems Laboratory (IISL) and the Predictive Science Laboratory, for sharing your knowledge and insights, your company made my time at Purdue really enjoyable and pleasant. I sincerely appreciate the trust given by professors Robert Jacko, Donald Meinheit, Roger Reckers, and Ayhan Irfanoglu, with the teaching assistant positions; and all the guidance and collaboration from Jenny Ricksy, Molly Stetler, and Bailey Ritchie.

I would like to thank my family for all their love and support, my children, Sebastian and Emily, for all the happiness and joy that you bring to my life. Most importantly, special gratitude is owed to my wife, Natalia, for her undoubted faith in me, unconditional support and love. There are no enough words to express how thankful I am to her. The culmination of this doctoral degree is entirely because of her sacrifices and continuous encouragement.

The financial support of this research is provided by the Colombian government, through the scholarship 568-2012 from Colciencias (Departamento Administrativo de Ciencia, Tecnología e Innovación), the Colombia-Purdue Institute for Advanced Scientific Research, and the teaching assistantship from the Lyles school of Civil Engineering at Purdue University.

## TABLE OF CONTENTS

	Page
LIST OF TABLES . . . . .	viii
LIST OF FIGURES . . . . .	ix
ABBREVIATIONS . . . . .	xi
ABSTRACT . . . . .	xiii
<b>1 INTRODUCTION . . . . .</b>	<b>1</b>
1.1 Literature Review . . . . .	5
1.1.1 <i>History</i> . . . . .	6
1.1.2 <i>Classification of FFs</i> . . . . .	9
1.1.3 <i>Uncertainty in FFs</i> . . . . .	11
1.1.4 <i>Computation of FFs</i> . . . . .	19
1.2 Remaining Challenges . . . . .	25
1.3 Objective and Scope . . . . .	25
1.4 Overview . . . . .	28
<b>2 AVAILABLE DATA . . . . .</b>	<b>30</b>
2.1 Case Study Structure . . . . .	30
2.2 Generating synthetic broadband ground motions . . . . .	31
<b>3 INPUT-OUTPUT PARAMETERS . . . . .</b>	<b>35</b>
3.1 Problem definition . . . . .	35
3.2 Structural response parameter: Engineering demand parameter (EDP) . . . . .	36
3.3 Structural response threshold: Limit States - LS . . . . .	37
3.4 Input parameter: Intensity Measure - IM . . . . .	39
<b>4 BAYESIAN MODEL SELECTION . . . . .</b>	<b>48</b>
4.1 Monte Carlo approximation of the FF . . . . .	48
4.2 Learning the FF from limited data . . . . .	49
4.2.1 <i>Conventional model of FF</i> . . . . .	49
4.2.2 <i>Prior knowledge about the model</i> . . . . .	50
4.2.3 <i>Models of FFs</i> . . . . .	51
4.2.4 <i>Bayesian inference of model parameters</i> . . . . .	52
4.2.5 <i>Predicting fragility under a given model</i> . . . . .	54
4.3 <i>Bayesian model selection</i> . . . . .	55
4.3.1 <i>Model Validation</i> . . . . .	56
4.3.2 Epistemic uncertainty of FF . . . . .	58

	Page
4.4 Numerical Results . . . . .	59
4.4.1 Selecting the best model . . . . .	59
4.4.2 Quantification of the epistemic uncertainty in the FFs . . . . .	67
4.5 Conclusions . . . . .	69
5 GROUND MOTION SELECTION TECHNIQUE . . . . .	71
5.1 Effect of a hypothetical observation on the reduction in epistemic uncertainty	72
5.2 Sequential selection based on single-dimension IM . . . . .	73
5.3 Sequential selection based on two-dimensional IM . . . . .	73
5.4 Process to implement the sequential selection strategy . . . . .	75
5.5 Numerical Results . . . . .	77
5.5.1 Sequential selection strategies . . . . .	77
5.5.2 Statistical model for secondary IM . . . . .	78
5.5.3 Results of selected selection strategies . . . . .	80
5.6 Conclusions . . . . .	85
6 PRACTICES THAT UNINTENTIONALLY BIAS FRAGILITY FUNCTIONS . . . . .	87
6.1 Effect of using different ground motion databases . . . . .	87
6.2 Effect of including historical ground motions . . . . .	90
6.3 Effect of scaling ground motions . . . . .	92
6.4 Conclusions . . . . .	93
7 SUMMARY AND CONCLUSIONS . . . . .	96
REFERENCES . . . . .	99
A DERIVATION OF FRAGILITY FUNCTION UNDER LOGNORMAL DISTRIBUTION ASSUMPTION . . . . .	108
B SELECTION OF INPUT PARAMETERS . . . . .	110
C EARTHQUAKE DATABASE . . . . .	115
D UPDATED PARTICLE APPROXIMATION . . . . .	118
E MODEL SELECTION RESULTS . . . . .	120
F SEQUENTIAL SELECTION OF GROUND MOTIONS RESULTS . . . . .	122
VITA . . . . .	124

## LIST OF TABLES

Table	Page
1.1 Classification of fragility functions . . . . .	10
1.2 Examples of uncertainty on modeling parameters . . . . .	12
3.1 Commonly used critical values of building response per limit state [98–101] . .	39
4.1 Average (COV) log-value for the evidence, $\log(Z_M)$ , for different number of basis functions . . . . .	62
4.2 Comparison metrics of the models . . . . .	63
5.1 Comparison metrics of the models . . . . .	80
E.1 Model selection results for $N = 500$ , $\mathcal{X}_1 : PGV$ , and $\mathcal{Y}$ . . . . .	121
F.1 Model selection results for $N = 1000$ , $\mathcal{X}_1 : PGV$ , and $\mathcal{X}_2 : S_v$ . . . . .	123

## LIST OF FIGURES

Figure	Page
1.1 <i>Representation of fragility function</i> . . . . .	6
1.2 <i>Overview of the fragility computation challenge</i> . . . . .	26
2.1 <i>Case study structure: 20-story nonlinear benchmark building</i> . . . . .	32
2.2 <i>Earthquake fault (rectangle), grid of stations (triangles), and hypocenter locations (stars) considered in this study.</i> . . . . .	33
2.3 <i>Example of synthetic acceleration and velocity time histories for the station denoted by the filled triangle in Fig. 2.2 and for <math>M_w = 6.5</math>, a hypocenter on the edge of the fault, and NEHRP Site Class C.</i> . . . . .	33
3.1 <i>Distribution of drift parameter <math>\mathcal{Y}</math></i> . . . . .	38
3.2 <i>Initial dataset of <math>\mathbf{X}</math> vs. <math>\mathcal{Y}</math></i> . . . . .	43
3.3 <i>Scatter matrix plot for <math>\mathbf{X}</math></i> . . . . .	44
3.4 <i>Linear regression for <math>\mathbf{X}</math> vs. <math>\mathcal{Y}</math> (<math>\mu, \sigma</math>) in log-space</i> . . . . .	46
3.5 <i>Distribution of PGV vs. <math>S_v</math></i> . . . . .	47
4.1 <i>Data set of 36,000 observations</i> . . . . .	60
4.2 <i>Considered basis functions</i> . . . . .	60
4.3 <i>Variation of <math>\log(Z_{M2})</math> (for <math>q_m = q_s = 1</math>) due to number of particles</i> . . . . .	62
4.4 <i>Numerical results for models (a) <math>\mathcal{M}_1</math>, (b) <math>\mathcal{M}_2</math>, (c) <math>\mathcal{M}_3</math>, (d) <math>\mathcal{M}_4</math>, and (e) <math>\mathcal{M}_5</math></i> . . . . .	64
4.5 <i>Comparison metrics of models (a) Bayesian model selection, (b) Predictive interval, (c) K-S distance, and (d) Q-Q error</i> . . . . .	65
4.6 <i>Expected FFs for life safety limit state</i> . . . . .	66
4.7 <i>Prior/posterior comparison (a) Exponent <math>\gamma_{\mathcal{M}}</math>, (b) Variance <math>\mathbf{S}_{\mathcal{M}}^2</math>, and (c) Coefficients <math>\mathbf{c}_{\mathcal{M}}</math></i> . . . . .	66
4.8 <i>Comparison of FFs for: (a) 10, (b) 40, and (c) 500 observations</i> . . . . .	68
4.9 <i>Reduction of epistemic uncertainty as function of <math>N</math></i> . . . . .	69
5.1 <i>Data set of <math>\mathcal{X}_1</math> vs. <math>\mathcal{X}_2</math> observations</i> . . . . .	78

Figure	Page
5.2 Numerical results for $(X_1, X_2)$ models (a) $\mathcal{M}_1$ , (b) $\mathcal{M}_2$ , (c) $\mathcal{M}_3$ , (d) $\mathcal{M}_4$ , and (e) $\mathcal{M}_5$ . . . . .	81
5.3 Comparison metrics of $(X_1, X_2)$ models (a) Bayesian model selection, (b) Predictive interval, (c) KS-distance, and (d) Q-Q error . . . . .	82
5.4 Prior/posterior comparison (a) Exponent $\gamma_{\mathcal{M}}$ , (b) Variance $\varsigma_{\mathcal{M}}^2$ , and (c) Coefficients $\mathbf{c}_{\mathcal{M}}$ . . . . .	83
5.5 Numerical representation of $X_1$ - $X_2$ lognormal model . . . . .	83
5.6 Input domain division for $X_1$ -sel (left) and $(X_1, X_2)$ -sel (right) . . . . .	84
5.7 Comparison of expected fragility functions using different datasets with 100 observations and different selection strategies . . . . .	84
5.8 Comparison of epistemic uncertainty performance as function of $N$ for the strategies (a) random, (b) $X_1$ -based and (c) $(X_1, X_2)$ -based with respect to uniform (blue) . . . . .	85
6.1 Representation of different ground motion databases . . . . .	89
6.2 Fragility analysis for 50 observations from different ground motion databases: (a) California, (b) Stiff-Soil, and (c) SAC-Combined . . . . .	91
6.3 Fragility analysis for 50 observations combining synthetic and historic ground motion records . . . . .	92
6.4 Observations from scaled and original ground motions from SAC-Combined database . . . . .	93
6.5 Fragility analysis for (a) $N = 720$ scaled, (b) $N = 60$ scaled, and (c) $N = 60$ original records . . . . .	94
A.1 Relationship between normal and lognormal RV (Courtesy of [123]) . . . . .	109
B.1 Scatter matrix plot for $\mathbf{X}$ (top left portion of Fig. 3.3) . . . . .	111
B.2 Scatter matrix plot for $\mathbf{X}$ (bottom left portion of Fig. 3.3) . . . . .	112
B.3 Scatter matrix plot for $\mathbf{X}$ (top right portion of Fig. 3.3) . . . . .	113
B.4 Scatter matrix plot for $\mathbf{X}$ (bottom right portion of Fig. 3.3) . . . . .	114
C.1 Ground motion database in terms of earthquake magnitude . . . . .	115
C.2 Ground motion database in terms of NEHRP soil classification . . . . .	116
C.3 Ground motion database in terms of wave direction . . . . .	116
C.4 Ground motion database in terms of epicentral distance . . . . .	117

## ABBREVIATIONS

ATC	Applied Technology Council
BIF	Bayesian Inference Framework
CCDF	Complementary cumulative distribution function
CDF	Cumulative distribution function
CLT	Central limit theorem
CP	Collapse prevention
DC	Damage control
DM	Damage measure
EDP	Engineering demand parameter
FF	Fragility function/curve
FS	Fragility surface
FY	First yield
IDA	Incremental dynamic analysis
IM	Intensity measure
IO	Immediate occupancy
KS	Kolmogorov-Smirnov
LHS	Latin-Hypercube sampling
LS	Limit state
LSf	Life safety
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
MISD	Maximum inter-story drift
MLE	Maximum likelihood estimation
MRF	Moment-resisting frame
NEHRP	National Earthquake Hazards Reduction Program

PMI	Plastic mechanism initiation
PDF	Probability density function
PGA	Peak ground acceleration
PGD	Peak ground displacement
PGV	Peak ground velocity
RQ	Research question
RV	Random variable
SBM	Specific barrier model
SDOF	Single-degree-of-freedom
SHM	Structural health monitoring
SMC	Sequential Monte Carlo
SR	Limited safety range
SSMRP	Seismic safety margins research program

## ABSTRACT

Author: Peña, Francisco. Ph.D.

Institution: Purdue University

Degree Received: May 2019

Title: Efficient Computation of Accurate Seismic Fragility Functions Through Strategic Statistical Selection.

Major Professors: Shirley J. Dyke & Ilias Bionis.

A fragility function quantifies the probability that a structural system reaches an undesirable limit state, conditioned on the occurrence of a hazard of prescribed intensity level. Multiple sources of uncertainty are present when estimating fragility functions, e.g., record-to-record variation, uncertain material and geometric properties, model assumptions, adopted methodologies, and scarce data to characterize the hazard. Advances in the last decades have provided considerable research about parameter selection, hazard characteristics and multiple methodology for the computation of these functions. However, there is no clear path on the type of methodologies and data to ensure that accurate fragility functions can be computed in an efficient manner. Fragility functions are influenced by the selection of a methodology and the data to be analyzed. Each selection may lead to different levels of accuracy, due to either increased potential for bias or the rate of convergence of the fragility functions as more data is used. To overcome this difficulty, it is necessary to evaluate the level of agreement between different statistical models and the available data as well as to exploit the information provided by each piece of available data. By doing this, it is possible to accomplish more accurate fragility functions with less uncertainty while enabling faster and widespread analysis. In this dissertation, two methodologies are developed to address the aforementioned challenges. The first methodology provides a way to quantify uncertainty and perform statistical model selection to compute seismic fragility functions. This outcome is achieved by implementing a hierarchical Bayesian inference framework in conjunction with a sequential Monte Carlo technique. Using a finite amount of simulations,

the stochastic map between the hazard level and the structural response is constructed using Bayesian inference. The Bayesian approach allows for the quantification of the epistemic uncertainty induced by the limited number of simulations. The most probable model is then selected using Bayesian model selection and validated through multiple metrics such as the Kolmogorov-Smirnov test. The subsequent methodology proposes a sequential selection strategy to choose the earthquake with characteristics that yield the largest reduction in uncertainty. Sequentially, the quantification of uncertainty is exploited to consecutively select the ground motion simulations that expedite learning and provides unbiased fragility functions with fewer simulations. Lastly, some examples of practices during the computation of fragility functions that results in undesirable bias in the results are discussed. The methodologies are implemented on a widely studied twenty-story steel nonlinear benchmark building model and employ a set of realistic synthetic ground motions obtained from earthquake scenarios in California. Further analysis of this case study demonstrates the superior performance when using a lognormal probability distribution compared to other models considered. It is concluded by demonstrating that the methodologies developed in this dissertation can yield lower levels of uncertainty than traditional sampling techniques using the same number of simulations. The methodologies developed in this dissertation enable reliable and efficient structural assessment, by means of fragility functions, for civil infrastructure, especially for time-critical applications such as post-disaster evaluation. Additionally, this research empowers implementation by being transferable, facilitating such analysis at community level and for other critical infrastructure systems (e.g., transportation, communication, energy, water, security) and their interdependencies.

## 1. INTRODUCTION

The successful operation of a community relies on the preservation of normal functionality of the built environment. However, normal performance is unfortunately interrupted by major disruptions, e.g., natural or man-made hazard events. Even though the livelihood of a community after a strong event may be disrupted by extensive damage in the built infrastructure, additional factors like the lack of knowledge about the safety and integrity of the existing assets may further impede the recovery process (e.g., loss of functionality, inability to use designated shelter structures, inadequate classification of damage levels for critical infrastructure). As a result, it is very important for any community to be aware of the structural integrity of its assets before, but most importantly after, a major disruption. By knowing the conditions of the built environment after a disaster, it is possible to make informed decisions and an adequately allocate resources. Herein, the terms "built environment", or "infrastructure", refer to the set of assets and facilities that provide essential services and commodities to the inhabitants of a community, such as buildings, hospitals, roads, bridges, power and communication networks.

Current practices to evaluate the integrity of a structural asset include on-site inspection and structural health monitoring (SHM). The former requires a team of experts that are available and willing to participate in the assessment, while the latter requires deploying a set of sensors and other computational resources to monitor the dynamic response of a single asset. Both practices have certain limitations, precluding their application at large scale (e.g., community-level, city-level, or even state-level). On the one hand, human inspection is expensive, time-consuming, and it may be inconsistent. Typically, there is a limited number of qualified teams available to perform a subjective evaluation of the integrity for an extensive number of large-scale assets [1]. On the other hand, SHM can be expensive, time-consuming, and it is not a widely-accepted technique among stakeholders (e.g., owners, regulatory agencies). A large network of sensors, data acquisition systems, and other

computational resources in conjunction with trained staff are needed for the implementation of a robust, redundant, and well maintained system to extract the necessary information. Furthermore, these systems may require supplementary investments for maintenance or tuning due to time-dependent disturbances that weather conditions and transient loads impose on the structure. Lastly, it may be difficult to identify local damage because these systems are often trained to recognize patterns associated with global anomalies on the structure. This is partly due to the fact that SHM systems often operate in an unsupervised learning mode because data from damaged structures is scarce [2].

Consequently, special interest has been given to predict, in a more accurate and rapid manner than human inspection or SHM, the structural integrity of an asset after a strong event. Here the terms, hazard, event, or disturbance, refer to ground motion. By increasing the level of knowledge about the effects that a ground motion may impose on the built environment, stakeholders would be able to make better decisions to improve community resilience. Thus, critical infrastructure would endure less damage and communities will face minimal loss of functionality and number of casualties after a hazard event. For example, it is possible to direct a larger amount of resources to the critical infrastructure that requires immediate intervention. Similarly, response and reconnaissance missions can better prepare action plans to safely mount rescue and response activities. Because of this need to assess the structural integrity, the concept of *fragility function* (FF) emerged in the mid 1970s. Originally FFs were established to address the need for evaluating assets in the nuclear power industry. Recently, these methods obtained recognition for a broad range of elements in the built environment.

Initially the term fragility was used to refer to the threshold value of the seismic capacity of a nuclear power plant before failure occurs [3–5]. It was in 1980 when the term fragility became recognized as the probability of failure of systems and components in the nuclear power industry [6]. FFs are employed as an alternative solution to assess the structural integrity of the built environment. A FF quantifies the probability of a system to reach an undesirable *limit state* (LS), e.g., collapse, yielding, conditioned on the occurrence

of a hazard event of given magnitude [7, 8]. Thus, a FF is a crucial component for the quantification of economic loss and casualties, vulnerability, and risk assessment [7–11].

Although quite informative, FFs are sensitive to the choice of data used for their estimation. The uncertainty in a FF can be decomposed into aleatory (a.k.a., fundamental, frequency, probabilistic, random variability) and epistemic uncertainties (other names given in the literature are incomplete knowledge, probability, statistical, systematic, uncertainty and modeling variation [12]). The former corresponds to the intrinsic variability (true randomness) that may characterize the system (e.g., real structure, numerical model) and the ground motion (e.g., record-to-record variation), while the latter refers to the lack of knowledge about the problem (e.g., uncertainty induced by limited data). Since aleatory uncertainty is irreducible, efforts should address reducing epistemic uncertainty.

The last decades have been dedicated substantially to increasing the level of knowledge about all of the factors that contribute into the uncertainty of FFs [13–15]. For instance, several methodologies have been implemented (although some of them were not originally intended) for the computation of FFs, such as: Safety Factor Method [12, 16], Linear Regression in Logarithmic Space [17, 18], numerical simulation using Maximum Likelihood Estimation [19, 20], fitting model parameters using Moment Matching [21], Sum-of-Squared-Errors [21, 22], Least-Squares [21, 22], Gaussian Kernel Smoothing [23], *Incremental Dynamic Analysis* (IDA) [24, 25], Neural Networks [25], Capacity Spectrum Method [26, 27], Modal Pushover Analysis [28], Bayesian Inference [29, 30], among others. In addition to the notable differences in the time that a fragility analysis may take among all the different methodologies, the shape of the resulting FFs may exhibit differences as well [22, 31]. Similarly, multiple probability distribution functions have been used to describe the dispersion in the structural response, the lognormal distribution being the most commonly adopted case for seismic excitation [8]. Other distributions studied are the Weibull [15], generalized extreme value [15], and non-parametric distributions [32]. For other natural hazard, such as turbulent wind forces, it is common practice to use lognormal, Weibull, gamma, normal, and Cauchy distributions [33–35]. Additionally, other factors have been studied for their influence on uncertainty, such as material and geometric

properties [36, 37], record-to-record variations [19], damping [36], input/output parameters [31, 35], aging and deterioration [38], modeling calibration [39], and even the quantity of analyzed ground motions [22, 25, 30].

Consequently, it is possible to find an extensive number of guidelines, frameworks, and books dedicated almost exclusively to discuss all of these findings [7, 8, 40–45]. Despite the tremendous advances made towards increasing the understanding about ground motions, structural response, and FFs, there is a need to evaluate all the assumptions (most of them adopted since the late 1970s when the computational resources were limited), the different methodologies, and the intrinsic variation between fragility analyses. By being more selective in the methodologies and type of data used for the analysis, it is possible to achieve more accurate and realistic FF, supporting a better informed-decision-making process and optimizing the allocation of community resources.

Given that FFs have potential to produce important supporting information in the decision process that transforms the built environment and correspondingly impacts community livelihood, the following inquiries need to be properly addressed:

- Is this FF derived using the appropriate model?
- Are the data used representative and applicable for the case study?
- Is the amount of data used sufficient for the FF to be conclusive?

Although some of the techniques found in the literature partially address these questions, primarily for the dynamic response of the structures, the development of a systematic approach centered on seismic FFs remains unaddressed. The methodologies developed in this dissertation leverage the use of Bayesian inference which is used to estimate the model parameters while quantifying uncertainty. Similarly, this approach enables the evaluation of the level of agreement between models and actual data as a means to acquire more accurate FFs with a minimum expenditure of computational resources. Additionally, a methodology for the selection of seismic data is developed to expedite the computation of accurate FFs. Furthermore, this dissertation demonstrates that some of the common

practices in the computation of FF may lead to biased results, incorporating in some cases substantial levels of uncertainty.

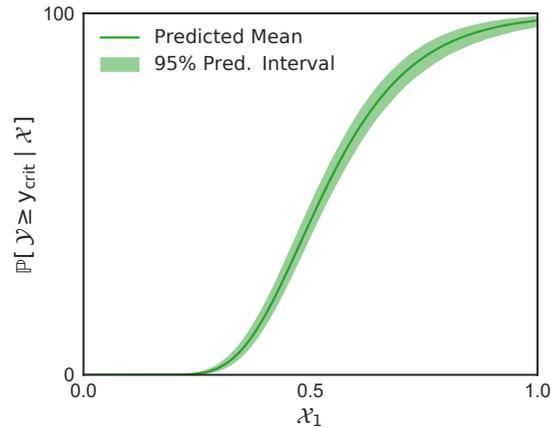
## 1.1 Literature Review

The concept of a seismic FF is defined as the conditional probability of a structure or component to reach a prescribed LS, given the occurrence of a ground motion with *intensity measure* (IM)  $\mathcal{X}_1 = x$ . Here, the undesirable LS does not necessarily signify the complete failure or collapse of a structure. It only refers to a given condition of functionality that is associated with a selected structural response (a.k.a., *damage measure*, DM, or *engineering demand parameter*, EDP), denoted by  $\mathcal{Y}$ . It is possible to establish a critical level or threshold  $y_{\text{crit}}$  in the structural response for each LS. The fragility or probability of reaching a LS refers to  $\mathcal{Y} \geq y_{\text{crit}}$  and the most general mathematical expression is:

$$\begin{aligned} F(x; y_{\text{crit}}; \mathbb{I}) &:= \mathbb{P}[\mathcal{Y} \geq y_{\text{crit}} \mid \mathcal{X}_1 = x, \mathbb{I}] = \int_{y_{\text{crit}}}^{\infty} f_{\mathcal{Y}}(y \mid \mathcal{X}_1 = x, \mathbb{I}) dy \\ &= \mathbb{E} \left[ 1_{[y_{\text{crit}}, +\infty)}(\mathcal{Y}) \mid \mathcal{X}_1 = x, \mathbb{I} \right] \end{aligned} \quad (1.1)$$

where  $\mathbb{P}[A \mid B]$  is the probability of event  $A$  occurring given that  $B$  already occurred,  $f_{\mathcal{Y}}(y)$  is the *probability density function* (PDF) for  $\mathcal{Y} = y$ ,  $\mathbb{E}[\cdot \mid \cdot]$  is the conditional expected value, and  $1_A$  is the characteristic function of set  $A$ , i.e.,  $1_A(y) = 1$ , if  $y \in A$ , and  $1_A(y) = 0$ , otherwise. Additionally, the term  $\mathbb{I}$  corresponds to all known (or assumed to be known) information about the structure and its surroundings. For instance,  $\mathbb{I}$  includes information such as the geographical location and orientation of the structure, specific site conditions, soil characteristics, distance to seismically active faults, type of seismic fault, weather conditions that may affect the soil properties, among many other factors. Including this information in the fragility analysis is translated into implementing a realistic database of ground motions that are representative for the system. From herein, the information  $\mathbb{I}$  is assumed to be an implicit property of the FF and it will be omitted to simplify the mathematical notation. A graphical representation of a FF is shown in Fig. 1.1. The line

represents the expected FF while the are corresponds to the 95% predictive interval. FFs have



**Fig. 1.1.** Representation of fragility function

become one of the principal tools for structural, loss, vulnerability, and risk assessments [7–11, 40]. Given the tight relationship between the concepts of hazard, vulnerability and risk, it is possible to unintentionally interchange their definitions. For this reason, it is convenient to briefly introduce proper definitions based on the detailed explanation prepared by [8]. In summary, hazard or hazard probability refers to the relationship between a given IM value and the frequency in which events of this magnitude or larger are expected to occur at a specific geographical location. Vulnerability differs from fragility because it quantifies the impact of reaching an undesirable LS in terms of a variable that measures the loss (e.g., casualties, dollars). And, the seismic risk of an asset represents the potential consequences (vulnerabilities) that ground motions can generate in a structure with a specific geographical location during a certain period of time.

### 1.1.1 History

The term FF was initially introduced in the late 1970s for the design and analysis of facilities from the nuclear power industry. Although FFs are extensively applied nowadays for the built environment, especially buildings and bridges, it was only in the late 1990s when the concept started to be applied to such types of structures [15]. As was already

stated, the term fragility was associated with the margin  $y_{\text{crit}}$  in 1975 [3] and later became the probability of failure in 1980 [6]. The surge of scientific research that precipitated the advances in fragility during the late 1970s was triggered in response to the Reactor Safety Study (a.k.a., the Rasmussen Report or WASH-1400) [46]. This controversial work concluded that "earthquake-induced accidents should not contribute significantly to reactor accident risk" along with multiple unsupported assumptions, probability distributions, and methodologies [47]. According to [48], three programs followed that exposed deficiencies of the Rasmussen Report and contributed remarkably with the advances in seismic fragility analysis, including: (i) The Diablo canyon seismic risk evaluation [4], (ii) the *Seismic Safety Margins Research Program* (SSMRP) [6, 16, 49–52], and (iii) the Oyster creek probabilistic safety analysis [12].

The first program exposed the necessity to almost double the structural capacity (denoted by  $R$  and measured in the same units as the ground motion) of the nuclear power plant to preserve a similar level of reliability, after the discovery of a seismic fault line (Hosgri) just 6 km away. Two major contributions resulted from this program: (i) an expression to define the probability of failure,  $P_F \in [0, 1]$  (see Eq. (1.2)), and (ii) the use of the lognormal distribution to model the seismic capacity (see Eq. (1.3)), which was initially proposed by [53]. These are expressed as

$$\mathbb{P}[X_1 \geq y_{\text{crit}}] := \int_0^\infty F_R(x) f_{X_1}(x) dx = \int_0^\infty [1 - F_{X_1}(x)] f_R(x) dx. \quad (1.2)$$

$$F_R(x) := \Phi\left(\frac{\log(x/\bar{r})}{\sigma_R}\right) \quad (1.3)$$

where  $F_R(x)$  is the *cumulative distribution function* (CDF) for  $R$  evaluated at  $x$ ,  $\Phi(\cdot)$  represents the CDF for the standard normal distribution,  $\bar{r}$  is the median seismic capacity of the system, and  $\sigma_R$  corresponds to the standard deviation of the natural logarithm of  $R$ . Assuming that  $R$  can be expressed as the product of a series of positive *random variables* (RVs), the *central limit theorem*, CLT, (in particular Gibra's law) states that  $f_R$  can be approximated using a lognormal distribution [53]. Thus, the capacity becomes

$R := c_1 c_2 \cdots c_k \hat{R}$ , in which  $\hat{R}$  corresponds to the calculated estimate of the capacity, and the coefficients  $\{c_k\}_{k=1}^K$  are correction terms to account for the variations induced by site conditions, dynamic amplification, damping, modeling discrepancies, among other factors.

The SSMRP included a compendium of projects to enhance the prediction of the behavior of nuclear facilities disturbed by seismic excitation [6]. One of these projects exclusively focused on fragility. In particular, this fragility project centered its attention on augmenting a database of FFs and safety margins (conservative values for  $y_{\text{crit}}$  with approximately 95% confidence) for elements and components from nuclear power plants [6]. Additionally, the program separated the correction terms  $c_k$  into two ensembles: aleatory and epistemic uncertainties.

The Oyster creek project centered its attention on the importance of incorporating confidence intervals around FFs given the scarceness on available data and the considerable dependency on engineering judgment for their derivation [12]. Before this project, FFs were presented as deterministic functions representing the "best estimate" (median FF), whose shape was labelled as unwarranted in accordance with the lack of incorporating uncertainty [12].

Although [12, 16] supported the use of lognormal distributions, the applicability of low probability events into real life was brought into question. The use of the lognormal distribution for such rare events was said to result in a conservative estimate, giving lower capacity and/or larger demand values [12]. Confidence intervals were obtained after defining the ground motion IM for a prescribed probability of failure  $0 \leq P_F \leq 1$  as:

$$\mathcal{X}_1 := \check{\mathcal{X}}_1 \epsilon_R \epsilon_U \quad (1.4)$$

where  $\check{\mathcal{X}}_1$  is the "best estimate" (for  $P_F = 0.5$ ) of the median ground motion IM,  $\epsilon_R$  represents the random (aleatory) variability while  $\epsilon_U$  corresponds to the uncertainty (epistemic). Both  $\epsilon_R$  and  $\epsilon_U$  were modeled using lognormal RVs with median value of one and logarithmic standard deviations of  $\sigma_R$  and  $\sigma_U$ , respectively. This approach is the most common methodology used worldwide for seismic probability risk assessment in the nuclear power

industry, and it is known as the Safety Factor Method [31] (further explanation is provided in Section 1.1.4). Its name was given due to the implementation of a set of safety factors describing the uncertainty in terms of capacity for multiple components with the purpose of computing the entire system's median capacity  $\check{\chi}_1$ . This method is still used nowadays because of its simplicity which requires inferring just three parameters:  $\check{\chi}_1$ ,  $\sigma_R$ , and  $\sigma_U$ , but the method also has the complexity to include different sources of uncertainty as safety factors [15, 54].

### 1.1.2 *Classification of FFs*

FFs are often classified into four categories, according to the source of the data used for their derivation. These four categories are: empirical, judgmental or expert-based, analytical, and hybrid [7, 11, 14, 39, 55]. A description of the source of data used, and the advantages and disadvantages for each category can be found in Table 1.1. Examples of empirical FFs can be found in [25, 27, 35]; judgmental FFs are commonly found in building and rehabilitation codes [55–57]; analytical FF are becoming more common due to the growing expertise and widespread use of computational modeling tools (e.g., Opensees, SAP2000, Abaqus) [15, 17, 18, 22, 28, 30, 58]; and some of the major applications of hybrid FF include [39, 56, 57]. In particular, the use of hybrid FF results is fascinating since it may allow for the strengths of certain categories to compensate for the weaknesses of other sources of data. For instance, [39] presented the differences between analytical FFs for a calibrated/uncalibrated numerical model of a bridge that were subsequently updated by implementing real experimental data for a scaled version of the center pile model. The results show that experimental data enables convergence of the FF regardless of the differences in the models due to calibration. Another example of hybrid FFs are the judgemental functions that were corrected by incorporating data from the San Fernando (1971) and Northridge (1994) earthquakes, and were finally presented by the *American Technology Council* (ATC) in ATC-13 [56] and ATC-40 [57], respectively. Using numerical models to derive analytical FFs has been widely adopted due to increased availability and access

**Table 1.1.** Classification of fragility functions

Classification	Source	Advantages	Disadvantages
<b>Empirical</b>	Post-earthquake assessments	<ul style="list-style-type: none"> <li>+ Most realistic class</li> <li>+ Incorporates complex pragmatic information (e.g., soil-structure interaction effects, topography, location, orientation of the structure, path and source characteristics, construction process, material and geometric properties, actual boundary conditions)</li> </ul>	<ul style="list-style-type: none"> <li>- Available data is scarce and clustered in the low-damage/low-intensity region</li> <li>- Requires the occurrence of a major earthquake</li> </ul>
<b>Judgmental</b>	Panel of experts with extensive experience in earthquake eng.	<ul style="list-style-type: none"> <li>+ No requirement for expensive computations</li> </ul>	<ul style="list-style-type: none"> <li>- Easily biased by the opinion, knowledge, expertise and recognition of each expert</li> </ul>
<b>Analytical</b>	Simulation of numerical model	<ul style="list-style-type: none"> <li>+ No requirement for real data nor panel of experts</li> <li>+ Artificial ground motions can be used</li> <li>+ Most accessible option -&gt; widely used</li> </ul>	<ul style="list-style-type: none"> <li>- Accuracy depends on numerical model and its capability to mimic real dynamic behavior</li> <li>- A large ensemble of ground motions may be required</li> <li>- Biased by earthquake characteristics and model assumptions</li> </ul>
<b>Hybrid</b>	Two or more different sources	<ul style="list-style-type: none"> <li>+ Same as the ones from the selected sources</li> <li>+ Strength of one class may compensate for weakness of a different class</li> </ul>	<ul style="list-style-type: none"> <li>- Requires enough data to counteract each source's disadvantage</li> </ul>

to computing clusters and supercomputers [9, 11, 14, 18, 22, 30, 40, 55, 58, 59]. Researchers are not only able to perform faster fragility analysis but they are also able to quantify the uncertainty associated with each function. For this reason, the methodologies explained in this dissertation are discussed in the context of analytical FFs. However the proposed methodologies of the subsequent chapters are independent of the category of FF and the selected method.

### 1.1.3 *Uncertainty in FFs*

The propagation of the uncertainty from specific parameters into the FF has been major concern since the early studies in the 1980s [12]. In the case of analytical FFs, it is possible to identify three major ingredients that contribute to the uncertainty of such functions: (i) the numerical model, (ii) the available data, and (iii) the methodology used for its derivation. The first one refers to the variations in the material, geometry, structural and dynamic properties as well as the modeling assumptions adopted to represent the original structure. The second considers the uncertainty in the ground motions and the selection of the input/output parameters for the analysis. And, the third includes the discrepancies between different methods, probability distributions, and algorithms for computing the FFs.

George Box stated in 1976 that "all numerical models are wrong but some are useful" [60]. Consequently, a key aspect is implementing a realistic numerical model able to represent the dynamics of the actual structure for the computation of analytical FFs. Otherwise, the FFs will not be applicable to the current state of the structure and they should not be used for making important decisions. For this reason, the influence of modeling parameters on the FFs is perhaps the first consideration when dealing with uncertainty in these functions. Uncertainty in the numerical model includes the variations due to geometrical and material properties, implementation of non-linear behavior, release, and partial-fixity between the structural members at the connection level, among other factors. This uncertainty is typically integrated into the fragility analysis after assigning certain probability distributions to each of the parameters that affect the response of the structure. Then, the fragility of an ensemble of numerical models, each with a specific sample of parameters, is analyzed and then combined into a unique FF for a given type of structure. The ensemble can be obtained by randomly sampling parameters from each distribution. However, the use of more efficient statistical samplers (than uniform sampling) has gained popularity. For instance, one of the most used samplers is the Latin-Hypercube sampling (LHS) technique [19, 25, 29, 61]. LHS enables the creation of an efficient and varied ensemble by distributing the samples along the entire domain. Examples of the probability distributions

used in the literature to describe uncertainty in the model parameters are presented in Table 1.2.

**Table 1.2.** Examples of uncertainty on modeling parameters

Parameter	Distribution	Mean	COV
Concrete Compressive strength ( $f'_c$ )	Normal	29.53 MPa	0.15 [36]
		4.500 psi	0.2 [19, 62]
	Lognormal	38.80 MPa	0.2 [37]
Concrete Young's modulus ( $E_c$ )	Normal	25727.35 MPa	0.15 [36]
Masonry Compressive strength ( $f'_m$ )	Normal	18.40 MPa	0.15 [36]
Masonry Young's modulus ( $E_m$ )	Normal	20700.00 MPa	0.15 [36]
Steel yielding strength ( $f_y$ )	Lognormal	336.72 MPa	0.11 [19, 36, 62]
		384.80 MPa	0.1 [37]
Steel Young's modulus ( $E_s$ )	Lognormal	200 GPa	0.05 [37]
Viscous damping ratio ( $\zeta$ )	Uniform	3 %	0.19 [36]
	Lognormal	0.4	1.00 [63]
Eigen-frequency ( $f_0$ )	Lognormal	5 Hz	0.04 [63]
Cover depth	Lognormal	3.81 cm	0.20 [38]
Diffusion Coefficient	Lognormal	1.29 cm/year	0.1 [38]
Rate of corrosion	Lognormal	0.127 mm/year	0.3 [38]
General dimension	Normal	-	0.05 [37]

In 2000, [19] presented a statistical analysis of the uncertainty in analytical FFs for bridges representative of the Memphis (Tennessee) area. Within this work, the authors included in the analysis uncertainties in the material properties,  $f'_c$  and  $F_y$ , the response associated with different types of bridge, and the influence of the characteristics of the time-histories in  $\mathcal{Y}$  by analyzing different ground motions datasets. The FFs were computed using the maximum likelihood method and a lognormal distribution to describe the dispersion in the DM. In summary, the work done by [19] provided evidence of the uncertainty between realizations of FF and the importance of following an appropriate statistical procedure for their derivation. Similarly, [19] showed the importance of including confidence or predictive intervals with FFs as a measure to represent the uncertainty. The predictive intervals in

a FF can be obtained comparably from the quantiles of the lognormal distribution used to model the dispersion in the DM, as is the case in the work done by [11]. The authors reduced the simulation time of the recursive analysis of ground motions by implementing an equivalent *single-degree-of-freedom* (SDOF) response database. By characterizing a structure in terms of its stiffness, strength and ductility, it was possible to extract polynomial functions describing the mean response and the corresponding variance using the database. Likewise, the study presented in [38] includes the variation in seismic FFs due to aging and deterioration in steel girder bridges. The resulting FFs presented in [38] exhibits the evolution of the moderate damage FF for the bridge from its construction until an age of 100 years, with aging and the corresponding deterioration being considered. As expected, fragility of the bridge increases with the passage of time when no rehabilitation nor maintenance is included in the analysis. But what it is not obvious is that the change in fragility occurs in a nonlinear fashion.

In terms of sensitivity analysis, [64] and [63] ranked the influence of modeling parameters in the resulting uncertainty of FFs. A detailed sensitivity analysis for the uncertainties associated with models of multispan simply supported steel girder bridges was presented in [64]. Three different probability distributions, uniform, normal, and lognormal, were used to describe the variability in the parameters describing the structural capacity for each component of the bridge, including foundation, piers, abutments, bearings, girders, and deck. This work concluded that a careful definition of modeling parameters might be done in a deterministic manner with the intention of focus all the efforts on dealing with the ground motion uncertainty which significantly influences the FF. The work performed in [63] consisted of ranking the influence of multiple modeling parameters contributing to the uncertainty in FFs. A total of 14 factors with their corresponding parameters variation were considered and categorized into three groups: capacity (e.g., strength, ductility), equipment response (e.g., damping, ground motion component), and structural response (e.g., spectral shape, eigen-frequencies and eigen-values, mode shapes). Within the most influential factors, one finds the spectral shape, the ductility, and the equipment response damping. Additionally, the variation in the eigen-frequency, damping, ground motion

variability and model uncertainty were modeled using lognormal distribution. In [35] a quantification of the epistemic uncertainty in empirical FF using post-tornado measurements of a dataset of 1241 residential structures was presented. Four different sources of uncertainty were evaluated: IM values (scarceness on the network of sensors distributed in a given geographical area), damage classification (mis-classification), finite sampling (bias due to reconnaissance missions assessing most damaged structures), and fitting method (probability distribution). The results showed that the implementation of these sources of uncertainty generated a variation of approximately  $\pm 6.5$  m/s around the median wind speed (fragility of 50 %), the lognormal distribution fitted adequately for all the damage cases except the collapse condition, and FFs were sensitive to the uncertainty in the IM.

As it can be observed from the previous studies and from Table 1.2, the civil engineering community is partial towards the use of lognormal distribution [8, 11, 18, 19, 22, 39, 41, 55, 58]. This is understandable since most of the variables/parameters used within the community have positive values, the shape of the capacity of some materials follow this distribution [12], and the parameters can be written as a product or quotient of RVs (resulting in lognormal according to the CLT theorem). According to [8], the reasons for using the lognormal distribution are: it is a simple two-parameter distribution that takes only positive values, sometimes it fits the data well, it has been used for decades, and it uses the fewest assumptions (i.e., maximum entropy). Multiple probability distributions have been analyzed to describe the dispersion in the the structural response, the lognormal distribution being the most commonly adopted case for seismic excitation [8, 65, 66]. Other distributions are the Weibull [15], generalized extreme value [15], and non-parametric distributions [32]. However, it is common to use other probability distributions to describe the dispersion in the response of structures under other classes of hazards. For instance, the variation in the response of structures under turbulent wind forces like tornadoes is typically described using a lognormal distribution, but other distributions, such as Weibull, gamma, normal, and Cauchy distributions, are also used [33–35]. The work presented by [35] provides evidence that the lognormal distribution is not the most suitable option for the collapse LS due to tornadoes. Although, the FFs using normal, lognormal, gamma, and

Weibull distributions exhibit similar values for the low magnitude wind speeds, considerable differences are observed between these functions after a wind speed of approximately 75 m/s, which corresponds to a probability of failure of approximately 35 % for all the considered probability distributions. The Weibull, normal, and gamma distribution show better performance to describe the available data for the collapse LS. For the case of non-parametric distributions, [32] presented discrepancies when using the lognormal distribution to describe seismic FFs. The FFs computed with the lognormal distribution show larger deformation than its non-parametric counterpart. This statement agrees with the initial statement mentioned by [12] about the conservatism of the lognormal distribution. The use of the lognormal distribution needs to be managed with care because it represents conservative estimates for the low probability events.

When determining an analytical FF, it is important to avoid biasing practices that may change the results. The specific ground motions used, as well as their number, play a significant role in obtaining an unbiased and accurate outcome. Although there is no consensus on the specific number of simulations that is adequate, some authors have found the error in FFs is considerably reduced when using between 30 and 40 records for the analysis [22, 67, 68].

Ground motions can be characterized in terms of several different IMs. An intensity measure provides a quantitative description of the magnitude of the motion, but each ground motion contains a great deal of information that is impossible to capture in a single parameter. Even if two ground motions share the same IM value, it is expected that each motion will affect a given structure in a different way. This observation is evident in the work developed by [19], who studied the influence of selecting sets of ground motions randomly on the resulting FF and the uncertainty in that FF. Although subsets of records were selected randomly from a larger set of similar ground motions, each dataset yielded a different realization of the FF. According to [69], the use of different ground motion databases for FF generation results in similar trends for the median structural response, but with different standard deviation. [31] examined the selection of IM and its effect on a fragility analysis. They found that the magnitude of the standard deviation obtained in the

data after performing a linear regression analysis in the logarithmic space is proportional to the level of uncertainty in the FF. Additionally, this study concluded that it is nearly impossible to accurately estimate the appropriate number of simulations a priori, as such number depends on the structure, the selection of the IM parameter(s) used to characterize the ground motion, and the structural response [31,70]. [63] performed a sensitivity analysis with FFs and determined that the parameters with the largest influence on the resulting uncertainty are the spectral shape, the ductility, and the damping.

Then, [25] presents a comparison of the FFs computed using IDA [24] with different number of ground motions. Different sizes of dataset result in different representations of FF. However, the number of ground motions not necessarily correlates with overestimating or underestimating a FF. For the five-story case study structure in [25] and assuming that 100 records represents the true FF, the use of 40 and 60 records leads to overestimate the FF while using 20 and 80 ground motions underestimates the FF.

Later, [22] found that the algorithm used for fitting the parameters to obtain the FF also plays a significant role in the variability of FF. For instance, a similar level of accuracy (5% of error against the true FF) was observed using 40 ground motions when using the least-squares regression for the output parameter, 150 records for the maximum likelihood estimation on the probability of exceedance, and 300 motions for the sum-of-squared-errors on the probability of reaching a limit state.

Another approach is to express the FF in terms of multiple intensity measures. In 2016, [70] developed a framework for the computation of FF under multiple sources of uncertainty. It was found that a conditional FF, using more than a single IM, enables correlation between the IM and the structural response in spite of large variations. Considering the analysis of multiple IMs, [66] introduced a methodology to build a generalized conditional intensity measure using a multivariate lognormal distribution to describe the relationships between the IM values. With this approach it is possible to identify ground motion records that are representative of a given region while discarding those that may bias the CDF of the response. In 2018, [71] developed a methodology to sequentially select the next observation to enable a better estimation of the PDF for a quantity of interest. The selected observation

is the one that provides the maximum expected reduction in the uncertainty of the PDF. It was shown that by sequentially selecting a reduced set of additional observation, it is possible to more efficiently generate a close estimate for the quantity of interest, even for the low probability values (tails of the distribution).

For convenience, fragility has been typically expressed in the literature using Eq. (1.1) instead of using the appropriate notation  $F(x; y_{\text{crit}}) = \mathbb{P}[\mathcal{Y} \geq y_{\text{crit}} \mid \mathcal{X}_1 = x, I]$  by including all the known information  $I$ . The information  $I$  includes key aspects as: the structure with all its characteristics, properties, materials, etc; the geographical location; the orientation of the structure; the seismic hazard; the seismic faults that increase the likelihood of ground motions to occur; among other factors. Consequently, different FFs should be expected for different structures with a similar geographical location, or the "same" structure located in two different places, or even in a similar location but with different orientations. However, with the passage of time and the simplification of the notation (i.e., taking  $I$  out of the equation), the fragility of a structure has been erroneously assumed as an inherent characteristic that depends exclusively on the asset without being affected by the site conditions (i.e., some people assume that the fragility of a structure remains constant no matter its location or orientation). Conversely, reality is different. There is no such thing as ignoring site conditions when evaluating the fragility of a structure. It is possible to average out the site conditions to compute a FF that is independent of the site conditions. Obviously, this indicates that a large amount of uncertainty needs to be accounted for in the fragility analysis. For example, every single ground motion needs to be considered in the analysis, including those disastrous events that take place in remote areas, as well as each ground motion impacting the structure in all the possible orientations. Otherwise, the estimated structural response will be most likely underestimated and/or overestimated, increasing the potential of biasing the FF. Thus, performing this type of analysis is infeasible even with the computational capabilities and ground motion databases available now.

A representative numerical model of the structure is needed for accurate FFs, but the specific ground motions used also strongly influence the results. Both an accurate model of the structure and a sufficient number of appropriate ground motions are needed to compute

unbiased, realistic and applicable FFs for a given structure. However, the availability of realistic ground motions for a specific geographical location is scarce. When a sufficient number of realistic ground motions is not available for a given location, researchers augment their databases by exploiting other databases [72, 73], using scaling to amplify existing ground motions [24, 25], or adopting synthetic ground motions [30, 31, 74]. Each of these alternatives must be applied with care to reduce the chance of bias in the resulting FF [69, 75, 76].

According to [75], scaling of ground motions influences the median nonlinear response of a structural model, even when the records are associated with similar values of magnitude and epicentral distance. The bias in the structural response is proportional to the scaling factors, the natural period of the structure, the strength, higher modes contribution, magnitude and epicentral distance. Finally, it is recommended to implement a methodology known as spectrum matching [77] when ground motion scaling is necessary [67, 75, 76, 78, 79]. From the work presented in [67], the spectrum matching method shows superiority at covering a wider range of frequencies, but it underestimates structural response. However, it is recommended to use this approach for far-fault conditions or near-fault events that do not exhibit the presence of pulses in the ground velocity [76]. The scaling processes of matching spectral acceleration values and modal push-over techniques have the inconvenience of overestimating the acceleration response while underestimating the displacements [76].

Last, but not least, is the uncertainty imparted by the methodology chosen for the computation of the FFs. Some of the methodologies found in the literature include: Safety Factor Method [12, 16], Linear Regression in Logarithmic Space [17], numerical simulation using *Maximum Likelihood Estimation* (MLE) [20], Moment Matching [21], Least-Squares [21, 22], or Sum-of-Squared-Errors [21, 22], Gaussian Kernel Smoothing [23], IDA [24], Neural Networks [25], Capacity Spectrum Method [26, 27], Modal Pushover Analysis [28], Bayesian Inference [29, 30], among others. Multiple studies have made an effort to compare the resulting FFs derived by different methods [21, 22, 27, 31, 37]. These studies expose differences between the FFs, specially at large IM values. The work developed in [31] exhibits the FFs for a 3-story reinforced concrete structure computed using different

methods: the Safety Factor Method, Linear Regression, MLE, and IDA. Although the mean FFs of the different methods are similar, except for the Safety Factor method that underestimates the FF for large acceleration values, considerable differences are observed for the FFs at different percentiles (i.e., predictive intervals). These four techniques are the most widely used methods. For this reason, a brief description of these representative methods is presented in the following section.

#### 1.1.4 Computation of FFs

Several methodologies are available in the literature for the computation of FF. These alternatives have emerged in response to the necessity and the available computational resources at the time. Here, a brief introduction of some of the most relevant and widely used methodologies is presented. The reader is referred to specific citations given for further details about each technique.

##### *Safety Factor Method*

The Safety Factor Method is widely used in the nuclear power industry for the computation of the seismic risk of nuclear facilities [80]. The method was initially introduced in the late 1970s by [12] and remains applicable nowadays. The method is based on the incorporation of seismic margins (a.k.a., safety factors) to estimate the median seismic capacity  $\check{\chi}_1$  of the structure. The median capacity  $\check{\chi}_1$  of the nuclear facility is obtained as the product of the design-basis ground motion (i.e., maximum ground motion's IM at which the nuclear facility is able to remain functional and to assure a safe shut-down process) and the median value of the safety product,  $SF = \prod SF_i$ . The considered margin factors  $SF_i$  correspond to material strength, energy dissipation, spectral shape, damping, modeling, structural mode combination, combination of horizontal ground motion components, soil-structure-interaction, and spacial incoherence [31]. Each factor can be calculated as  $SF_i = \check{S}\check{F}_i \epsilon_R \epsilon_U$ , where  $\check{S}\check{F}_i$  is the median value for the safety factor  $SF_i$ ,  $\epsilon_R$  represents the random (aleatoric) variability while  $\epsilon_U$  corresponds to the uncertainty (epistemic) for

the aforementioned factor. Both  $\epsilon_{Ri}$  and  $\epsilon_{Ui}$  are modeled using lognormal RVs with a median value of one and a logarithmic standard deviation of  $\sigma_{Ri}$  and  $\sigma_{Ui}$ , respectively. The parameters  $S\check{F}_i$ ,  $\sigma_{Ri}$  and  $\sigma_{Ui}$  have to be estimated, although standard values can be found in [54].

Finally, the FF at percentile  $\rho$  (the median FF corresponds to  $\rho = 0.5$ ) is expressed using a lognormal distribution with shifted mean as:

$$F(x) = \Phi \left( \frac{\log \left( x / \check{X} \right) + \sigma_U \Phi^{-1}(\rho)}{\sigma_R} \right) \quad (1.5)$$

where  $\sigma_R$  and  $\sigma_U$  are the logarithmic standard deviation associated with the aleatoric and epistemic uncertainty, respectively, for the seismic capacity (see Eq. (1.4)). Further information of this method can be found in [12, 16, 31].

#### *Maximum Likelihood Estimator - MLE*

This method has the characteristic that it is able to compute the FF without the intermediate step of estimating the response of the structure. Instead, the only information required is the ground motion IM values and binary information about the exceedance of a limit state. As the name of the method indicates, the objective is to determine the parameters that dictate the shape of the probability distribution used to describe the FF, and maximize the likelihood function:

$$\mathcal{L} = \prod_{n=1}^N F \left( x^{(n)} \right)^{z^{(n)}} \left( 1 - F \left( x^{(n)} \right) \right)^{1-z^{(n)}} \quad (1.6)$$

where  $N$  is the total number of ground motions to be analyzed,  $x^{(n)}$  is the IM for the  $n$ -th ground motion,  $z^{(n)}$  is equal to 1 or 0 depending if the structure reached the LS with the occurrence of the corresponding ground motion, and  $F(x)$  represents the FF for the

analyzed LS and evaluated at  $\mathcal{X}_1 = x$ . Assuming that the seismic FF can be expressed using a lognormal distribution, the fragility takes the form:

$$F(x) = \Phi\left(\frac{\log(x) - \mu}{\sigma}\right) \quad (1.7)$$

where  $\mu$  and  $\sigma$  represent the mean and standard deviation, respectively, of the natural logarithm of  $\mathcal{X}_1$ . Finally,  $\mu$  and  $\sigma$  are computed such that the following equation is satisfied:

$$\frac{d(\log \mathcal{L})}{d\mu} = \frac{d(\log \mathcal{L})}{d\sigma} = 0. \quad (1.8)$$

Further information of this method can be found in [19, 20, 31, 81].

### *Linear regression in logarithmic space*

This method and the IDA are perhaps the most common approaches used in the literature for the computation of FF. As the name of the method indicates, the idea is to fit the best line within a set of  $N$  data points after performing a logarithmic transformation  $\mathcal{D}_N = \{\log(x^{(n)}), \log(y^{(n)})\}_{n=1}^N$ . It is a common practice to assume that the median value of the response, denoted as  $\exp(\mu(x))$ , (i.e.,  $\mu(x)$  corresponds to the mean value of the RV  $\log(\mathcal{Y})$ ) evaluated at  $\mathcal{X}_1 = x$  takes the following form:

$$\exp(\mu(x)) = c x^\gamma \quad \xrightarrow{\log} \quad \mu(x) = \log(c) + \gamma \log(x) \quad (1.9)$$

where  $c \in \mathbb{R}_+$  is a positive coefficient and  $\gamma \in \mathbb{R}_+$  is a positive exponent [17, 18, 40, 58]. The right-hand side of Eq. (1.9) represents the equation of a line with intercept  $\log(c)$  and coefficient  $\gamma$ . Now the objective is to find the parameters  $c$  and  $\gamma$  that minimize the error between the data  $\mathcal{D}_N$  and the estimated line. Then, the fragility equation is computed as (see Appendix A for the step-by-step derivation from Eq. (1.1)):

$$F(x; y_{\text{crit}}) = 1 - \Phi\left(\frac{\log(y_{\text{crit}}) / \mu(x)}{\sigma}\right) \quad (1.10)$$

where  $\sigma$  is the logarithmic standard deviation of the response  $\mathcal{Y}$  to be computed as:

$$\sigma = \sqrt{\sigma_D^2 + \sigma_C^2 + \sigma_M^2}. \quad (1.11)$$

Here,  $\sigma_D = \sqrt{\log(1 + s^2)}$  is the uncertainty in the demand, the standard error of the sample population is denoted by  $s = \sum_n \left( \log \left( y^{(n)} \right) - \mu(x^{(n)}) \right)^2 / (N - 2)$ ,  $\sigma_C$  is the uncertainty in the capacity, and  $\sigma_M$  corresponds to the modeling uncertainty.  $\sigma_C$  and  $\sigma_M$  are typically taken as 0.3 according to [17, 18, 40, 82, 83] (see these references for further information on the method). Also, notice that both parameters  $\mu(x)$  and  $\sigma(x)$  can be written as a function of  $\mathcal{X}_1$  including other mathematical expressions. For instance, these parameters can be described in terms of polynomials (making necessary to perform a polynomial regression). Lastly, note that although Eq. (1.7) and Eq. (1.10) both represent fragility, they are expressed differently. The difference is the result of the two alternative perspectives about the uncertainty: treating the variability as dispersion in the IM (Eq. (1.7)) or contrarily as dispersion in the DM (Eq. (1.10)).

### *Incremental Dynamic Analysis - IDA*

Another well-known method is IDA, introduced by [24]. This method is analogous to static pushover, but is the dynamic version. The method consists of a progressive scaling process of the value of IM used to characterize each available ground motion. For each ground motion  $n = 1, \dots, N$ , the magnitude of the motion is scaled in an incremental fashion until plastic deformation is achieved (i.e., monotonically increasing deformation for constant or minimal increment in the IM value). The values at which these ground motions are scaled can be simply determined to be the same for all the ground motions or selected by implementing an appropriate and efficient scaling algorithm to reduce the number of iterations [25]. For the latter, given that the IM values for each ground motion are different, it is necessary to perform an additional step in which the results are interpolated into specific IM values to carry on the statistical analysis of the variation for the  $N$  structural responses  $\{y^{(n)} \mid \mathcal{X}_1 = x\}_{n=1}^N$ .

The objective of the method is to estimate the statistics of the response  $\mathcal{Y}$  at the IM value  $\mathcal{X}_1 = x$ , in particular the PDF  $f_{\mathcal{Y}}(y | x)$ . This can be achieved by using the ensemble of  $\{y^{(n)} | \mathcal{X}_1 = x\}_{n=1}^N$  to estimate  $f_{\mathcal{Y}}(y | x)$ , and then compute the FF using Eq. (1.1). It is important to clarify that to minimize the error, it is preferable to quantify the fragility by estimating the shape for  $f_{\mathcal{Y}}$  (fitting a probability distribution) instead of counting the percentage of responses  $y^{(n)} \geq y_{\text{crit}}$ . This is because the number of ground motions implemented in IDA,  $N$ , is typically kept on the order of a few dozen (to limit the excessive use of computational resources during the iterative evaluation of the numerical model subjected to several IM values) which becomes insufficient to estimate the *complementary cumulative distribution function* (CCDF) by merely counting exceedances. Further explanation of the method can be found in [24, 25].

#### *Bayesian Inference Framework (BIF)*

This method is based on the Bayes' rule. The fragility analysis is carried out after defining a statistical model to describe  $f_{\mathcal{Y}}$  in terms of a set of RVs  $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_m\}$ , i.e.,  $f_{\mathcal{Y}}(y | x, \boldsymbol{\theta})$ . For instance, for the case when a lognormal distribution is used to model  $f_{\mathcal{Y}}$ , it is possible to define  $\mu(x)$  and  $\sigma(x)$  in terms of a set of basis functions and the parameters  $\boldsymbol{\theta}$  may be the coefficients for these functions.

Given that  $\boldsymbol{\theta} \in \mathbb{R}^m$  is a multivariate RV, the values that  $\boldsymbol{\theta}$  may take are dictated by a probability distribution  $p(\boldsymbol{\theta})$ , a.k.a, *prior distribution*. However, there is no or little knowledge about  $\boldsymbol{\theta}$  *a priori* (before observing actual seismic data). Therefore,  $p(\boldsymbol{\theta})$  contains the belief or hypothesis about  $\boldsymbol{\theta}$ . Now, the objective is to test this hypothesis against the seismic data  $\mathcal{D}_N = \{x^{(n)}, y^{(n)}\}_{n=1}^N$  to update the prior into the *posterior distribution*  $p(\boldsymbol{\theta} | \mathcal{D}_N)$ . This can be achieved using Bayes' theorem as follows:

$$p(\boldsymbol{\theta} | \mathcal{D}_N) = \frac{p(\mathcal{D}_N | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{D}_N)} \quad (1.12)$$

where  $p(\mathcal{D}_N | \boldsymbol{\theta})$  is known as the *likelihood* and  $p(\mathcal{D}_N)$  as the *evidence*. The likelihood is the measurement process for the statistical model, and it quantifies the probability of

observing the data  $\mathcal{D}_N$  for a particular choice of  $\theta$ . And the evidence is a normalization constant that quantifies the likelihood among all the possible values that  $\theta$  may take:

$$p(\mathcal{D}_N) = \int p(\mathcal{D}_N | \theta) p(\theta) d\theta. \quad (1.13)$$

Finally, once the posterior distribution is determined, the PDF  $f_{\mathcal{Y}}(y | x, \theta)$  is known and the FF can be computed using Eq. (1.1). An addendum of the Bayesian approach is that it is able to quantify the epistemic uncertainty induced by limited data which can be used to drive experimental design. The Bayesian approach has been implemented in the field of civil engineering for applications of structural health monitoring [84, 85], to integrate real post-earthquake data in analytic fragility functions [39], and quantify the epistemic uncertainty model selection for FF [30].

#### *Other approaches*

In general, other methods available for the computation of FF have a degree of similarity with the already discussed methods. Mostly, the objective is to apply a novel computational technique to determine the parameters of the probability distribution that describes the dispersion in the response or to estimate the actual distribution. For instance, the use of neural networks for estimating the response  $\mathcal{Y}$  presented in [25], the implementation of non-parametric probability distributions as occurred in [23] with the use of kernel smoothing methods, describing the response in terms of a set of parameters and basis functions and then estimate these parameters deterministically by minimizing an error function, e.g., moment matching, sum-of-squared-errors, least-squares [21, 22], or stochastically with Bayesian inference [29, 30].

## 1.2 Remaining Challenges

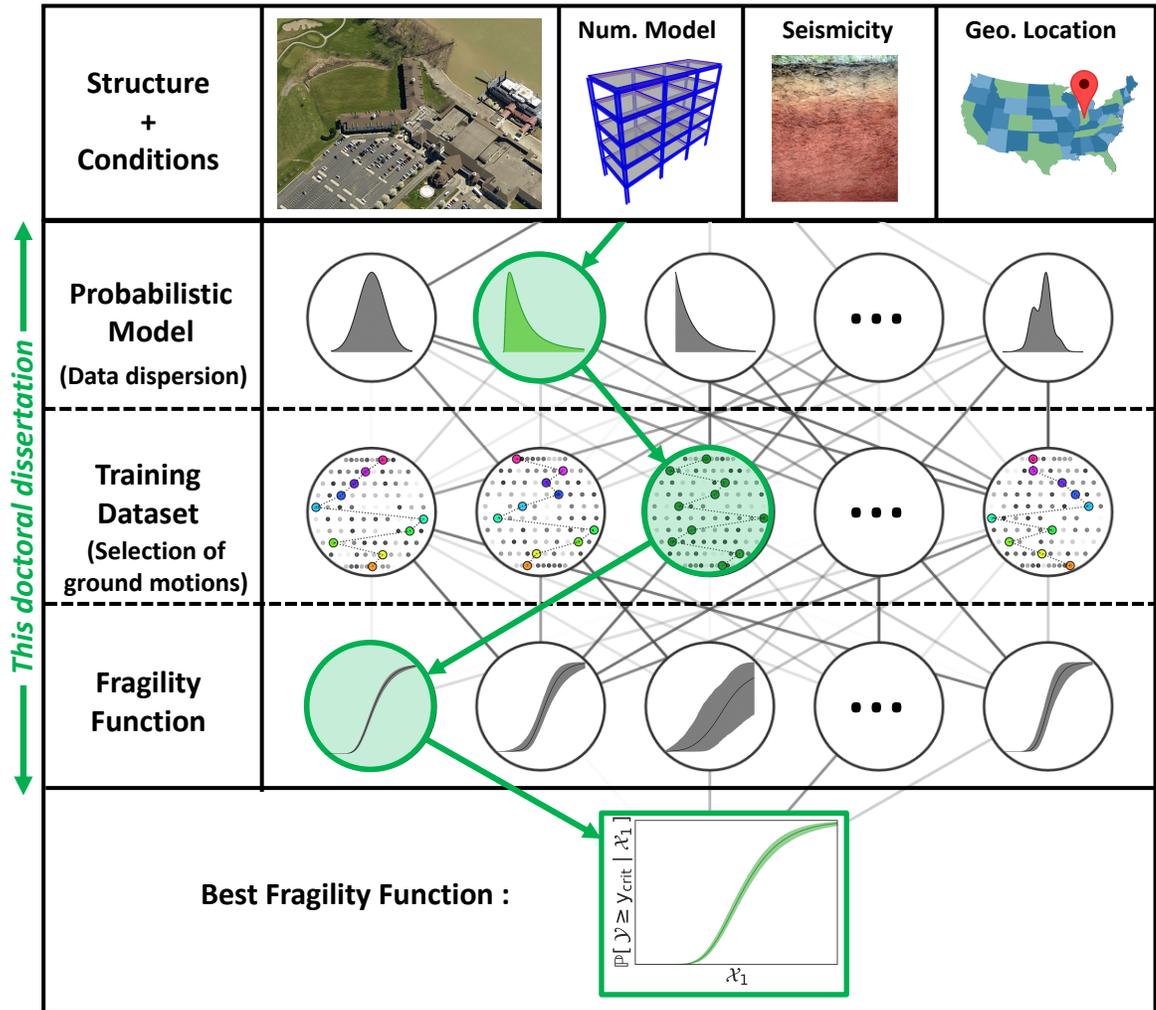
Despite the many advances made in quantifying uncertainty in FFs and understanding how sensitive the shape of a FF is to variations in its parameters, there are still limitations in determining the best alternative and data to compute unbiased FFs in the presence of uncertainty. Analytical FFs are sensitive to the presence of uncertainty in the parameters used to describe the numerical model, the hazard, the structural response, the probability distribution, and even the method used to compute such functions. Additionally, there is only a vague idea about the number of ground motion simulations that is adequate for a fragility analysis and almost no knowledge about the influence of selecting a particular set of ground motions on the uncertainty of FFs. If all of this were not enough, the repeated evaluation of an expensive numerical model is not feasible under time constraints and limited computational resources. As a result, the analysis of all the assets at a community-level becomes inconceivable.

The methodologies developed in this dissertation are focused on breaking down these barriers and gaps by creating: (i) a comparative methodology to implement a Bayesian model selection approach and (ii) a ground motion selection strategy that predicts the consequences of including a new ground motion record in the analysis. Additionally, some of the common practices performed in a fragility analysis that result in bias in the FFs are discussed.

## 1.3 Objective and Scope

In this dissertation, the overarching objective is to develop a methodology to expedite the computation of accurate seismic FFs, accounting for the intrinsic uncertainty in the data used for their derivation. The developed methods will enable automated, reliable, and efficient computation of seismic FF. Rather than considering the FF as a deterministic function and working with the best estimate, the perspective taken here is Bayesian. With this viewpoint, it is possible to better understand the quality of the resulting FF by analyzing the uncertainty associated with the dataset used for their derivation. By leveraging statistical

techniques, such as Bayesian inference and *sequential Monte Carlo* (SMC), it is possible to quantify the epistemic uncertainty associated with a given choice of dataset, methodology, model selection, and to determine what additional data would be needed for more accurate and unbiased FFs.



**Fig. 1.2.** Overview of the fragility computation challenge

The principal contribution of this research is to establish a new paradigm for the computation of accurate FFs. By following the methodologies developed in this dissertation, readers will be able to establish a clear understanding of the statistical model and the seismic data that yields the most accurate FF. Fig. 1.2 presents an overview of the problem of

computing the FF for a specific structure with given conditions in the presence of multiple alternatives of statistical models as well as multiple or extensive datasets. The identification of this optimal path, marked in green in Fig. 1.2, enables the widespread use of the fragility analysis. This is achieved by addressing the aforementioned *research questions* (RQ):

- *RQ 1: What is the most appropriate model for FFs?*

The objective is to determine the statistical model that enables the most accurate FFs. The approach is to perform the Bayesian model selection methodology developed in this dissertation, to decide which model has the best representation according with the available data. The results are compared with the ones derived from an exhaustive analysis as it is the case of the Kolmogorov-Smirnov (KS) test. The conclusion is that the proposed Bayesian model selection methodology is able to determine the model with the largest level of agreement without requiring a large dataset, which is the disadvantage of methodologies as the KS test.

- *RQ 2: How to select the ground motions?*

The objective is to determine the database that yields accurate FFs. The approach is to demonstrate the importance of using representative and applicable ground motions to the specific case study structure by presenting examples of commonly used practices with the potential to bias the FFs. In particular three examples of common practices that may result in biased FFs are discussed: (i) using random ground motion databases; (ii) augmenting ground motion databases with famous historical time histories; and (iii) augmenting ground motion databases by scaling available ground motion records. The conclusion is obvious but important to disseminate, that the use of representative ground motions is imperative to accomplish accurate FFs and augmenting ground motion databases is a process that needs to be done carefully to avoid bias in the FFs.

- *RQ 3: How to compute accurate FFs with fewer simulations?*

The objective is to carefully select the ground motion records that enable and expedite the computation of accurate FFs. The approach is to follow a ground motion sequential selection strategy yielding the largest expected reduction in epistemic uncertainty

based on up to two IM parameters. The anticipated effect of specific values of IM associated with future observations are estimated with respect to the potential changes in both the already inferred model parameters and the reduction in epistemic uncertainty. The conclusion is that the ground motion selection strategy developed in this dissertation enables rapid estimation of the FFs for a given structure, while reducing the number of evaluations of that structure's numerical model that are needed.

The implementation of these methodologies will greatly improve the ability to perform rapid computations to obtain unbiased FFs.

#### 1.4 Overview

This dissertation is organized as follows. Chapter 2 introduces the available data to create the dataset to validate the methodologies developed in subsequent chapters. Two major ingredients are discussed within this chapter (i) a well-studied 20-story nonlinear benchmark building and its structural characteristics that are used as the case study, and (ii) synthetic ground-motions generated for this study that consider site-specific response effects for sites of classes C and D, according to the *National Earthquake Hazards Reduction Program* (NEHRP). In Chapter 3 several options are discussed as possible parameters used to characterize both the intensity of a ground motion and the response of the structure. Additionally, the limit states and the corresponding critical values of the response that are applicable to the case study structure are discussed. A selection of these values, denoted as, IM, EDP/DM, LS, and  $y_{\text{crit}}$  is established and the reasoning for each choice is provided. The Bayesian model selection methodology developed in this dissertation is presented in Chapter 4. Then, the mathematical definition of the FF and its statistical model using the BIF is presented, and the epistemic uncertainty measure is defined. This chapter concludes with the Bayesian model selection methodology and the corresponding model validation approach used to corroborate the model selection by means of a predictive interval analysis and the KS test. Chapter 5 introduces the methodology developed in this dissertation for the

sequential selection of ground motions. The extension of the concept of FF into a *fragility surface* (FS) under the BIF is explained. Two possible strategies for sequential selection are presented within this section. The selection of the next ground motion observation to be used is performed based on an evaluation of the characteristics of all ground motions to identify the optimal region, in terms of up to two IM values, that will lead to the largest reduction in epistemic uncertainty. Chapter 6 is dedicated to exposing commonly used practices that may lead towards bias in the resulting FFs and discussing and demonstrating the importance of using realistic ground motions. Some conclusions of the methodologies, the numerical results, and discussion of possible future work are presented in Chapter 7. A few of the derivations used here in as well as some of the results derived in the application of the methodologies are also contained in the Appendix.

## 2. AVAILABLE DATA

This chapter presents the data needed to carry out the application of the methodologies developed in this dissertation. The necessary elements are a case study structure, with its corresponding numerical model, and a sufficient amount of ground motion data to perform the fragility analyses. A well-known and widely-studied benchmark building model has been selected for the analyses and a set of 78,000 synthetic ground motions obtained for earthquake scenarios in California provide the available input data. The methodologies developed in this dissertation are independent of the numerical model, type of structure, and source of data to be analyzed. The only required data to implement the methodologies is a dataset of  $\mathcal{X}_1$  vs.  $\mathcal{Y}$  observations. Consequently, the structural model and the ground motions are exclusively employed to create a dataset of observations.

### 2.1 Case Study Structure

The necessary data to develop and demonstrate the methodology is derived from a widely-studied nonlinear benchmark model of a twenty-story building, originally designed for the SAC steel project [86]. Although it was never constructed, the building was designed to satisfy the seismic requirements for Los Angeles, California. The structure is composed of two basement levels (with floor-to-floor height of 3.65 m), a ground level (with height equal to 5.49 m), and nineteen typical floors (each with height of 3.96 m) for a total height of 80.73 m above ground level. The building was designed to sustain both vertical and lateral loads through a system of steel *Moment-Resisting Frames* (MRFs). The structure has five bays in the N-S direction and six in the E-W direction, with column-to-column distance of 6.10 m for both directions. The entire floor system is composed of A36 steel (248 MPa) while the columns are made of A50 steel (345 MPa).

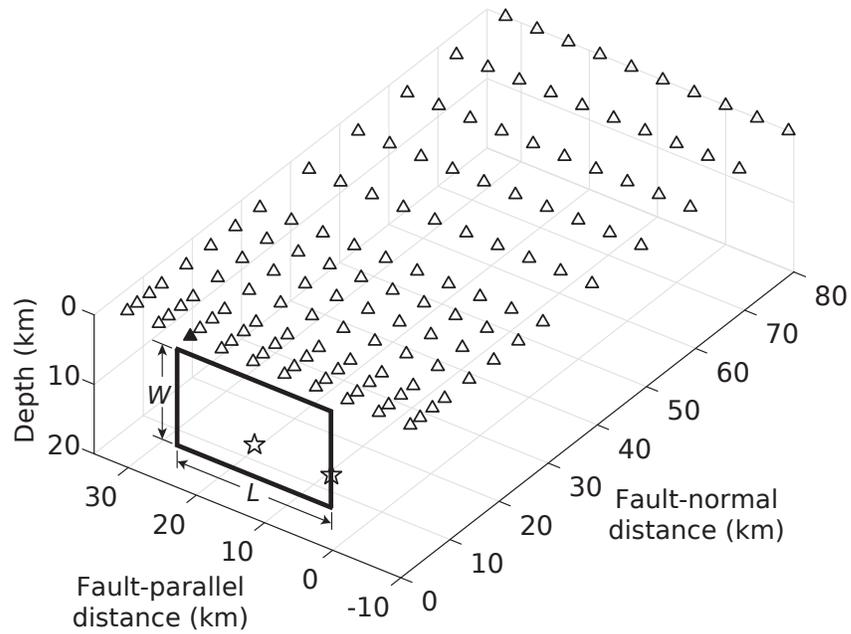
The numerical model of the benchmark building was developed by [87] and it was made available to the public by [88]. The dynamic analysis of the building is performed in the N-S direction (see Fig. 2.1). Additional information of the structural configuration about the building is presented in detail by [88]. The non-linearity is implemented through bi-linear hysteretic behavior that takes place at the plastic hinges (column-beam and column-column connections). The remainder of the structure is idealized as a twenty story shear model [88]. The first three natural frequencies (and damping ratios) of the model are 0.261 Hz (2.00%), 0.753 Hz (1.19%), and 1.30 Hz (1.34%).

## 2.2 Generating synthetic broadband ground motions

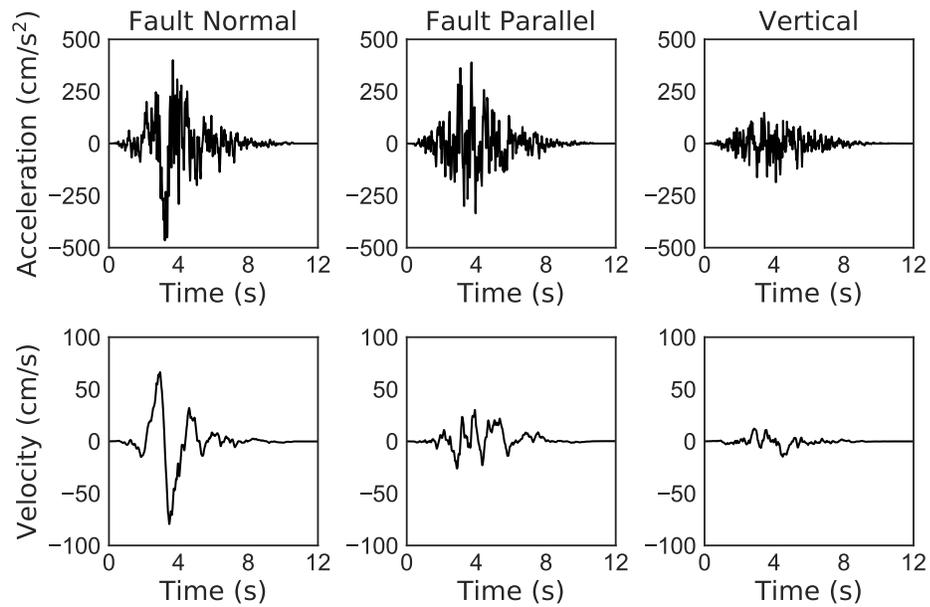
The broadband ground motions used in this study are generated using the simplified method proposed by [89]. The incoherent (high-frequency) component of ground motion is synthesized using the *specific barrier model* (SBM) [90] in the context of the stochastic modeling approach [91]. The SBM is a physical model of the seismic source that applies both in the near-field and far-field regions, allowing for consistent ground-motion simulations over a wide frequency range and for all distances of engineering interest. The SBM has been calibrated to shallow crustal earthquakes of three different tectonic regions: interplate, intraplate, and extensional regimes [92]. The coherent (long-period) component of ground motion (affected by forward rupture directivity) is generated using the simplified pulse model proposed by [89], which involves input parameters:  $A_P$ ,  $f_P$ ,  $\nu_P$ ,  $\gamma_P$ , and  $t_0$  that control the amplitude, prevailing frequency, phase, oscillatory character, and time shift of the pulse, respectively. In addition, the pulse period  $T_P$  is defined as the inverse of the prevailing frequency  $f_P$ .

The broadband ground motions are simulated for hypothetical earthquakes of five different moment magnitudes,  $M_w = 5.5, 6.0, 6.5, 7.0, 7.5$ . The causative fault is assumed to be a vertical, rectangular strike-slip fault with length  $L$ , width  $W$ , and burial depth of 0.1 km located in an interplate region such as California. Two hypocenter locations are considered: one at  $0.5L$  along strike and  $0.7W$  down-dip and the other at the same





**Fig. 2.2.** Earthquake fault (rectangle), grid of stations (triangles), and hypocenter locations (stars) considered in this study.



**Fig. 2.3.** Example of synthetic acceleration and velocity time histories for the station denoted by the filled triangle in Fig. 2.2 and for  $M_w = 6.5$ , a hypocenter on the edge of the fault, and NEHRP Site Class C.

are considered for site characterization. To account for the uncertainty in the amplitude of the white Gaussian noise used in the high-frequency simulations, fifteen realizations are considered for each station and for every combination of magnitude, hypocenter location, and site characterization. The long-period ground motion pulses are generated using the empirical relationships and practical guidelines proposed by [89,94,95]. Namely, amplitude  $A_P$  is determined by  $A_P = 0.9PGV$ , where the *Peak Ground Velocity* (PGV) is provided by  $\log(PGV) = 2.040 - 0.032r_{rup}$  ( $\sigma_{\log(PGV)} = 0.187$ ). The distance measure  $r_{rup}$  is defined as the closest fault-to-station distance. The pulse period  $T_P$  scales self-similarly with earthquake magnitude and is determined by  $\log(T_P) = -2.9 + 0.5M_w$  ( $\sigma_{\log(T_P)} = 0.12$ ). Parameter  $\gamma_P$  follows a normal distribution with a mean value of 1.93 and a standard deviation of 0.47. This distribution is left-truncated to 1.0 to ensure that all  $\gamma_P$  values are greater than 1. In addition, phase angle  $\nu_P$  is set equal to  $90^\circ$ . *Monte Carlo* (MC) sampling is used to generate input values for  $A_P$ ,  $f_P$ , and  $\gamma_P$ .

Not all 130 stations shown in (Fig. 2.2) experience long-period pulses caused by forward rupture directivity. [96] proposed a model for predicting the probability of observing a directivity pulse at a near-fault station for a given source-to-station geometry. Their model is adopted herein to evaluate whether the long-period component of ground motion should be included in a realization. If included, the long-period component is combined with the high-frequency component in the fault-normal direction, with the initiation of the pulse aligned with a time instance corresponding to 1% of the Arias intensity of the high-frequency component. Fig. 2.3 shows an example of synthetic acceleration and velocity time histories (with the long-period component of ground motion included) for a station affected by forward rupture directivity (displayed by a filled triangle in Fig. 2.2). It is noted that the broadband ground motions generated in this study do not include permanent displacements in the fault-parallel direction caused by dislocation across the fault surfaces, nor do they incorporate irregular subsurface basin effects or topographic amplification effects.

### 3. INPUT-OUTPUT PARAMETERS

This chapter presents the parameters needed to perform a fragility analysis and discusses the process for selecting such parameters. The chapter is divided into four sections. Section 3.1 defines the seismic problem and the parameters that control the FF. Although the normal procedure is to start by defining the input parameter, IM, it is actually more convenient to start defining the output parameter, DM or EDP, first (see Section 3.2) and use this selection to make a decision on the choice of IM. In Section 3.3, the definition of qualitative and quantitative LSs are discussed. Then, some examples of critical values associated with different LSs are discussed. Various parameters that may be used to characterize the ground motion are studied in Section 3.4, along with a procedure to select the most appropriate IM. Finally, the reader is directed to Appendix C to understand the ground motion database used here in terms of the selected parameters and the initial classification presented in Chapter 2 (magnitude, epicentral distance, ground motion orientation, and type of soil).

#### 3.1 Problem definition

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space composed of a sample space  $\Omega$ , a  $\sigma$ -algebra of subsets denoted by  $\mathcal{F}$ , and a probability measure  $\mathbb{P}$ . Any seismic event can be described using an acceleration record measured at the ground surface level in the geographical location of interest. This ground acceleration is modeled as a stochastic process  $\{a_t : \Omega \rightarrow \mathbb{R}\}_{0 \leq t \leq T}$ , where the RV  $T : \Omega \rightarrow \mathbb{R}_+$  is the duration of the event. For convenience, the shorthand notation  $a_t$  will be used throughout this work.

The seismic event  $a_t$  disturbs a structure of interest or an appropriate numerical model. As a result, a dynamic response  $\mathbf{Z}_{z,t} : \Omega \rightarrow \mathbb{R}^{\tilde{z} \times \tilde{t}}$  occurs, where the dimension  $\tilde{z}$  depends on the number of degrees-of-freedom and the number of responses considered (e.g., absolute acceleration, relative displacement), while  $\tilde{t}$  depends on the duration  $T$  and the time step of

the signal  $a_t$ . The response  $\mathbf{Z}_{z,t}$  is deterministically specified given the ground motion  $a_s$ ,  $s \leq t$ , when modeling uncertainty is not considered.

Let  $\mathbf{X} : \Omega \rightarrow \mathbb{R}^{n_x}$  be an  $n_x$ -dimensional set of IMs characterizing  $a_t$ . De Biasio [97] presented an extensive list of IM options that are commonly used to describe an ground motion. The list contains peak-based parameters like the *peak ground acceleration* (PGA) or velocity, PGV, duration-based terms such as the *Arias intensity* ( $I_a$ ), and frequency-based parameters as the *spectral acceleration* ( $S_a$ ) or *velocity* ( $S_v$ ). Further information about the selection of these IMs is discussed later in Section 3.4. Similarly, let  $Y : \Omega \rightarrow \mathbb{R}$  be a single-valued characteristic of the response of the structure  $\mathbf{Z}_{z,t}$ , e.g., relative displacement, relative velocity, absolute acceleration.

From this range of possibilities, only one IM, denoted by  $\mathcal{X}_1 \subseteq \mathbf{X}$ , where  $\mathcal{X}_1 \in \mathbb{R}$ , and one EDP,  $\mathcal{Y}$ , are used for presenting the FF for a given LS with a critical value  $y_{\text{crit}}$ :

$$F(x; y_{\text{crit}}) := \mathbb{P}[\mathcal{Y} \geq y_{\text{crit}} \mid \mathcal{X}_1 = x]. \quad (3.1)$$

Later, a second IM,  $\mathcal{X}_2$ , is also considered for the sequential selection strategy in two dimensions where is necessary to compute fragility surfaces.

### 3.2 Structural response parameter: Engineering demand parameter (EDP)

The selection of  $\mathcal{Y}$  is made on a case-by-case basis because it depends on the type of analysis to be performed and the quantity that is most relevant (e.g., damage on local members, entire structure, or non-structural components). For instance, structural features expressed in terms of forces, stresses, and deformation for building components are associated with local effects, while features such as roof relative displacement or absolute acceleration are typically linked to the building global behavior [97]. Additionally, the selection of the structural feature  $\mathcal{Y}$  may consider additional factors related to regions of interest within the structure. For example, the *maximum inter-story drift* (MISD) or simply *drift* of a building, defined as the ratio between maximum relative displacement of adjacent floors to the associated floor height (see Eq. (3.2)) is typically used to describe global

damage among structural members, while floor absolute acceleration is used to characterize damage in non-structural components or content as office and laboratory equipment [97].

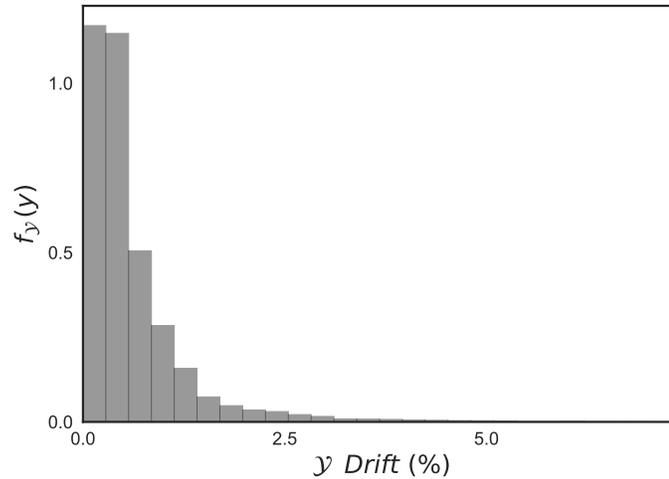
Given that there is no information about the content nor the non-structural components within structure, the focus here is in determining the FFs that represent the global damage that the building may suffer. For this reason, the structural response is characterized here in terms of the MISD. Additionally, multiple studies have determined that this parameter has high correlation with the presence of damage for multiple types of structures, materials and structural configurations (see Table 3.1). Assuming that  $Z_{i,t}$  represents the displacement of the  $i$ -th floor of the building at time  $t$ , the drift can be expressed as:

$$\mathcal{Y} = \max_{1 \leq i \leq L} \max_{1 \leq t \leq T} \left| \frac{Z_{i,t} - u_{i-1,t}}{H_i} \right| \quad (3.2)$$

where  $Z_{0,t} = 0$  (thus, there is no relative displacement between the base of the building and the ground surface),  $L$  is the number of stories, and  $H_i$  is the height of the columns for the  $i$ -th floor. Fig. 3.1 presents the distribution of the response  $\mathcal{Y}$  from the selected ground motions. From the figure it is observed that most of the ground motions analyzed generate low values of drift response in the structure. This behavior is linked to the large number of low magnitude ground motions included in database, as is often the case of real ground motions as well.

### 3.3 Structural response threshold: Limit States - LS

A LS refers to a particular level of structural performance or functionality for a given asset. This criterion is associated with a specific threshold value in the structural response  $\mathcal{Y}$  denoted as  $y_{\text{crit}}$ . Either qualitative or quantitative approaches are utilized to classify the performance of structures. The former measures specific changes in the material or geometric properties, such as *plastic mechanism initiation* (PMI) and *first yield* (FY), i.e., story deformation where a member starts to yield under an imposed lateral load [40]. The latter approach is based on specific operational states of the structure after the ground motion. For instance, *immediate occupancy* (IO), *life safety* (LSf) and *collapse prevention*



**Fig. 3.1.** *Distribution of drift parameter  $\mathcal{Y}$*

(CP). IO refers to a state in which the asset has sustained no significant structural damage after the occurrence of the ground motion. This means that the structure is able to continue its normal functionality. LSf denotes a state of the structure in which the life of the occupants is preserved at the cost of the asset sustaining certain damage, however there is no expectation for collapse in the structure. Finally, CP is associated with a state in which the seismic excitation has generated excessive damage to the structure and there is a risk of collapse. The asset is barely able to sustain only gravity loads [17,40,98]. Additionally, [98] establishes some intermediate levels such as *damage control* (DC) and *limited safety range* (SR). The order of increasing damage is the following IO < DC < LSf < SR < CP.

In the case of non-structural components, there are other LS cases although the IO and LSf states are also applicable for these types of components. The additional states are *operational* (components operate in the same manner as they did prior to the event) and *hazard reduced range* (existence of falling hazard of minor components, but not from major elements) [40,98].

To associate a LS with a particular value of the structural response, [99] proposed the implementation of a damage index for RC structural components. The computation of this index involves factors such as the maximum deformation that the structure suffered

during the ground motion, the ultimate deformation observed in monotonic loading, yield and maximum strength, and the hysteretic energy per loading cycle [99]. Subsequent studies [98, 100, 101] included classification of damage in terms of the MISD for multiple materials and different structural configurations. The work developed by [101] included some acceptable ranges to identify the damage levels for steel MRF. All the limit values associated with each equivalent LS are presented in Table 3.1.

**Table 3.1.** Commonly used critical values of building response per limit state [98–101]

Limit State	Damage index [99]	Drift (%) [101]	Drift (%) [98, 100]							
			Concrete		Steel MRF		Masonry			Wood
			MRF	Wall	Unbraced	Braced	Unreinf (infill)	Unreinf (no-infill)	Reinf	Stud Wall
None	0.01-0.1	0-0.2	-	-	-	-	-	-	-	-
Slight	0.01-0.2	0.2-0.5	-	-	-	-	-	-	-	-
Light (IO)	-	0.5-0.7	1.0	0.5	0.7	0.5	0.1	0.3	0.2	1.0
Moderate	0.2-0.5	0.7-1.5	-	-	-	-	-	-	-	-
<b>Heavy (LSf)</b>	-	<b>1.5-2.5</b>	<b>2.0</b>	<b>1.0</b>	<b>2.5</b>	<b>1.5</b>	<b>0.5</b>	<b>0.6</b>	<b>0.6</b>	<b>2.0</b>
Major	0.5-0.85	2.5-5.0	-	-	-	-	-	-	-	-
Collapse(CP)	0.85-1.15	>5	4.0	2.0	5.0	2.0	0.6	1.0	1.5	3.0

The methods developed in this dissertation are applicable to other states, but are demonstrated using the three main classification IO, LSf, CP. For this fragility analysis, the LSf state has been selected as the controlling LS to compute the FFs and demonstrate the methodologies developed. This LS represents intermediate structural damage without compromising the life safety of the occupants [40]. Considering that the case study structure is an unbraced steel MRF and the FF will be computed for the life safety condition, Table 3.1 specifies that the appropriate threshold value of the response is  $y_{crit} = 2.5\%$  [98, 100].

### 3.4 Input parameter: Intensity Measure - IM

The selection of the IMs  $\mathcal{X}_1$  and  $\mathcal{X}_2$  to characterize the ground motion can be more challenging due to the wide range of possibilities. In 2014, [97] completed a literature review of different options for IMs used in seismic probability risk analysis. In that study, the author presented a list of several IMs used to describe the structural demand under seismic load. The author presented a detailed review of each IM found in the literature and

introduced a new IM known as *average spectral acceleration* ( $S_{a-R\%}$ , see the description in the following lines). According to [97], ground motion IMs can be classified into three big groups which are briefly described below:

- *Peak-based IMs*: This group of parameters reflect the maximum absolute values from the ground motion's time history records. Some examples of these IMs include:
  - Peak Ground Acceleration:  $PGA := \max | a_t |$  where  $| \cdot |$  represents the absolute value operator.
  - Peak Ground Velocity:  $PGV := \max | v_t |$  where  $v_t$  corresponds to the velocity that the ground experiences during  $a_t$ .
  - Peak Ground Displacement:  $PGD := \max | d_t |$  where  $d_t$  corresponds to the displacement that the ground experiences during  $a_t$ .
- *Frequency-based IMs*: This collection of parameters relates to the seismic response of a SDOF system. For this particular case, the SDOF considered system has the same dynamic properties as the first dynamic mode of the structure: natural frequency  $f_n = 0.26$  Hz and damping ratio  $\zeta_n = 2.0$  %. A few examples of these IMs are included:
  - Spectral Acceleration:  $S_a(f_n, \zeta_n) := \max | \ddot{u}(t) |$  for  $t \in [0, T]$  where  $\ddot{u}(t)$  represents the absolute acceleration response for a SDOF due to  $a_t$ .
  - Spectral Velocity:  $S_v(f_n, \zeta_n) := \max | \dot{u}(t) |$  for  $t \in [0, T]$  where  $\dot{u}(t)$  is the relative velocity response of a SDOF due to  $a_t$ .
  - Spectral Displacement:  $S_d(f_n, \zeta_n) := \max | u(t) |$  for  $t \in [0, T]$  where  $u(t)$  is the relative displacement response of a SDOF due to  $a_t$ .
  - Average spectral acceleration:  $S_{a-R\%}(f_n, \zeta_n) = (1/R f_n) \int_{(1-R)f_n}^{f_n} S_a(f, \zeta_n) df$ . This IM was proposed by [97] as spectral measure that accounts for inflicted damage in the structure. Such damage is associated with an  $R\%$  of reduction of the initial natural frequency  $f_n$ .

For convenience in the notation, these frequency-based IMs are described as  $S_a$ ,  $S_v$ ,  $S_d$ , and  $S_{a-R\%}$  from now on.

- *Duration-based IMs*: This set of parameters captures not only the maximum magnitude, but also the cumulative effect of the ground motion during its entire duration (typically represented by integrating the whole duration of the acceleration record). Here are some examples of these IMs:

- Arias Intensity:  $I_A = (\pi/2g) \int_0^T a_t^2 dt$ .

- Cumulative Absolute Velocity:  $CAV = \int_0^T |a_t| dt$ .

- Root-mean-squared acceleration:  $a_{RMS} = \sqrt{(1/T) \int_0^T a_t^2 dt}$

In the analysis of IMs provided by [97] the following conclusions were provided: frequency-based IMs show superior correlation with damage in the structure, making of them a better choice for risk analysis (similar results were found in [102]);  $S_a$  is typically more efficient than PGA, but  $S_a$  does not account for the contribution of higher dynamic modes in the response nor the elongation of the natural period of the structure;  $S_a$  shows good correlation when the materials of the structure remain linear; and the use of PGA and PGV is recommended for structures with intermediate values of  $f_n$  while PGD is recommended for low frequency values.

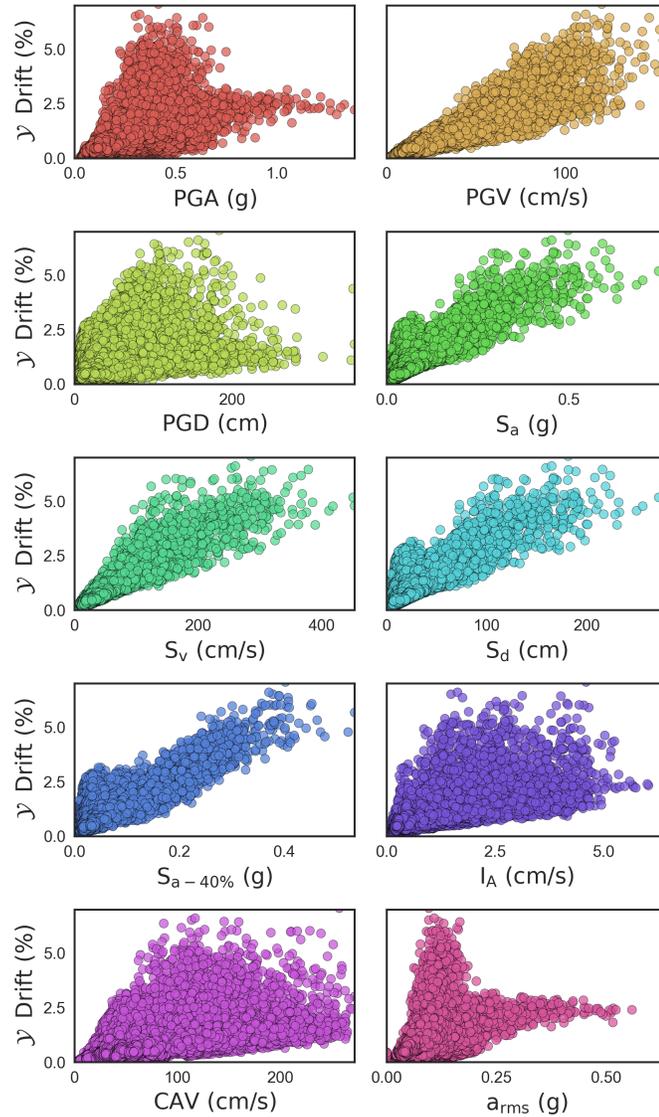
Although there is a wide range of IM alternatives, most of the studies found in the literature use PGA and  $S_a$  [11, 15, 16, 19, 25, 55]. However, it is important to consider all possible IMs so as not to bias the analysis with other researchers' assumptions or preferences. For this investigation, the ground motion database is composed of 600 realizations of ground motions measured at 130 different stations distributed in a perpendicular distance between 3-80 km from the fault surface (see Fig. 2.2), for a total of 78,000 records. Nonetheless, the records originated from the stations located up to 20 km away from the fault (measured perpendicularly) are accounted for implementation of the model and ground motion selection methodologies (600 motions  $\times$  60 stations = 36,000 records). The reason for this decision is that a reduced area needs to be considered in order to narrow the analysis into a specific region in order to obtain more realistic results. Later the remaining ground

motions are used to demonstrate the importance of using realistic ground motion for a specific geographical location.

Here, the initial vector of IMs is composed of all the aforementioned parameters  $\mathbf{X} := \{\text{PGA}, \text{PGV}, \text{PGD}, S_a, S_v, S_d, S_{a-40\%}, I_A, \text{CAV}, a_{\text{RMS}}\}$ . Fig. 3.2 presents the dataset obtained for the different IMs and the simulation of the structural response using the various ground motions. It is important to clarify that this figure is presented herein with the purpose of displaying how the dataset to be analyzed may be affected by the selection of IM. However, the figure is not used to select the IM. An assumption in this dissertation is that the potential users do not have access to a large database and/or there are no sufficient computational resources to carry out such volume of simulations.

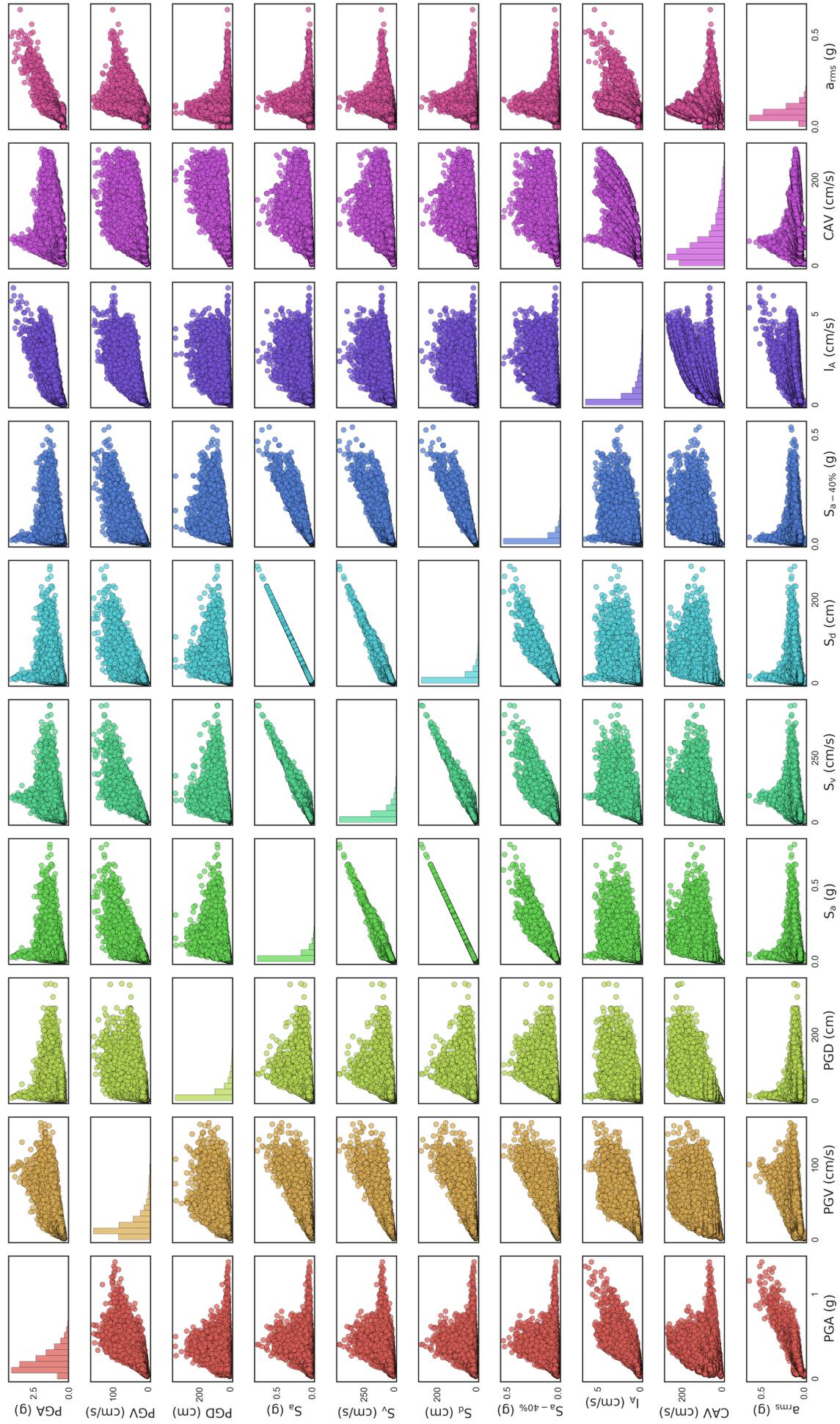
In the first analysis performed the goal is to determine which parameters are directly correlated with each other. The procedure is to select  $\mathcal{X}_1$  and  $\mathcal{X}_2$  and retain only the independent IMs and discard those that can be expressed in terms of others. Fig. 3.3 presents a scatter-matrix plot for all of the IMs contained in  $\mathbf{X}$ . The subplot located in the  $j$ -th row and the  $i$ -th column corresponds to the scatter plot  $X_i$  vs.  $X_j$ , where  $X_i$  is the  $i$ -th component of  $\mathbf{X}$ . Additionally, the main diagonal in Fig. 3.3 presents the normalized histogram for each of the elements in  $\mathbf{X}$  to be understood as an estimate of the PDF  $f_X$ . Fig. 3.3 has been divided into Figures B.1 to B.4 (see pages 111 to 114 from Appendix B) to see the amplified subplots. From the scatter-matrix plot, the following conclusions are drawn:

- Parameters  $S_a$ ,  $S_v$ , and  $S_d$  are highly correlated. This means that there is a clear, almost-linear dependency between these parameters. Thus, only one of the three parameters should remain in the analysis, and  $S_v$  is selected. Parameters  $S_a$  and  $S_d$  are discarded because there is a clear linear relationship between them, another form of spectral acceleration is kept  $S_{a-40\%}$ , and the  $S_v$  values contain more samples in the mid-range section (see histograms).



**Fig. 3.2.** Initial dataset of  $\mathbf{X}$  vs.  $\mathcal{Y}$

- Some patterns in the data points are observed for the  $I_A$ , CAV, and  $a_{RMS}$  cases. No conclusion can be drawn, but a further investigation is needed to understand these cases.
- The  $a_{RMS}$  IM presents some discontinuities at low intensity values. For this reason, this IM is removed from the analysis.



**Fig. 3.3.** Scatter matrix plot for X

The array of IMs is redefined with the remaining parameters,  $\mathbf{X} := \{\text{PGA}, \text{PGV}, \text{PGD}, S_v, S_{a-40\%}, I_A, \text{CAV}\}$ . For the next analysis, a sample population of 40 observations (ground motions) is chosen for making a decision about the parameters  $\mathcal{X}_1$  and  $\mathcal{X}_2$ . Ground motions are selected for NEHRP site class C and include one sample from all the possible combinations of the following parameters (i) Magnitude  $\rightarrow M_w = (5.5, 6.0, 6.5, 7.0, 7.5)$ ; (ii) Epicentral (normal) distance  $\rightarrow (3.0, 5.0, 7.5, 10.0)$  in km; and (iii) motion orientation to be either parallel or perpendicular to the fault. Additionally, a sample of size 40 is selected because [22,67,68] have agreed that this number represents a lower bound for an acceptable number of observations in fragility analysis. In this way, it is possible to assume that a representative sample of the entire population of ground motions is selected. The following analysis requires to evaluate the magnitude of the standard deviation of the IM vs. EDP plot in the logarithmic space. According to [31] this standard deviation is proportional to the imposed uncertainty in the FF. The value of the logarithm standard deviation  $\sigma$  is determined following the process described in Section 1.1.4.

The results of the linear regression are presented in Fig. 3.4. For each IM, two plots are presented: the linear regression results before and after applying a logarithmic transformation. The black dashed line represents the median function (mean in the log-space)  $\exp(\mu(x)) = c x^\gamma$ , the gray area corresponds to 95% predictive intervals (a.k.a., confidence intervals) computed as  $\mu \pm 2\sigma$ , and the dots are the 40 observations. The bar plot located in the bottom right corner represents the magnitude of  $\sigma$  for each IM. It can be observed that the IM alternatives that should reduce the amount of uncertainty are PGV and  $S_v$ , with  $\sigma$  values of 0.246 and 0.222, respectively.

Now, the objective is to choose the principal and the secondary IMs,  $\mathcal{X}_1$  and  $\mathcal{X}_2$ , respectively. Fig. 3.5 represents the distribution of the ground motions in terms of the two selected IMs. It is important to clarify that this figure can be used to make decisions because it is derived without performing any simulations on the expensive structural model. The red dashed lines represents the IM threshold values at which the observations start to become limited for each IM. The threshold values are  $\text{PGV} \approx 125$  cm/s and  $S_v \approx 100$  cm/s. These lines divide the entire domain into three zones. Zone (1) represents the region

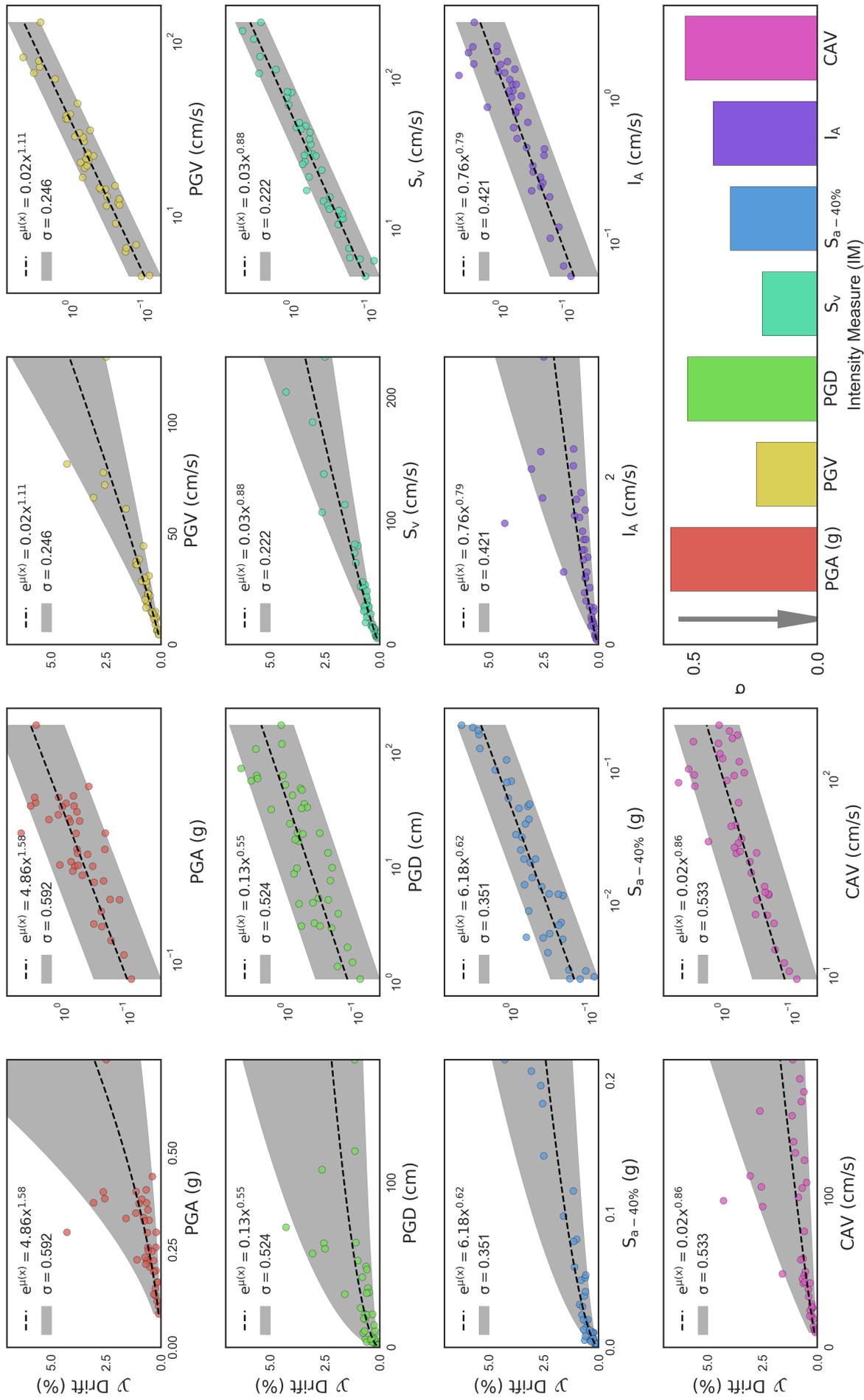
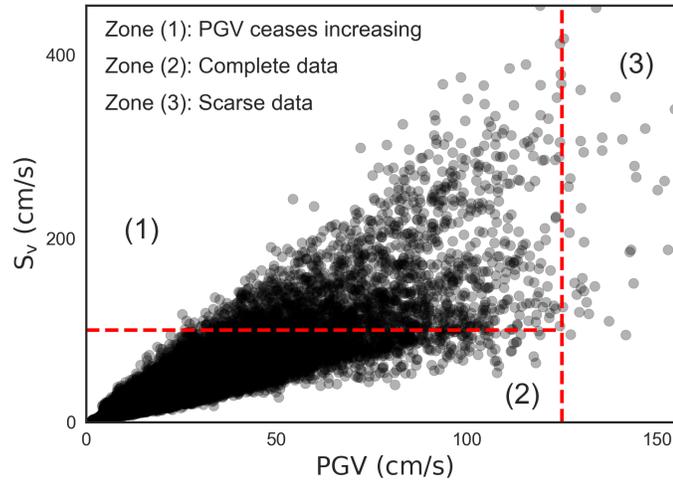


Fig. 3.4. Linear regression for  $X$  vs.  $Y$  ( $\mu$ ,  $\sigma$ ) in log-space



**Fig. 3.5.** *Distribution of PGV vs.  $S_v$*

in which  $S_v$  continues increasing but  $PGV$  has reached its threshold. This means that an estimation of  $f_{S_v}(x|PGV)$  is more reliable than  $f_{PGV}(x|S_v)$ , which is one of the fundamental components of the sequential selection strategy from Chapter 5. Zone (2) corresponds to the region in which both IMs exhibit increasing behavior. And, the available data in zone (3) is very limited which results in a large uncertainty when one wants to infer any type of information in this region. According to this reasoning, it is determined that  $\mathcal{X}_1 \rightarrow PGV$  since it is possible to account for a larger number of observations in the analysis and a reliable estimation of  $f_{S_v}(x|PGV)$ . Correspondingly, the secondary IM  $\mathcal{X}_2 \rightarrow S_v$ . To illustrate the effect of ground motion features such as earthquake magnitude, NEHRP soil class, wave orientation, and epicentral distance, plots of the database ( $\mathcal{X}_1$  vs.  $\mathcal{Y}$  data points) with respect to these parameters can be observed in Appendix C (see page 115).

## 4. BAYESIAN MODEL SELECTION

This chapter presents the Bayesian model selection methodology developed to quantify the level of agreement between a given dataset and different statistical models with the purpose of determining the model that enables the most accurate FFs. The methodology developed in this dissertation has been already introduced in [30, 103] for a different ground motion database. The results presented in this chapter exhibits quite similar trends in the results and conclusions. This chapter begins by presenting the Monte Carlo approximation of a FF for seismic hazards and its challenges. The conventional model of FF is introduced, followed by the required conditions that the models must satisfy. Then, a sample set of statistical models of FF for seismic data and their corresponding parameters are presented. This set of models is presented to facilitate an understanding of the model selection methodology, but it could be complemented with additional models. Next, Bayesian inference and the corresponding calibration of model parameters with respect to the observed data are presented in detail. The concept of epistemic uncertainty for this study is also defined for predicting fragility with a given model. The methodology discussion concludes with the Bayesian model selection method and the corresponding model validation approach. The latter corroborates the model selection by means of a predictive interval analysis and a *Kolmogorov-Smirnov* (K-S) *test* [104]. An application to illustrate the methodology is presented using data derived from a well-known twenty-story nonlinear benchmark building [86, 88] subjected to a ground excitation.

### 4.1 Monte Carlo approximation of the FF

The simplest approximation of Eq. (1.1) is via MC sampling. This approach performs a point-wise approximation of the function  $F(x; y_{\text{crit}})$ . Specifically for a given ground motion characteristic  $\mathcal{X}_1 = x$ ,  $S$  possible ground motion are generated  $\{a_t^{(s)} : 0 \leq t \leq T^{(s)}\}_{s=1}^S$

compatible with  $x$ . For each one of these ground motion  $a_t^{(s)}$ , the structural response  $y^{(s)}$  is computed evaluating the expensive numerical model to obtain  $\{y^{(s)}\}_{s=1}^S$ , i.e., samples of  $\mathcal{Y}$  conditioned on  $\mathcal{X}_1 = x$ . Then, Eq. (1.1) can be approximated by:

$$F(x; y_{\text{crit}}) \approx \frac{1}{S} \sum_{s=1}^S \mathbb{1}_{[y_{\text{crit}}, +\infty)}(y^{(s)}). \quad (4.1)$$

The challenges of the MC approximation are that: (i) it has to be applied to each  $x$  at which the fragility value is required; (ii) it requires the ability to sample that are compatible with  $\mathcal{X}_1 = x$ ; (iii) it requires a large number of samples  $S$  in order to yield convergent statistics (especially for low probability events). These challenges, in combination with the computational cost of non-linear structural simulations, make MC impractical for calculating the FF in complex cases.

## 4.2 Learning the FF from limited data

It is possible to learn  $F(x; y_{\text{crit}})$  from a finite number of observations. Specifically, let  $\{a_t^{(n)}\}_{n=1}^N$  be  $N$  independently-drawn sample time series, e.g., randomly picked from a larger dataset of real or synthetic ground motions. For each  $n = 1, \dots, N$ , the ground motion characteristic  $x^{(n)}$  is evaluated, the corresponding structural response is computed, and the response characteristic  $y^{(n)}$  is determined.

The goal is to use the observed seismic data  $\mathcal{D}_N = \left\{ \left( x^{(n)}, y^{(n)} \right) \right\}_{n=1}^N$  (a.k.a., *training dataset*) to learn the PDF of  $\mathcal{Y}$  conditioned on  $\mathcal{X}_1 = x$ , denoted  $f_{\mathcal{Y}}(y | x)$ . This is sufficient according with Eq. (1.1).

### 4.2.1 Conventional model of FF

Consider a set of  $M$  models  $\mathcal{M}_1, \dots, \mathcal{M}_M$  from which it is desired to determine the one with the best representation of  $\mathcal{D}_N$ . Each model  $\mathcal{M} \in \{\mathcal{M}_1, \dots, \mathcal{M}_M\}$  defines a possible functional form for  $f_{\mathcal{Y}\mathcal{M}}(y | x) := f_{\mathcal{Y}}(y | x, \mathcal{M})$  in terms of a set of unknown statistical parameters (e.g., location and shape parameter). For example, the most common model

selection in seismic fragility studies uses the lognormal distribution with linear trend and constant variance [11,41], and the model's functional form is:

$$\begin{aligned} f_{\mathcal{Y}\mathcal{M}}(y | x, \mathbf{c}_{\mathcal{M}}) &= \log \mathcal{N}(y | \mu_{\log}(x; \mathbf{c}_{\mathcal{M}}), \sigma_{\log}(x; \mathbf{c}_{\mathcal{M}})) \\ &= \frac{1}{y\sigma_{\log}(x; \mathbf{c}_{\mathcal{M}})\sqrt{2\pi}} \exp\left(\frac{-(\log(y) - \mu_{\log}(x, \mathbf{c}_{\mathcal{M}}))^2}{2\sigma_{\log}^2(x; \mathbf{c}_{\mathcal{M}})}\right) \end{aligned} \quad (4.2)$$

where  $\mathbf{c}_{\mathcal{M}} \in \mathbb{R}^{q_{\mathcal{M}}}$  is the vector of unknown coefficients to be estimated from the data,  $\mu_{\log}(x; \mathbf{c}_{\mathcal{M}}) = c_1 + c_2 x$  is the mean and  $\sigma_{\log}^2(x; \mathbf{c}_{\mathcal{M}}) = c_3$  is the variance of the natural logarithm for parameter  $\mathcal{Y}$ . For this conventional model,  $\mathbf{c}_{\mathcal{M}} = (c_1, c_2, c_3)$  and  $q_{\mathcal{M}} = 3$ .

#### 4.2.2 Prior knowledge about the model

For the case of undamaged structures subjected to seismic input, it is certain that  $\mathbb{E}[\mathcal{Y} | \mathcal{X}_1 = 0] = 0$  and  $\mathbb{V}[\mathcal{Y} | \mathcal{X}_1 = 0] = 0$ , where  $\mathbb{V}[\cdot | \cdot]$  represents the conditional variance. Similarly, previous studies have shown that the expectation in the response can be assumed to be

$$\mathbb{E}[\mathcal{Y} | \mathcal{X}_1 = x] = m(x) = x^{\gamma_m} h_m(x) \quad (4.3)$$

where  $\gamma_m$  is a non-negative exponent and  $h_m(x)$  is a positive correction term typically assumed to be a constant value [11,39,41,58]. Likewise, it is possible to define the standard deviation as

$$\sqrt{\mathbb{V}[\mathcal{Y} | \mathcal{X}_1 = x]} = s(x) = x^{\gamma_s} h_s(x). \quad (4.4)$$

Moreover,  $m(x)$  and  $s(x)$  need to be monotonically increasing functions because it is expected that both the average value and the variance of  $\mathcal{Y}$  increases when magnitude  $\mathcal{X}_1$  increases (i.e.,  $dm(x)/dx \geq 0 \leq ds(x)/dx$ ).

Additionally, the correction terms  $h_m(x)$  (Eq. (4.3)) and  $h_s(x)$  (Eq. (4.4)), can be expanded using a set of  $q_m$  and  $q_s$  basis functions, respectively. For convenience the subscript  $m - s$  is used to simplify notation of about the different number of basis functions  $q_m$  and

$q_s$ . In particular,  $\log(h_{m-s}(x))$  can be expanded in a set of  $q_{m-s}$  basis functions  $\{\varphi_i(x)\}_{i=1}^{q_{m-s}}$ , which is equivalent to

$$h_{m-s}(x) = \exp\left(\sum_{i=1}^{q_{m-s}} c_{\mathcal{M}i} \varphi_i(x)\right). \quad (4.5)$$

The selection of the basis functions needs to be performed on a case-by-case basis to keep their number to a minimum. Alternatively, a generic set of basis function can be used, e.g., step, radial basis, polynomial, or sine functions. For reasons of numerical stability, a set of orthogonal polynomials has been selected to represent  $\{\varphi_i(x)\}_{i=1}^{q_{m-s}}$ . The polynomials  $\varphi_i(x)$  and  $\varphi_j(x)$  are orthogonal if the inner product  $\langle \varphi_i, \varphi_j \rangle = 0$  for all  $i \neq j$ , where the inner product is defined as

$$\langle \varphi_i, \varphi_j \rangle = \int_{\mathbb{R}} \varphi_i(x) \varphi_j(x) g(x) dx \quad (4.6)$$

and  $g(x)$  is a weight function describing the  $\mathcal{X}_1$ -domain. The orthogonal polynomials can be estimated numerically using the ORTHPOL package [105] via the Python interface developed by [106].

### 4.2.3 Models of FFs

For this study, five models, each with a different but well-known PDFs, have been adopted to explain the methodology, including

- $\mathcal{M}_1$  (normal) :  $f_{Y\mathcal{M}_1}(y | x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$
- $\mathcal{M}_2$  (lognormal) :  $f_{Y\mathcal{M}_2}(y | x, \mu_{\log}, \sigma_{\log}^2) = \frac{1}{y\sigma_{\log}\sqrt{2\pi}} \exp\left(\frac{-(\log(y)-\mu_{\log})^2}{2\sigma_{\log}^2}\right)$
- $\mathcal{M}_3$  (log-student T) :  $f_{Y\mathcal{M}_3}(y | x, \mu_{\log}, \sigma_{\log}^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma_{\log}y} \left(1 + \frac{1}{\nu} \left(\frac{\log(y)-\mu_{\log}}{\sigma_{\log}}\right)^2\right)^{-\frac{\nu+1}{2}}$
- $\mathcal{M}_4$  (gamma) :  $f_{Y\mathcal{M}_4}(y | x, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$
- $\mathcal{M}_5$  (beta) :  $f_{Y\mathcal{M}_5}(y | x, \alpha, \beta) = x^{\alpha-1} (1-x)^{\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$

where the parameters  $\mu, \mu_{\log} \in \mathbb{R}$  and  $\sigma, \sigma_{\log}, \nu, \alpha, \beta \in \mathbb{R}^+$  are functions of  $x$  and  $\Gamma(\cdot)$  is the gamma function. Defining all of the statistical parameters as a function of  $x$  enables

$f_{y_M}(y | x)$  to take multiple forms. For instance,  $f_{y_M}(y | x)$  may adopt the form of the common model of linear trend and constant variance in Eq. (4.2) or more complex forms. In order to satisfy the conditions from equations (4.3) and (4.4), it is necessary to define the model parameters as follows:

- $\mathcal{M}_1$  :  $\mu = m(x)$ ,  $\sigma = s(x)$
- $\mathcal{M}_2$  :  $\mu_{\log} = \log \left( \frac{m(x)^2}{\sqrt{m(x)^2 + s(x)^2}} \right)$ ,  $\sigma_{\log}^2 = \log \left( \frac{m(x)^2 + s(x)^2}{m(x)^2} \right)$
- $\mathcal{M}_3$  : Similar to  $\mathcal{M}_2$  plus  $\nu = 2 + x^{\gamma_\nu} h_\nu(x) > 2$  (to have a finite variance)
- $\mathcal{M}_4$  :  $\alpha = m(x)^2/s(x)^2$ ,  $\beta = m(x)/s(x)^2$
- $\mathcal{M}_5$  :  $\alpha = \frac{m(x)^2(1-m(x))}{s(x)^2} - m(x)$ ,  $\beta = \alpha \frac{1-m(x)}{m(x)}$

The vector  $\mathbf{c}_M$  contains a total of  $q_M = 2q$  coefficients for the case of  $f_{y_M}(y | x)$  defined by two parameters (models  $\mathcal{M}_1$ ,  $\mathcal{M}_2$ ,  $\mathcal{M}_4$ , and  $\mathcal{M}_5$ ) or  $q_M = 3q$  when three parameters are used (model  $\mathcal{M}_3$ ).

#### 4.2.4 Bayesian inference of model parameters

Model  $\mathcal{M}$  is able to represent the data  $\mathcal{D}_N$  by calibrating the vector of coefficients  $\mathbf{c}_M$  and the exponents  $\boldsymbol{\gamma}_M = (\gamma_m, \gamma_s, \gamma_\nu)$ , where the exponent  $\gamma_\nu$  only applies for model  $\mathcal{M}_3$ . The vector  $\mathbf{c}_M$  corresponds to a  $q_M$ -dimensional RV with normal *prior probability density* with zero mean and unknown covariance matrix:

$$\mathbf{c}_M | \boldsymbol{\tau}_M \sim \mathcal{N}_{q_M}(\mathbf{0}_{q_M}, \boldsymbol{\Sigma}_M) \quad (4.7)$$

where  $\mathbf{0}_{q_M}$  is a  $q_M$ -dimensional zero vector,  $\boldsymbol{\Sigma}_M$  is a square diagonal matrix with  $\boldsymbol{\varsigma}_M^2$  as the main diagonal (assuming no *a priori* correlation between the parameters  $\mathbf{c}_M$ ), and  $\boldsymbol{\varsigma}_M^2 \in \mathbb{R}_+^{q_M}$  is the variance vector. Then, the magnitude for each element in  $\boldsymbol{\varsigma}_M^2$  is inferred from the data. To this end, an exponential *hyper-prior* is assigned to each component of  $\boldsymbol{\varsigma}_M^2$ :

$$\boldsymbol{\varsigma}_M^2 \sim E(\lambda_\varsigma = 0.1) \quad (4.8)$$

where  $E(\lambda_\zeta)$  is an exponential distribution with rate parameter  $\lambda_\zeta \in \mathbb{R}^+$  and represents the prior belief that the expected variance is *a priori* equal to 10. Additionally, the prior for the vector of exponents is

$$\boldsymbol{\gamma}_M \sim E(\lambda_\gamma = 1) \quad (4.9)$$

and its expected value is *a priori* equal to 1. The values for the rate parameters  $\lambda_\zeta$  and  $\lambda_\gamma$  were selected for the case of  $\mathcal{Y}$  corresponding to the maximum inter-story drift response, in which  $\mathcal{Y} \ll 1$ . But the rate parameters can be adjusted according with the magnitude of the response  $\mathcal{Y}$ . The vector  $\boldsymbol{\gamma}_M$  does not require a hyper-prior because its components can take any value due to the infinity support of the exponential distribution.

The objective is to find the parameter  $\boldsymbol{\theta}_M = (\boldsymbol{\gamma}_M, \mathbf{c}_M, \boldsymbol{\tau}_M)$  with prior probability density

$$p(\boldsymbol{\theta}_M) = p(\boldsymbol{\gamma}, \mathbf{c}_M, \boldsymbol{\tau}_M) = p(\mathbf{c}_M | \boldsymbol{\tau}_M) p(\boldsymbol{\tau}_M) p(\boldsymbol{\gamma}_M). \quad (4.10)$$

Since the ground motion observations are independent, the likelihood of the data under model  $\mathcal{M}$  is:

$$p(\mathcal{D}_N | \boldsymbol{\theta}_M) = \prod_{n=1}^N f_{\mathcal{Y}_M} \left( y^{(n)} | x^{(n)}, \boldsymbol{\theta}_M \right). \quad (4.11)$$

Using Bayes rule, the *posterior probability density* of the model parameters is:

$$p(\boldsymbol{\theta}_M | \mathcal{D}_N) = \frac{p(\mathcal{D}_N | \boldsymbol{\theta}_M) p(\boldsymbol{\theta}_M)}{p(\mathcal{D}_N)} \quad (4.12)$$

where the denominator is a normalization constant (a.k.a., the *evidence*)

$$Z_M = \int_{\boldsymbol{\theta}_M} p(\mathcal{D}_N | \boldsymbol{\theta}_M) p(\boldsymbol{\theta}_M) d\boldsymbol{\theta}_M. \quad (4.13)$$

The posterior of the model parameters is not analytically available and it needs to be obtained via sampling. The most common sampling algorithm is the *Markov-Chain Monte Carlo* (MCMC) [107, 108]. Unfortunately, MCMC is not sufficient for producing a robust estimate of  $Z_M$  (which is essential for implementing the Bayesian model selection technique) because all known estimators have an infinite variance [109]. SMC remedies this challenge

[110–113]. SMC is an algorithm that allows the gradual transformation of a particle approximation (weighted samples) from the prior to a particle approximation of the posterior.

This is achieved by following a one-parameter family of distributions of increasing complexity that continuously links the prior to the posterior. The resulting particle approximation of the posterior  $\left\{ \left( w_{\mathcal{M}}^{(s)}, \boldsymbol{\theta}_{\mathcal{M}}^{(s)} \right) \right\}_{s=1}^S$  is interpreted as

$$p(\boldsymbol{\theta}_{\mathcal{M}}) \approx \sum_{s=1}^S w_{\mathcal{M}}^{(s)} \delta \left( \boldsymbol{\theta}_{\mathcal{M}} - \boldsymbol{\theta}_{\mathcal{M}}^{(s)} \right) \quad (4.14)$$

where  $w_{\mathcal{M}}^{(s)}$  are normalized weights of the parameter samples  $\boldsymbol{\theta}_{\mathcal{M}}^{(s)}$ ,  $\delta$  is the Dirac delta function and  $S$  is the number of particles or samples. Eq. (4.14) is to be interpreted in the sense that for any sufficiently smooth function  $g$  of  $\boldsymbol{\theta}_{\mathcal{M}}$ ,

$$\sum_{s=1}^S w_{\mathcal{M}}^{(s)} g \left( \boldsymbol{\theta}_{\mathcal{M}}^{(s)} \right) \rightarrow \mathbb{E} [g(\boldsymbol{\theta}_{\mathcal{M}}) \mid \mathcal{D}_N] \quad (4.15)$$

almost surely as  $S \rightarrow +\infty$  [114]. In addition, SMC can produce an estimate of  $Z_{\mathcal{M}}$ . [115] presents a detailed explanation of the SMC algorithm along with the estimator of  $Z_{\mathcal{M}}$ .

#### 4.2.5 Predicting fragility under a given model

In this section, it is shown how the particle approximation of the posterior can be exploited to predict the FF with quantified epistemic uncertainty intervals. The FF under model  $\mathcal{M}$  is fully specified if  $\boldsymbol{\theta}_{\mathcal{M}}$  is known:

$$F_{\mathcal{M}|\mathbf{c}_{\mathcal{M}}}(x; y_{\text{crit}}) = \int_{y_{\text{crit}}}^{+\infty} f_{\mathcal{Y}_{\mathcal{M}}}(y \mid x, \boldsymbol{\theta}_{\mathcal{M}}) dy. \quad (4.16)$$

Using Eq. (4.16) and the particle approximation of the posterior, it is possible to obtain a particle approximation of the FF  $\left\{ \left( w_{\mathcal{M}}^{(s)}, F_{\mathcal{M}|\theta_{\mathcal{M}}^{(s)}}^{(s)}(x, y_{\text{crit}}) \right) \right\}_{s=1}^S$ , to be interpreted as

$$F_{\mathcal{M}|\mathcal{D}_N}(x; y_{\text{crit}}) = p(F(x; y_{\text{crit}}) | \mathcal{D}_N, \mathcal{M}) \approx \sum_{s=1}^S w_{\mathcal{M}}^{(s)} \delta \left( F(x, y_{\text{crit}}) - F_{\mathcal{M}|\theta_{\mathcal{M}}^{(s)}}^{(s)}(x, y_{\text{crit}}) \right). \quad (4.17)$$

The best prediction of the FF under model  $\mathcal{M}$  conditioned on the data  $\mathcal{D}_N$  is its expected value:

$$F_{\mathcal{M}|\mathcal{D}_N}(x, y_{\text{crit}}) = \mathbb{E} [F_{\mathcal{M}|\theta_{\mathcal{M}}}^{(s)}(x, y_{\text{crit}})] \approx \sum_{s=1}^S w_{\mathcal{M}}^{(s)} F_{\mathcal{M}|\theta_{\mathcal{M}}^{(s)}}^{(s)}(x, y_{\text{crit}}) \equiv F_{\mathcal{M}|\mathcal{D}_N}^{(S)}(x; y_{\text{crit}}). \quad (4.18)$$

### 4.3 Bayesian model selection

The objective of the Bayesian model selection approach [116] is to find the model  $\mathcal{M}$  with the best representation of  $\mathcal{D}_N$ . This can be understood as the model with the largest probability to replicate the data  $\mathcal{D}_N$ . It is possible to assign a prior  $p(\mathcal{M})$  to each model. Assuming there is no preference towards any of the  $M$  models results in a non-informative prior  $p(\mathcal{M}) = 1/M \quad \forall \mathcal{M} \in \{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ . Using Bayes rule, the posterior distribution of the model is:

$$p(\mathcal{M} | \mathcal{D}_N) = \frac{p(\mathcal{D}_N | \mathcal{M})p(\mathcal{M})}{p(\mathcal{D}_N)} = \frac{p(\mathcal{D}_N | \mathcal{M})p(\mathcal{M})}{\sum_{m=1}^M p(\mathcal{D}_N | \mathcal{M}_m)p(\mathcal{M}_m)}. \quad (4.19)$$

For the case when a non-informative prior is selected, Eq. (4.19) simplifies into

$$p(\mathcal{M} | \mathcal{D}_N) = \frac{p(\mathcal{D}_N | \mathcal{M})}{\sum_{m=1}^M p(\mathcal{D}_N | \mathcal{M}_m)} = \frac{Z_{\mathcal{M}}}{\sum_{m=1}^M Z_{\mathcal{M}_m}}. \quad (4.20)$$

Consequently, the best model is the one with the largest evidence  $Z_{\mathcal{M}}$  when there is no preference towards a subset of models.

### 4.3.1 Model Validation

Bayesian model selection is a useful technique, especially when there are constraints on the size of  $\mathcal{D}_N$ . However, models should be validated empirically. For this reason, an additional set of  $K$  observations is isolated from the design stage. The unobserved data  $\overline{\mathcal{D}}_K = \left\{ \left( x^{(k)}, y^{(k)} \right) \right\}_{k=1}^K$  (a.k.a., *testing dataset*) is exclusively used for the subsequent validation of the Bayesian model selection results.

#### Predictive interval validation

The selected model can be validated evaluating the testing dataset with some predictive intervals. The  $\rho$ -percentile of response  $\mathcal{Y} \mid \mathcal{X}_1 = x$  is defined such that

$$\mathbb{P} \left[ \mathcal{Y}(x) \leq \mathcal{Y}_{\mathcal{M}|\mathcal{D}_N}^{\rho \times 100\%}(x) \mid \mathcal{M}, \mathcal{D}_N \right] = \rho \quad (4.21)$$

where  $\mathcal{Y}_{\mathcal{M}|\mathcal{D}_N}^{\rho \times 100\%}(x)$  is the corresponding  $\rho$ -percentile which can be estimated empirically from a set of samples derived from the model. Similarly, the percentile can be obtained analytically using  $f_{\mathcal{Y}|\mathcal{M}}(y \mid x)$ . The 95% predictive interval for the response  $\mathcal{Y} \mid \mathcal{X}_1 = x$  is defined as  $\left[ \mathcal{Y}_{\mathcal{M}|\mathcal{D}_N}^{2.5\%}(x), \mathcal{Y}_{\mathcal{M}|\mathcal{D}_N}^{97.5\%}(x) \right]$ . Additionally, it is possible to determine the actual percentage of observations enclosed in this interval by counting the number of observations enclosed by the interval

$$\rho_{\overline{\mathcal{D}}_K, \mathcal{M}|\mathcal{D}_N} = \frac{1}{K} \sum_{k=1}^K \mathbb{1}_{\left[ \mathcal{Y}_{\mathcal{M}|\mathcal{D}_N}^{2.5\%}(x^{(k)}), \mathcal{Y}_{\mathcal{M}|\mathcal{D}_N}^{97.5\%}(x^{(k)}) \right]} \left( y^{(k)} \right). \quad (4.22)$$

A good choice of model corresponds to have a small error between the actual and the desired number of observations inside the predictive interval, which is equivalent to minimizing  $\left| \rho_{\overline{\mathcal{D}}_K, \mathcal{M}|\mathcal{D}_N} - 95\% \right|$ .

## Statistical agreement validation

Numerical validation requires determining if model  $\mathcal{M}$  statistically agrees with the testing dataset. This can be achieved by performing a K-S test [104]. The K-S test consists of comparing the samples from the testing dataset with the probability distribution defined by  $f_{\mathcal{Y}\mathcal{M}}(y | x)$ . This comparison is quantified in terms of the error between two CDFs, the theoretical CDF of the reference probability distribution using  $f_{\mathcal{Y}\mathcal{M}}(y | x)$  and the empirical distribution function using  $\overline{\mathcal{D}}_K$ . In order to perform the K-S test, it is necessary to transform all the observations into a common space for each particle of the posterior distribution  $\theta_{\mathcal{M}}^{(s)}$ . Let's define the  $\Upsilon$ -transformation of observation  $(x^{(k)}, y^{(k)})$  as the CDF of  $\mathcal{Y}$  evaluated at particle  $\theta_{\mathcal{M}}^{(s)}$

$$\Upsilon(x^{(k)}, y^{(k)}, \theta_{\mathcal{M}}^{(s)}) := \mathbb{P} \left( \mathcal{Y} \leq y^{(k)} \mid x^{(k)}, \theta_{\mathcal{M}}^{(s)} \right). \quad (4.23)$$

It is expected that the variable  $\Upsilon$  follows a standard uniform distribution  $U([0, 1])$ . Consequently, the K-S test is performed between the empirical CDF for  $\Upsilon$ ,  $\hat{F}_{\Upsilon}(v)$  and the CDF for  $U([0, 1])$ . The results from the K-S test can be illustrated through a *quantile-quantile* (Q-Q) plot [117].

Effectively, if the Bayesian model selection technique works properly, the selected model should be the one with the smallest difference between both CDFs,  $\hat{F}_{\Upsilon}$  and  $F_U(v) = v$ . This difference can be quantified in terms of either the peak error between the functions (a.k.a., *K-S distance*), or an overall or accumulated error between both CDFs, e.g.,  $L_2$  norm.

## Validation metrics

The most probable model, the one with largest  $Z_{\mathcal{M}}$ , can be validated through the following metrics:

- $\mathcal{J}_1$  : The magnitude of the error in the percentage of observations enclosed by the 95% predictive interval,  $|\rho_{\overline{\mathcal{D}}_K, \mathcal{M} | \mathcal{D}_N} - 95\%|$ .

- $\mathcal{J}_2$  : The K-S distance,  $\max_{0 \leq v \leq 1} |\hat{F}_Y(v | \overline{\mathcal{D}}_K, \theta_M) - v|$  (i.e., maximum error in the Q-Q plot).
- $\mathcal{J}_3$  : The Euclidean distance in the Q-Q plot,  $\sqrt{\sum \left( \hat{F}_Y(v | \overline{\mathcal{D}}_K, \theta_M) - v \right)^2}$  for  $v \in [0, 1]$ .

Finally, the objective is to determine if the selected model simultaneously minimizes  $\mathcal{J}_1$ ,  $\mathcal{J}_2$ , and  $\mathcal{J}_3$ .

#### 4.3.2 Epistemic uncertainty of FF

The Bayesian approach enables the quantification of all sorts of epistemic statistics for the FF induced by the limited size of  $\mathcal{D}_N$ . To this end, think of  $F(x; y_{\text{crit}})$  as a purely epistemic (lack of knowledge) RV and let  $p(F(x; y_{\text{crit}}) | \mathcal{D}_N, \mathcal{M})$  be the state of knowledge about it under model  $\mathcal{M}$  and after seeing the data  $\mathcal{D}_N$ . It is possible to estimate the epistemic uncertainty for  $F(x; y_{\text{crit}})$  using its particle approximation  $\left\{ \left( w_{\mathcal{M}}^{(s)}, F_{\mathcal{M}|\theta_{\mathcal{M}}^{(s)}}^{(s)}(x, y_{\text{crit}}) \right) \right\}_{s=1}^S$ .

Let  $A_{\mathcal{M}|\theta_{\mathcal{M}}^{(s)}}^{(s)}(y_{\text{crit}}) : \Omega \rightarrow \mathbb{R}$  be a single-valued characteristic of the FF evaluated at  $y_{\text{crit}}$  and associated with particle  $s$ . For instance,  $A_{\mathcal{M}|\theta_{\mathcal{M}}}$  can be the fragility at a given value of interest  $\mathcal{X}_1 = x_{\text{crit}}$  (i.e.,  $F_{\mathcal{M}|\theta_{\mathcal{M}}}(x_{\text{crit}}; y_{\text{crit}})$ ). Alternatively, this study adopts for this chapter  $A_{\mathcal{M}|\theta_{\mathcal{M}}}$  to be the area below the FF on its entire domain

$$A_{\mathcal{M}|\theta_{\mathcal{M}}^{(s)}}^{(s)}(y_{\text{crit}}) = \int_{x_{\min}}^{x_{\max}} F_{\mathcal{M}|\theta_{\mathcal{M}}^{(s)}}^{(s)}(x, y_{\text{crit}}) dx \quad (4.24)$$

where  $x_{\min}$  and  $x_{\max}$  are the minimum and maximum  $\mathcal{X}_1$  value in  $\mathcal{D}_N$ , respectively. The reason for this particular selection is that there is interest in quantifying the variation of the entire FF and not on a single data point. The epistemic uncertainty  $\Delta_{\mathcal{M}|\mathcal{D}_N}$  represents the

variation between the fragility samples. Thus,  $\Delta_{\mathcal{M}|\mathcal{D}_N}$  can be quantified using the variance for the characteristic  $A_{\mathcal{M}|\theta_{\mathcal{M}}}$

$$\begin{aligned} \Delta_{\mathcal{M}|\mathcal{D}_N}(y_{\text{crit}}) &:= \mathbb{V} \left[ A_{\mathcal{M}|\theta_{\mathcal{M}}}(y_{\text{crit}}) \mid \mathcal{D}_N \right] \\ &= \mathbb{E} \left[ A_{\mathcal{M}|\theta_{\mathcal{M}}}^2(y_{\text{crit}}) \mid \mathcal{D}_N \right] - \mathbb{E} \left[ A_{\mathcal{M}|\theta_{\mathcal{M}}}(y_{\text{crit}}) \mid \mathcal{D}_N \right]^2 \\ &\approx \sum_{s=1}^S w_{\mathcal{M}}^{(s)} A_{\mathcal{M}|\theta_{\mathcal{M}}^{(s)}}^2(y_{\text{crit}}) - \left( \sum_{s=1}^S w_{\mathcal{M}}^{(s)} A_{\mathcal{M}|\theta_{\mathcal{M}}^{(s)}}(y_{\text{crit}}) \right)^2. \end{aligned} \quad (4.25)$$

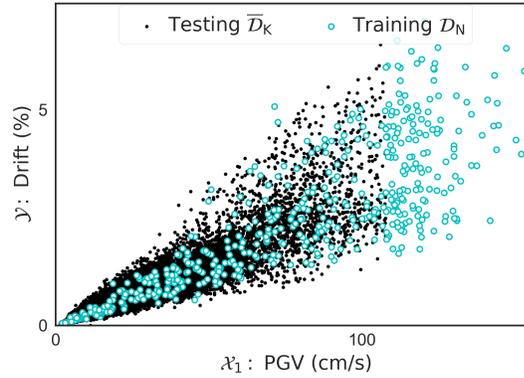
Finally, the epistemic uncertainty  $\Delta_{\mathcal{M}|\mathcal{D}_N}$  can be computed for datasets with different size and content (i.e., different ground motions) to determine the uncertainty induced by  $\mathcal{D}_N$ .

#### 4.4 Numerical Results

This section presents an illustrative application of the methodology section using data derived from case study and the ground motion database. This section starts by presenting the results for the five statistical models, Bayesian model selection, and the corresponding validation through a set of relevant metrics. Finally, the quantification of the epistemic uncertainty in FFs section shows some examples of FFs for the case study and the corresponding quantification of the epistemic uncertainty with respect to the size and content of the training data.

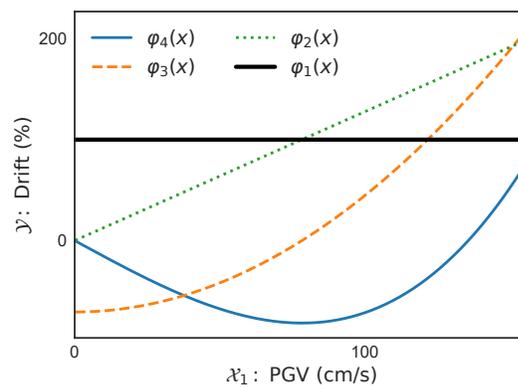
##### 4.4.1 Selecting the best model

Gehl et al. [22] found that the regression method with the lowest convergence rate requires at least 300 observations, a total of  $N = 500$  observations uniformly distributed in the  $\mathcal{X}_1$ -domain are selected for the analysis to avoid significant variation among the FFs. Fig. 4.1 shows all the available observations. The 500 light dots represents the training dataset,  $\mathcal{D}_N$ , while the remaining  $K = 35,500$  black dots constitutes the testing dataset  $\overline{\mathcal{D}}_K$ .



**Fig. 4.1.** *Data set of 36,000 observations*

The statistical models are defined by the functional forms described previously in the methodology section and their parameters vary according to the coefficients  $\mathbf{c}_M$ , exponents  $\gamma_M$ , and basis functions  $\{\varphi_i(x)\}_{i=1}^{q_m-s}$ . Considering the small variation and the level of smoothness observed in Fig. 4.1, a total of  $q_m = 4$  and  $q_s = 3$  polynomial functions were determined more than appropriate for the analysis. These polynomials are computed using the Python interface ORTHPOL [105] developed by [106]. The selected weight function  $g(x)$  for this analysis corresponds to the PDF of a normal RV with zero mean and standard deviation of 78.3 cm/s, which corresponds to half of the  $\mathcal{X}_1$ -domain,  $(0, 156.7]$  (cm/s). The basis functions implemented in the analysis are shown in Fig. 4.2.



**Fig. 4.2.** *Considered basis functions*

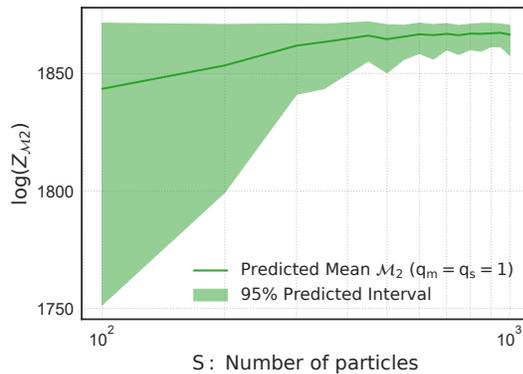
Before performing Bayesian model selection, it is necessary to determine the minimum number of basis function that each model requires. Starting with the basis function  $\varphi_1(x) = 1$ , polynomial functions of increasing order are added. Then, the coefficients  $\mathbf{c}_M$  and exponents  $\boldsymbol{\gamma}_M$  can be estimated after performing an SMC analysis with 750 particles. SMC analysis is performed using the PYSMC package developed by [112]. The performance of the models is initially evaluated in terms of the logarithmic value of the evidence,  $\log(Z_M)$ . Given that one of the outcomes from the SMC analysis is the approximation of the evidence, this analysis is repeated twenty times.

Table 4.1 depicts the results of this initial analysis in terms of the sample mean (and COV) for  $\log(Z_M)$  after repeating the analysis 20 times. The bold values represent the cases with the maximum evidence or the best representation of  $\mathcal{D}_N$ . Since models of increasing complexity are nested, the evidence value exhibits asymptotic behavior after a specific number of basis functions is included [118]. Similarly, Table 4.1 depicts the increasing variation in the evidence when more basis functions are used in the analysis. Therefore, it is determined that on average a constant correction term (i.e., only using the function  $\varphi_1(x) = 1$ ,  $q_m = q_s = 1$ ) yields to the best results for all the models but  $\mathcal{M}_5$  which requires an additional basis function to describe the standard deviation ( $q_m = 0$  while  $q_s = 3$ ). Additionally, it is necessary to validate that having  $S = 750$  particles generate accurate results. The *Coefficient of Variation* (COV) is a good estimate of the uncertainty in a measure since it represents the ratio of the standard variation and the mean value. The COV values presented in Table 4.1 show that the maximum COV for the best case of each probability distribution is 1.79 % corresponding to model  $\mathcal{M}_4$ . Nonetheless, SMC analysis is performed with different numbers of particles for the best model, the one with the largest evidence value,  $\mathcal{M}_2$ . The results are included in Fig. 4.3, where a variation of  $\pm 11$  units (0.6%) around the mean of  $\log(Z_M)$  found for model  $\mathcal{M}_2$  with a  $q_m = q_s = 1$  basis function and  $S = 750$  particles. Besides this level of variation, it is clear that  $\mathcal{M}_2$  is the best model for having the largest evidence. As a result, the Bayesian model selection technique shows that  $\mathcal{M}_2$  is almost surely the best model to describe the data  $\mathcal{D}_N$ . Additionally,  $\mathcal{M}_5$  shows

the second best correlation with  $\mathcal{D}_N$ . There is less than 1% of probability that any of the other models to have a better fit using this specific training dataset.

**Table 4.1.** Average (COV) log-value for the evidence,  $\log(Z_M)$ , for different number of basis functions

Likelihood	$q_s$	Number of basis functions: $q_m$			
		1	2	3	4
$\mathcal{M}_1$ : Normal	1	<b>1826.8 (0.34%)</b>	1826.0 (0.34%)	1816.9 (0.58%)	1814.2 (0.81%)
	2	1825.4 (0.41%)	1820.4 (0.51%)	1815.0 (0.64%)	1809.6 (0.67%)
	3	1824.4 (0.37%)	1805.3 (1.52%)	1809.8 (0.45%)	1803.8 (0.89%)
$\mathcal{M}_2$ : Lognormal	1	<b>1866.8 (0.14%)</b>	1861.4 (0.26%)	1845.0 (0.68%)	1824.0 (1.22%)
	2	1859.7 (0.40%)	1848.7 (0.46%)	1835.2 (1.20%)	1810.1 (1.73%)
	3	2836.8 (1.18%)	1801.2 (1.36%)	1803.0 (1.92%)	1764.3 (2.83%)
$\mathcal{M}_3$ : Log St. T	1	<b>1847.0 (0.54%)</b>	1824.8 (0.86%)	1811.6 (1.31%)	1803.7 (1.52%)
	2	1840.1 (0.87%)	1827.1 (1.15%)	1805.7 (1.11%)	1787.7 (1.06%)
	3	1827.6 (0.86%)	1813.0 (0.90%)	1811.0 (0.98%)	1770.1 (1.67%)
$\mathcal{M}_4$ : Gamma	1	<b>1834.0 (1.79%)</b>	1737.4 (7.45%)	1711.9 (5.97%)	1662.7 (7.65%)
	2	1756.4 (9.90%)	1711.7 (11.81%)	1532.2 (19.29%)	1251.7 (10.53%)
	3	1455.4 (33.66%)	1060.8 (39.04%)	998.7 (12.88%)	948.7 (13.01%)
$\mathcal{M}_5$ : Beta	1	1831.1 (2.09%)	1820.7 (2.75%)	1809.8 (2.41%)	1764.4 (3.63%)
	2	1794.5 (4.29%)	1841.9 (1.34%)	1840.8 (1.04%)	1837.0 (0.99%)
	3	<b>1852.6 (0.70%)</b>	1845.9 (1.13%)	1837.6 (1.08%)	1826.3 (1.44%)



**Fig. 4.3.** Variation of  $\log(Z_{M2})$  (for  $q_m = q_s = 1$ ) due to number of particles

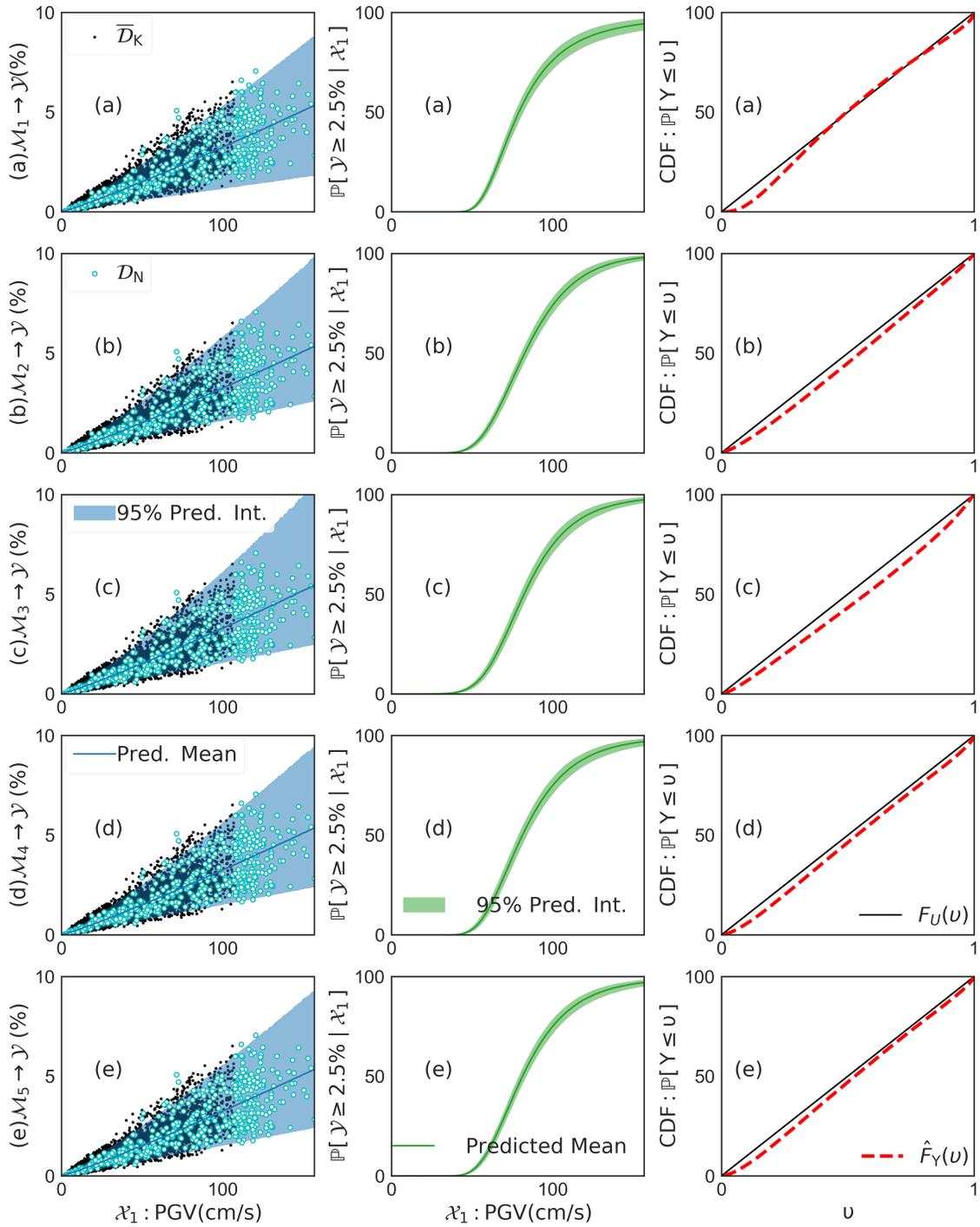
The methodology is illustrated using the five proposed models with one basis function (except  $\mathcal{M}_5$  that requires  $q_s = 3$ ) and 750 particles in the PYSMC package [112]. Fig. 4.4 depicts the results of this analysis. The left column of plots shows the entire dataset and the numerical predictions for the  $\mathcal{X}_1$ - $\mathcal{Y}$  response of each model. The line represents the

**Table 4.2.** Comparison metrics of the models

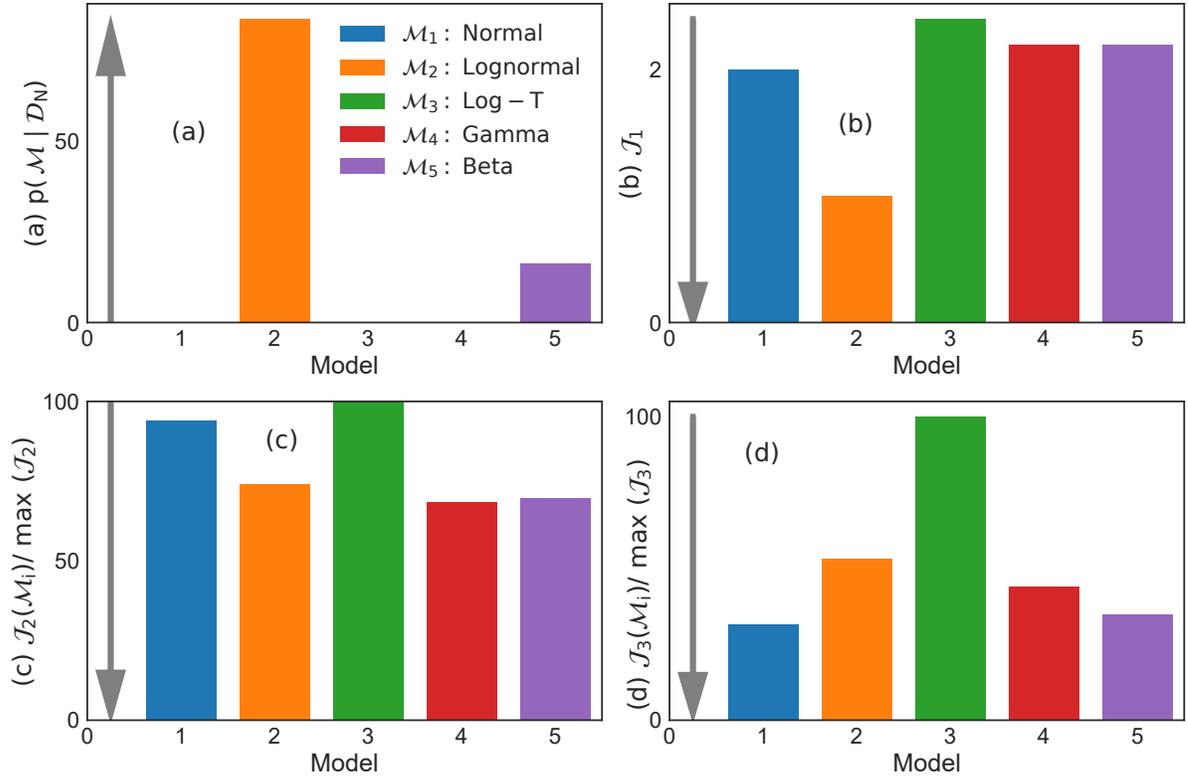
<b>Likelihood</b>	<b>Evidence <math>\log(Z)</math></b>	$\mathcal{J}_1$ [%]	$\mathcal{J}_2$	$\mathcal{J}_3$
$\mathcal{M}_1$ : Normal	1829.6	2.0	0.069@PGV=20.4 cm/s	<b>0.102</b>
$\mathcal{M}_2$ : Lognormal	<b>1867.2</b>	<b>1.0</b>	<b>0.054@PGV=48.6cm/s</b>	<b>0.171</b>
$\mathcal{M}_3$ : Log-T	1834.8	2.4	0.073@PGV=98.7cm/s	0.322
$\mathcal{M}_4$ : Gamma	1827.4	2.2	<b>0.050@PGV=28.2cm/s</b>	0.142
$\mathcal{M}_5$ : Beta	1865.6	2.2	0.051@PGV=26.6cm/s	0.112

expected value and the area is the 95% predictive interval. The predictive interval is numerically determined by finding the bounds that encloses 95% for a dataset of random samples generated for each model. A total of 45 millions random samples are generated to estimate the predictive intervals (200 values of  $\mathcal{X}_1 \times 300$  samples of  $\mathcal{Y}$  per value of  $\mathcal{X}_1 \times 750$  particles  $\theta_{\mathcal{M}}^{(s)}$ ) on each model. The middle column of the plots in Fig. 4.4 shows the corresponding FF for each model. Similar to the  $\mathcal{X}_1$ - $\mathcal{Y}$  response prediction, the line represents the expected FF and the area is the 95% predictive interval. Lastly, the right column of plots depicts the Q-Q plots from the K-S test. The black solid line is the theoretical CDF for the standard uniform distribution,  $F_U(v)$ , and the dashed line is the estimated CDF for the transformed parameter  $\Upsilon$ ,  $\hat{F}_{\Upsilon}(v | \overline{\mathcal{D}}_K, \theta_{\mathcal{M}})$ . This estimator is obtained after transforming the 35,500 observations from the testing dataset  $\overline{\mathcal{D}}_K$  for all the 750 particles of the posterior and using Eq. (4.23). In general (for Fig. 4.4) it is desired to have the 95% predictive interval enclosing as many observations as possible (left column), have a small predictive interval for the FFs (middle column), and have  $\hat{F}_{\Upsilon}(v | \overline{\mathcal{D}}_K, \theta_{\mathcal{M}})$  (dashed line on the right column) as close as possible to  $F_U(v)$  (black solid line).

Additionally, Table 4.2 summarizes the numerical results corresponding to the validation metrics  $\mathcal{J}_1$ ,  $\mathcal{J}_2$ , and  $\mathcal{J}_3$ , as well as  $\log(Z_{\mathcal{M}})$ . The results of all analyzed models (mean and standard deviation values for a population of twenty iterations in each model), for different probability distributions and number of basis function, can be found in Table E.1 in the Appendix (see page 121). Recall that a larger evidence represents a model with a better fit to  $\mathcal{D}_N$ . Similarly,  $\mathcal{J}_1 \rightarrow 0$  means that the percentage of observation inside the predictive interval (Fig. 4.4-left) is close to the desired 95%; it is preferred  $\mathcal{J}_2 \rightarrow 0$  because



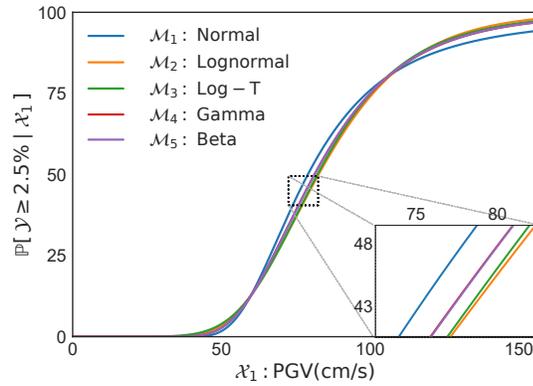
**Fig. 4.4.** Numerical results for models (a)  $M_1$ , (b)  $M_2$ , (c)  $M_3$ , (d)  $M_4$ , and (e)  $M_5$



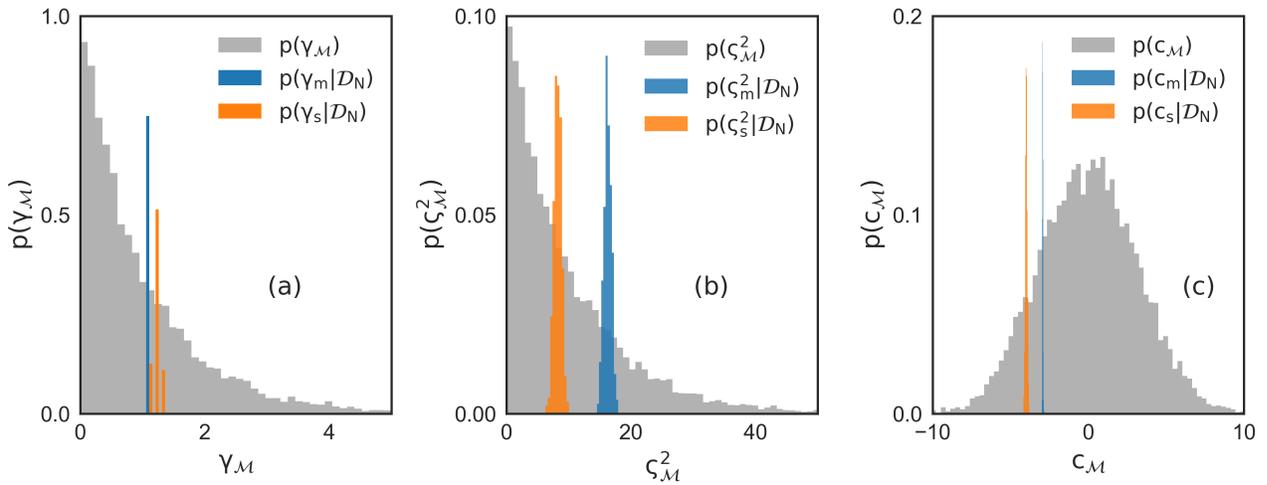
**Fig. 4.5.** Comparison metrics of models (a) Bayesian model selection, (b) Predictive interval, (c) K-S distance, and (d) Q-Q error

it represents a smaller K-S distance (i.e., maximum error in the Q-Q plot on the right column of Fig. 4.4); and it is desired that  $\mathcal{J}_3 \rightarrow 0$  since it is the Euclidean norm between  $F_U(\nu)$  and  $\hat{F}_Y(\nu | \overline{\mathcal{D}}_K, \theta_M)$  (proportional to the area between both CDFs). Similarly, these results are illustrated in Fig. 4.5 using bar plots. The top left (a) plot presents the results of the Bayesian model selection technique, the top right (b) plot shows the comparison metrics  $\mathcal{J}_1$ , the bottom left (c) presents the ratio of the K-S distance to respect to the maximum distance, and the bottom right (d) shows the ratio of the Q-Q error with respect to the maximum error. Just to point out, the gray arrow in each plot of Fig. 4.5 points to the desired magnitude for each metric. For instance, the gray arrow pointing up in plot (a) means that a larger value is desired because it represents a larger probability of being the most accurate model. Note that the evidence is determined with respect to training data  $\mathcal{D}_N$ , while the  $\mathcal{J}$  metrics are defined with respect to the testing data  $\overline{\mathcal{D}}_K$ . Additionally, Fig. 4.7 presents the comparison

of the prior and posterior distributions for the exponents  $\gamma_{\mathcal{M}}$ , the hyper-parameter for the variance  $\zeta_{\mathcal{M}}^2$ , and the coefficients  $\mathbf{c}_{\mathcal{M}}$  for model  $\mathcal{M}_2$ .



**Fig. 4.6.** Expected FFs for life safety limit state



**Fig. 4.7.** Prior/posterior comparison (a) Exponent  $\gamma_{\mathcal{M}}$ , (b) Variance  $\zeta_{\mathcal{M}}^2$ , and (c) Coefficients  $\mathbf{c}_{\mathcal{M}}$

The results presented in Table 4.2 and Fig. 4.5 show the superiority of model  $\mathcal{M}_2$  (lognormal). This model selection agrees with the assumptions from previous studies of lognormal dispersion and the capability of the constant term ( $q_m = q_s = 1$ ) to describe the seismic data [8, 11, 19, 41, 55]. It is important to mention that the Bayesian model selection determines that  $\mathcal{M}_2$  is undoubtedly superior to the rest of the models by only analyzing the

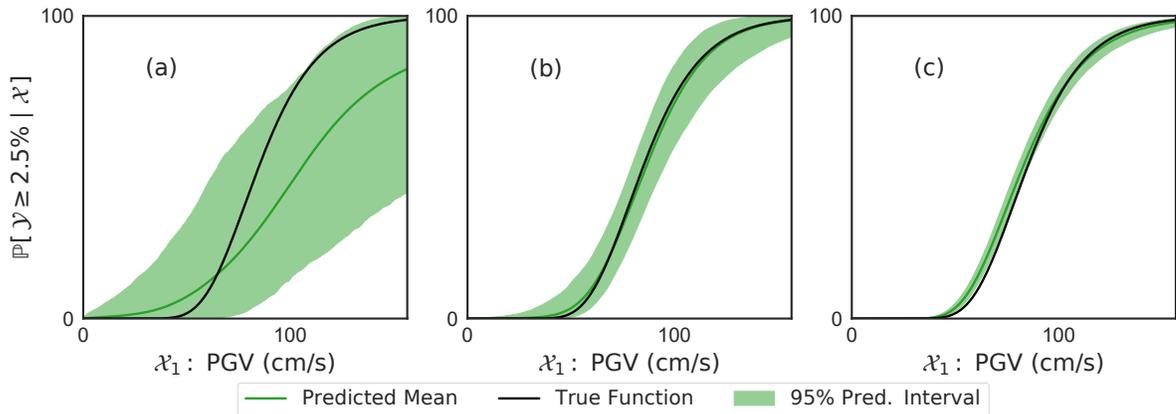
training dataset. The evidence of each model provides valuable information which can be used to determine the best model. For this illustrative application, it is possible to validate the Bayesian model selection because there is a sufficiently large dataset  $\overline{\mathcal{D}}_K$  available to perform a K-S test. However, this is not the case for most real-life applications. Instead reduced-size datasets are available. As a result, Bayesian model selection is a powerful technique that takes advantage of the information that every single data point provides.

Fig. 4.4 provides valuable information about this case study. It is possible to perceive small differences between the  $\mathcal{X}_1$ - $\mathcal{Y}$  responses of the different models (left). There are some differences between the median line and 95% predictive interval among models, which may cause one to think that the functional form  $f_{Y\mathcal{M}}(y | x)$  does not play a significant role. However, it is critical to recognize that these small differences in the  $\mathcal{X}_1$ - $\mathcal{Y}$  response do generate differences in the resulting FFs (middle column). Fig. 4.6 presents the mean of the FFs for each model and a closed-up plot to show the differences. For instance, the expected FF at ground motion intensity level PGV=83 cm/s (i.e.,  $\mathbb{P}[Y \geq 2.5\% | X = 0.17g]$ ) is approximately 57% for  $\mathcal{M}_1$  (normal), 50% for  $\mathcal{M}_2$  (lognormal), 51% for  $\mathcal{M}_3$  (log-T), 52% for  $\mathcal{M}_4$  (gamma), and 52% for  $\mathcal{M}_5$  (beta). This corresponds to a difference of up to 14% in the expected values, which may be even larger when considering the uncertainty of the predictive interval or a smaller number of observations. These differences between FFs may impact the decision-making process. For example, a stakeholder may decide to allocate too few resources to retrofit a structure based on an erroneously low fragility value compared to the true value. Such a situation motivates validating the analysis and suggests performing an additional analysis such as the K-S test, which confirms that the statistics of the model are really representative of the entire dataset.

#### 4.4.2 Quantification of the epistemic uncertainty in the FFs

Given the clear superiority of model  $\mathcal{M}_2$ , the uncertainty quantification of the FF is performed exclusively for this model. The SMC analysis gives a particle approximation for the FFs from which it is possible to estimate their 95% predictive intervals as an

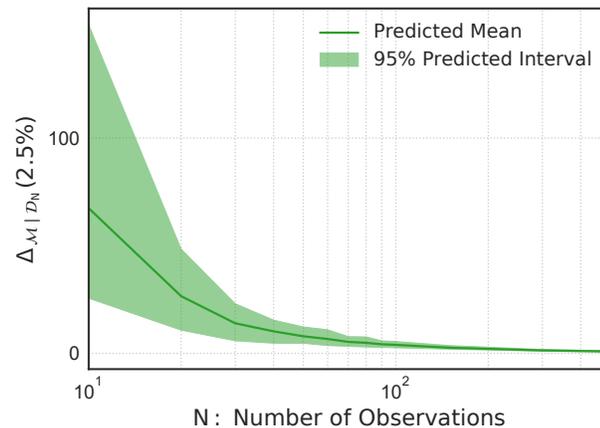
uncertainty measure. SMC analyses are performed using training datasets  $\mathcal{D}_N$  of different sizes,  $N \in \{10, 40, 500\}$ , in order to identify differences in the resulting FFs (see Fig. 4.8). Additionally, it is possible to characterize the convergence of these FFs (light lines) to the best estimate of FF (black line), which is the result of training the model with  $N = 10,000$  uniformly distributed observations. Fig. 4.8 demonstrates the variation in the FFs due to the size and the content of  $\mathcal{D}_N$ , which emphasizes the importance of quantifying their uncertainty. From the figure, it is possible to determine that using  $N = 10$  observations (see plot *a*) results in a predictive interval with big uncertainty and an expected FF that differs from the best estimate of the function. As a result, it can be concluded that using this set of  $N = 15$  observations is not sufficient to generate an accurate model. However, an expected FF similar to the best estimated function is observed for datasets with at least 40 observations (plot *b* and *c*).



**Fig. 4.8.** Comparison of FFs for: (a) 10, (b) 40, and (c) 500 observations

Here, the epistemic uncertainty  $\Delta_{\mathcal{M}|\mathcal{D}_N}$  in a particle approximation of FFs is quantified as the variance of the area below these functions. Consequently, each particular training dataset  $\mathcal{D}_N$  is associated with a specific value of  $\Delta_{\mathcal{M}|\mathcal{D}_N}$ . Therefore, the epistemic uncertainty is quantified only as a function of the size of  $\mathcal{D}_N$ . For this reason, the SMC analysis is repeated varying the size and the content of the training dataset. Different datasets with 10 to 500 observations are evaluated to characterize the variation in  $\Delta_{\mathcal{M}|\mathcal{D}_N}$  due to the size

$N$ . Similarly, the variation due to content is obtained by analyzing 100 different randomly selected datasets (i.e., using different ground motions) per each dataset of size  $N$ . Fig. 4.9 depicts the reduction in  $\Delta_{\mathcal{M}|\mathcal{D}_N}$  as  $N$  increases. The line shows the mean function for  $\Delta_{\mathcal{M}|\mathcal{D}_N}$  and the area represents the 95% predictive interval. There is approximately a 85% reduction in the expected  $\Delta_{\mathcal{M}|\mathcal{D}_N}$  when the number of observations is increased from 10 to 40 (from 67.41 to 10.24). This outcome also reinforces the prior research of [22] where it is stated that approximately 40 observations are associated with a maximum error of 5% in the K-S distance. However, more than 60 observations are required to achieve similar performance ( $\Delta_{\mathcal{M}|\mathcal{D}_{60}} = 10.96$  for the 97.5% percentile) when the variation is also considered (i.e., upper bound of predictive interval). Although these numerical results are only directly applicable to this case study, similar trends are expected for nonlinear multi-story building models [30].



**Fig. 4.9.** Reduction of epistemic uncertainty as function of  $N$

## 4.5 Conclusions

This chapter presents and validates a model selection methodology intended to enhance the computation of more accurate FFs. There are several proposed probability distributions in the literature to describe the dispersion on seismic data. However this does not mean that just any model is suitable to represent the available data. An appropriate model is

one that can represent the training dataset. Otherwise, making decisions based on its FF results is meaningless and it may even be counterproductive. The methodology established in this chapter addresses this issue. The key contribution of the developed methodology is to identify the probabilistic model with the best representation of the training dataset. Although there is no guarantee that the selected model has a perfect fit for the problem, the methodology developed here is able to identify the model that yields the most accurate results from a range of possibilities. Additionally, the methodology developed here is able to quantify the level of epistemic uncertainty that the scarceness in the numerical model inflicts on the resulting FF. The feasibility of the methodology is demonstrated using seismic response data derived from a well-known numerical model, thousands of synthetic ground motions, and a set of probabilistic models that have been previously promoted to describe seismic data. Ultimately, the methodology provides the means to identify the most representative model independently of the type of data.

## 5. GROUND MOTION SELECTION TECHNIQUE

This chapter presents the methodology developed for sequentially selecting the ground motions. The objective of developing this method is to expedite the computation of accurate FFs by carefully selecting the ground motion with the largest potential to achieving the most reduction in epistemic uncertainty.

The methodology developed in this chapter is in review at this time [119]. The central hypothesis is that simulations should be selected sequentially by considering their effect on reducing epistemic uncertainty. To quantify the epistemic uncertainty in the first place, the Bayesian approach in conjunction with the SMC approach presented in the previous chapter and in [30] is employed. The effect of specific values of IM for the additional observation, are anticipated with respect to changes in the already inferred model parameters and the reduction in epistemic uncertainty.

This chapter begins by discussing the effect of an additional observation on the posterior distribution and its corresponding particle approximation. Using the updated particle approximation of the posterior distribution it is possible to infer the epistemic uncertainty of the new dataset. Then, two selection strategies are presented, one in terms of a single IM characterization and the other with respect to two different IM for each ground motion. The former requires the computation of fragility curves while the latter is performed in terms of fragility surfaces which are projected onto a two-dimensional space for comparison purposes. The methodology section concludes by presenting the algorithm for sequential selection of ground motions, which is applicable to both strategies. Finally, an application of the methodology is presented to illustrate the efficiency of the sequential selection of ground motions strategies with respect to conventional sampling techniques (e.g., random selection of ground motions, uniform selection of IM).

### 5.1 Effect of a hypothetical observation on the reduction in epistemic uncertainty

Consider the expected effect of a hypothetical observation at  $\tilde{x}$  on reducing epistemic uncertainty. Let  $\tilde{y}$  correspond to the structural response at this  $\tilde{x}$ . The most straightforward way to quantify the effect of a hypothetical observation  $(\tilde{x}, \tilde{y})$  is to include it with the rest of the data and perform a new SMC analysis. However, this scheme would be prohibitively expensive since the analysis would have to be performed for various  $x$ 's and many possible  $y$ 's. In the Appendix (see page 115), it is shown that an updated particle approximation  $\left\{ \left( w_a^{(s)}(\tilde{x}, \tilde{y}), \boldsymbol{\theta}_M^{(s)} \right) \right\}_{s=1}^S$  of the hypothetical posterior  $p(\boldsymbol{\theta}_M | \mathcal{D}_N, \tilde{x}, \tilde{y})$  is:

$$p(\boldsymbol{\theta}_M | \mathcal{D}_N, \tilde{x}, \tilde{y}) = \frac{p(\tilde{y} | \tilde{x}, \boldsymbol{\theta}_M) p(\boldsymbol{\theta}_M | \mathcal{D}_N)}{p(\tilde{y} | \tilde{x})} \approx \sum_{s=1}^S w_a^{(s)}(\tilde{x}, \tilde{y}) \delta(\boldsymbol{\theta}_M - \boldsymbol{\theta}_M^{(s)}) \quad (5.1)$$

where  $w_a^{(s)}(\tilde{x}, \tilde{y})$  represents the updated normalized weight for particle  $\boldsymbol{\theta}_M^{(s)}$ , after the additional observation  $(\tilde{x}, \tilde{y})$  is included in the analysis:

$$w_a^{(s)}(\tilde{x}, \tilde{y}) = w^{(s)} \frac{p(\tilde{y} | \tilde{x}, \boldsymbol{\theta}_M^{(s)})}{\sum_{z=1}^S w^{(z)} p(\tilde{y} | \tilde{x}, \boldsymbol{\theta}_M^{(z)})}. \quad (5.2)$$

This updated approximation enables the calculation of the epistemic uncertainty after adding a hypothetical observation  $(\tilde{x}, \tilde{y})$ , denoted as  $\Delta(\tilde{x}, \tilde{y}, \mathcal{D}_N)$ , using Eq. (4.25) and replacing  $w^{(s)} \rightarrow w_a^{(s)}$ .

Finally, the ground motion database provides values of  $\mathcal{X}_1$  but there is no knowledge of their associated values of  $\mathcal{Y}$ . For this reason, the estimate of the epistemic uncertainty  $\Delta(\tilde{x}, \mathcal{D}_N)$  is used to select of the IM value associated with the next observation. The expected value for this epistemic uncertainty is expressed as

$$\mathbb{E}[\Delta(\tilde{x}, \mathcal{D}_N)] = \int \Delta(\tilde{x}, \mathcal{D}_N | y = \tilde{y}) p(\tilde{y} | \tilde{x}, \mathcal{D}_N) d\tilde{y} = \int \Delta(\tilde{x}, \tilde{y}, \mathcal{D}_N) p(\tilde{y} | \tilde{x}, \mathcal{D}_N) d\tilde{y}. \quad (5.3)$$

## 5.2 Sequential selection based on single-dimension IM

The first selection strategy consists of identifying a reduced interval, within the entire  $\mathcal{X}_1$ -domain  $[x_{min}, x_{max}]$ , that provides the largest expected reduction in epistemic uncertainty. This optimal interval is referred as the  $i^*$ -th interval and is the one such that:

$$i^* = \arg \max_i \mathbb{E} [\Delta(\mathcal{D}_N) - \Delta_i(\mathcal{D}_N, \mathcal{X}_1) \mid \mathcal{D}_N, \mathcal{X}_1] \text{ for } \mathcal{X}_1 \in [x_i, x_{i+1}] \quad (5.4)$$

where  $\Delta_i(\mathcal{D}_N, \mathcal{X}_1)$  represents the value of the epistemic uncertainty in the  $i$ -th interval  $[x_i, x_{i+1}]$ , the bound  $x_i := x_{min} + (x_{max} - x_{min})(i - 1)/n_1$ ,  $n_1$  is the number of adjacent and non-overlapping intervals of equal size, and  $x_{min}$  and  $x_{max}$  correspond to the lower and upper bound for the  $\mathcal{X}_1$ -domain, respectively.

## 5.3 Sequential selection based on two-dimensional IM

This sequential selection strategy is an extension of the aforementioned strategy (single-dimension IM) by including the additional information provided by a second IM,  $\mathcal{X}_2$ . As a result, the analysis involves a fragility surface instead of a fragility curve. The second IM,  $\mathcal{X}_2$ , is chosen to be one of the remaining elements of  $\mathbf{X}$ . The fragility surface is defined as

$$F(x_1; x_2; y_{crit}) = \mathbb{P}[\mathcal{Y} \geq y_{crit} \mid \mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2] = \int_{y_{crit}}^{+\infty} f_{\mathcal{Y}}(y \mid \mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2) dy. \quad (5.5)$$

The FF (one dimensional) obtained from this fragility surface results from a marginalization process along  $\mathcal{X}_2$ :

$$F(x_1; y_{crit}) = \int_{y_{crit}}^{+\infty} f_{\mathcal{Y}}(y \mid \mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2) p(x_2 \mid \mathcal{X}_1 = x_1) dx_2 dy \quad (5.6)$$

where the conditional probability  $p(x_2 \mid \mathcal{X}_1 = x_1)$  captures the relationship between the two IMs. This probability can be easily and rapidly inferred from the entire ground motion database since no structural response is required (i.e., there is no need to evaluate expensive computational models). It is important to point out that a particle approximation of a fragility

surfaces can readily be mapped to a particle approximation of the FF using Eq. (5.6). Thus, it is possible to compute the epistemic uncertainty and its expected reduction using the procedure defined already in Chapter 4.

The incorporation of the additional dimension  $\mathcal{X}_2$  requires defining a new model and its corresponding assumptions. In this work, the expected value of  $\mathcal{Y}$  takes the form:

$$\mathbb{E}[\mathcal{Y} \mid \mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2] = m_2(x_1, x_2) = x_1^{\gamma_{m1}} x_2^{\gamma_{m2}} h_{m_2}(x_1, x_2) \quad (5.7)$$

where  $\gamma_{m1}$  and  $\gamma_{m2}$  are non-negative exponents for  $\mathcal{X}_1$  and  $\mathcal{X}_2$ , respectively, while  $h_{m_2}$  is the correction term to be expanded in terms of  $q^*$  basis functions  $\{\varphi_i(x_1, x_2)\}_{i=1}^{q^*}$  as follows

$$h_{m_2}(x_1, x_2) = \exp\left(\sum_{i=1}^{q^*} c_{m_2, i} \varphi_i(x_1, x_2)\right). \quad (5.8)$$

Similarly, the standard deviation is

$$\sqrt{\mathbb{V}[\mathcal{Y} \mid \mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2]} = s_2(x_1, x_2) = \exp(c_{s_2}) x_1^{\gamma_{s1}} x_2^{\gamma_{s2}} \quad (5.9)$$

with the corresponding non-negative exponents  $\gamma_{s1}$ ,  $\gamma_{s2}$  and coefficient  $c_{s_2} \in \mathbb{R}$ . As in the previous model, the basis functions for this model are polynomial functions in two dimensions, although any other type of basis functions may be used.

The application of the Bayesian framework is similar to the single-dimension case, and the parameters can be inferred from the training dataset  $\mathcal{D}_N = \left\{ \left( x_1^{(n)}, x_2^{(n)}, y^{(n)} \right) \right\}_{n=1}^N$ . However, a larger number of parameters (i.e., exponents and coefficients) must be inferred from the data. The additional parameters to be inferred due the presence of  $\mathcal{X}_2$  are  $\gamma_{m2}$ ,  $\gamma_{s2}$ , as well as some additional basis functions that are required to define the polynomials of a particular degree.

The  $(\mathcal{X}_1, \mathcal{X}_2)$  sequential selection strategy consists of dividing the  $(\mathcal{X}_1, \mathcal{X}_2)$  space into a finite number, denoted by  $n_2$ , of blocks or intervals  $\{I_n\}_{n=1}^{n_2}$ . Then the optimal interval  $I_{l^*}$  is determined as the one in which the expectation of the reduction of epistemic uncertainty is maximized. Assuming that the most informative ground motion is the one defined

by the observation  $(\tilde{x}_1, \tilde{x}_2, \tilde{y})$ , the inclusions of this observation into  $\mathcal{D}_N$  results in an improved posterior distribution (Eq. (5.1)). This improvement yields the fragility function with the minimum expected value of epistemic uncertainty, which can be computed using Eq. (4.25) for the updated particle approximation (Eq. (5.2)). The challenge is that there is no knowledge about the optimal location  $(\tilde{x}_1, \tilde{x}_2, \tilde{y})$  so an exploratory analysis is required to investigate each interval and handpick  $I_{i^*}$ . To accomplish this, it is necessary to draw random observations  $(x_1, x_2)$ , enclosed in each box  $I_n$ , and determine their corresponding effects on reducing the epistemic uncertainty (see Eq. (5.3)). Notice that the intrinsic randomness in the  $\mathcal{Y}$  dimension is accounted for in the analysis by determining the average from a population of observations sampled from the posterior distribution (Eq. (4.12)) for each sample  $(x_1^{(s)}, x_2^{(s)})$ . Finally, the expected epistemic uncertainty reduction in the interval  $I_n$  is approximated by the average sample of the set of  $(x_1, x_2)$  observations.

#### 5.4 Process to implement the sequential selection strategy

Given that the idea adopted for both selection strategies is very similar (the only difference is the number of dimensions to be analyzed), it is helpful to now explain the general methodology for determining the  $(N + 1)$ -th observation. To discuss the methodology, the following notation is used:

- $I_i$  is the corresponding interval to be analyzed and represents a subset of the entire domain.
- $\mathcal{X}$  is an array of the coordinates to be analyzed, e.g.,  $\mathcal{X} = \mathcal{X}_1$  (one dimension),  $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2\}$  (two dimensions).
- $\mathcal{X}^{(k)}$  is the  $k$ -th sample (when performing inference) or observation (for the training dataset) for the chosen IM.
- $\mathcal{D}_N$  is the training dataset containing  $N$  observations of  $\mathcal{X}$ , i.e.,  $\mathcal{D}_N = \{\mathcal{X}^{(n)}\}_{n=1}^N$ .

In the sequential selection strategy, the process to select the  $(N + 1)$ -th observation is as follows:

1. Divide the  $\mathcal{X}$ -domain into  $n_*$  adjacent and disjoint subsets of approximately equal size.
  2. For each  $i = 1, \dots, n_*$  interval:
    - (a) Using a uniform distribution  $U(I_i)$ , sample  $K$  values from the array of coordinates,  $\{\mathcal{X}^{(k)}\}_{k=1}^K$ .
    - (b) For every  $\mathcal{X}^{(k)}$ :
      - i. Sample  $R$  structural response values  $\{\mathcal{Y}^{(r)}\}_{r=1}^R$  using the posterior distribution  $p(\boldsymbol{\theta} \mid \mathcal{D}_N)$ . Specifically, the samples are obtained using the point-predictive distributions  $f_{\mathcal{Y}}(y \mid \mathcal{X}^{(k)}, \mathcal{D}_N, \boldsymbol{\theta}^{(s)})$  for all  $\boldsymbol{\theta}$ -particles.
      - ii. For each and every  $\mathcal{Y}^{(r)}$ :
        - Compute the updated particle approximation  $w_a(\mathcal{X}^{(k)}, \mathbf{Y}^{(r)})$  using Equation (5.2).
        - Compute the expected value of epistemic uncertainty  $\Delta(\mathcal{X}^{(k)}, \mathcal{Y}^{(r)})$  using Eq. (4.25), replacing  $w^{(s)}$  with  $w_a^{(s)}$ .
      - iii. Evaluate the average value of the epistemic uncertainty at  $\mathcal{X}^{(k)}$ ,
 
$$\mathbb{E} \left[ \Delta(\mathcal{D}_N, \mathcal{X}^{(k)}) \right] = (1/R) \sum_{r=1}^R \Delta(\mathcal{X}^{(k)}, \mathcal{Y}^{(r)})$$
    - (c) Determine the mean value of the epistemic uncertainty for the  $i$ -th interval:
 
$$\mathbb{E} [\Delta_i(\mathcal{D}_N, \mathcal{X})] = (1/K) \sum_{k=1}^K \mathbb{E} \left[ \Delta(\mathcal{D}_N, \mathcal{X}^{(k)}) \right]$$
3. Determine the index  $i^*$  for the optimal interval using Eq. (5.4).
4. From the ground motion database, randomly select a sample with an IM  $\mathcal{X} \in I_{i^*}$  to be the  $(N + 1)$ -th observation.
5. END of the sequential selection algorithm.

## 5.5 Numerical Results

This section presents an implementation of the methodology used for the case study. First, a new parameter to quantify the uncertainty in the FFs and the considered selection strategies is presented. Second, the basis for the selection of the statistical model to describe the relationship between  $\mathcal{X}_1$  and  $\mathcal{X}_2$  is discussed. Then, the results of the fragility analysis for several different selection algorithms are presented.

### 5.5.1 Sequential selection strategies

To evaluate the performance of the different selection strategies, an initial training dataset with  $N = 10$  observations is used, and is sequentially complemented with 90 additional observations. This process is repeated at least a hundred times to evaluate the success of the selection strategy independently of random initialization of the training dataset. Four different strategies are compared, including:

1. Random: ground motions are selected in an aleatory fashion regardless of the IM  $\mathcal{X}_1$ .
2. Uniform: ground motions are selected to generate a uniform distribution in the IM coordinate  $\mathcal{X}_1$ .
3.  $\mathcal{X}_1$ -based: sequential selection strategy to choose ground motions in terms of the  $\mathcal{X}_1$ -coordinate.
4.  $(\mathcal{X}_1, \mathcal{X}_2)$ -based: sequential selection strategy to choose ground motions using both  $(\mathcal{X}_1, \mathcal{X}_2)$ -coordinates.

The uniform selection alternative is considered because it tends to be the preferred option among researchers, but implementation of the other strategies are expected to accelerate the process. Although the common practice for selecting ground motions is to use random samples, this strategy has quite a low convergence rate when using a realistic ground motion database because the records are typically clustered at the low-intensity, and thus low-

damage, region. As a result, any prediction at medium or high IM values may be associated with large values of uncertainty due to the scarceness of data at these IM magnitudes.

For the implementation of the  $\mathcal{X}_1$ -based and  $(\mathcal{X}_1, \mathcal{X}_2)$ -based selection strategies, the following change in the parameter to quantify the uncertainty in the FFs was found to yield better convergence rate:

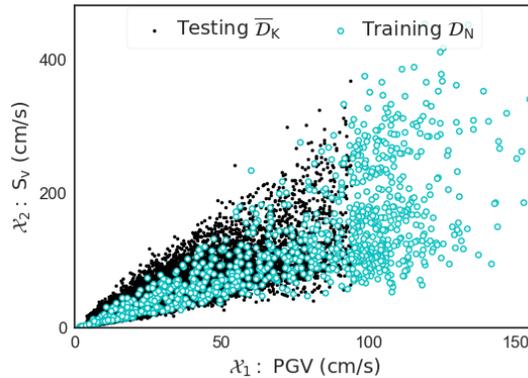
$$A_{\theta_M}(y_{\text{crit}}) = F(x_{\text{crit}}; y_{\text{crit}}) \quad (5.10)$$

where, the specific IM value  $x_{\text{crit}}$  is chosen as the location yielding the largest amount of uncertainty

$$x_{\text{crit}} = \arg \max_x \mathbb{V}[F(x; y_{\text{crit}})]. \quad (5.11)$$

It is important to clarify that the location of  $x_{\text{crit}}$  depends on multiple factors such as the training dataset, the number of particles, random initialization, among other, and for this reason varies at each iteration of the sequential selection process. This behavior is exactly what is to be exploited with this approach.

### 5.5.2 Statistical model for secondary IM



**Fig. 5.1.** Data set of  $\mathcal{X}_1$  vs.  $\mathcal{X}_2$  observations

To project the fragility surfaces computed in the  $(\mathcal{X}_1, \mathcal{X}_2)$ -based strategy onto the  $\mathcal{X}_1 - \mathcal{Y}$  plane, it is necessary to determine the appropriate functional form for  $p(\mathcal{X}_2 | \mathcal{X}_1 = x_1)$ . Applying the methodology proposed in [30] and described in detail in Chapter 4, it is

possible to determine this functional form. The only difference is that the output parameter  $\mathcal{Y}$  is replaced by  $\mathcal{X}_2/100$  in the analysis (the factor  $1/100$  is introduced to keep a similar scale and to use the same basis functions presented in Fig. 4.2), while the number of particles remains the same ( $S = 750$ ). A total of 1,000 ground motions uniformly distributed along  $\mathcal{X}_1$  are used as the training dataset while the remaining 35,000 motions as the testing dataset (see Fig. 5.1). Table 5.1 presents a summary of the model selection results for the number of basis functions  $q_m$  and  $q_s$  that yields to the best representation of the training dataset per likelihood function. This table includes the same comparison metrics described in Chapter 4,  $\mathcal{J}_1$ ,  $\mathcal{J}_2$ ,  $\mathcal{J}_3$ , and  $\log(Z_M)$ . The complete table for all the considered numbers of basis functions is included in the Appendix (see page 123). Fig. 5.2 depicts the results of this analysis for the selected models included in Table 5.1. The left column of plots shows the entire dataset and the numerical predictions for the  $\mathcal{X}_1$ - $\mathcal{X}_2$  response of each model. The line represents the expected value and the area the 95% predictive interval. The predictive intervals are determined numerically from 15 million random samples generated for each model (200 values of  $\mathcal{X}_1 \times 100$  samples of  $\mathcal{X}_2$  per value of  $\mathcal{X}_1 \times 750$  particles). The right column of plots corresponds to the Q-Q plots from the K-S test. The black solid line is the theoretical CDF for the standard and the dashed line is the estimated distribution from the testing dataset. Furthermore, the results presented in Table 5.1 are shown graphically in Fig. 5.3 using bar plots. The top left (a) plot presents the results of the Bayesian model selection technique, the top right (b) plot shows the comparison metric  $\mathcal{J}_1$ , the bottom left (c) plot corresponds to the ratio of the K-S distance for each model and the maximum distance among the models, and the bottom right (d) plot shows the ratio of the Q-Q error with respect to the maximum error. The gray arrow points the direction of the desired behavior of the magnitude for each metric.

The results from the model selection methodology show the superiority of the lognormal distribution to describe the dispersion of  $\mathcal{X}_2$  when it is conditioned on  $\mathcal{X}_1$ . The dependency of  $\mathcal{X}_2$  with  $\mathcal{X}_1$  can be modelled using a lognormal  $(\mu_{\log}(x_1), \sigma_{\log}(x_1))$ , where  $\mu_{\log}(x_1) = \log\left(\frac{m(x_1)^2}{\sqrt{m(x_1)^2 + s(x_1)^2}}\right)$  and  $\sigma_{\log}(x_1)^2 = \log\left(\frac{m(x_1)^2 + s(x_1)^2}{m(x_1)^2}\right)$ . The expected value  $m(x_1)$  and the standard deviation  $s(x_1)$  can be obtained from the particle approximation presented in

**Table 5.1.** Comparison metrics of the models

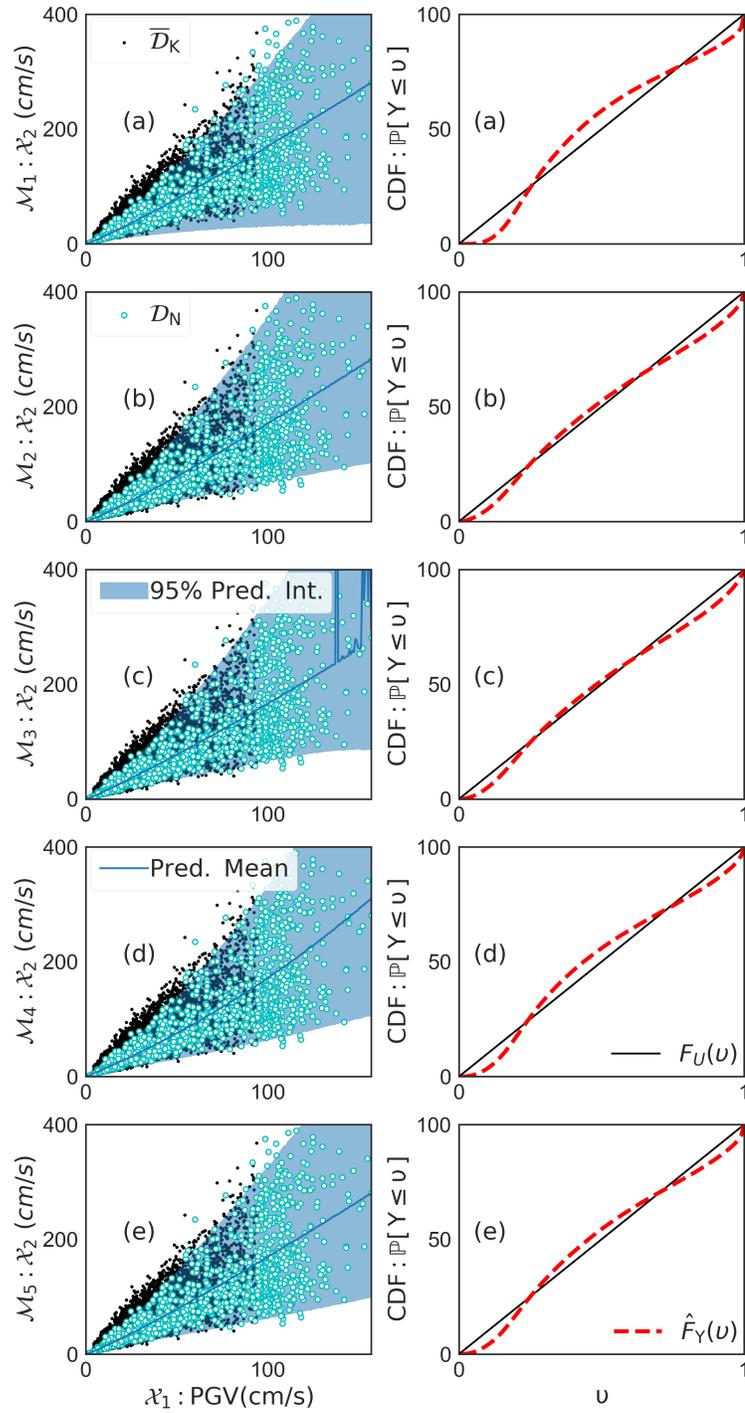
<b>Likelihood</b>	$q_m$	$q_s$	<b>Evidence</b> $\log(Z)$	$\mathcal{J}_1$ [%]	$\mathcal{J}_2$	$\mathcal{J}_3$
$\mathcal{M}_1$ : Normal	1	1	4195.0	3.1	0.091@PGV=18.4cm/s	0.309
$\mathcal{M}_2$ : <b>Lognormal</b>	1	1	<b>4305.2</b>	2.2	<b>0.054@PGV=141.0cm/s</b>	0.099
$\mathcal{M}_3$ : Log-T	4	1	4257.7	<b>0.9</b>	<b>0.052@PGV=139.4cm/s</b>	<b>0.092</b>
$\mathcal{M}_4$ : Gamma	2	2	4260.4	<b>1.4</b>	0.066@PGV=15.8cm/s	0.182
$\mathcal{M}_5$ : Beta	1	2	4284.2	2.3	0.062@PGV=17.3cm/s	0.141

Fig. 5.4, resulting in  $m(x_1) = x_1^{1.12} \exp(-3.567)$  and  $s(x_1) = x_1^{1.341} \exp(-4.288)$ . Finally, the numerical predictions and the training dataset (dots) are presented in Fig. 5.5, where the solid line corresponds to the expected value  $m(x_1)$  while the area is the 95% predictive interval.

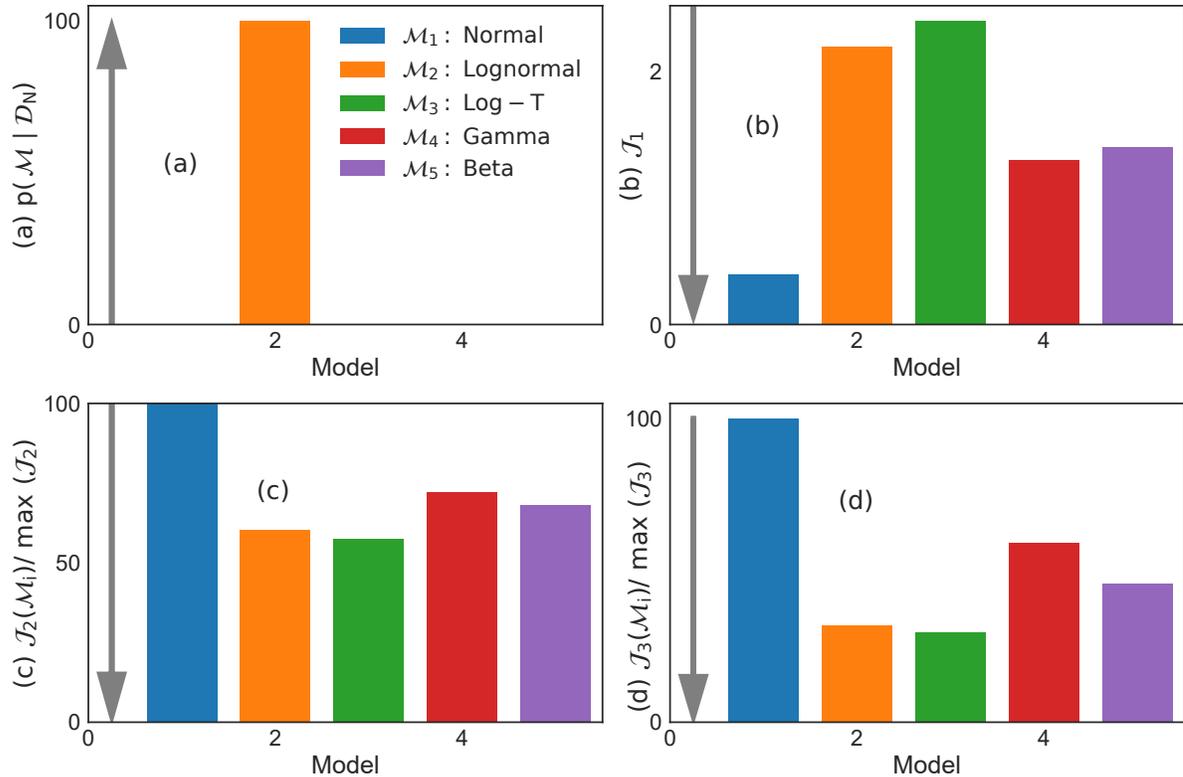
### 5.5.3 Results of selected selection strategies

Following the methodology proposed here, the performance of the four selection strategies is evaluated for an increasing number of observations (from 10 to 100 observations). For the  $\mathcal{X}_1$ -based strategy, the  $\mathcal{X}_1$  domain is divided into 22 intervals (see left plot in Fig. 5.6) and a total of  $\mathcal{K} = 300$  random samples are evaluated for each interval  $i = 1, \dots, 22$ . On the other hand, in the  $(\mathcal{X}_1, \mathcal{X}_2)$ -based strategy, the  $\mathcal{X}_1$  domain is divided into 15 intervals while the  $\mathcal{X}_2$  domain is divided into a number of intervals that is increasing as a function of  $\mathcal{X}_1$ . Thus, the  $\mathcal{X}_2$  domain uses just one interval for the lowest values of  $\mathcal{X}_1$  and is divided into eight intervals for the largest values of  $\mathcal{X}_1$  (see right plot in Fig. 5.6). A total of 83 boxes of similar size are evaluated for this selection strategy, and a total of  $\mathcal{K} = 1000$  samples are randomly distributed among the  $(\mathcal{X}_1, \mathcal{X}_2)$  subdomain for each box  $(i, j)$ . For each sample  $(\mathcal{X}_1^{(k)}, \mathcal{X}_2^{(k)})$ , a total of  $R = 250$  samples of  $\mathcal{Y}$  are derived to determine the box indicating the location of the next desired observation.

Fig. 5.7 shows the expected FF after sequentially selecting 90 observations (i.e, for a total of 100 observations if the initial 10 are accounted for) for the three selection strategies and different training datasets. In Fig. 5.7 the (a) (orange), (b) (blue), (c) (green)

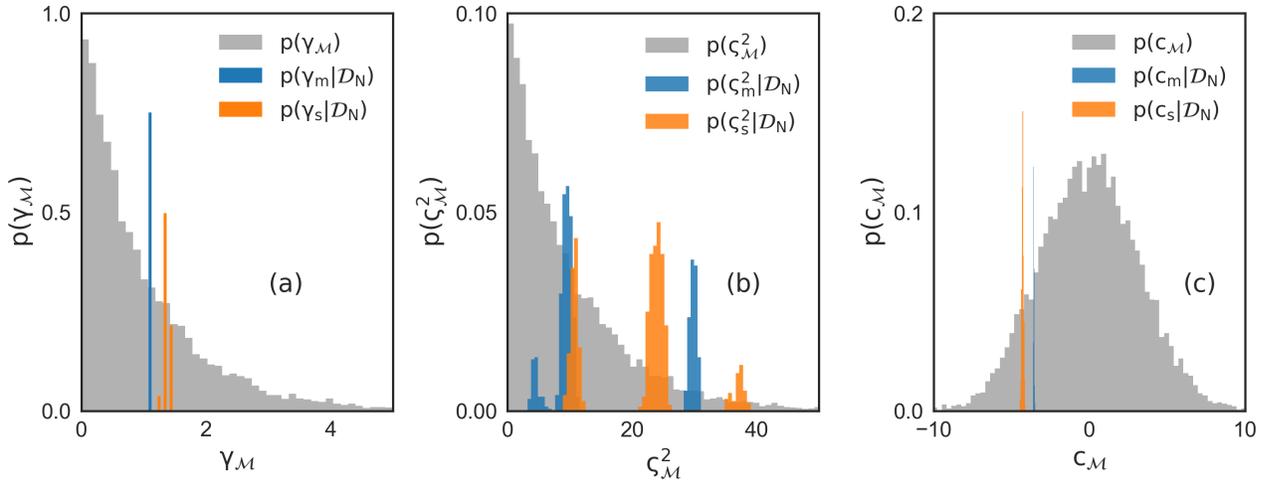


**Fig. 5.2.** Numerical results for  $(X_1, X_2)$  models (a)  $\mathcal{M}_1$ , (b)  $\mathcal{M}_2$ , (c)  $\mathcal{M}_3$ , (d)  $\mathcal{M}_4$ , and (e)  $\mathcal{M}_5$

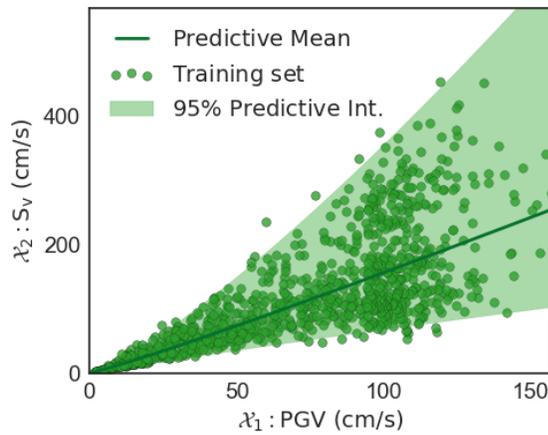


**Fig. 5.3.** Comparison metrics of  $(\mathcal{X}_1, \mathcal{X}_2)$  models (a) Bayesian model selection, (b) Predictive interval, (c) KS-distance, and (d) Q-Q error

and (d) (red) subplots correspond to the random, uniform,  $\mathcal{X}_1$ -based, and  $(\mathcal{X}_1, \mathcal{X}_2)$ -based strategies, respectively. In each plot the black dashed line represents the best estimate for the FF based on the available information, which is computed using 10,000 ground motions, uniformly distributed along the  $\mathcal{X}_1$  domain. The solid colored line in each plot represents the expected FF for a dataset of size 100 (regardless of the initial dataset), while the shaded area corresponds to the 95% predictive interval of the mean FF among all of the different training datasets. The results show that all strategies exhibit similar mean convergence, albeit with different degrees of performance uncertainty (performance uncertainty is induced by the random selection of the initial 10 observations). In particular, the  $\mathcal{X}_1$ -based strategy has is characterized by fastest performance in terms of uncertainty reduction, followed by the uniform strategy, and finally the random and  $(\mathcal{X}_1, \mathcal{X}_2)$ -based, both having the slowest performance uncertainty reduction. Given this evidence, it is concluded



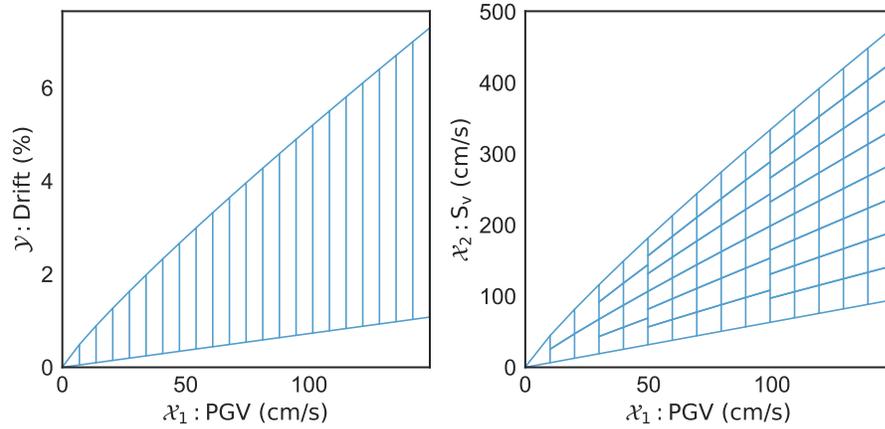
**Fig. 5.4.** Prior/posterior comparison (a) Exponent  $\gamma_M$ , (b) Variance  $\zeta_M^2$ , and (c) Coefficients  $\mathbf{c}_M$



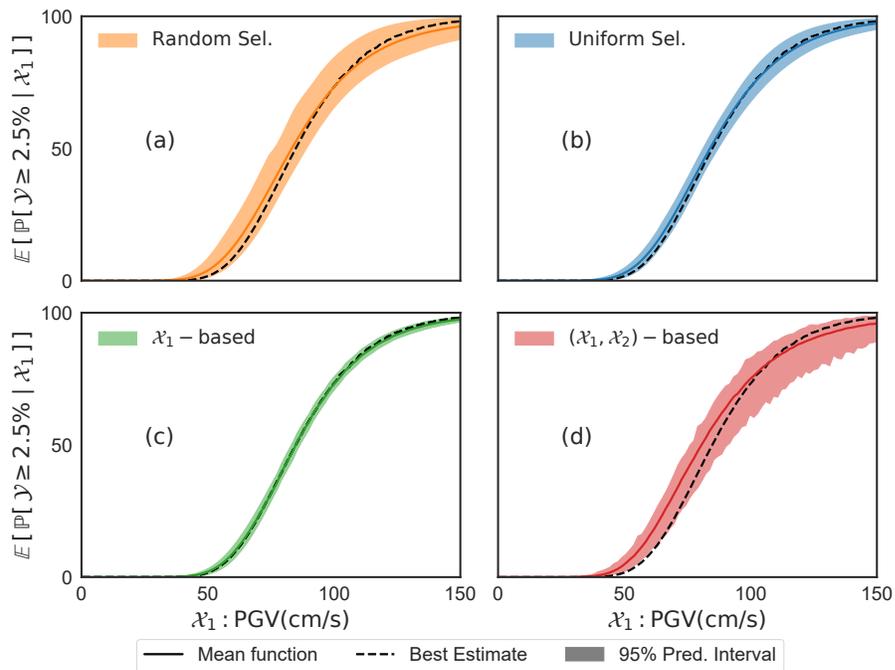
**Fig. 5.5.** Numerical representation of  $\mathcal{X}_1$ - $\mathcal{X}_2$  lognormal model

that the  $\mathcal{X}_1$ -based strategy should be preferred because its performance is similar to the uniform strategy but with lower performance uncertainty for the same number of samples.

Fig. 5.8 shows the epistemic uncertainty performance for these strategies as a function of the number of ground motions. In this figure, the left subplot (a) shows the reduction in this uncertainty as  $N$  increases after implementing the random selection strategy (orange), the subplot (b) corresponds to the  $\mathcal{X}_1$ -based strategy (green), while the subplot (c) contains the

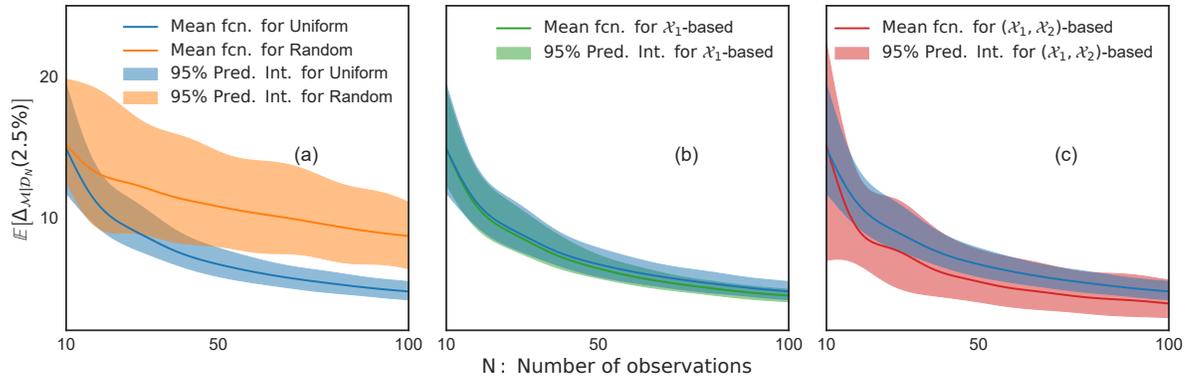


**Fig. 5.6.** Input domain division for  $X_1$ -sel (left) and  $(X_1, X_2)$ -sel (right)



**Fig. 5.7.** Comparison of expected fragility functions using different datasets with 100 observations and different selection strategies

results for the  $(X_1, X_2)$ -based strategy (red). To provide a common comparison, the results of the uniform selection strategy (blue) are included on all plots. Additionally, in each plot the line corresponds to the mean value while the shaded area indicates the 95% predictive



**Fig. 5.8.** Comparison of epistemic uncertainty performance as function of  $N$  for the strategies (a) random, (b)  $\mathcal{X}_1$ -based and (c)  $(\mathcal{X}_1, \mathcal{X}_2)$ -based with respect to uniform (blue)

interval for the epistemic uncertainty values. Although the mean epistemic uncertainty value is similar for  $N = 10$  observations among all the considered selection strategies, observe that both the descent rate and the variation are different for the cases shown in Fig. 5.8. Specifically, the uniform and the  $\mathcal{X}_1$ -based strategies have similar performance until approximately  $N = 50$  observations, but the latter exhibits lower values of both epistemic uncertainty and variation when additional observations are sequentially added to the dataset. For the  $(\mathcal{X}_1, \mathcal{X}_2)$ -based strategy, a larger variation in the epistemic uncertainty is observed, but with the same or better results than the other strategies. However, although there is small variation between the FFs, when the  $(\mathcal{X}_1, \mathcal{X}_2)$ -based strategy is implemented, Fig. 5.7 indicates that no convergence has been achieved in the FF at  $N = 100$  observations when the random initialization is considered. Furthermore, the random strategy exhibits the lowest descent rate among all the cases with an almost constant variation for  $N \in [10, 100]$  observations.

## 5.6 Conclusions

This chapter introduces and validates a methodology for sequentially selecting the ground motion records that yield more accurate FFs in an even more efficient manner. Current selection strategies are based on a random selection of ground motion records or

an uniform distribution of values for a selected IM. However, obtaining an accurate and consistent FF may require the analysis of a training dataset with an extensive number of ground motions. The key contribution of the developed methodology, the establishment of two sequential selection strategies, addresses this issue. The methodology is validated by implementing thousands of synthetic ground motion records based on earthquake scenarios for California and a widely used numerical model of a steel building designed for the same region. The results show that the proposed selection strategies achieve more accurate FF with a reduced number of ground motion records and less uncertainty than conventional selection algorithms. It is expected that these sequential selection strategies will offer even greater advantages with models of higher complexity that require basis functions of higher order than linear. Ultimately, implementation of these sequential selection strategies will empower the engineer to compute more accurate FFs more efficiently.

## 6. PRACTICES THAT UNINTENTIONALLY BIAS FRAGILITY FUNCTIONS

This chapter presents some common practices that may produce bias in the resulting FF when constructing or defining a database of ground motions. The first such practice is the aleatory selection of an existing ground motion database. Each ground motion record includes specific characteristics that triggers a unique response in a given structure. Consequently, a similar level of uniqueness can be found for the case of ground motion databases that have been constructed with specific parameters, such as: frequency content, target location, magnitude and distance, among others. The second such practice is the inclusion of historical ground motion records (e.g., El Centro 1940, Loma Prieta 1989, Kobe 1995, Chi Chi 1999) with the intention of creating a more "robust" or just bigger database. Lastly, the final such practice that is discussed here is the effect of scaling the ground motion records as it is the case of the well-known and widely-used IDA method. The introduction of bias to the FFs is exposed through a series of examples using different ground motion databases and the twenty-story nonlinear benchmark building [86,88].

### 6.1 Effect of using different ground motion databases

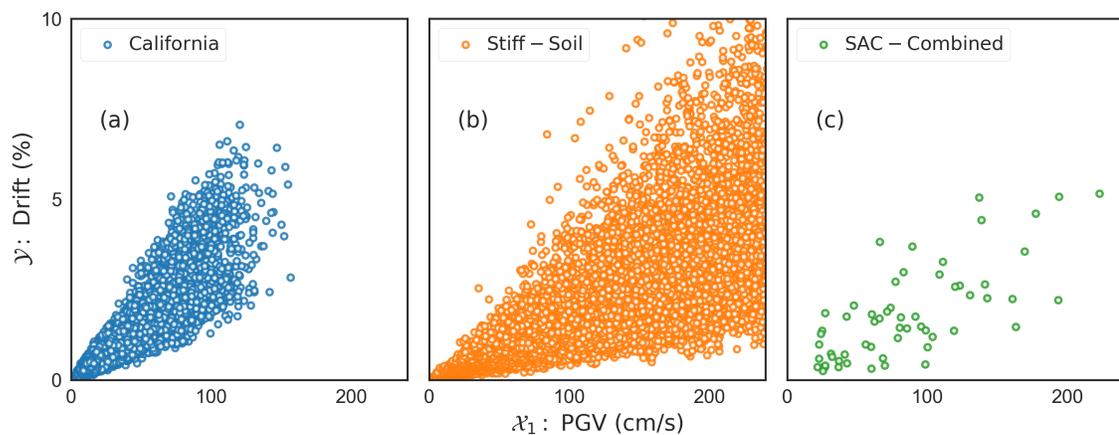
The first biasing practice is selecting random databases of ground motions when performing a fragility analysis. There is a misconception that FFs are an inherent property function of the studied structure. For example, it is commonly believed that the FF for a given structure is an invariant function with respect to the geographical location and the orientation of the building (which are factors that dictate the characteristics of the ground motions). The validity of such understanding is questionable since the set of ground motion records is used to train the probabilistic model. Here, it is clear that the inclusion of a ground

motion database for FF generation with particular characteristics yields certain dispersion in the response of the structure, and consequently a distinct FF. To illustrate this behavior, three different earthquake databases are considered for the computation of FFs:

1. California database: the synthetic ground motions introduced in Chapter 2 which are obtained from earthquake scenarios in California. The following are some of characteristics of the ground motions within this database:
  - Size: 36,000 (synthetic)
  - Epicentral distance: up to 20 Km
  - Magnitude: from 5.5 to 7.5
  - NHERP site class: C and D
  - Earthquake scenario location: California
  
2. Stiff-Soil database: synthetic ground motions obtained using a different technique, that is filtering a white noise signal through a Kanai-Tajimi filter [120,121] to simulate the ground motions for a stiff soil site [122]. Some of the characteristics of these ground motions are:
  - Size: 20,000 (synthetic)
  - Soil characteristic: Stiff soil
  - Soil natural frequency: 20.8 rad/s for the average value (7.33 rad/s of standard deviation)
  - Soil damping ratio: 0.4 for the average value (0.18 of standard deviation)
  - Acceleration maximum amplitude: up to 1.7g (uniform distribution)
  
3. SAC-Combined database: using a combination of synthetic and real ground motions [73] created for the SAC steel project [86], in which can be found ground motions such as El Centro (1940), San Fernando (1971), Imperial Valley (1979), Loma Prieta (1989), Northridge (1994), Kobe (1995), among others.

- Size: 60 (50 real motions and 10 synthetic records)
- Hazard level within a 50-year period: 50% for 20 records, 10% for 20 records, and 2% for 20 records (10 real and 10 synthetic)

Fig. 6.1 shows the three ground motion database in terms of the selected IM (PGV) and the structural response. The blue dots in the (a) subplot corresponds the ground motions from the California database, the orange dots in the (b) subplot to the Stiff-Soil ground motions, and the green dots in the (c) subplot to the SAC-Combined records.



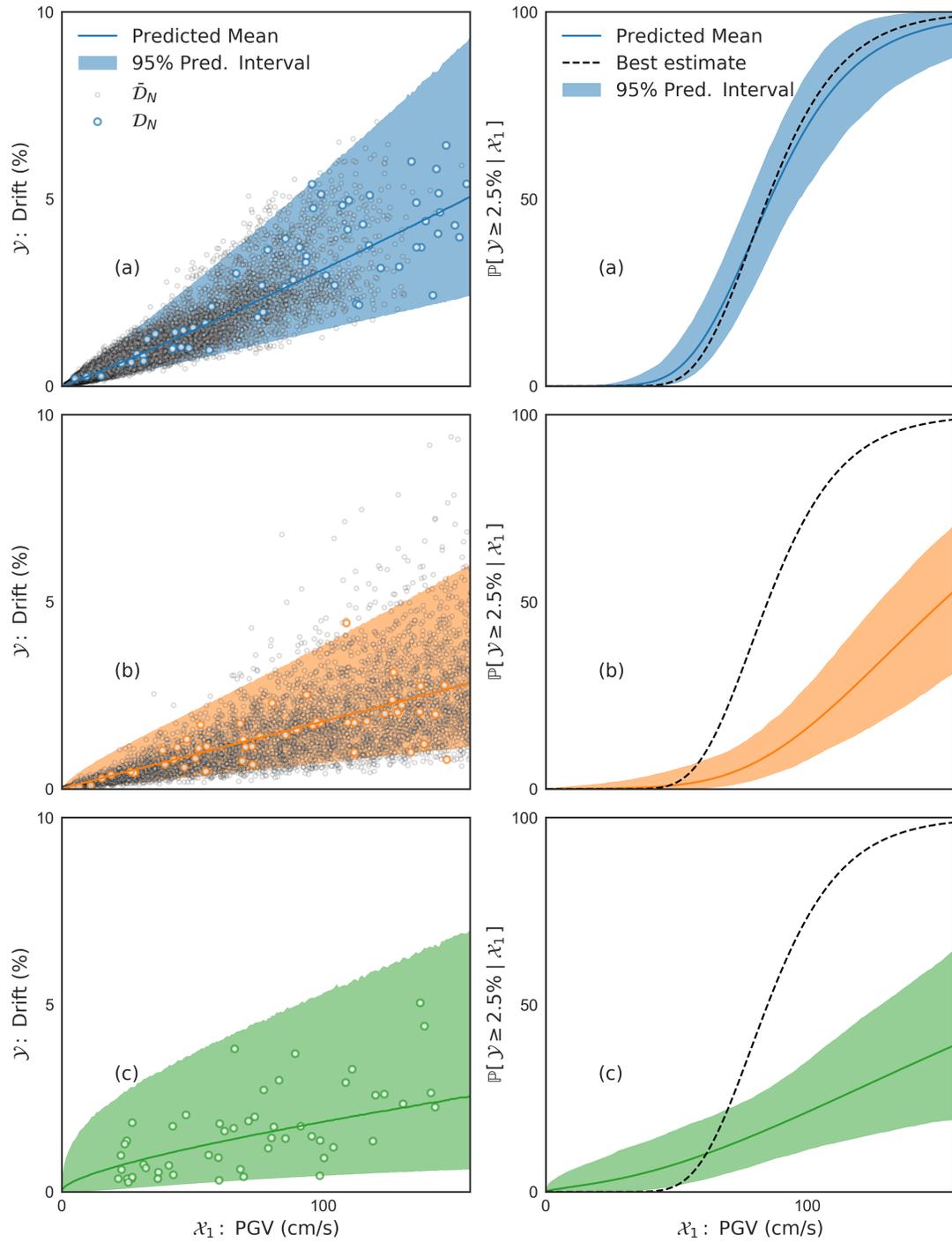
**Fig. 6.1.** Representation of different ground motion databases

Due to the size differences among the databases, the methodology proposed in Chapter 4 is implemented for  $N = 60$  records following an uniform selection strategy for the California and Stiff-Soil databases. The results of these analyses are presented in Fig. 6.2. The left subplots correspond to the structural response data: the colored dots correspond to the training dataset, the black dots are the unobserved data points from the database, and the solid line is the mean while the colored area represents the 95% predictive interval response from the probabilistic model. Additionally, the right subplots present the FFs for each training dataset. The dashed line corresponds to the best estimate previously derived using  $N = 10,000$  records from the California database. As it was previously stated, the California database of ground motions are treated in this dissertation as the reference, or the representative motion records, for the case study. From Fig. 6.2 it is clear that the ground

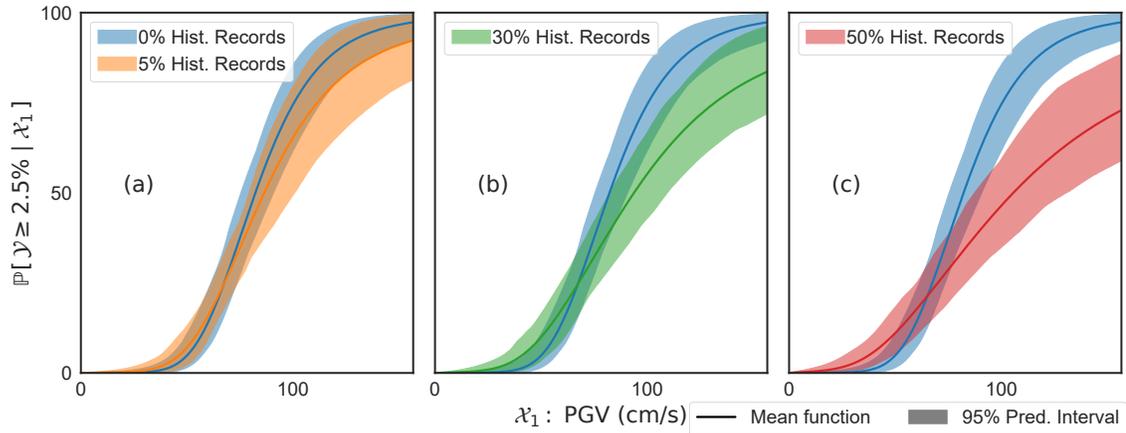
motion database needs to be carefully selected to be representative of the actual motions that the studied structure is prone experience based on its geographical location and orientation.

## 6.2 Effect of including historical ground motions

Another common practice when selecting the ground motion database is to augment an existing database, perhaps containing low-intensity or with pure synthetic records, with real historic ground motions [73]. As was previously shown, the use of different databases may bias the resulting FF due the differences in the ground motion characteristics. Consequently, similar behavior can be expected when the training dataset of ground motions is derived from a combination of records of two or more databases. To exemplify this, a fragility analysis is carried out using a training dataset of  $N = 50$  records derived from the California database of ground motions and a reduced number of real records from the SAC-Combined database. Four different cases are considered for this demonstration, the inclusion of none (0%), 5 (10%), 15 (30%), and 25 (50%) real ground motion records. In order to account for the randomness of the aleatory initialization, the analysis for each case is repeated for a total of 100 different training datasets. Fig. 6.3 shows the results of this analysis. In the figure, the solid lines represent the mean FF and the colored area its 95% predictive interval. The (a) subplot corresponds to the FF when 5 observations (10%) of the training dataset are derived from historic records (orange), the (b) subplot presents the case with 15 observations (30%) in green, and the (c) subplot is the 25 observations case (50%) in red. For comparison purposes, the case with pure synthetic ground motions (0%) is presented in all the subplots with the FF in blue. From Fig. 6.3, it can be observed that including ground motions from different databases, as it is the case of the historic records on this example, introduces bias in the computed FFs. Furthermore, this bias increases proportionally with the number of ground motions added from different databases.



**Fig. 6.2.** Fragility analysis for 50 observations from different ground motion databases: (a) California, (b) Stiff-Soil, and (c) SAC-Combined

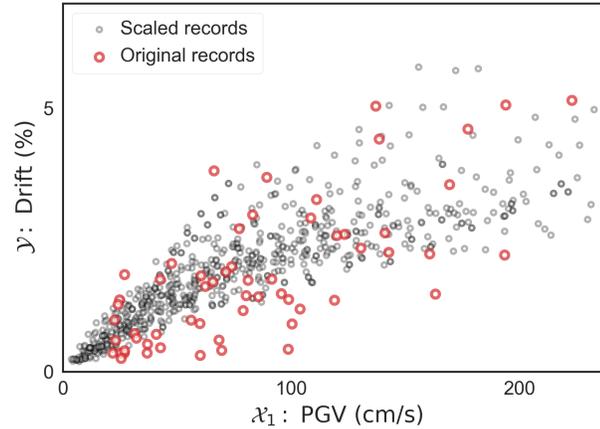


**Fig. 6.3.** Fragility analysis for 50 observations combining synthetic and historic ground motion records

### 6.3 Effect of scaling ground motions

The final biasing practice to discuss in this dissertation is the effect of scaling the ground motions to increase the size of the dataset. A common example of this approach is the well-known IDA method [24]. Although this method results in an effective way to identify the probability distribution that describes the dispersion of the structural response for a given value of IM, the numerical evaluation of the scaled ground motions can result computationally expensive. Previous studies have demonstrated that scaling the ground motion introduces bias into the MISD response [67, 75, 76]. Likewise, the assumption is that scaling the ground motions should introduce bias in the FF. To illustrate this, a fragility analysis is carried out for an increased ground motion database obtained after scaling the records from the SAC-Combined set. The values of spectral acceleration  $S_a(f_n, \zeta_n)$  for each of the 60 records from the SAC-Combined database is scaled for 12 different values of acceleration and then the  $\mathcal{X}_1$  IM is evaluated. The observations from the scaled ground motions are presented in Fig. 6.4.

To expose the bias in the FF, the fragility analysis is carried out for the original ground motions (SAC-Combined), the augmented database with 720 records, and a dataset from this augmented database of similar size than the SAC-Combined set (60 ground motions).

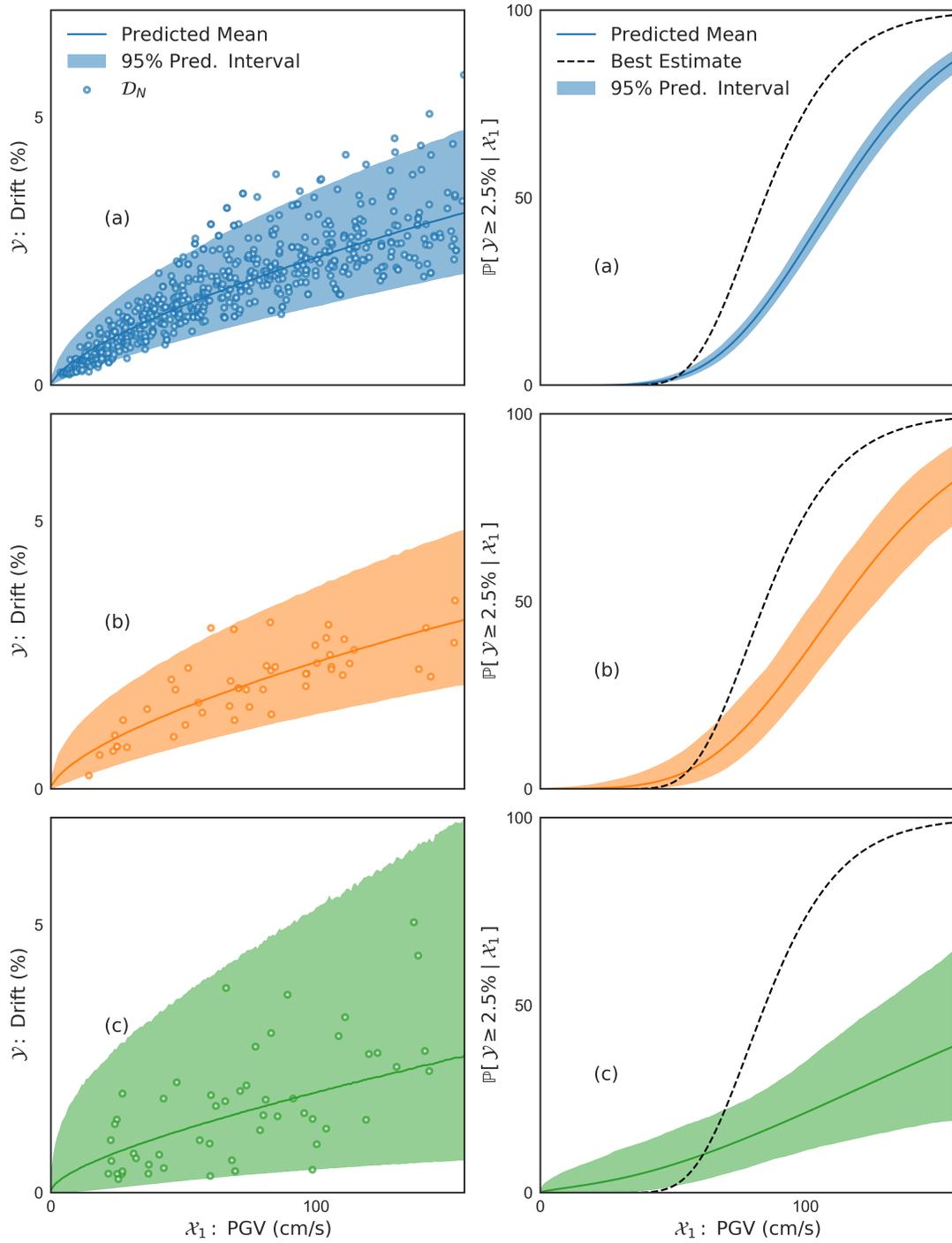


**Fig. 6.4.** Observations from scaled and original ground motions from SAC-Combined database

The results are included in Fig. 6.5 In the figure, the left subplots correspond to the structural response data: the colored dots correspond to the training dataset, the solid line is the mean while the colored area represents the 95% predictive interval for the structural response from the asset's probabilistic model. Additionally, the right subplots present the FFs for each training dataset. The dashed line corresponds to the best estimate previously derived using  $N = 10,000$  records from the California database. From Fig. 6.5, it is observed that the cases (a) 720 scaled ground motions and (b) 60 scaled records generate FFs that are completely different from the best estimate. This behavior is expected since the training datasets are derived from a totally different database, however the best estimate is included only for reference. But, what it is really important to realize is that the FFs from cases (a) and (b) differ significantly from the FF of the original dataset in case (c).

## 6.4 Conclusions

This chapter illustrates some of the common practices that introduces bias in the FF when defining or creating a training dataset of ground motions. The examples presented in this chapter show that the aleatory selection of ground motion databases, the combination between different databases, and the amplitude scaling of the ground motion records are



**Fig. 6.5.** Fragility analysis for (a)  $N = 720$  scaled, (b)  $N = 60$  scaled, and (c)  $N = 60$  original records

practices that can introduce bias in the resulting FF. For this reason, the ground motion database needs to be representative to the geographical location for the studied structure. For instance, the direction in which these motions are recorded or created needs to coincide with the rotation of the structure, among other site-specific conditions are important factors to perform an accurate fragility analysis, the one that can be used to make important decisions within a community.

## 7. SUMMARY AND CONCLUSIONS

In this dissertation, strategic statistical selection methodologies have been developed for civil engineering applications. These methodologies focus on closing the gaps that impede the extensive use of fragility analyses for real-world applications: unfamiliarity with the fragility functions, the potentially large uncertainty in such functions, the lack of quantification of this uncertainty, and the high computational demand needed for the analysis during the repetitively evaluation of ground motions. The implementation of the methodologies developed herein enable the **selection of the best model**, the **acceleration of the computation**, and the **quantification of the epistemic uncertainty** of more accurate fragility functions. Consequently, this implementation would provide the means to compute more efficient and reliable fragility functions, and thus enable much more extensive use of such analysis for structural applications in civil infrastructure. For instance, a possible application of the developed methodologies is in the field of structural assessment evaluation that is mostly dominated by two main alternatives: on-site inspection and structural health monitoring. Both alternatives require a large amount of valuable resources (e.g., qualified personnel, time, and money). For this reason, the implementation of more efficient assessment techniques, such as efficient fragility analysis, offers a cost-effective option for the development of a community.

To investigate the efficiency of the methodologies developed, a twenty-story benchmark building model and a large dataset of synthetic ground motions are used to demonstrate and validate the methodologies. Next, a brief summary of the topics and conclusions covered in this dissertation are summarized:

- In chapter 3, an analysis is conducted to identify the most representative features for the ground motion and structural response in the context of earthquake engineering. Rather than simply adopting the common input parameters of acceleration (e.g., peak

ground acceleration, spectral acceleration), a comprehensive study is carried out to determine the intensity measure that yields the lowest value of uncertainty.

- In chapter 4, a methodology is developed for selecting the best statistical model to describe the dispersion in the structural response. The methodology is based on a Bayesian model selection algorithm and validated through multiple comparison metrics. In addition, a quantification procedure for the epistemic uncertainty in fragility functions induced by finite datasets is introduced.
- In chapter 5, a methodology is developed for sequentially selecting the ground motions that produce the lowest values of uncertainty in fragility functions. This is achieved by estimating the value(s) of intensity measure yielding the largest reduction in epistemic uncertainty.
- In chapter 6, some widely-used practices that unintentionally bias fragility functions are discussed and evaluated. Within these practices, three cases are illustrated: selecting the ground motion database randomly, combining records from multiple databases, and scaling the ground motions.

The implementation of the selection methodologies developed herein will empower engineers to efficiently compute more accurate fragility functions, opening a venue for the fragility analysis to become the first option when dealing with structural assessments in the world. Additionally, the quantification and targeted reduction of the epistemic uncertainty of such function empower a better understanding of the community assets performance, an essential step towards an informed decision-making process, which may lead a more efficient allocation and mobilization of resources.

Although the methodologies developed here have been designed and illustrated in the context of buildings and ground motions, these methodologies are: **independent of the source of data** (e.g., numerical simulation, real world observations, hybrid), hazard, type of structure or system; and **transferable** to other critical infrastructure systems such as transportation, communication, energy, etc. However, it is important to clarify that these

methodologies have been developed under the assumption that the database of ground motions and the nonlinear benchmark building model are representative and realistic. This means that no uncertainty on these two aspects have been considered in this dissertation. It would make no sense to implement the model selection and sequential selection of ground motion methodologies proposed herein if the basis of the analysis, the ground motion database and the numerical model, has been poorly defined. For this reason, it is important to recall Eq. (1.1):

$$F(x; y_{\text{crit}}; \mathbb{I}) := \mathbb{P}[\mathcal{Y} \geq y_{\text{crit}} \mid \mathcal{X}_1 = x, \mathbb{I}] \quad (1.1)$$

and the importance of the information  $\mathbb{I}$  in the analysis. In fact,  $\mathbb{I}$  includes all what is considered to be known about the structure and the surrounding area that impacts the selection of the ground motion database. For this reason, it is crucial to incorporate all known information about the system conditions in  $\mathbb{I}$ . In addition, every unknown aspect that affects the fragility function should be averaged out. For example, the epicentral distance, type of seismic fault, soil classification, building orientation, geographical location, presence of saturated soil, occupancy, among other factors must be considered in the analysis or otherwise averaged out.

And, all the uncertain information needs to be averaged in the analysis in order to generate the best unbiased estimate of the fragility function. Ultimately, what we think we know but may not be true is something that should scare us.

Some recommendations for future studies extending the contributions of the methodologies developed in this dissertation include:

- Updating the fragility functions automatically to account for non-stationary processes such as the aging and degradation of the structural materials.
- Investigating the possibility of scaling these methodologies to encompass a larger number of structures to facilitate the analysis at a community level.
- Exploring how this implementation facilitates fundamental aspects at a community level, such as life-cycle management, decision-making process, and efficient allocation and mobilization of resources.

## REFERENCES

- [1] Chul Min Yeum and Shirley J Dyke. Vision-based automated crack detection for bridge inspection. *Computer-Aided Civil and Infrastructure Engineering*, 30(10):759–770, 2015.
- [2] Charles R Farrar and Keith Worden. An introduction to structural health monitoring. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365(1851):303–315, 2007.
- [3] NM Newmark. Probability of predicted seismic damage in relation to nuclear reactor facility design. *Consulting Engineering Services*, 1975.
- [4] Alfredo Hua-Sing Ang and Nathan Mortimore Newmark. *A probabilistic seismic safety assessment of the Diablo Canyon Nuclear Power Plant*. NM Newmark Consulting Engineering Services, 1977.
- [5] C Allin Cornell and Nathan Mortimore Newmark. Seismic reliability of nuclear power plants. In *Probabilistic analysis of nuclear reactor safety*. 1978.
- [6] James E Richardson, Goutam Bagchi, and Rutlage J Brazee. The seismic safety margins research program of the US Nuclear Regulatory Commission. *Nuclear Engineering and Design*, 59(1):15–25, 1980.
- [7] K Pitilakis, H Crowley, and AM Kaynia. Syner-g: typology definition and fragility functions for physical elements at seismic risk. *Geotechnical, Geological and Earthquake Engineering*, 27, 2014.
- [8] K. Porter. Beginner’s guide to fragility, vulnerability, and risk. *Encyclopedia of Earthquake Engineering*, 2015.
- [9] Bruce R Ellingwood, David V Rosowsky, Yue Li, and Jun Hee Kim. Fragility assessment of light-frame wood construction subjected to wind and earthquake hazards. *Journal of Structural Engineering*, 130(12):1921–1930, 2004.
- [10] YK Wen and BR Ellingwood. The role of fragility assessment in consequence-based engineering. *Earthquake Spectra*, 21(3):861–877, 2005.
- [11] Seong-Hoon Jeong and Amr S Elnashai. Probabilistic fragility analysis parameterized by fundamental response quantities. *Engineering Structures*, 29(6):1238–1251, 2007.
- [12] Robert P Kennedy, CA Cornell, RD Campbell, S Kaplan, and HF Perla. Probabilistic seismic safety study of an existing nuclear power plant. *Nuclear Engineering and Design*, 59(2):315–338, 1980.
- [13] G Michele Calvi, Rui Pinho, Guido Magenes, Julian J Bommer, L Fernando Restrepo-Vélez, and Helen Crowley. Development of seismic vulnerability assessment methodologies over the past 30 years. *ISET journal of Earthquake Technology*, 43(3):75–104, 2006.

- [14] Murat Erberik. Seismic fragility analysis. *Encyclopedia of Earthquake Engineering*, pages 1–10, 2015.
- [15] Irmela Zentner, Max Gündel, and Nicolas Bonfils. Fragility analysis methods: Review of existing approaches and application. *Nuclear Engineering and Design*, 323:245–258, 2017.
- [16] RP Kennedy and MK Ravindra. Seismic fragilities for nuclear power plant risk studies. *Nuclear Engineering and Design*, 79(1):47–68, 1984.
- [17] E. Taylor. The development of fragility relationships for controlled structures, 2007.
- [18] A. Wilbee, F. Peña, J. Condori, and S. Dyke. Fragility analysis of structures incorporating control systems. In *Proceedings 6th International Conference on Advances in Experimental Structural Engineering, 11th International Workshop on Advanced Smart Materials and Smart Structures Technology*. University of Illinois, 2015.
- [19] Masanobu Shinozuka, Maria Q Feng, Jongheon Lee, and Toshihiko Naganuma. Statistical analysis of fragility curves. *Journal of Engineering Mechanics*, 126(12):1224–1231, 2000.
- [20] JW Baker. Fitting fragility functions to structural analysis data using maximum likelihood estimation. *Technical note, Stanford University*, 2011.
- [21] David Lallemand, Anne Kiremidjian, and Henry Burton. Statistical procedures for developing earthquake damage fragility curves. *Earthquake Engineering & Structural Dynamics*, 44(9):1373–1389, 2015.
- [22] Pierre Gehl, John Douglas, and Darius M Seyed. Influence of the number of dynamic analyses on the accuracy of structural response estimates. *Earthquake Spectra*, 31(1):97–113, 2015.
- [23] Hae Young Noh, David Lallemand, and Anne S Kiremidjian. Development of empirical and analytical fragility functions using kernel smoothing methods. *Earthquake Engineering & Structural Dynamics*, 44(8):1163–1180, 2015.
- [24] Dimitrios Vamvatsikos and C Allin Cornell. Incremental dynamic analysis. *Earthquake Engineering & Structural Dynamics*, 31(3):491–514, 2002.
- [25] Chara Ch Mitropoulou and Manolis Papadrakakis. Developing fragility curves based on neural network ida predictions. *Engineering Structures*, 33(12):3409–3421, 2011.
- [26] Sigmund A Freeman. The capacity spectrum method. In *Proceedings of the 11th European conference on earthquake engineering, Paris*, 1998.
- [27] Masanobu Shinozuka, Maria Q Feng, Ho-Kyung Kim, and Sang-Hoon Kim. Non-linear static procedure for fragility curve development. *Journal of Engineering Mechanics*, 126(12):1287–1295, 2000.
- [28] Sang Whan Han, Ki-Hoon Moon, and Anil K Chopra. Application of mpa to estimate probability of collapse of structures. *Earthquake Engineering & Structural Dynamics*, 39(11):1259–1278, 2010.
- [29] Bryant G Nielson and Reginald DesRoches. Seismic fragility methodology for highway bridges using a component level approach. *Earthquake Engineering & Structural Dynamics*, 36(6):823–839, 2007.

- [30] Francisco Pena, Ilias Bilonis, and Shirley Dyke. Model selection and uncertainty quantification of seismic fragility functions (accepted). *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 2019. doi:10.1061/AJRUA6.0001014.
- [31] I Zentner, N Humbert, S Ravet, and E Viallet. Numerical methods for seismic fragility analysis of structures and components in nuclear industry-application to a reactor coolant system. *Georisk*, 5(2):99–109, 2011.
- [32] B Sudret, CV Mai, and K Konakli. Computing seismic fragility curves using non-parametric representations. *Earthquake Eng. Struct. Dyn.*, 2014.
- [33] Yue Li and Bruce R Ellingwood. Hurricane damage to residential construction in the US: Importance of uncertainty modeling in risk assessment. *Engineering structures*, 28(7):1009–1018, 2006.
- [34] Marra A Smith and Luca Caracoglia. A monte carlo based method for the dynamic fragility analysis of tall buildings under turbulent wind loading. *Engineering Structures*, 33(2):410–420, 2011.
- [35] David B Roueche, David O Prevatt, and Franklin T Lombardo. Epistemic uncertainties in fragility functions derived from post-disaster damage assessments. *ASCE-ASME journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 4(2):04018015, 2018.
- [36] Howard HM Hwang and Jun-Rong Huo. Generation of hazard-consistent fragility curves. *Soil Dynamics and Earthquake Engineering*, 13(5):345–354, 1994.
- [37] Yutao Pang, Xun Wu, Guoyu Shen, and Wancheng Yuan. Seismic fragility analysis of cable-stayed bridges considering different sources of uncertainties. *Journal of Bridge Engineering*, 19(4):04013015, 2013.
- [38] Jayadipta Ghosh and Jamie E Padgett. Aging considerations in the development of time-dependent seismic fragility curves. *Journal of Structural Engineering*, 136(12):1497–1511, 2010.
- [39] J. Li, B. Spencer, and A. Elnashai. Bayesian updating of fragility functions using hybrid simulation. *journal of Structural Engineering*, 139(7):1160–1171, 2012.
- [40] Y. Wen, B. Ellingwood, and J. Bracci. Vulnerability function framework for consequence-based engineering. *MAE Center Report 04-04*, April 2004.
- [41] Keith Porter, Robert Kennedy, and Robert Bachman. Creating fragility functions for performance-based earthquake engineering. *Earthquake Spectra*, 23(2):471–489, 2007.
- [42] Amir M Kaynia, Fabio Taucer, and Ufuk Hancilar. Guidelines for deriving seismic fragility functions of elements at risk: Buildings, lifelines, transportation networks and critical facilities. Technical report, European Commission-Joint Research Centre, 2013.
- [43] T Rossetto, I Ioannou, and DN Grant. Existing empirical fragility and vulnerability relationships: compendium and guide for selection. *Pavia, Italy: GEM Foundation*, 2013.

- [44] Fernando Moreu and Billie F Spencer Jr. Framework for consequence-based management and safety of railroad bridge infrastructure using wireless smart sensors (wss). Technical report, Newmark Structural Engineering Laboratory. University of Illinois at Urbana-Champaign., 2015.
- [45] Fadzli Mohamed Nazari. *Seismic Fragility Assessment for Buildings Due to Earthquake Excitation*. Springer, 2018.
- [46] Nuclear Regulatory Commission et al. Reactor safety study. an assessment of accident risks in US commercial nuclear power plants. executive summary: main report. Technical report, Nuclear Regulatory Commission, 1975.
- [47] Harold Walter Lewis, Robert J Budnitz, WD Rowe, HJC Kouts, F Von Hippel, WB Loewenstein, and F Zachariasen. Risk assessment review group report to the US nuclear regulatory commission. *IEEE Transactions on Nuclear Science*, 26(5):4686–4690, 1979.
- [48] PP Zemanick. Evolution of structural reliability methodology as applied in major nuclear risk assessments. *Nuclear Engineering and Design*, 60(1):163–168, 1980.
- [49] RP Kennedy, RD Campbell, G Hardy, and H Banon. Subsystem fragility: Seismic safety margins research program (phase i). Technical report, Lawrence Livermore National Lab., CA (USA), 1981.
- [50] PD Smith, RG Dong, DL Bernreuter, MP Bohn, TY Chuang, GE Cummings, JJ Johnson, RW Mensing, and JE Wells. Seismic safety margins research program. phase i final report-overview. Technical report, Lawrence Livermore Laboratory, 1981.
- [51] LE Cover. Equipment fragility data base, seismic safety margins research program. Technical report, Lawrence Livermore National Lab., CA (USA), 1983.
- [52] LE Cover, MP Bohn, RD Campbell, and DA Wesley. Handbook of nuclear power plant seismic fragilities, seismic safety margins research program. Technical report, Lawrence Livermore National Lab., CA (USA), 1983.
- [53] NM Newmark. Comments on conservatism in earthquake resistant design. *Report to US Nuclear Regulatory Commission, NM Newmark Consulting Engineering Services, Urbana, IL (September 1974)*, 1974.
- [54] EPRI. Seismic probabilistic risk assessment implementation guide. Technical report, Electric Power Research Institute, Palo Alto, CA., 2003.
- [55] T. Rossetto and A. Elnashai. Derivation of vulnerability functions for European-type RC structures based on observational data. *Engineering Structures*, 2003.
- [56] ATC. Earthquake damage evaluation data for california. Technical report, American Technology Council, Report ATC-13, Redwood City, CA., 1985.
- [57] ATC. Seismic evaluation and retrofit of concrete buildings. Technical report, American Technology Council, Report ATC-40, Redwood City, CA., 1996.
- [58] A. Wilbee, F. Peña, S. Dyke, I. Billionis, and P. Pandita. Quantifying recovery if structures incorporating control. In *Proceedings International Conference on Smart Infrastructure and Construction (ICSIC)*. University of Cambridge, 2016.

- [59] Oh-Sung Kwon and Amr Elnashai. Sensitivity of analytical vulnerability functions to input and response parameter randomness. In *Proceedings of the 13th World Conference on Earthquake Engineering*, 2004.
- [60] George E. P Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.
- [61] Jon C Helton and Freddie Joe Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety*, 81(1):23–69, 2003.
- [62] JB Jernigan and HM Hwang. Inventory and fragility analysis of memphis bridges. *Center for Earthquake Research and Information, University of Memphis, Memphis TN, Sept*, 15, 1997.
- [63] E Borgonovo, Irmela Zentner, A Pellegrini, Stefano Tarantola, and Etienne de Rocquigny. On the importance of uncertain factors in seismic fragility assessment. *Reliability Engineering & System Safety*, 109:66–76, 2013.
- [64] Jamie Ellen Padgett and Reginald DesRoches. Sensitivity of seismic response and fragility to parameter uncertainty. *Journal of Structural Engineering*, 133(12):1710–1718, 2007.
- [65] Nirmal Jayaram and Jack W Baker. Statistical tests of the joint distribution of spectral acceleration values. *Bulletin of the Seismological Society of America*, 98(5):2231–2243, 2008.
- [66] Brendon A Bradley. A generalized conditional intensity measure approach and holistic ground-motion selection. *Earthquake Engineering & Structural Dynamics*, 39(12):1321–1342, 2010.
- [67] CB Haselton, AS Whittaker, A Hortacsu, JW Baker, J Bray, and DN Grant. Selecting and scaling earthquake ground motions for performing response-history analyses. In *Proceedings of the 15th World Conference on Earthquake Engineering*, pages 4207–4217. Earthquake Engineering Research Institute, 2012.
- [68] Ting Lin, Curt B. Haselton, and Jack W. Baker. Conditional spectrum-based ground motion selection. part ii: Intensity-based assessments and evaluation of alternative target spectra. *Earthquake Engineering & Structural Dynamics*, 42(12):1867–1884, 2013.
- [69] Katsuichiro Goda. Record selection for aftershock incremental dynamic analysis. *Earthquake Engineering & Structural Dynamics*, 44(7):1157–1162, 2015.
- [70] Luis Sousa, Vitor Silva, Mário Marques, and Helen Crowley. On the treatment of uncertainties in the development of fragility functions for earthquake loss estimation of building portfolios. *Earthquake Engineering & Structural Dynamics*, 45(12):1955–1976, 2016.
- [71] Mustafa A Mohamad and Themistoklis P Sapsis. A sequential sampling strategy for extreme event statistics in nonlinear dynamical systems. *arXiv preprint arXiv:1804.07240*, 2018.
- [72] Paul G Somerville. *Development of ground motion time histories for phase 2 of the FEMA/SAC steel project*. SAC Joint Venture, 1997.

- [73] Prishati Raychowdhury. Seismic response of low-rise steel moment-resisting frame (SMRF) buildings incorporating nonlinear soil-structure interaction (SSI). *Engineering Structures*, 33(3):958–967, 2011.
- [74] Audrey Olivier and Andrew W Smyth. Trade offs between statistical agreement and data reproduction in the generation of synthetic ground motions. *Probabilistic Engineering Mechanics*, 43:36–49, 2016.
- [75] Nicolas Luco and Paolo Bazzurro. Does amplitude scaling of ground motion records result in biased nonlinear structural drift responses? *Earthquake Engineering & Structural Dynamics*, 36(13):1813–1835, 2007.
- [76] YN Huang, DV Ha, and A Samanta. Scaling ground motions for response-history analysis of tall buildings. *Proceedings, Paper*, (4577), 2012.
- [77] Jonathan Hancock, Jennie Watson-Lamprey, Norman A Abrahamson, Julian J Bommer, Alexandros Markatis, EMMA McCOY, and Rishmila Mendis. An improved method of matching response spectra of recorded earthquake ground motion using wavelets. *Journal of earthquake engineering*, 10(spec01):67–89, 2006.
- [78] YeongAe Heo, Sashi K Kunnath, and Norman Abrahamson. Amplitude-scaled versus spectrum-matched ground motions for seismic performance assessment. *Journal of Structural Engineering*, 137(3):278–288, 2010.
- [79] Damian N Grant and Riccardo Diaferia. Assessing adequacy of spectrum-matched ground motions for response history analysis. *Earthquake Engineering & Structural Dynamics*, 42(9):1265–1280, 2013.
- [80] Zhiyi Wang, Irmela Zentner, and Enrico Zio. A bayesian framework for estimating fragility curves based on seismic damage data and numerical simulations by adaptive neural networks. *Nuclear Engineering and Design*, 338:232–246, 2018.
- [81] Jack W Baker. Efficient analytical fragility function fitting using dynamic structural analysis. *Earthquake Spectra*, 31(1):579–599, 2015.
- [82] Jong-Wha Bai. *Seismic fragility and retrofitting for a reinforced concrete flat-slab structure*. PhD thesis, Texas A&M University, 2004.
- [83] Waleed T Barnawi and Shirley J Dyke. Seismic fragility relationships of a cable-stayed bridge equipped with response modification systems. *Journal of Bridge Engineering*, 19(8):A4013003, 2013.
- [84] Iman Behmanesh and Babak Moaveni. Probabilistic identification of simulated damage on the dowling hall footbridge through bayesian finite element model updating. *Structural Control and Health Monitoring*, 22(3):463–483, 2015.
- [85] Iman Behmanesh, Babak Moaveni, Geert Lombaert, and Costas Papadimitriou. Hierarchical bayesian model updating for structural identification. *Mechanical Systems and Signal Processing*, 64:360–376, 2015.
- [86] Technical Office SAC Steel Project. SAC Steel Project, 1994.
- [87] Y Ohtori and BF Spencer. A matlab-based tool for nonlinear structural analysis. In *Proc. of the 13th Engineering Mechanics Conf*, pages 13–16, 1999.

- [88] Y Ohtori, RE Christenson, BF Spencer Jr, and SJ Dyke. Benchmark control problems for seismically excited nonlinear buildings. *Journal of Engineering Mechanics*, 130(4):366–385, 2004.
- [89] George P Mavroeidis and Apostolos S Papageorgiou. A mathematical representation of near-fault ground motions. *Bulletin of the seismological society of America*, 93(3):1099–1131, 2003.
- [90] Apostolos S Papageorgiou and Keiiti Aki. A specific barrier model for the quantitative description of inhomogeneous faulting and the prediction of strong ground motion. i. description of the model. *Bulletin of the Seismological Society of America*, 73(3):693–722, 1983.
- [91] David M Boore. Stochastic simulation of high-frequency ground motions based on seismological models of the radiated spectra. *Bulletin of the Seismological Society of America*, 73(6A):1865–1894, 1983.
- [92] Benedikt Halldórsson and Apostolos S Papageorgiou. Calibration of the specific barrier model to earthquakes of different tectonic regions. *Bulletin of the Seismological society of America*, 95(4):1276–1300, 2005.
- [93] Donald L Wells and Kevin J Coppersmith. New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bulletin of the seismological Society of America*, 84(4):974–1002, 1994.
- [94] Benedikt Halldórsson, George P Mavroeidis, and Apostolos S Papageorgiou. Near-fault and far-field strong ground-motion simulation for earthquake engineering applications using the specific barrier model. *Journal of Structural Engineering*, 137(3):433–444, 2011.
- [95] Timothy G Cork, Jung Han Kim, George P Mavroeidis, Jae Kwan Kim, Benedikt Halldórsson, and Apostolos S Papageorgiou. Effects of tectonic regime and soil conditions on the pulse period of near-fault ground motions. *Soil Dynamics and Earthquake Engineering*, 80:102–118, 2016.
- [96] Shrey K Shahi and Jack W Baker. An efficient algorithm to identify strong-velocity pulses in multicomponent ground motions. *Bulletin of the Seismological Society of America*, 104(5):2456–2466, 2014.
- [97] Marco De Biasio, Stephane Grange, Frederic Dufour, Frederic Allain, and Ilie Petre-Lazar. A simple and efficient intensity measure to account for nonlinear structural behavior. *Earthquake Spectra*, 30(4):1403–1426, 2014.
- [98] FEMA. *FEMA 356 Prestandard and commentary for the seismic rehabilitation of buildings*. Federal Emergency Management Agency (FEMA), Washington, D.C., 2000.
- [99] Young-Ji Park and Alfredo H-S Ang. Mechanistic seismic damage model for reinforced concrete. *Journal of structural engineering*, 111(4):722–739, 1985.
- [100] ASCE. *Seismic Rehabilitation of Existing Buildings*. Reston, Va. : American Society of Civil Engineers, 2007.
- [101] Y. Cha, A. Agrawal, B. Phillips, and B. Spencer. Direct performance-based design with 200kn MR dampers using multi-objective cost effective optimization for steel MRFs. *Engineering structures*, 2014.

- [102] A Tan and A Irfanoglu. Correlation between ground motion based shaking intensity estimates and actual building damage. In *15th World Conference on Earthquake Engineering, Lisbon, Portugal*, 2012.
- [103] Francisco Pena, Ilias Bilonis, and Shirley Dyke. Model selection and uncertainty quantification of seismic fragility functions. *DesignSafe-CI*, 2019. doi:10.17603/DS2498Z.
- [104] Frank J Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.
- [105] Walter Gautschi. Algorithm 726: Orthpol—a package of routines for generating orthogonal polynomials and gauss-type quadrature rules. *ACM Transactions on Mathematical Software (TOMS)*, 20(1):21–62, 1994.
- [106] Ilias Bilonis. py-orthpol: Construct orthogonal polynomials in python, 2013.
- [107] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.
- [108] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.
- [109] Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.
- [110] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.
- [111] Jiang Wan and Nicholas Zabarar. A bayesian approach to multiscale inverse problems using the sequential monte carlo method. *Inverse Problems*, 27(10):105004, 2011.
- [112] Ilias Bilonis. pySMC 1.0 sequential monte carlo in python, 2014.
- [113] Ilias Bilonis, Beth A Drewniak, and Emil M Constantinescu. Crop physiology calibration in the clm. *Geoscientific Model Development*, 8(4):1071–1083, 2015.
- [114] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.
- [115] Adam Dachowicz, Siva Chaitanya Chaduvula, Mikhail J Atallah, Ilias Bilonis, and Jitesh H Panchal. Strategic information revelation in collaborative design. *Advanced Engineering Informatics*, 36:242–253, 2018.
- [116] Adrian E Raftery. Bayesian model selection in social research. *Sociological methodology*, pages 111–163, 1995.
- [117] Mary Natrella. Nist/sematech e-handbook of statistical methods. 2010.
- [118] Rohit Tripathy, Ilias Bilonis, and Marcial Gonzalez. Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. *journal of Computational Physics*, 321:191–223, 2016.

- [119] Francisco Pena, Ilias Bilonis, Shirley Dyke, Yenan Cao, and George Mavroeidis. Efficient seismic fragility functions through sequential selection (under review). *Probabilistic Engineering Mechanics*, 2019.
- [120] Kiyoshi Kanai. Semi-empirical formula for the seismic characteristics of the ground. 1957.
- [121] Hiroshi Tajimi. Statistical method of determining the maximum response of building structure during an earthquake. *Proc. of the 2nd WCEE*, 2:781–798, 1960.
- [122] Jeen-Shang Lin and Jaung-Yann Tyan. Equivalent stationary motion and average response spectra. *Earthquake engineering & structural dynamics*, 14(2):267–279, 1986.
- [123] Stijin De Vuyst. Relation between normal and lognormal distribution. [https://upload.wikimedia.org/wikipedia/commons/4/4e/Lognormal\\_Distribution.svg](https://upload.wikimedia.org/wikipedia/commons/4/4e/Lognormal_Distribution.svg), 12 2016. Wikipedia.

## APPENDICES

## A. DERIVATION OF FRAGILITY FUNCTION UNDER LOGNORMAL DISTRIBUTION ASSUMPTION

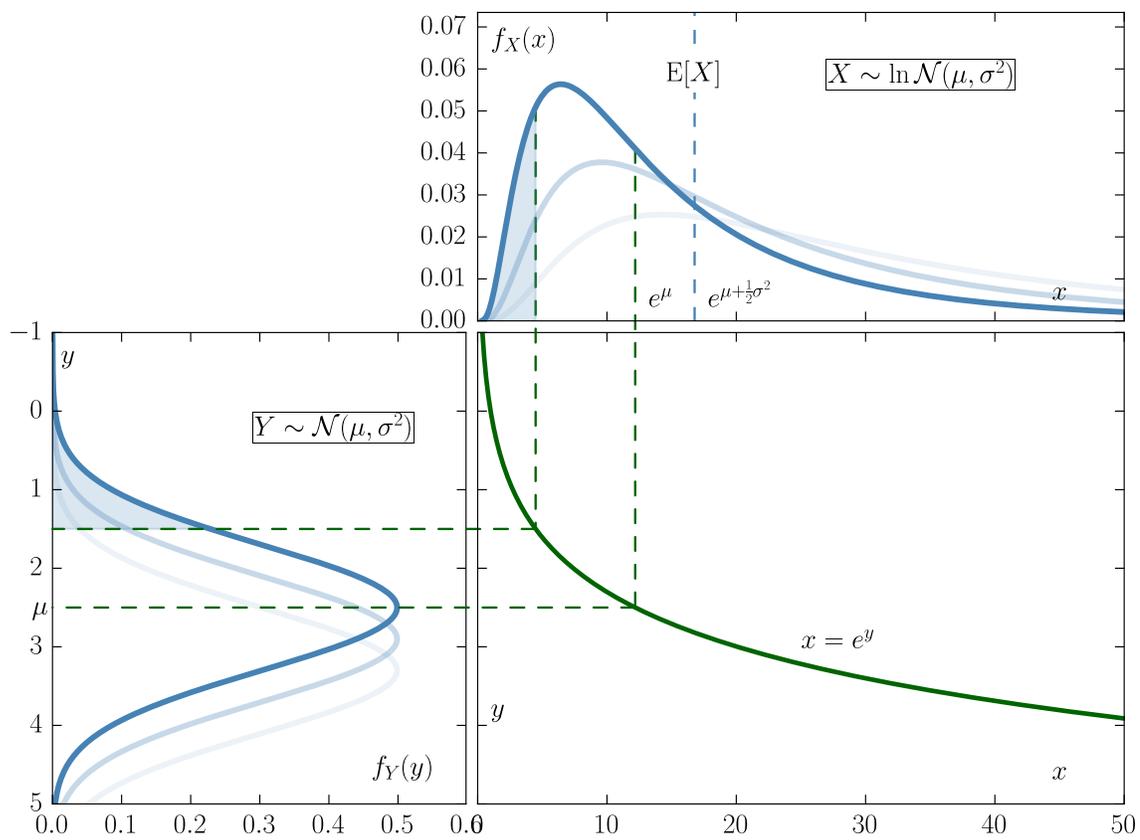
Here, it is assumed that the structural response  $\mathcal{Y}$  follows a lognormal distribution (i.e.,  $\mathcal{Y} \sim \text{lognormal}(\mu, \sigma^2)$ ). This means that the natural logarithm of  $\mathcal{Y}$  can be modelled using a normal distribution, i.e.,  $\log(\mathcal{Y}) \sim \mathcal{N}(\mu, \sigma^2)$ . Before the FF is calculated, it is necessary to define the standard normal variable  $\mathbf{Z} \sim \mathcal{N}(0, 1)$  to fully understand the derivation.

The probability of the structural response to exceed the critical level  $y_{\text{crit}}$  after the occurrence of an earthquake with intensity  $\mathcal{X}_1 = x$  is the FF and it can be expressed as:

$$\begin{aligned}
 F(x; y_{\text{crit}}) &:= \mathbb{P}[\mathcal{Y} > y_{\text{crit}}] && \text{: Using CCDF} \\
 &= 1 - \mathbb{P}[\mathcal{Y} \leq y_{\text{crit}}] && \text{: Using CDF} \\
 &= 1 - \mathbb{P}[\log(\mathcal{Y}) \leq (y_{\text{crit}})] && \text{: Logarithmic transformation} \\
 &= 1 - \mathbb{P}[\log(\mathcal{Y}) - \mu \leq (y_{\text{crit}}) - \mu] && \text{: Translation} \\
 &= 1 - \mathbb{P}\left[\frac{\log(\mathcal{Y}) - \mu}{\sigma} \leq \frac{(y_{\text{crit}}) - \mu}{\sigma}\right] && \text{: Scaling} \\
 &= 1 - \mathbb{P}\left[\mathbf{Z} \leq \frac{(y_{\text{crit}}) - \mu}{\sigma}\right] && \text{: Standard normal variable} \\
 &= 1 - \Phi\left(\frac{(y_{\text{crit}}) - \mu}{\sigma}\right) && \text{: Standard normal variable}
 \end{aligned}$$

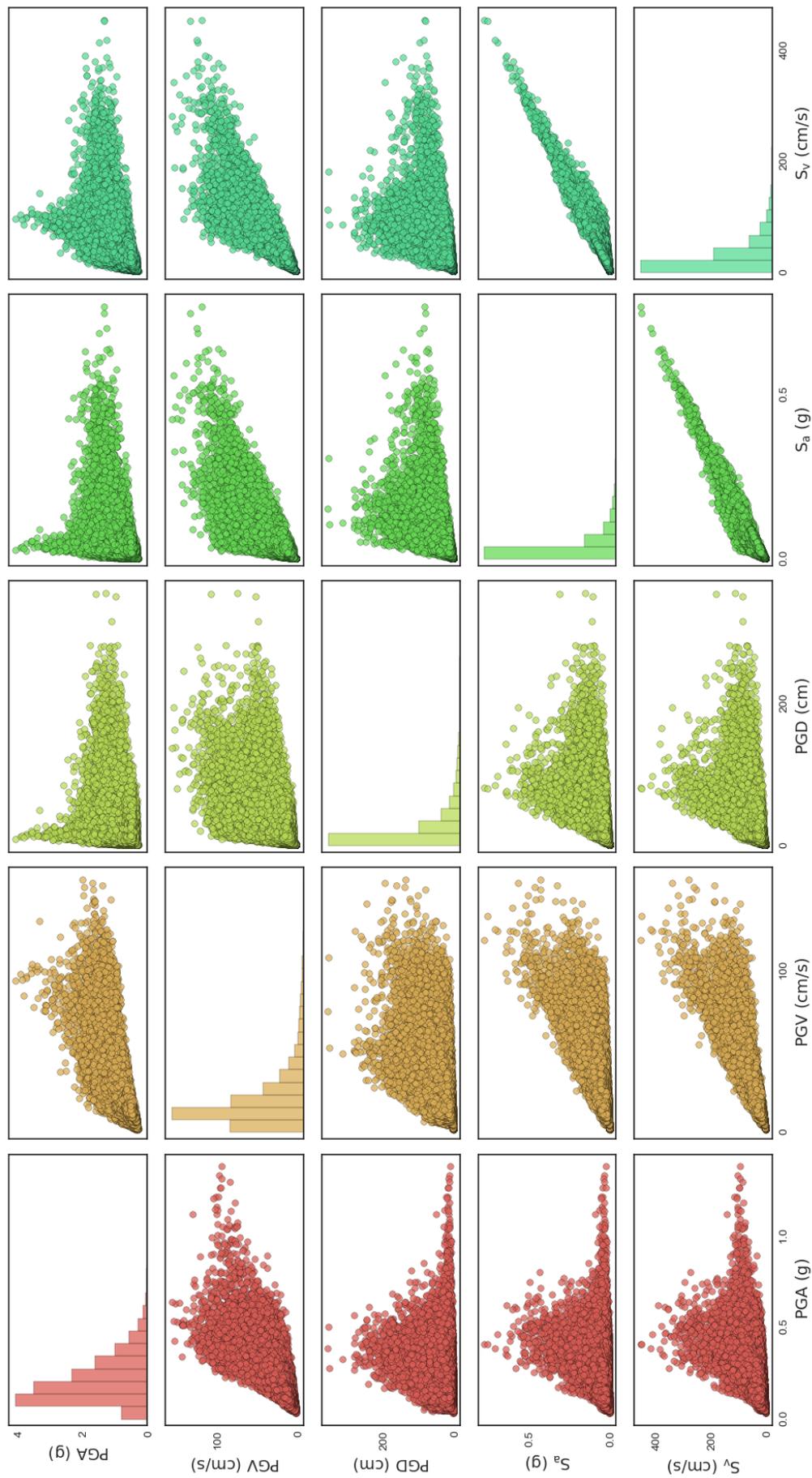
where  $\Phi$  is the CDF of the standard normal distribution. It is important to clarify that the value of the CDF remains constant after applying transformation, translation, and scaling to the analyzed random variable. Perhaps it is easier to understand this in a graphical manner. Fig. A.1 shows the relationship between a normal and a lognormal random variable, where the blue area represents the CDF at a random point  $y^*$  (normal), equivalent to  $x^* = \exp(y^*)$

(lognormal). The ratio between the blue area and the total area is equivalent to the CDF and it remains constant after the exponential/logarithmic transformation.

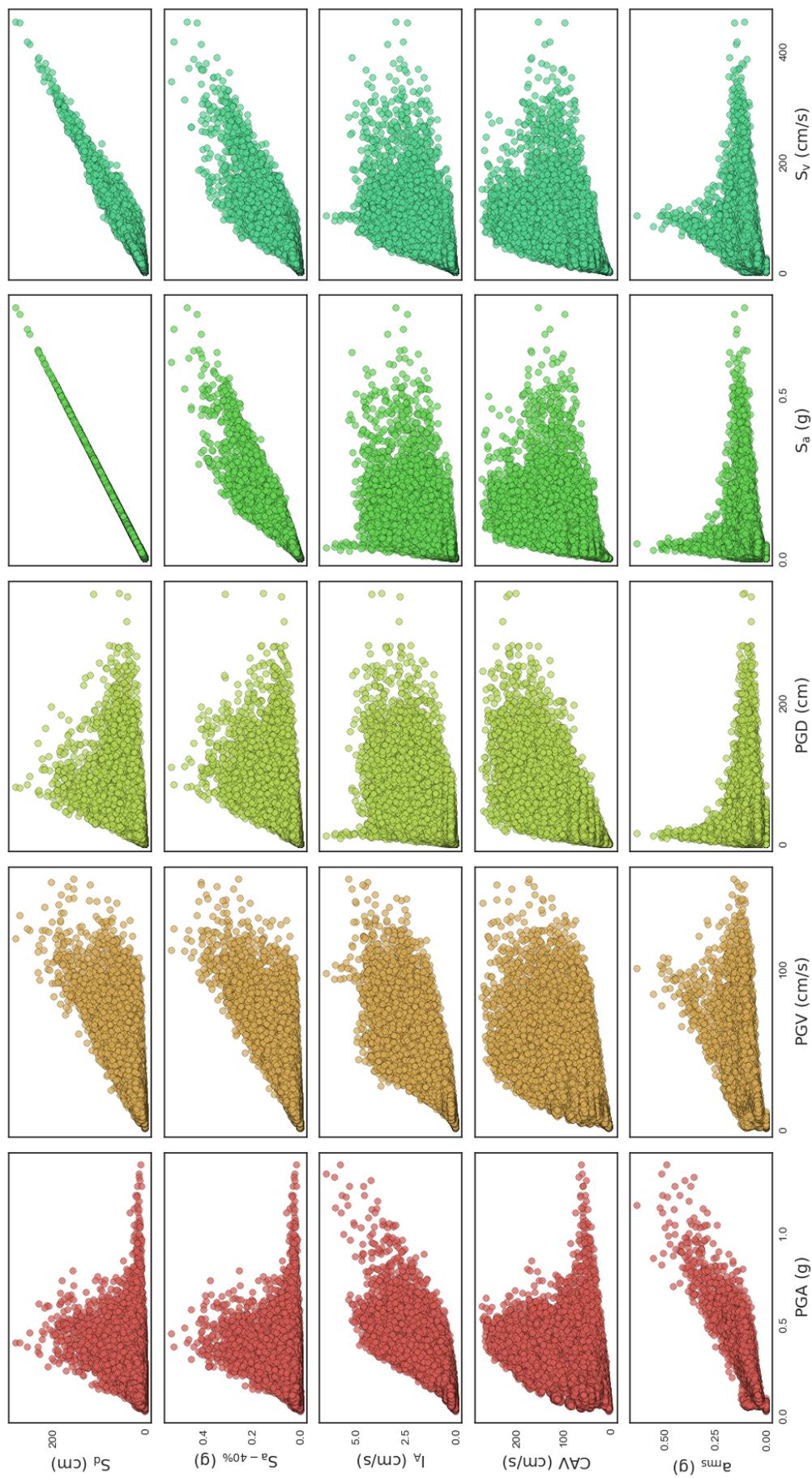


**Fig. A.1.** Relationship between normal and lognormal RV (Courtesy of [123])

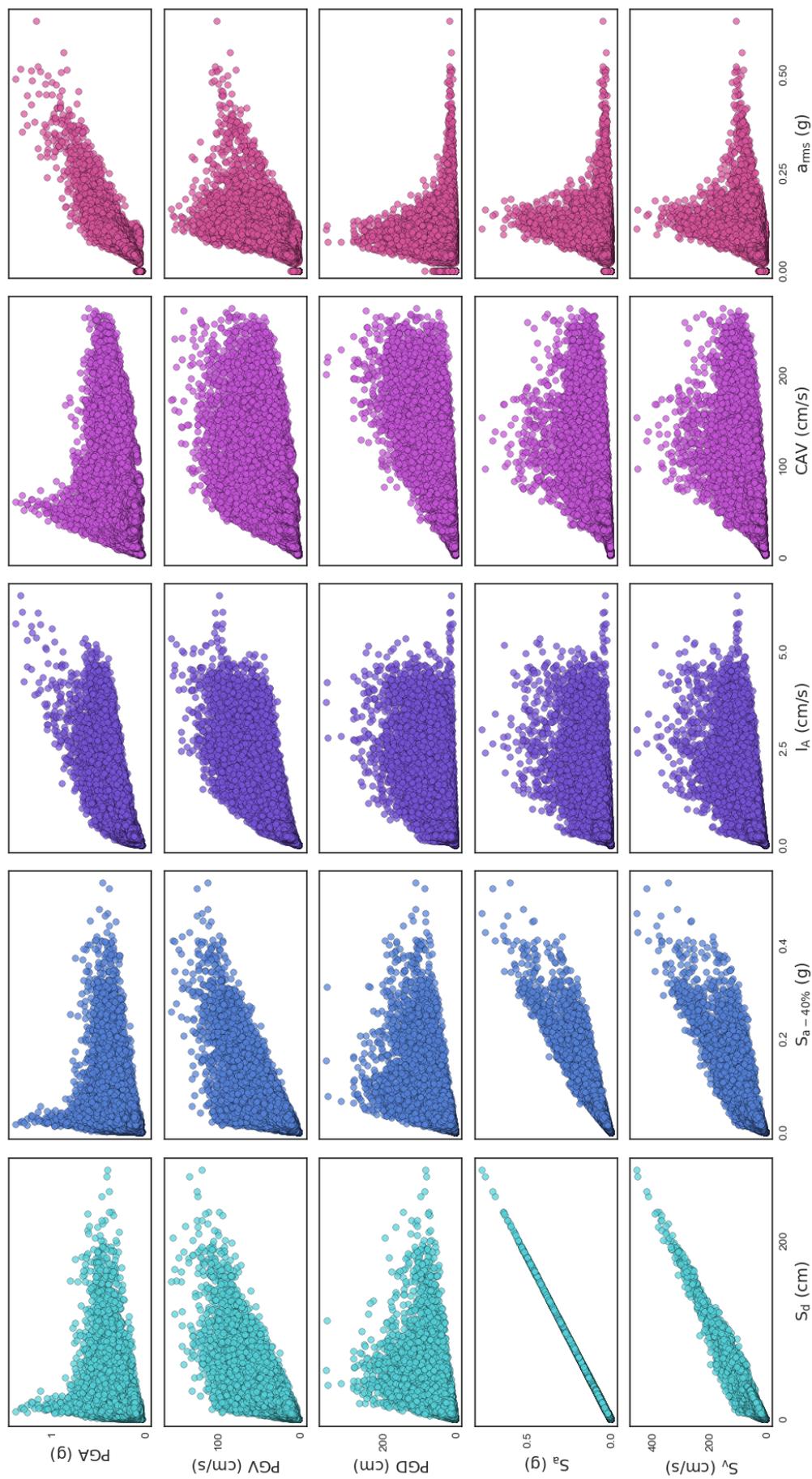
## B. SELECTION OF INPUT PARAMETERS



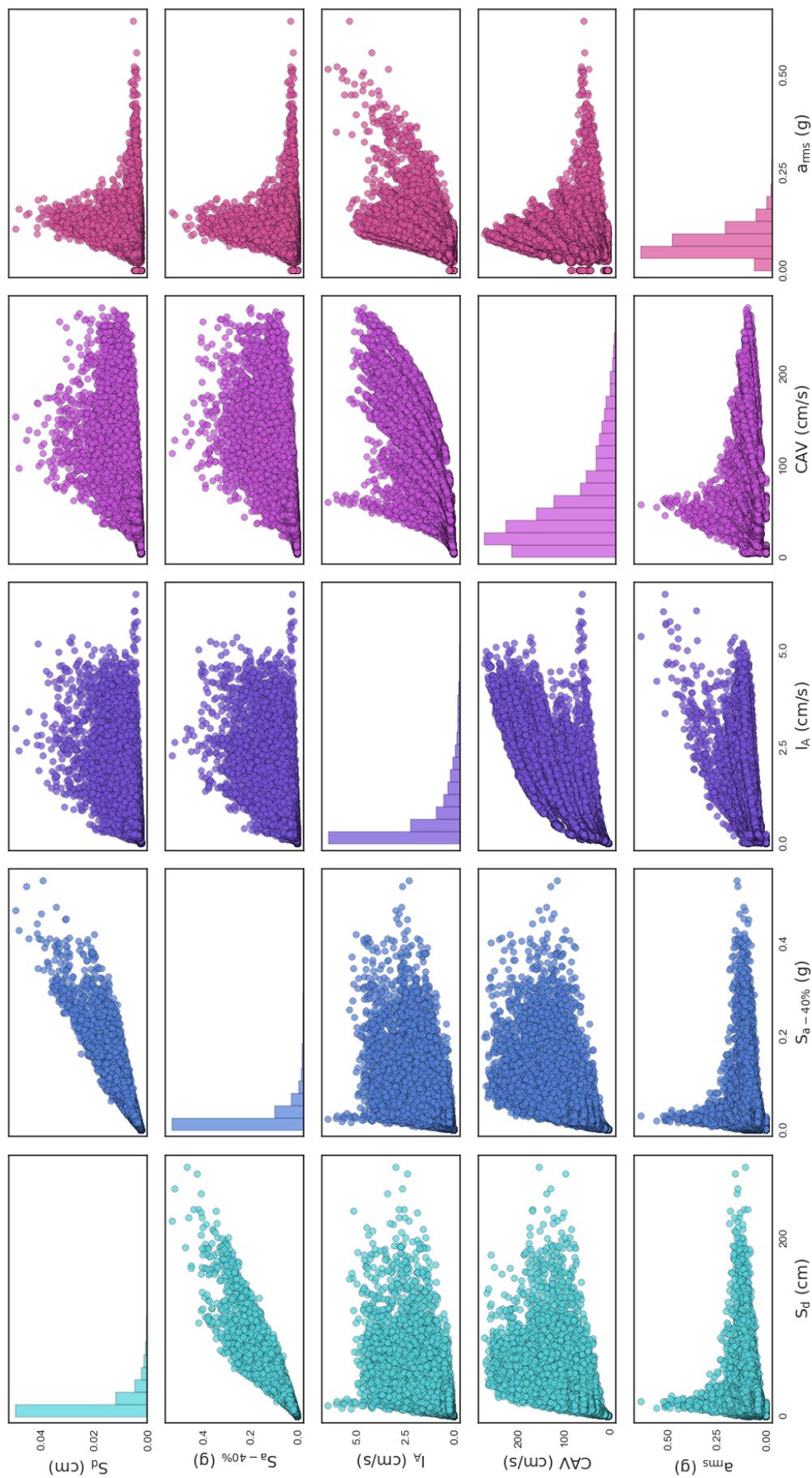
**Fig. B.1.** Scatter matrix plot for  $\mathbf{X}$  (top left portion of Fig. 3.3)



**Fig. B.2.** Scatter matrix plot for **X** (bottom left portion of Fig. 3.3)

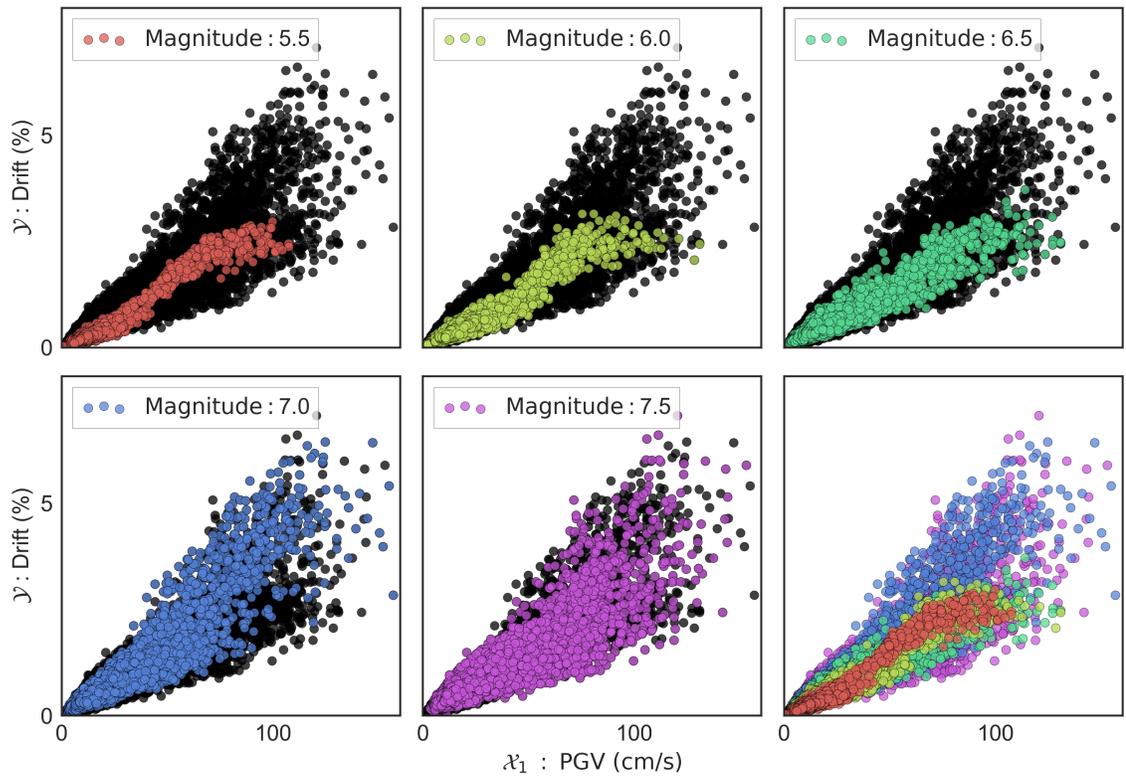


**Fig. B.3.** Scatter matrix plot for X (top right portion of Fig. 3.3)

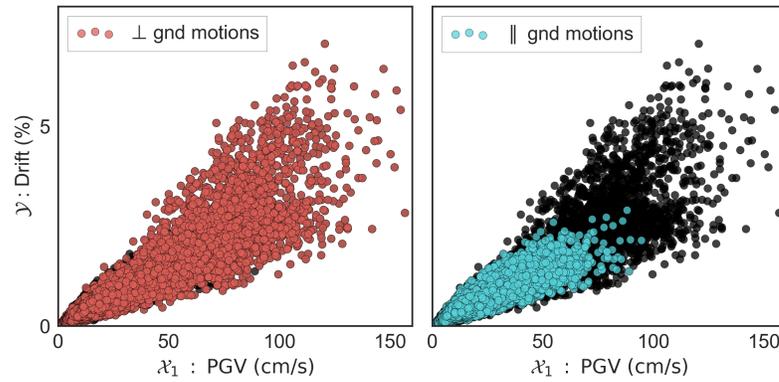


**Fig. B.4.** Scatter matrix plot for  $\mathbf{X}$  (bottom right portion of Fig. 3.3)

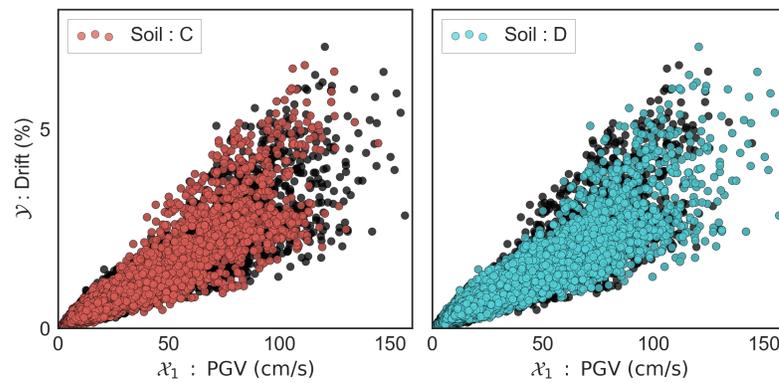
## C. EARTHQUAKE DATABASE



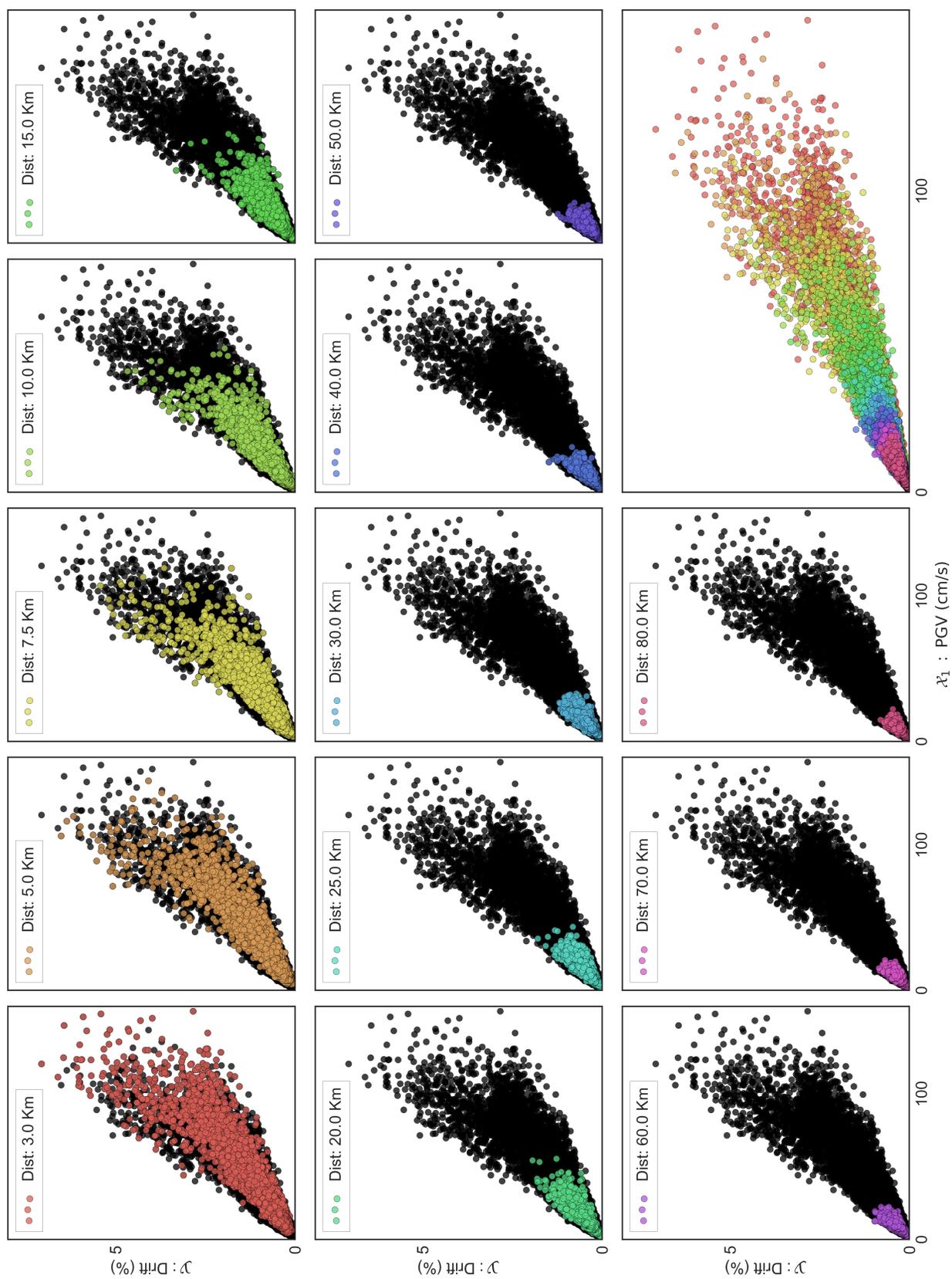
**Fig. C.1.** *Ground motion database in terms of earthquake magnitude*



**Fig. C.2.** *Ground motion database in terms of NEHRP soil classification*



**Fig. C.3.** *Ground motion database in terms of wave direction*



**Fig. C.4.** Ground motion database in terms of epicentral distance

## D. UPDATED PARTICLE APPROXIMATION

An important ingredient of this methodology is the estimation of the updated particle approximation  $\left\{ \left( w_a^{(s)}, \boldsymbol{\theta}_M^{(s)} \right) \right\}_{s=1}^S$  when the observation  $(\tilde{x}, \tilde{y})$  is added to the training dataset. Here, it is assumed that the value that  $\mathcal{Y}$  takes depend on the magnitude of independent coordinate  $X_1$ , and any prediction for  $\mathcal{Y} \mid X_1$  must be conditioned on the observed data  $\mathcal{D}_N$ . Although the particle approximation can be derived after performing an SMC analysis of the new training dataset, this is infeasible there is a large testing dataset and the numerical model is expensive to evaluate. Alternatively, it is possible to estimate this updated particle approximation as:

$$\begin{aligned}
 p(\boldsymbol{\theta}_M \mid \tilde{x}, \tilde{y}, \mathcal{D}_N) &= \frac{p(\tilde{x}, \tilde{y} \mid \boldsymbol{\theta}_M, \mathcal{D}_N) p(\boldsymbol{\theta}_M \mid \mathcal{D}_N)}{p(\tilde{x}, \tilde{y} \mid \mathcal{D}_N)} \\
 &= \frac{p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}_M, \mathcal{D}_N) p(\tilde{x} \mid \boldsymbol{\theta}_M, \mathcal{D}_N) p(\boldsymbol{\theta}_M \mid \mathcal{D}_N)}{p(\tilde{y} \mid \tilde{x}, \mathcal{D}_N) p(\tilde{x} \mid \mathcal{D}_N)} \\
 &= \frac{p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}_M, \mathcal{D}_N) p(\boldsymbol{\theta}_M \mid \mathcal{D}_N)}{\int p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}_M^*, \mathcal{D}_N) p(\boldsymbol{\theta}_M^* \mid \mathcal{D}_N) d\boldsymbol{\theta}_M^*} \\
 &\approx \frac{p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}_M, \mathcal{D}_N) p(\boldsymbol{\theta}_M \mid \mathcal{D}_N)}{\sum_{z=1}^S p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}_M^{(z)}, \mathcal{D}_N) w^{(z)}} \\
 &\approx \frac{p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}_M, \mathcal{D}_N) \left( \sum_{s=1}^S w^{(s)} \delta(\boldsymbol{\theta}_M - \boldsymbol{\theta}_M^{(s)}) \right)}{\sum_{z=1}^S p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}_M^{(z)}, \mathcal{D}_N) w^{(z)}} \\
 &= \sum_{s=1}^S \left( \frac{p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}_M, \mathcal{D}_N) w^{(s)}}{\sum_{z=1}^S p(\tilde{y} \mid \tilde{x}, \boldsymbol{\theta}_M^{(z)}, \mathcal{D}_N) w^{(z)}} \right) \delta(\boldsymbol{\theta}_M - \boldsymbol{\theta}_M^{(s)}) \\
 &= \sum_{s=1}^S w_a^{(s)}(\tilde{x}, \tilde{y}) \delta(\boldsymbol{\theta}_M - \boldsymbol{\theta}_M^{(s)})
 \end{aligned}$$

where  $w_a^{(s)}(\tilde{x}, \tilde{y})$  is the weighting factor associated to the particle  $\boldsymbol{\theta}_{\mathcal{M}}^{(s)}$  from the updated particle approximation  $\left\{ \left( w_a^{(s)}, \boldsymbol{\theta}_{\mathcal{M}}^{(s)} \right) \right\}_{s=1}^S$  and to be understood as a function of  $\tilde{x}$  and  $\tilde{y}$ :

$$w_a^{(s)}(\tilde{x}, \tilde{y}) = w^{(s)} \frac{p(\tilde{y} | \tilde{x}, \boldsymbol{\theta}_{\mathcal{M}})}{\sum_{z=1}^S w^{(z)} p(\tilde{y} | \tilde{x}, \boldsymbol{\theta}_{\mathcal{M}}^{(z)})}.$$

It is important to clarify that the expected value for the updated posterior distribution remains constant along the entire derivation almost surely as  $S \rightarrow \infty$  [114]. Equivalently, the implementation of a sufficiently large number of particles along the computation should not affect the expected value.

## E. MODEL SELECTION RESULTS

Table E.1. Model selection results for  $N = 500$ ,  $\mathcal{X}1 : PGV$ , and  $\mathcal{Y}$ .

Model	$q_m$	$q_s$	log(Z)			$\Delta(\mathcal{D}_N)$			$\mathcal{J}_2$ (KS dist.)			$\mathcal{J}_3$ (Q-Q error)			$\mathcal{J}_1$ (CI observ.)			
			$\mu$	$\sigma$	COV	$\mu$	$\sigma$	COV	$\mu$	$\sigma$	COV	$\mu$	$\sigma$	COV	$\mu$	$\sigma$	COV	
$\mathcal{M}_1$ (Normal)	<b>1</b>	<b>1</b>	<b>1826.8</b>	<b>6.2</b>	<b>0.34%</b>	<b>1.04</b>	<b>0.18</b>	<b>17.73%</b>	<b>6.66E-02</b>	<b>2.31E-03</b>	<b>3.48%</b>	<b>0.098</b>	<b>0.007</b>	<b>7.50%</b>	<b>0.09</b>	<b>0.13</b>	<b>144%</b>	
	2	1	1826.0	6.2	0.34%	1.10	0.14	13.05%	7.11E-02	4.89E-03	6.88%	0.146	0.060	41.21%	0.27	0.25	92.6%	
	3	1	1816.9	10.5	0.58%	1.15	0.16	13.97%	7.10E-02	6.99E-03	9.85%	0.185	0.121	65.43%	0.29	0.31	107%	
	4	1	1814.2	14.7	0.81%	1.16	0.22	18.88%	8.43E-02	3.65E-02	43.31%	0.391	0.643	164.59%	0.35	0.36	103%	
	1	2	1825.4	7.5	0.41%	1.01	0.12	11.89%	6.45E-02	5.31E-03	8.23%	0.098	0.012	12.07%	0.50	0.61	122%	
	2	2	1820.4	9.2	0.51%	1.16	0.27	23.69%	6.84E-02	1.73E-02	25.34%	0.161	0.211	131.12%	0.76	0.68	89.5%	
	3	2	1815.0	11.6	0.64%	1.17	0.15	13.08%	6.57E-02	7.95E-03	12.10%	0.127	0.057	44.95%	0.76	0.50	65.8%	
	4	2	1809.6	12.1	0.67%	1.23	0.32	26.45%	6.78E-02	8.30E-03	12.25%	0.159	0.125	78.57%	0.99	0.51	51.5%	
	1	3	1824.4	6.8	0.37%	1.09	0.15	13.56%	6.57E-02	3.42E-03	5.21%	0.098	0.010	10.47%	0.45	0.37	82.2%	
	2	3	1805.3	27.4	1.52%	1.44	0.36	24.86%	1.11E-01	9.49E-02	85.91%	0.974	2.078	213.30%	0.63	0.37	58.7%	
	3	3	1809.8	8.2	0.45%	1.28	0.18	14.10%	6.87E-02	5.10E-03	7.43%	0.125	0.043	34.78%	0.42	0.55	131%	
	4	3	1803.8	16.1	0.89%	1.32	0.21	15.80%	7.83E-02	1.87E-02	23.86%	0.250	0.304	121.68%	0.32	0.61	191%	
	$\mathcal{M}_2$ (Lognormal)	<b>1</b>	<b>1</b>	<b>1866.8</b>	<b>2.7</b>	<b>0.14%</b>	<b>0.82</b>	<b>0.12</b>	<b>14.06%</b>	<b>5.51E-02</b>	<b>2.16E-03</b>	<b>3.92%</b>	<b>0.180</b>	<b>0.020</b>	<b>11.25%</b>	<b>0.71</b>	<b>0.22</b>	<b>31.0%</b>
		2	1	1861.4	4.9	0.26%	0.89	0.14	15.85%	5.81E-02	1.01E-02	17.44%	0.206	0.071	34.35%	0.80	0.31	38.8%
		3	1	1845.0	12.5	0.68%	0.92	0.15	16.78%	6.79E-02	1.56E-02	23.03%	0.293	0.146	49.89%	0.60	0.47	78.3%
		4	1	1824.0	22.2	1.22%	0.92	0.24	26.27%	7.75E-02	2.24E-02	28.96%	0.391	0.227	58.07%	0.32	0.47	147%
1		2	1859.7	7.5	0.40%	0.90	0.13	14.52%	5.79E-02	5.66E-03	9.79%	0.198	0.028	14.08%	0.60	0.85	142%	
2		2	1848.7	8.5	0.46%	0.93	0.17	17.72%	5.44E-02	1.28E-02	23.63%	0.207	0.110	52.95%	0.04	0.69	1725%	
3		2	1835.2	22.0	1.20%	0.97	0.21	21.49%	6.09E-02	1.35E-02	22.12%	0.240	0.106	44.13%	0.03	1.35	4500%	
4		2	1810.1	31.3	1.73%	1.03	0.23	22.82%	7.49E-02	3.87E-02	51.69%	0.412	0.388	94.18%	0.54	1.77	385%	
1		3	1836.8	21.6	1.18%	0.83	0.11	12.84%	5.90E-02	6.12E-03	10.37%	0.212	0.036	16.89%	0.13	0.70	538%	
2		3	1801.2	24.5	1.36%	0.94	0.31	33.43%	6.01E-02	3.45E-02	57.44%	0.284	0.462	162.73%	0.34	0.77	226%	
3		3	1803.0	34.6	1.92%	1.02	0.23	22.09%	6.98E-02	5.06E-02	72.49%	0.390	0.741	189.94%	0.53	1.33	251%	
4		3	1764.3	49.9	2.83%	1.33	0.50	37.44%	8.33E-02	6.01E-02	72.13%	0.580	0.904	155.81%	0.25	1.06	424%	
$\mathcal{M}_3$ (Logstudent T)		<b>1</b>	<b>1</b>	<b>1847.0</b>	<b>9.9</b>	<b>0.54%</b>	<b>0.80</b>	<b>0.08</b>	<b>10.15%</b>	<b>6.64E-02</b>	<b>1.15E-02</b>	<b>17.26%</b>	<b>0.283</b>	<b>0.081</b>	<b>28.57%</b>	<b>1.44</b>	<b>1.31</b>	<b>91.0%</b>
		2	1	1824.8	15.7	0.86%	0.95	0.16	16.37%	9.23E-02	1.84E-02	19.98%	0.489	0.207	42.33%	2.75	1.42	51.6%
		3	1	1811.6	23.7	1.31%	0.99	0.28	27.89%	1.04E-01	3.68E-02	35.48%	0.589	0.373	63.32%	3.37	1.45	43.0%
		4	1	1803.7	27.4	1.52%	1.01	0.27	26.69%	1.25E-01	3.67E-02	29.40%	0.851	0.491	57.66%	3.23	1.41	43.7%
	1	2	1840.1	16.0	0.87%	0.80	0.10	12.97%	8.37E-02	1.40E-02	16.71%	0.370	0.086	23.33%	2.15	2.16	100%	
	2	2	1827.1	21.0	1.15%	0.86	0.15	17.28%	8.30E-02	2.56E-02	30.91%	0.361	0.146	40.57%	1.08	2.03	188%	
	3	2	1805.7	20.1	1.11%	0.92	0.19	20.24%	9.89E-02	3.52E-02	35.60%	0.511	0.327	63.97%	0.87	2.73	314%	
	4	2	1787.7	18.9	1.06%	1.08	0.31	28.43%	1.05E-01	2.84E-02	26.96%	0.544	0.241	44.24%	1.27	1.67	132%	
	1	3	1827.6	15.6	0.86%	0.85	0.10	11.65%	8.54E-02	1.69E-02	19.85%	0.391	0.104	26.64%	1.81	1.77	97.8%	
	2	3	1813.0	16.4	0.90%	0.94	0.16	16.76%	7.89E-02	1.88E-02	23.77%	0.335	0.129	38.63%	2.16	1.83	84.7%	
	3	3	1811.0	17.8	0.98%	1.13	0.22	19.42%	6.64E-02	2.45E-02	36.87%	0.286	0.213	74.47%	2.26	1.46	64.6%	
	4	3	1770.1	29.6	1.67%	1.48	0.67	45.08%	9.90E-02	4.96E-02	50.09%	0.604	0.659	109.19%	3.13	1.35	43.1%	
	$\mathcal{M}_4$ (Gamma)	<b>1</b>	<b>1</b>	<b>1834.0</b>	<b>32.7</b>	<b>1.79%</b>	<b>0.86</b>	<b>0.10</b>	<b>11.89%</b>	<b>5.17E-02</b>	<b>2.92E-03</b>	<b>5.64%</b>	<b>0.101</b>	<b>0.023</b>	<b>23.02%</b>	<b>0.86</b>	<b>0.42</b>	<b>48.8%</b>
		2	1	1737.4	129.4	7.45%	1.02	0.25	24.46%	1.33E-01	1.36E-01	102.87%	1.222	1.971	161.32%	2.45	1.84	75.1%
		3	1	1711.9	102.2	5.97%	1.24	0.76	60.97%	9.76E-02	9.31E-02	95.30%	0.681	1.226	179.93%	2.04	1.75	85.8%
		4	1	1662.7	127.2	7.65%	1.67	1.11	66.35%	9.15E-02	6.85E-02	74.95%	0.538	0.789	146.84%	2.02	1.96	97.0%
1		2	1756.4	173.9	9.90%	0.85	0.11	12.43%	6.21E-02	1.41E-02	22.75%	0.157	0.058	36.96%	1.60	1.47	91.9%	
2		2	1711.7	202.2	11.81%	0.75	0.21	28.20%	8.19E-02	4.97E-02	60.69%	0.403	0.601	149.17%	1.90	2.93	154%	
3		2	1535.2	296.1	19.29%	1.89	2.62	138.92%	8.20E-02	9.45E-02	115.14%	0.586	1.655	282.48%	2.16	1.60	74.1%	
4		2	1251.7	131.8	10.53%	2.70	1.27	47.11%	1.01E-01	4.74E-02	46.72%	0.515	0.467	90.59%	3.65	0.72	19.7%	
1		3	1455.4	489.8	33.66%	2.22	1.91	86.32%	1.96E-01	1.79E-01	91.04%	2.352	2.804	119.23%	2.23	2.35	105%	
2		3	1060.8	414.2	39.04%	4.49	2.43	54.07%	3.44E-01	1.69E-01	49.31%	4.715	2.786	59.10%	3.77	2.89	76.7%	
3		3	998.7	128.7	12.88%	8.41	3.94	46.87%	3.61E-01	1.00E-01	27.69%	5.214	2.162	41.47%	4.75	0.38	8.0%	
4		3	948.7	123.5	13.01%	9.55	3.70	38.74%	3.68E-01	8.43E-02	22.92%	5.231	1.896	36.25%	4.85	0.31	6.4%	
$\mathcal{M}_5$ (Beta)		1	1	1831.1	38.3	2.09%	0.91	0.15	16.93%	5.08E-02	2.50E-03	4.93%	0.099	0.029	28.96%	0.91	0.64	70.3%
		2	1	1820.7	50.1	2.75%	0.86	0.12	13.45%	6.53E-02	6.99E-02	107.01%	0.311	0.913	293.31%	1.24	1.15	92.7%
		3	1	1809.8	43.7	2.41%	0.97	0.23	23.22%	5.41E-02	1.93E-02	35.66%	0.146	0.177	121.07%	1.08	0.91	84.3%
		4	1	1796.4	65.2	3.63%	0.98	0.34	34.67%	6.88E-02	6.73E-02	97.78%	0.342	0.947	276.47%	1.73	1.14	65.9%
	1	2	1794.5	77.0	4.29%	0.82	0.11	13.18%	5.79E-02	1.11E-02	19.20%	0.136	0.038	28.04%	0.95	1.73	182%	
	2	2	1841.9	24.8	1.34%	0.77	0.16	21.45%	5.58E-02	8.97E-03	16.06%	0.130	0.050	38.46%	1.74	1.37	78.7%	
	3	2	1840.8	19.1	1.04%	0.93	0.20	21.61%	6.33E-02	2.72E-02	42.91%	0.222	0.225	101.26%	1.23	1.53	124%	
	4	2	1837.0	18.3	0.99%	0.94	0.29	30.95%	7.08E-02	1.83E-02	25.80%	0.241	0.145	60.26%	1.78	1.25	70.2%	
	<b>1</b>	<b>3</b>	<b>1852.6</b>	<b>13.0</b>	<b>0.70%</b>	<b>0.83</b>	<b>0.14</b>	<b>16.67%</b>	<b>5.14E-02</b>	<b>5.45E-03</b>	<b>10.60%</b>	<b>0.119</b>	<b>0.021</b>	<b>17.44%</b>	<b>0.04</b>	<b>0.89</b>	<b>2225%</b>	
	2	3	1845.9	20.8	1.13%	0.86	0.12	14.57%	5.45E-02	1.68E-02	30.91%	0.138	0.119	86.11%	1.38	1.15	83.3%	
	3	3	1837.6	19.8	1.08%	0.93	0.17	17.95%	5.47E-02	1.10E-02	20.12%	0.149	0.096	64.09%	0.93	0.95	102%	
	4	3	1826.3	26.3	1.44%	1.10	0.26	23.45%	6.53E-02	1.94E-02	29.77%	0.216	0.170	78.48%	0.80	1.39	174%	

\*The values presented in the table corresponds to the results of 20 iterations

## F. SEQUENTIAL SELECTION OF GROUND MOTIONS RESULTS

Table F.1. Model selection results for  $N = 1000$ ,  $\mathcal{X}_1 : PGV$ , and  $\mathcal{X}_2 : S_V$ .

Model	$q_\mu$	$q_\sigma$	log(Z)			$\mathcal{J}_2$ (KS dist.)			$\mathcal{J}_3$ (Q-Q error)			$\mathcal{J}_1$ (CI observ.)		
			$\mu$	$\sigma$	COV	$\mu$	$\sigma$	COV	$\mu$	$\sigma$	COV	$\mu$	$\sigma$	COV
$\mathcal{M}_1$ (Normal)	1	1	<b>4194.6</b>	<b>6.0</b>	<b>0.14%</b>	<b>9.11E-02</b>	<b>1.33E-03</b>	<b>1.46%</b>	<b>0.309</b>	<b>0.008</b>	<b>2.46%</b>	<b>3.13</b>	<b>0.08</b>	<b>2.7%</b>
	2	1	4190.6	7.3	0.17%	8.40E-02	1.12E-03	1.33%	0.269	0.028	10.34%	3.00	0.07	2.4%
	3	1	4181.2	13.2	0.31%	7.55E-02	6.25E-03	8.27%	0.205	0.059	28.60%	2.38	0.55	23.3%
	4	1	4173.3	19.1	0.46%	7.61E-02	2.93E-03	3.85%	0.200	0.063	31.54%	2.36	0.31	13.3%
	1	2	4187.5	7.3	0.17%	8.71E-02	2.80E-03	3.21%	0.279	0.022	7.78%	3.46	0.29	8.4%
	2	2	4188.4	11.0	0.26%	8.13E-02	2.94E-03	3.62%	0.262	0.022	8.48%	3.24	0.19	5.9%
	3	2	4188.6	6.4	0.15%	7.97E-02	9.74E-03	12.22%	0.255	0.067	26.30%	2.55	0.49	19.3%
	4	2	4168.2	17.7	0.42%	8.00E-02	9.08E-03	11.35%	0.240	0.062	25.67%	2.66	0.69	26.0%
	1	3	4192.9	8.9	0.21%	9.12E-02	1.14E-03	1.25%	0.276	0.013	4.74%	3.14	0.09	3.0%
	2	3	4183.1	10.1	0.24%	8.71E-02	1.86E-03	2.14%	0.271	0.048	17.78%	2.67	0.26	9.7%
	3	3	4185.1	9.5	0.23%	8.86E-02	6.45E-03	7.28%	0.301	0.044	14.74%	2.56	0.56	21.9%
	4	3	4166.5	22.6	0.54%	8.03E-02	5.85E-03	7.29%	0.207	0.074	35.74%	1.78	0.57	32.0%
$\mathcal{M}_2$ (Lognormal)	1	1	<b>4306.3</b>	<b>2.2</b>	<b>0.05%</b>	<b>5.45E-02</b>	<b>1.27E-03</b>	<b>2.33%</b>	<b>0.100</b>	<b>0.001</b>	<b>0.83%</b>	<b>2.14</b>	<b>0.14</b>	<b>6.4%</b>
	2	1	4291.4	12.7	0.30%	5.00E-02	1.34E-03	3.59%	0.089	0.012	13.21%	1.86	0.19	10.3%
	3	1	4287.9	12.5	0.29%	4.61E-02	7.23E-03	15.69%	0.073	0.020	27.95%	1.49	0.50	33.5%
	4	1	4255.8	37.8	0.89%	4.57E-02	7.72E-03	16.91%	0.080	0.039	48.27%	0.96	0.46	48.0%
	1	2	4298.8	5.5	0.13%	5.60E-02	3.65E-03	6.53%	0.100	0.005	5.38%	2.17	0.63	29.3%
	2	2	4269.1	22.7	0.53%	4.89E-02	3.41E-03	6.98%	0.091	0.013	13.96%	2.05	0.52	25.4%
	3	2	4260.5	14.0	0.33%	4.96E-02	9.70E-03	19.57%	0.091	0.025	27.51%	2.00	0.90	45.2%
	4	2	4242.0	38.4	0.91%	4.80E-02	6.98E-03	14.55%	0.086	0.026	30.14%	1.09	1.00	91.4%
	1	3	4268.0	29.8	0.70%	5.81E-02	3.09E-03	5.32%	0.105	0.006	5.85%	2.55	0.39	15.3%
	2	3	4233.2	56.6	1.34%	5.37E-02	1.92E-02	35.68%	0.119	0.092	77.44%	0.96	1.22	128%
	3	3	4233.9	39.5	0.93%	4.96E-02	2.10E-02	42.35%	0.109	0.127	116.26%	0.03	1.73	5182%
	4	3	4110.4	158.5	3.86%	1.01E-01	7.95E-02	78.81%	0.797	1.365	171.35%	0.75	3.77	507%
$\mathcal{M}_3$ (Logstudent T)	1	1	4254.6	45.0	1.06%	6.65E-02	1.02E-02	15.32%	0.156	0.070	45.11%	0.51	2.64	514%
	2	1	4248.4	32.9	0.77%	6.74E-02	1.34E-02	19.89%	0.138	0.053	38.11%	0.19	2.52	1343%
	3	1	4217.5	52.0	1.23%	6.26E-02	1.49E-02	23.87%	0.142	0.081	57.40%	1.23	2.00	162%
	4	1	<b>4256.4</b>	<b>21.5</b>	<b>0.50%</b>	<b>5.19E-02</b>	<b>1.06E-02</b>	<b>20.50%</b>	<b>0.104</b>	<b>0.041</b>	<b>39.85%</b>	<b>0.13</b>	<b>1.60</b>	<b>1241%</b>
	1	2	4261.8	18.7	0.44%	7.49E-02	6.08E-03	8.12%	0.165	0.044	26.67%	1.39	1.99	143%
	2	2	4251.4	20.1	0.47%	6.32E-02	1.16E-02	18.37%	0.129	0.039	30.13%	0.43	1.74	408%
	3	2	4200.5	41.7	0.99%	6.11E-02	1.83E-02	29.98%	0.150	0.109	72.57%	0.88	2.50	285%
	4	2	4268.8	237.3	5.56%	2.02E-01	2.34E-01	115.90%	3.797	7.551	198.87%	1.47	2.94	200%
	1	3	4260.7	16.3	0.38%	6.74E-02	3.97E-03	5.89%	0.132	0.013	9.55%	2.69	0.90	53.1%
	2	3	4217.5	35.6	0.84%	6.04E-02	1.66E-02	27.08%	0.133	0.086	64.77%	0.14	2.01	1455%
	3	3	4217.9	32.9	0.78%	5.29E-02	1.24E-02	23.41%	0.087	0.037	43.18%	1.33	1.82	137%
	4	3	4165.9	32.4	0.78%	7.04E-02	2.21E-02	31.45%	0.225	0.178	79.41%	2.56	1.59	61.9%
$\mathcal{M}_4$ (Gamma)	1	1	4099.5	183.7	4.48%	2.21E-01	1.52E-01	68.78%	2.499	2.362	94.55%	2.44	3.36	234%
	2	1	3845.1	154.0	4.01%	3.65E-01	2.60E-02	7.13%	4.688	0.736	15.70%	4.80	0.01	0.1%
	3	1	3856.8	82.5	2.14%	3.78E-01	3.47E-02	9.19%	5.415	0.851	15.72%	4.80	0.01	0.2%
	4	1	3825.2	244.2	6.38%	2.71E-01	1.30E-01	47.92%	3.408	2.250	66.02%	3.26	2.47	75.8%
	1	2	3992.1	324.1	8.12%	1.68E-01	1.07E-01	63.99%	1.445	1.330	92.02%	0.89	3.77	423%
	2	2	<b>4113.8</b>	<b>372.7</b>	<b>9.06%</b>	<b>8.75E-02</b>	<b>3.54E-02</b>	<b>40.42%</b>	<b>0.350</b>	<b>0.274</b>	<b>78.47%</b>	<b>0.09</b>	<b>2.25</b>	<b>2547%</b>
	3	2	3664.6	383.9	10.48%	9.36E-02	1.81E-02	19.36%	0.320	0.105	32.65%	0.63	0.69	108%
	4	2	3550.1	479.6	13.51%	1.01E-01	3.35E-02	33.08%	0.414	0.245	59.13%	1.01	1.11	110%
	1	3	2862.3	934.2	32.64%	3.64E-01	2.04E-01	56.05%	5.250	3.510	66.86%	2.60	3.67	142%
	2	3	2198.5	24.2	1.10%	4.79E-01	6.10E-02	12.74%	7.108	1.296	18.23%	5.00	0.00	0.00%
	3	3	3048.8	821.6	26.95%	2.88E-01	1.63E-01	56.66%	3.688	3.092	83.83%	3.31	2.11	63.8%
	4	3	2556.6	205.1	8.02%	3.50E-01	1.00E-01	28.67%	4.721	4.721	42.67%	4.46	0.63	14.2%
$\mathcal{M}_5$ (Beta)	1	1	4243.9	99.2	2.34%	1.02E-01	9.37E-02	92.00%	0.645	1.412	218.96%	1.18	2.12	180%
	2	1	4103.6	180.3	4.39%	2.27E-01	1.60E-01	70.59%	2.763	2.633	95.30%	1.37	3.43	251%
	3	1	4114.1	153.7	3.74%	1.57E-01	1.55E-01	98.43%	1.799	2.579	143.36%	0.59	2.79	473
	4	4086.2	135.7	3.32%	2.08E-01	1.58E-01	75.93%	2.576	2.720	105.57%	1.63	3.07	189%	
	1	2	<b>4282.7</b>	<b>8.0</b>	<b>0.19%</b>	<b>6.21E-02</b>	<b>2.75E-03</b>	<b>4.42%</b>	<b>0.141</b>	<b>0.008</b>	<b>5.92%</b>	<b>2.49</b>	<b>0.53</b>	<b>21.1%</b>
	2	2	4246.9	45.8	1.08%	8.14E-02	2.49E-02	30.62%	0.268	0.164	61.40%	0.54	1.90	351%
	3	2	4269.0	20.5	0.48%	6.00E-02	1.42E-02	23.69%	0.141	0.069	49.06%	0.98	0.94	95.5%
	4	2	4241.6	30.3	0.71%	7.47E-02	1.92E-02	25.72%	0.215	0.104	48.30%	0.35	1.24	355%
	1	3	4270.2	15.5	0.36%	6.40E-02	5.90E-03	9.23%	0.152	0.031	20.64%	2.36	1.01	42.7%
	2	3	4162.6	284.3	6.83%	8.20E-02	2.53E-02	30.81%	0.272	0.181	66.58%	0.79	1.51	191%
	3	3	4201.9	67.7	1.61%	1.03E-01	4.22E-02	40.80%	0.448	0.324	72.34%	0.46	1.27	279%
	4	3	4236.6	36.8	0.87%	7.42E-02	2.68E-02	36.13%	0.226	0.169	74.46%	0.37	0.94	257%

\*The values presented in the table corresponds to the results of 10 iterations

## VITA

Francisco Peña was born in Cali, Colombia on August 25th, 1989. He graduated *summa cum laude* from Universidad del Valle, Colombia with a B.S. in Civil Engineering in April 2011.

In 2012, the Colombian government, through COLCIENCIAS (Departamento Administrativo de Ciencia, Tecnología e Innovación), awarded Francisco a scholarship to enroll in a graduate program abroad. He started his graduate studies in August 2013 at Purdue University where he earned his M.S. in Civil Engineering in August 2015 and completed his Ph.D. degree in May 2019. While attending Purdue University, Francisco received the *SE Solutions Structures Scholarship* from the Lyles School of Civil Engineering, and the *Estus H. and Vashti L. Magoon Award for Excellence in Teaching* from the College of Engineering in recognition of his outstanding collaboration in the Civil Engineering Design course (CE 49800).

In January 2019, Francisco joined Wiss, Janney, Elstner Associates in Indianapolis, Indiana.