A physics-informed Bayesian framework for characterizing ground motion process in the presence of missing data

Yu Chen[a], Edoardo Patelli*[b], Benjamin Edwards[a], Michael Beer[a,c,d]

[a] Institute for Risk and Uncertainty, University of Liverpool, Liverpool, UK
[b] Department of Civil and Environmental Engineering, University of Strathclyde, Glasgow, UK
[c] Institute for Risk and Reliability, Leibniz Universit¨at Hannover, Hannover, Germany
[d] International Joint Research Center for Resilient Infrastructure & International Joint Research Center for Engineering Reliability and Stochastic Mechanics, Tongji University, Shanghai, China

Abstract

A Bayesian framework to stochastically characterize ground motions even in the presence of missing data is developed. This approach features the combination of seismological knowledge (*a priori knowledge*) with empirical observations (even incomplete) via Bayesian inference. At its core is a Bayesian neural network model that probabilistically learns temporal patterns from ground motion data. Uncertainties are accounted for throughout the framework. Performance of the approach has been quantitatively demonstrated via various missing data scenarios. This framework provides a general solution to dealing with missing data in ground motion records by providing various forms of representation of ground motions in a probabilistic manner, allowing it to be adopted for numerous engineering and seismological applications. Notably, it is compatible with the versatile Monte Carlo simulation scheme, such that stochastic dynamic analyses are still achievable even with missing data. Furthermore, it serves as a complementary approach to current stochastic ground-motion models in data-scarce regions under the growing interests of PBEE (performance-based earthquake engineering), mitigating the data-model dependence dilemma due to the paucity of data, and ultimately, as a fundamental solution to the limited data problem in data scarce regions.

Keywords: Missing data, Stochastic variational inference, Bayesian model updating, Evolutionary power spectra, Uncertainty quantification, Earthquake ground motion

1. Introduction

The random nature of earthquake ground motions is well appreciated. Various research efforts and progress, based on stochastic process formulation, have been made towards the problem of characterization, simulation and response evaluation (Narayana Iyengar and Sundara Raja Iyengar, 1969; Shinozuka and Deodatis, 1988; Kiureghian and Fujimura, 2009). In recent years, the growing interest in performance-based earthquake engineering (PBEE), which requires ground motions of various hazard levels to consider the entire range of structural response, including nonlinear behaviour and even collapse (Kiureghian and Fujimura, 2009), has driven the need for simulating ground motions of various earthquake scenarios. Stochastic simulations are further utilised for evaluation of future seismic demand and seismic reliability assessment (Comerford et al., 2017), nonlinear stochastic dynamic analyses (Vlachos et al., 2018b), developing ground motion prediction equations (GMPEs) (Atkinson and Boore, 2006), or seismic hazard characterization and simulation-based seismic risk assessment (Vetter and Taflanidis, 2014; Tsioulou et al., 2018).

However, their applicability is not without questioning. Empirical ground motions are responsible for developing and calibrating stochastic ground motion models. However, the paucity of recordings (especially strong motions) in data scarce regions leads to a bottleneck that observational data are lacking in the first place to justify modelling and calibration. For instance, in characterizing seismic hazard, a category of predictive-relation based stochastic ground-motion models (see e.g. Rezaeian and Der Kiureghian (2010); Laurendeau et al. (2012); Vlachos et al. (2018a)) is gaining increasing attention for its ability to generate a suite of nonstationary time-histories, given specific earthquake scenarios. The core component of these models is an underlying empirical regression between model parameters and earthquake characteristics over a selected (sometimes limited) subset of records. However these empirical relations are largely bounded by the scope of data being regressed. Significant epistemic uncertainties are expected on further uses of these underlying empirical regressions as *extrapolation* than *interpolation*. Similarly, such uncertainty also applies to those empirical GMPEs developed using stochastic simulations calibrated from small to moderate earthquakes often due to a lack of strong motions (Atkinson and Boore, 2006; Edwards and Fäh, 2013). Concerns have been raised over the subsequent stochastic simulations from these biased models, as the underlying regression are typically not well-constrained by empirical data and their extrapolation may therefore not even be physically realistic (Baker et al., 2021).

Therefore, for data-scarce regions, where there are stronger needs of synthetic ground motions for abundant earthquake scenarios, however, the paucity of data poses a causality dilemma concerning the dependence between observations and the extracted knowledge/information for the development of models. This raises difficulties, in data scarce regions, in the characterization of ground motions for the seismic risk assessment as well as researches of regional seismicity and Earth regional structures.

As such, a method to make the most of existing data (even where incomplete), robustly characterizing the underlying physical processes from bad measurements (e.g. incomplete), could enrich the observational database, whereby one is able to progressively update the development and calibration of ground motion models, producing more realistic stochastic simulations in the otherwise data scarce regions, for hazard characterization and risk assessment. It serves as a complementary approach to stochastic ground-motion models under the growing interests of PBEE, and ultimately a fundamental solution to the limited data problem. This may be of particular interest to studies of historical earthquakes which may potentially provide strong-motion records but many of them are discarded due to the presence of data gaps (Maranò et al., 2017). Furthermore, missing data exist in both historical and modern earthquake time histories due to intermittent instrumentation or data-transmission failure. For instance, old mechanical, short-period high-sensitivity or broadband seismometers are vulnerable to clipping during local strong motions. In addition, sensor malfunctions, instrument tilt, or data contamination, may lead to missing or incorrect values, or waveform clipping around the peak motion (Smith-Boughner and Constable, 2012; Maranò et al., 2017; Zhang et al., 2016). With the recent use of low-cost temporary instruments, deployed at scale, sometimes in harsh conditions, the fidelity and continuity of recording is also not as reliable as traditional permanent seismological stations, which itself can be understood as a bad- or missing-data problem.

The characterization of ground motions and accounting for their random nature is challenging when only limited and partial recordings are available (Zhang et al., 2016; Comerford et al., 2016; Zhang et al., 2017). Pioneering works for analysis in the presence of missing data, such as the Lomb-Scargle periodogram (Scargle, 1982), iterative deconvolution CLEAN (Roberts et al., 1987), are acknowledgedly to have deficiencies such as bias issue and periodic content limitation (Bos et al., 2002; Wang et al., 2005; Babu and Stoica, 2010; Smith-Boughner and Constable, 2012). With different assumptions (hence limitations), many other methods have been proposed in recent years. Notably, a compressive sensing approach is exploited with the sparsity assumption of the underlying spectral representation (Comerford et al., 2016). By assuming the same frequency contents between the missing portion and the observations, a projection onto convex sets (POCS) method can be used to reconstruct clipped waveforms (Zhang et al., 2016). Parametric models are also developed based on various formulations, such as autoregressive modeling methods (Bos et al., 2002; Broersen et al., 2004; Hung, 2008), with parameterized assumptions on the structure of the underlying stochastic processes. Similarly, Maranò et al. (2017) proposed a method to fit a parametric seismological model to earthquake recordings with missing gaps.

Alternatively, a variety of methods are available that explicitly or implicitly transform spectral analysis with missing data into the imputation of missing values, followed by standard full-data spectral analysis (Stoica et al., 2000; Kondrashov and Ghil, 2006; Kondrashov et al., 2014; Comerford et al., 2015a; Musial et al., 2011). This strain of methods provides reconstructed waveforms in a straightforward manner, whereby extensive established spectral analyses, developed on equidistant data, whether stationary or nonstationary, can still be universally harnessed.

Two main challenges are identified in dealing with missing data. First, most current approaches fail to address the uncertainties related to the missing data properly (Comerford et al., 2015b; Zhang et al., 2017). For reconstruction based methods, inaccuracies of the imperfect reconstruction will be propagated to spectral estimates owing to the convolutional nature of Fourier transform. Similarly, for parametric modelling methods that results in a parametric form of spectrum, parameter uncertainties due to the incomplete data are not well captured. More importantly, despite existing approaches that handle uncertainties (notably Bayesian spectral analyses (Tobar, 2018; Christmas, 2013)), they are still constrained by the significantly limited information from the very incomplete signal.

Therefore, to exploit additional information besides the incomplete recording and to appropriately quantify the uncertainties brought by the missing data, we propose a novel Bayesian framework that aims to robustly combine prior seismological knowledge with empirical observations (even incomplete). A Bayesian neural network (BNN) model that probabilistically learns the temporal dynamics from earthquake time histories forms the key component of the framework. In particular, it is initially trained from physics-informed simulated ground motions given the event metadata (e.g. magnitude, epicentral distance, V_{s30}, etc.), as geological *a-priori*, and subsequently updated via Bayesian inference utilising the partial empirical observations. Importantly, uncertainty has been accounted for throughout the framework. Variability of the physics-informed simulations are considered. Epistemic uncertainties on model parameters of the BNN are learnt through stochastic variational inference, whereby an ensemble of reconstructed time histories are obtained by marginalizing over the posterior distribution of model parameters. Furthermore, uncertainties of the spectral representations (e.g. evolutionary power spectral density) of the underlying stochastic process are quantified, with the spectral density values represented by probability distributions. As a result, sample realizations associated with the stochastic process can be further simulated for stochastic dynamic analysis through the spectral representation method, even with incomplete recordings.

Details of the framework are discussed first, then the performance of the proposed method is demonstrated with various missing data scenarios based on an earthquake strong motion recording.

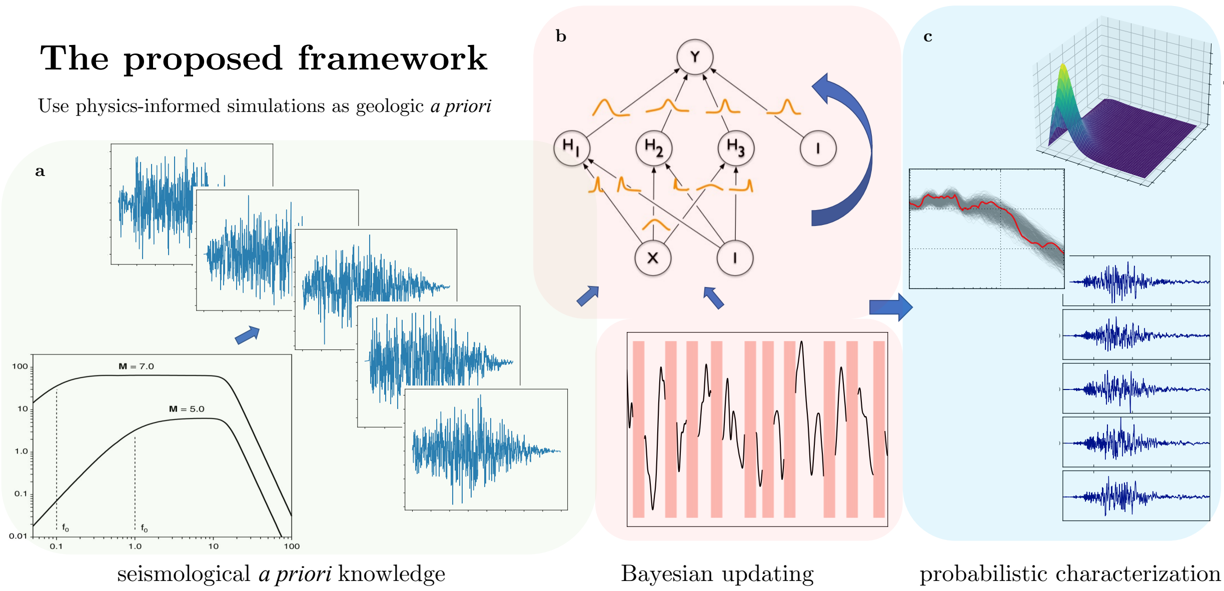


Figure 1: A stochastic framework characterizing ground motion process in the presence of missing data. Three components are presented: **a**. a seismological model generating physics-informed stochastic simulations with *a-priori* seismological knowledge; **b**. a Bayesian neural network model initially trained from physics-informed stochastic simulations and later updated by empirical partial observations; **c**. a host of model-based probabilistic representations of ground motions (e.g. evolutionary power spectral density EPSD, elastic response spectra, ensemble reconstructed time histories etc.)

2. A Bayesian framework for characterization of ground motion with missing data

We build on the premise that *a priori* seismological knowledge can provide a general, yet insightful, prior expectation of the ground motions of the certain earthquake scenario, which can be combined with the information extracted from empirical observations (even when incomplete).

2.1. Physics-informed stochastic simulations as a geological prior

A stochastic representation that encapsulates the physics of the earthquake process and wave propagation plays the central role, from the seismological perspective, in characterizing the ground motions (see e.g. Zeng et al. (1994); Boore (2003)). One of the most desired advantage is that such representations, is to explicitly distill the knowledge of various factors affecting ground motions (e.g. source, path, and site effects) into a parametric formulation. In this study, we have adopted a well-validated stochastic seismological model (Boore, 2003), as given below, whereby source process, attenuation, and site effects are encapsulated in a parameterized form of the Fourier amplitude spectrum. A finite fault strategy is particularly employed to represent the geometry of larger ruptures for large earthquakes (Atkinson and Boore, 2006; Edwards et al., 2019).

$$A(f;\mathbf{\Theta}) = \frac{CM_0}{1+(f/f_0)^2} Z(R)\exp[-\pi f R/Q(f)\beta]G(f) \tag{1}$$

where $\Theta = (\Theta_e, \Theta_g)$ represents the event parameters (Θ_e) that are still accessible from the metadata of an incomplete recording, such as seismic moment M_0 and hypocentral distance R, and region-specific seismological parameters (Θ_g) that embody the source, path and site effects. Specifically, f_0 is the earthquake's source corner frequency given by $f_0 = 0.4906\beta(\Delta\sigma/M_0)^{1/3}$ (in SI units); $R = \sqrt{r^2+d^2}$ where r and and d are the epicentral distance and depth to a given sub-fault; $\Delta\sigma$ is referred to as the stress drop, and β represents the shear wave velocity in the vicinity of the source. The constant C is given by: $C = R_{\theta\Phi}VF/(4\pi\rho_s\beta^3 R_0)$, where $R_{\theta\Phi}$ is the radiation pattern; V represents the partition of total shear-wave energy into horizontal components; F accounts for the free-surface effect; R_0 is the a reference distance and ρ is the density in the vicinity of the source. $Z(R)$ is the geometrical spreading function defined by a piece-wise series of segments in the form of R^{b_n}, where b_n defines the geometrical-spreading coefficient in the nth segment. The quality factor $Q(f)$ is an inverse measure of anelastic attenuation. The site effect $G(f) = \exp(-\pi f\kappa_0)10^{\upsilon}$ is given by the counteraction of a high-cut filter, $\exp(-\pi f\kappa_0)$, accounting for the diminution of the high-frequency motions and an amplification factor υ in log units. The specific values for each of the model terms used in this model can be taken from the existing literature, or directly through spectral modelling of waveform data (e.g. Edwards and Fäh (2013)).

In particular, the variability of model parameters in the spectral formulation, and hence the uncertainty in stochastic simulations, are represented by probability distribution over the input parameters Θ_g as proposed by Atkinson and Boore (2006); Vetter and Taflanidis (2012). Note that the above stochastic simulation procedures are distinct from those comprehensive deterministic numerical models that solve the complex 3D equations governing seismic wave propagation. Those models are typically referred to as physics-based numerical models in the literature, see e.g. McCallen et al. (2021a,b); Paolucci et al. (2021) among others.

2.2. Sequential modeling

In recent years, neural network models have become established in learning complex and nonlinear relations. Most recently, successes have been seen for neural networks to learn the temporal dynamics in sequential data (e.g. time series) under an autoregressive setting (Salinas et al., 2020; Beer and Spanos, 2009; Comerford et al., 2015a; Gatti and Clouteau, 2020). They model the data generating process by formulating the conditional distribution, $p(y_t|\mathbf{x}_t,\mathbf{w})$, of the value y_t based on a window of past lagged values $([y_{t-1},\ldots,y_{t-p}])$, as given by:

$$y_t = f(\mathbf{x}_t;\mathbf{w}) + \epsilon, \text{with } \mathbf{x}_t = [y_{t-1},\ldots,y_{t-p}] \tag{2}$$

where ϵ denotes the noise term; $f(\cdot)$ represents the neural network model, parameterized by $\mathbf{w}$, which learns complex nonlinear temporal dependence in the time series, as opposed to a linear combination of fixed coefficients in a classic autoregressive AR(p) model. y_t and $\mathbf{x}_t$ represent the prediction and the lagged window pair. In practice, training with maximum likelihood estimation (MLE) gives rise to a probabilistic interpretation of the data generating process. The likelihood function, assuming Gaussian noise with variance σ^2, is given by (Williams and Rasmussen, 2006):

$$p(y_t|\mathbf{x}_t,\mathbf{w}) = \mathcal{N}(y_t|f(\mathbf{x}_t,\mathbf{w}),\sigma^2) \tag{3}$$

Model parameters $\mathbf{w}$, collectively the weights and biases of the neural network model (referred as weights hereafter), are estimated during training by optimizing with the likelihood as the objective as follows:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \sum_t \log p(y_t|\mathbf{x}_t, \mathbf{w}) \tag{4}$$

Once trained, its generative power could be employed to generate sequences (Graves, 2013), forecast time series future values (Salinas et al., 2020), and impute missing values (Comerford et al., 2015a). However, despite accounting for the aleatoric uncertainty using Gaussian noise, the above MLE strategy ignores the uncertainties of the model parameters (i.e. epistemic uncertainties) that can explain the observed data (especially in the context of limited data and missing data) as well as the resulting predictive uncertainties regarding the imputation. Significant uncertainties exist on the model configurations that may have explained the limited data. Consequently, such uncertainties further compromise the generalization power of learned models in that predictions from uncertain/unrepresentative models can still be unreliable and over confident (Blundell et al., 2015; Gal and Ghahramani, 2016).

2.3. Bayesian updating on partial observations

In order to capture the model uncertainty, probability distributions are applied to the neural net model parameters (see Fig. 1). Bayesian inference hence formulates the update of the neural network modelling the underlying generating process, when new observations (even incomplete) become available, as given below:

$$p(\mathbf{w}|\mathcal{D}) = p(\mathcal{D}|\mathbf{w})p(\mathbf{w})/p(\mathcal{D}) \tag{5}$$

where $p(\mathbf{w})$ represents the prior probability distribution of weights learnt from the physics-informed simulations; $p(\mathcal{D}|\mathbf{w})$ stands for the likelihood and $\mathcal{D}$ specifically refers to the partial and incomplete observations. $p(\mathbf{w}|\mathcal{D})$ is the posterior distribution, in which both the prior seismological knowledge and the real-world empirical observations are collectively considered. The posterior predictive distribution for the prediction of the missing value y_t^*, based on the lagged window, can be made for each possible configuration of the weights, by marginalizing over the posterior distribution, as shown below:

$$\begin{aligned} p(y_t^*|\mathbf{x}_t, \mathcal{D}) &= \int p(\mathbf{w}|\mathcal{D}) p(y_t^*|\mathbf{x}_t, \mathbf{w}) d\mathbf{w} \\ &= \mathbb{E}_{p(\mathbf{w}|\mathcal{D})}\left[p(y_t^*|\mathbf{x}_t, \mathbf{w})\right] \end{aligned} \tag{6}$$

As a result of considering uncertainties within the neural network, an ensemble of reconstructed time-histories, based on Monte Carlo sampling of the posterior distributions of weights, can be obtained. Subsequently, an ensemble of spectral estimates (e.g. evolutionary power spectral density EPSD, response spectra, etc.) can be computed from the ensemble reconstructions using established spectral analysis methods. Performing such analyses for many incomplete recordings in the otherwise data scarce region produces an enriched database, which could be further adopted to update the development or calibration of ground motion models (including both stochastic ground-motion models and empirical GMPEs). This scheme is interpreted as an escape from the model-data dependence dilemma, as highlighted earlier, by making the most of the observed data (even when incomplete).

2.4. Stochastic variational inference

A key challenge in Eq. (5) is the approximation of the posterior distribution. Analytic Bayesian inference to the true posterior $p(\mathbf{w}|\mathcal{D})$ is intractable and Markov Chain Monte Carlo (MCMC) based sampling approaches generally have difficulties in scaling to the huge dimensions of neural networks (Hernández-Lobato and Adams, 2015; Gal and Ghahramani, 2016). Alternatively, stochastic variational inference (see e.g. Graves (2011); Kingma and Welling (2013); Blei et al. (2017)) approximates the posterior distribution $p(\mathbf{w}|\mathcal{D})$ efficiently, by turning such inference problem into an optimization problem. It optimizes the parameters of a proposed variational distribution, such that the Kullback-Leibler (KL) divergence between the approximate distribution and the true posterior distribution is minimised: $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}[q(\mathbf{w}|\boldsymbol{\theta}) \parallel p(\mathbf{w}|\mathcal{D})]$. This minimization objective is indeed equivalent to the following cost function (Graves, 2011):

$$\mathcal{J}(\mathcal{D}, \boldsymbol{\theta}) = \mathrm{KL}[q(\mathbf{w}|\boldsymbol{\theta}) \parallel p(\mathbf{w})] - \mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta})} \log p(\mathcal{D}|\mathbf{w}) \tag{7}$$

Eq. (7) hence represents the new cost function to which optimization on $\boldsymbol{\theta}$ is taken. Directly taking derivatives is computationally prohibitive. However it could be further re-arranged into the form of an expectation, lending itself to known approximate solutions such as Monte Carlo estimator of expectation on samples (see Appendix B). Specifically, prior to rearranging into an expectation, if assuming the variational posteriors have diagonal Gaussian distributions, the KL divergence term of Eq. (7) can be further analytically integrated (Kingma and Welling, 2013), as given below, leaving only the likelihood-dependent part to be computed by a Monte Carlo estimator:

$$\mathrm{KL}[q(\mathbf{w}|\boldsymbol{\theta}) \parallel p(\mathbf{w})] = \frac{1}{2}\sum_{j}(\sigma_j^2 + \mu_j^2 - \log\sigma_j^2 - 1) \tag{8}$$

where μ_j, σ_j denote the j-th element of the vectors that represent the variational distribution of weights, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$. Subsequently, a reparameterization operation (see e.g. Kingma and Welling (2013)) is used to remove the dependence on the distribution to which the expectation is taken (i.e. $q(\mathbf{w}|\boldsymbol{\theta})$) in the likelihood-dependent part, whereby unbiased Monte Carlo gradients can be obtained, as given below:

$$\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta})} \log p(\mathcal{D}|\mathbf{w}) = \mathbb{E}_{\boldsymbol{\epsilon}\sim r(\boldsymbol{\epsilon})}\Big[f(g(\boldsymbol{\epsilon}, \boldsymbol{\theta}))\Big] \simeq \frac{1}{L}\sum_{l=1}^{L} f(g(\boldsymbol{\epsilon}^{(l)}, \boldsymbol{\theta})) \tag{9}$$

where $f(\mathbf{w}, \boldsymbol{\theta}) = \log p(\mathcal{D}|\mathbf{w})$; L is the number of samples drawn for the Monte Carlo estimator; $g(\cdot)$ is a differentiable function that transforms a parameter free noise sample, $\boldsymbol{\epsilon}^{(l)} \sim r(\boldsymbol{\epsilon})$, into a sample of the variational posterior: $\mathbf{w}^{(l)} = g(\boldsymbol{\epsilon}^{(l)}, \boldsymbol{\theta}) = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}^{(l)}$, where $r(\boldsymbol{\epsilon})$ is often modelled as standard Gaussian distribution. Otherwise, when the KL divergence term in Eq. (8) is not analytically solvable, the reparameterization operation will then instead be applied to the full expectation from the cost function Eq. (7), given as: $\mathcal{J}(\mathcal{D}, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{w}\sim q(\mathbf{w}|\boldsymbol{\theta})}[\log q(\mathbf{w}|\boldsymbol{\theta}) - \log p(\mathbf{w}) - \log p(\mathcal{D}|\mathbf{w})]$.

In practice, when training in mini-batches (i.e. mini-batch optimization), the above implementation should be re-scaled before derivation is taken:

$$\mathcal{J}^M(\mathcal{D}_M, \boldsymbol{\theta}) = \frac{1}{N}\mathrm{KL}[q(\mathbf{w}|\boldsymbol{\theta}) \parallel p(\mathbf{w})] - \frac{1}{M}\mathbb{E}_{r(\boldsymbol{\epsilon})} \log p(\mathcal{D}_M|g(\boldsymbol{\epsilon}, \boldsymbol{\theta})) \tag{10}$$

where M and N are the size of the mini batch and whole training data, respectively. Reparameterization enables the cost function to be differentiated with respect to $\boldsymbol{\theta}$, whereby the resulting gradients can still be employed using standard stochastic optimization pipelines (e.g. stochastic gradient descent (Bottou, 2012)) :

$$\boldsymbol{\theta}^{\tau+1} = \boldsymbol{\theta}^{\tau} - \eta\nabla_{\boldsymbol{\theta}}\mathcal{J}^M(\mathcal{D}_M, \boldsymbol{\theta}) \tag{11}$$

where the variational parameters are sequentially updated by mini-batches during training; η represents the learning rate.

2.5. Stochastic process representation

For stochastic dynamic response analyses and reliability assessment, in which ground motions are represented as stochastic excitation inputs to engineering structural systems, a Monte Carlo simulation scheme plays a central part (see e.g. Shinozuka and Deodatis (1991, 1988); Spanos and Kougioumtzoglou (2012); Jalayer and Beck (2008); Kiureghian and Fujimura (2009); Rezaeian and Luco (2012); Vlachos et al. (2018b)). Sample realizations are generated, provided by the evolutionary power spectral density (EPSD) of the underlying stochastic process, whose estimation is challenging in the presence of missing data (Comerford et al., 2017; Zhang et al., 2017). Our framework is dedicated to solving this problem. Particularly, the EPSD of the process is estimated from the ensemble average over reconstructions imputed by Eq. (6) and the uncertainty on the spectral density estimates is represented by probability distributions.

Established spectral density estimation approaches, either for stationary cases or non-stationary cases, can be employed in this regard (see e.g. Spanos and Failla (2004); Liang et al. (2007); Spanos and Kougioumtzoglou (2012) for a review). Given the EPSD, sample realizations can hence be generated via a spectral representation method SRM (Liang et al., 2007):

$$m(t) = \sqrt{2} \sum_{n=0}^{N-1} \sqrt{2S_Y(t,\omega_n)\Delta\omega} \cos(\omega_n t + \Phi_n) \tag{12}$$

where $S_Y(t,\omega)$ is the two-sided EPSD of the underlying stochastic process $\{Y(t)\}$; $m(t)$ is the simulation, ϕ_n is the independent random phase angle distributed uniformly over the interval $[0, 2\pi]$; N and $\Delta\omega$ relate to the discretization of the frequency domain. This enables the proposed approach to be able to characterise the stochastic excitations for engineering simulation analyses, capturing the non-stationary characteristics of earthquake ground motions, even when the source load data are incomplete. This is of great engineering importance when the associated earthquake scenarios are of interest to the seismic assessment of engineering structures, under the PBEE practice.

3. Application examples

In this section we demonstrate the performance of the proposed framework using an accelerogram from the ESM (Engineering Strong Motion) database (Lanzano et al., 2021). Note that when working with recorded time-histories, one can generally have a single observed seismic recording as a realization of a stochastic process, where the true power spectrum of the underlying process is typically unknown (Narayana Iyengar and Sundara Raja Iyengar, 1969). Therefore, the spectral estimates from the otherwise complete recording could then serve as the reference for comparison. Given a ground motion time-history record, power spectral density (PSD) estimates are derived using the Welch method (Welch, 1967) (stationary case), and the evolutionary power spectra (EPSD) are estimated from short time Fourier transform (Liang et al., 2007) (nonstationary case).

Region specific parameters to the seismological model (see Eq. (1)) are inferred from seismographic studies of the region (Bindi and Kotha, 2020; Razafindrakoto et al., 2021), coupled with the event information associated with the target recording (i.e. $M_w = 6.5$, normal faulting, $R = 18.6km$, recorded at a class A site in Italy). To consider the variability of ground motions, some key input parameters of significance are modelled as probability distributions, as shown in Table 1, while other deterministic ones are listed in the Appendix in Table C.5. In generating ground motions, the slip distribution and hypocenter location are modelled as random. Specifically, 100 physics-informed simulations with parameter variability are obtained, from which we have trained a Bayesian neural network model with 2 hidden layers. Under the autoregressive modelling scheme, as suggested by Eq. (2), the input layer is specified by the lagged width p while the output layer has 1 output node. Each hidden layer is composed of 16 hidden units, activated by the rectified linear function. This architecture is the result of comprehensive hyperparameter tuning (including the learning rate η) based on a 20% hold-out validation set from these simulations.

Table 1: Statistical parameters of the stochastic finite fault model

Parameter	Distribution	mean	s.t.d	min	max
$\log \Delta\sigma$	Gaussian	1.96	0.31		
κ_0	Uniform			0.002	0.008
d	Gaussian	9.2	10	2	30
b_1 (0 – 70km)	Gaussian	-1.35	0.1		
b_2 (70 – 140km)	Gaussian	-0.57	0.5		
υ	Uniform			-0.15	0.15

3.1. Missing gaps at random locations

In this study we focus on the effect of missing gaps, which suggest a variable length of unknown samples consecutively grouped together from an otherwise continuous set of measurements, significantly decreasing the number of usable empirical records. This situation is of particular interest to studies of historical earthquakes which may potentially provide strong-motion records but many of them are discarded due to the presence of missing gaps (Church et al., 2013; Palombo and Pino, 2013). For example, in a study of an Italian earthquake in 1930 (Vannoli et al., 2015), only 11 out of the 113 seismograms recovered from seismological observatories across Europe were employed mostly due to the inability to analyze incomplete seismograms (Maranò et al., 2017). Moreover, the presence of gaps is also common in modern seismograms subject to serious clipping in which consecutive points are clipped during peak motions (Yang and Ben-Zion, 2010; Zhang et al., 2016). Instrumentation malfunction or incompetence, or loss of communications may also lead to missing data. Other examples include instrument bandwidth limitations, low-cost temporary instruments in harsh conditions, or data contamination etc. (Smith-Boughner and Constable, 2012; Comerford et al., 2015a, 2016; Zhang et al., 2017). To comprehensively investigate the effects of data gaps, various scenarios where different combinations of gap sizes (i.e. the number of missing samples) and gap number (i.e. the number of gaps) are randomly removed in the strong motion phase, are conducted in this analysis, as listed in Table A.4.

3.2. *Quantitative metrics to compare the performance*

To evaluate uncertainties and accuracy under different configurations of missing data, three quantitative metrics are designed. These metrics are reported on the power spectral densities for characterizing the input stochastic process and on pseudo spectral accelerations (5% damped) for characterizing responses of engineering systems. P_{95} corresponds to an interval coverage probability measure that reflects the percentage of target PSD values being captured by the estimated credible intervals (Pearce et al., 2018), given as:

$$P_{95} = \frac{c_f}{n_f} \tag{13}$$

where c_f represents the number of frequencies in which the target spectral density is captured within the 95% credible interval. Upon denoting the predicted lower and upper bound as y_L and y_U, c_f is defined by a variable k_i of length n_f (total number of frequency bins) that indexes a frequency value captured by the estimated credible interval:

$$c_f = \sum_{i=1}^{n} k_i \tag{14}$$

$$k_i = \begin{cases} 1 & y_{Li} \leq y_i \leq y_{Ui} \\ 0 & \text{else} \end{cases} \tag{15}$$

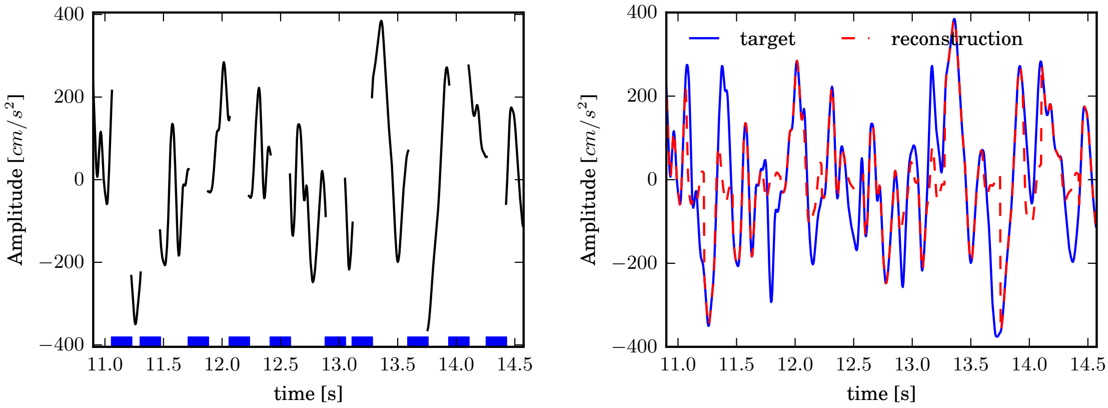
In addition, A_{LU} represents the area between the lower y_U and upper bounds y_L across the frequency range, which illustrates the magnitude of uncertainty levels. e denotes the mean absolute error of the PSD estimates, which evaluates the accuracy of the mean estimation:

$$e = \frac{1}{n} \sum_{i=1}^{n} y_i - \hat{y}_i \tag{16}$$

3.3. *A detailed scenario case*

Of all the scenarios considered (see Table A.4), one serious scenario case corresponding to 10 gaps of size 32, in total equivalent to 44% missing data within the strong motion phase, is specifically demonstrated herein in details for conciseness (see Fig. 2 - Fig. 7). Fig. 2a shows such incomplete recording with gaps indicated by the blue bar at the bottom. Fig. 2b then shows one reconstructed time-history from the ensemble collection of 500 reconstructions by the updated BNN model, which largely resemble the waveform of the original recording. Past studies have suggested the difficulty in restoring the waveform in the time domain with missing values consecutively grouped (as in gaps), compared to missing values scattered across the signal (Maranò et al., 2017; Comerford et al., 2017; Christmas, 2013). In fact, this difficulty further justifies the importance of uncertainty quantification due to the propagation of imperfect reconstruction error.

Based on the ensemble reconstructions, the uncertainties over the power spectrum can further be seen in Fig. 3a. Despite a significant portion of data missing (44%), the ensemble-averaged PSD agrees well with the target PSD from the otherwise complete recording, whose target spectral values across the whole frequency range are generally captured in the 95% credible interval bounds. The heteroscedasticity of variances with respect to frequencies is observed. As a comparison, significant power loss is seen from the result by a simple zero-padded approach. In more details, Fig. 3b illustratively displays the probability distribution shape of spectral density estimates with respect to



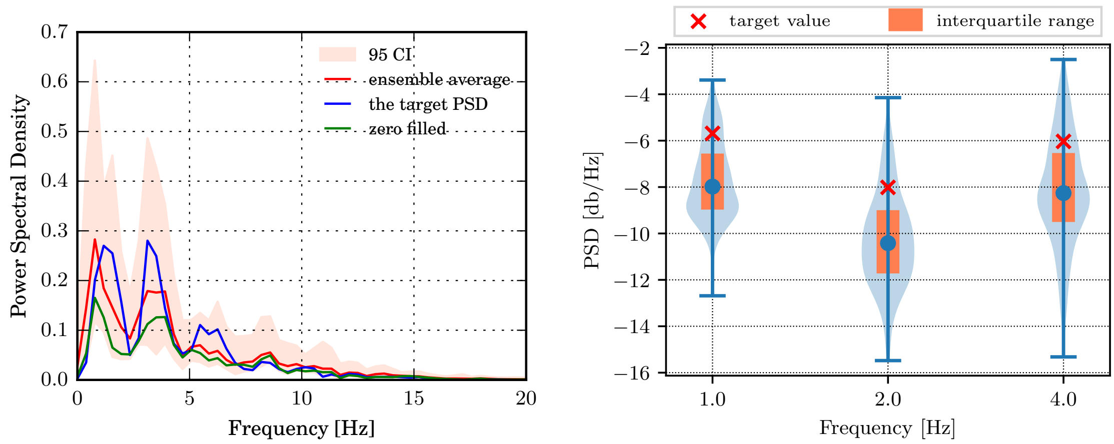
(a) Incomplete record with missing gaps

(b) One of the ensemble reconstructions

Figure 2: Gapped type of missing data and one reconstruction from the ensemble. Missing percentage 44%

frequency. In addition, descriptive statistics regarding the ensemble-averaged PSD estimates are also depicted. The box within represents the regular box plot showing the statistics corresponding to quantiles such as 25%, median and 75%. The blue circle represents the median value while the red cross represents the target i.e., the PSD value from the full recording.

In addition, results from another baseline method, in which missing values are filled with samples from standard Gaussian distribution (Comerford et al., 2015b), are shown in Fig. 4. By contrast, our ensemble-average estimate has better approximated the target result and our interval bounds have better covered the target, as clearly seen in Fig. 3b and Fig. 4b. This superior performance could be attributed to our updated BNN's ability to learn the temporal dependence of the underlying process. While the "white noise" imputation approach respects the basic property of a stochastic process, it can hardly know the variance with respect to the random variable at each time stamp and also the covariance structure.

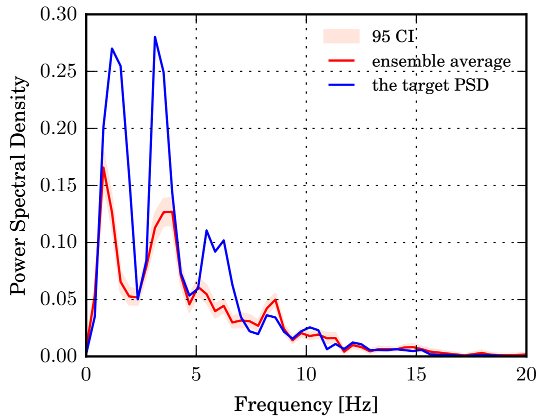


(a) Global power spectral density estimates of the ensemble reconstructions. The ensemble average and its 95% confidence interval are compared to the target and a time history with zero-filled gaps

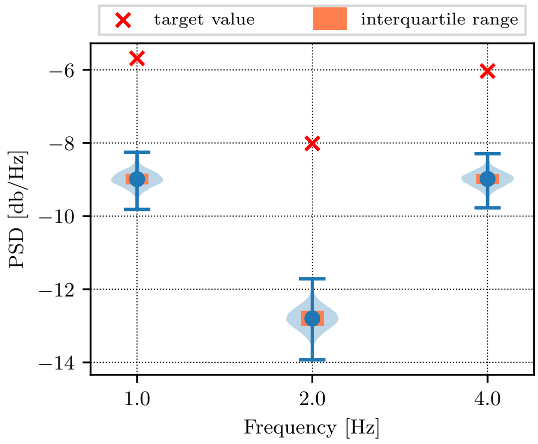
(b) The distribution of spectral density values with respect to frequency. The box plot shows reconstructed PSD quantiles at 25% (box), median (circle) and 75% (whisker). The red cross represents the target value

Figure 3: Uncertainties in the power spectral density estimates. Missing percentage 44%

It should be noted that the stationary (global) PSD estimates provide the spectral distribution in an average sense, without time information. But engineering interests, driven by PBEE, are increasingly focused on the time-varying



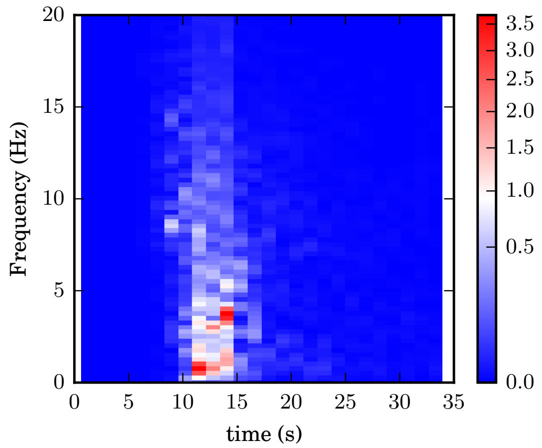
(a) Global power spectral density estimates of the baseline approach



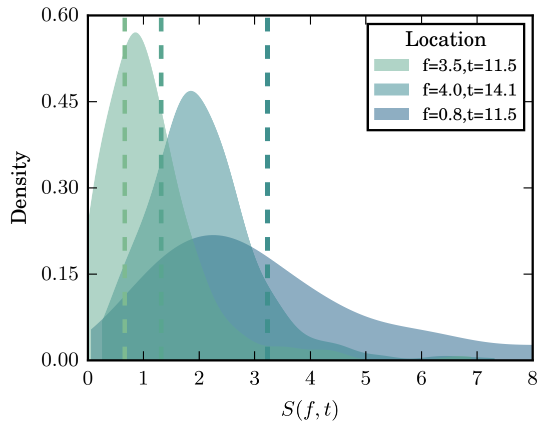
(b) The distribution of spectral density values with respect to frequency

Figure 4: An baseline approach for comparison with the proposed approach

spectral representation due to the "moving resonance" effect of nonlinear structural analysis. As such, an ensemble of estimates of the evolutionary power spectrum are computed, with the averaged EPSD shown in Fig. 5a; more importantly, the distribution of spectral density values, $S(f,t)$, at selected time instants and frequency bins are displayed in Fig. 5b for illustration. Several representative combinations of time instants and frequency bins are selected to show the variance of spectral estimates. The corresponding target values are shown by the vertical lines, which are well captured by the estimated probability distributions.



(a) Ensemble averaged evolutionary power spectral density

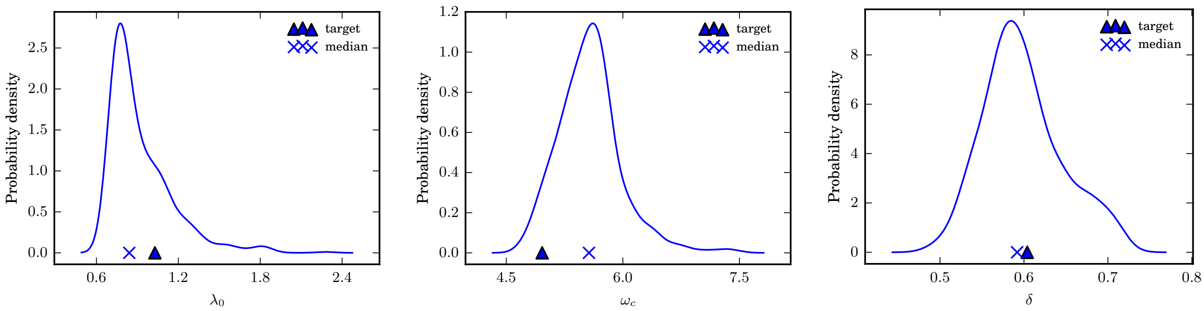


(b) Probability distributions of the evolutionary power spectral density estimates at selected time instant and frequency

Figure 5: Evolutionary power spectral density estimate and its uncertainty

Fig. 6 further displays the distribution of spectral moments (see definition in Appendix D), the key parameters of spectral representation of stochastic seismic inputs (Lai, 1982; Zhang et al., 2017). Uncertainties due to the incomplete data are shown, indicating that the target values from the full recording are well captured even with a missing percentage of 44%. Spectral moments can be used to calibrate parameterized stochastic process models, e.g. the established Kanai Tajimi model via a spectral moment method (see e.g. Lai (1982) for details). Indeed more complex models (e.g. Conte and Peng (1997); Vlachos et al. (2018a)) that reflect the nonstationary characteristics of ground motions could also similarly be calibrated with the ensemble reconstructions through, for example, spectral fitting.

Importantly, it suggests that parameter uncertainties could thus be accounted for when characterising ground motions using parameterized models.



(a) variance of excitation λ_0 (b) central frequency ω_c (c) shape factor δ

Figure 6: The distribution of spectral moments due to incomplete data

Relying on the Monte Carlo simulation approach (Shinozuka and Deodatis, 1988), powered by the spectral representation method SRM (Eq. (12)), sample realizations compatible with the given stochastic process can be simulated for stochastic nonlinear dynamic analyses, (see e.g. Jalayer and Beck (2008); Kiureghian and Fujimura (2009); Rezaeian and Luco (2012); Vlachos et al. (2018b)). As a result, Fig. 7 illustrates, side by side, the sample generation based on the ensemble averaged EPSD estimates, along with the reconstruction directly from our updated BNN model. It suggests that, even in the presence of a significant number of data gaps, both the reconstruction and the generation resemble the target recording very well.

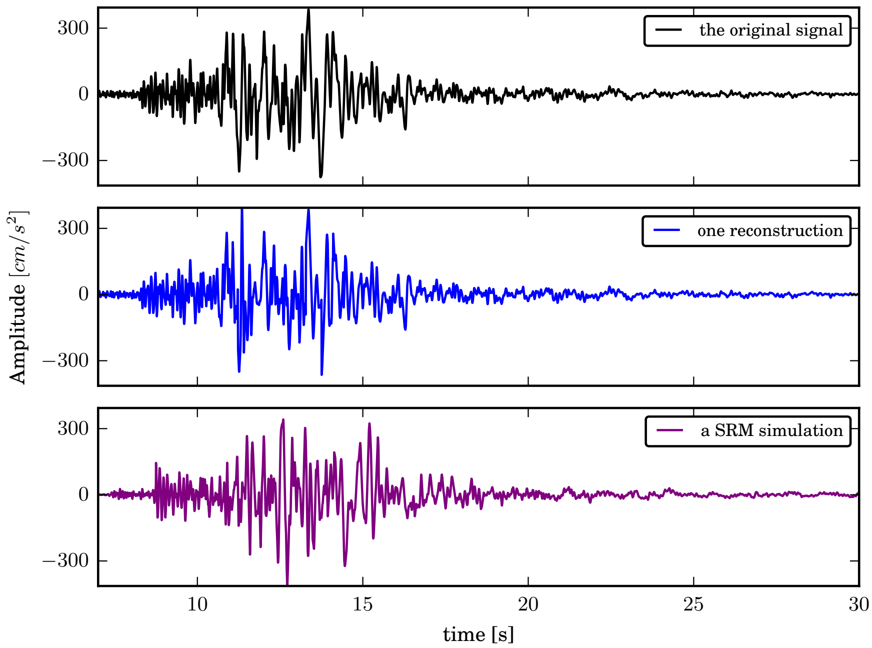
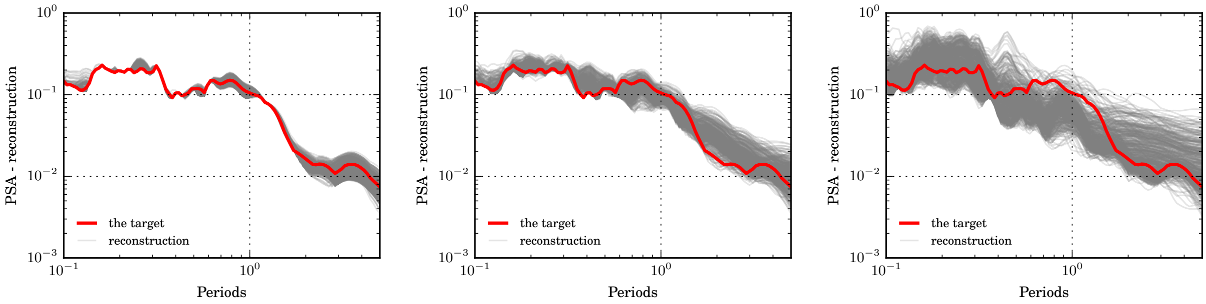


Figure 7: Target recording (top) compared with a direct reconstruction from the updated Bayesian neural network model (middle) and a sample generation of the underlying stochastic process by the stochastic representation method (SRM) from the ensemble-averaged EPSD (bottom)

3.4. *Performance comparison of many scenarios*

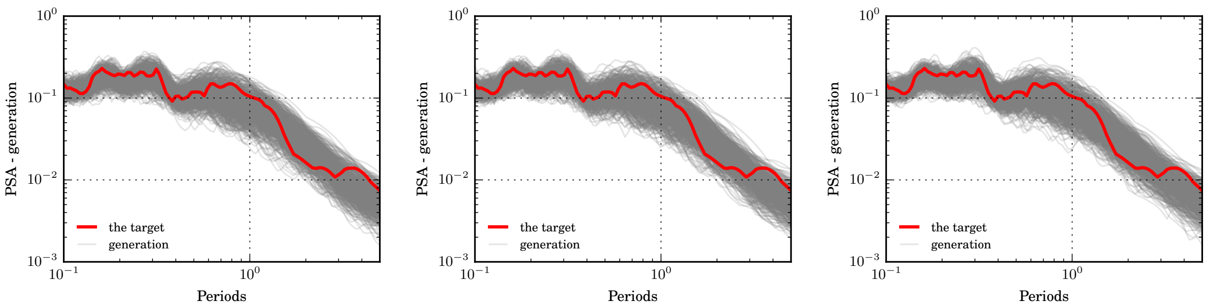
In earthquake engineering, accelerograms are also frequently characterized by the pseudo-acceleration (5% damped) elastic response spectra. Fig. 8 illustratively shows the variability of spectral amplitudes of the reconstructions associated with three representative levels of missing gaps. The target response spectrum is shown in thick line, together with response spectra of 500 reconstructions from the ensemble. While larger uncertainty is found with increasing levels of missing data, the extreme case with roughly 70% of missing gaps still captures the target spectra to a large extent. For less extreme cases, the target response spectra is well contained within the suite of reconstructed response spectra across the full range of spectral periods. This reflects the ability of the proposed approach to quantify uncertainty in our reconstructions in response to the missing data and suggests the validity for the reconstructions to be used for seismic structural analyses.



(a) size16 × num2 (MP=4%)   (b) size32 × num4 (MP=17%)   (c) size64 × num8 (MP=70%)

Figure 8: Response spectrum of reconstructions from BNN: three representative missing gap scenarios with increasing missing percentages. The target response spectrum is shown by the thick line, together with response spectra of 500 reconstructions from the ensemble

On the other hand, the response spectra of our sample generations from the EPSD, along with the target response spectra, are displayed in Fig. 9. All the sample realizations have captured the target spectra quite well. Little differences can be seen between the three data-loss scenarios, suggesting the robustness of the ensemble-averaged EPSD even under serious missing data (of up to 70%). This, therefore, validates the representation of the ground motion using estimated evolutionary power spectra by the presented approach and demonstrates its ability to make stochastic dynamic analyses still achievable in the presence of serious missing data. This result furthermore highlights the usefulness of the proposed method within a Monte Carlo simulation scheme.



(a) size16 × num2 (MP=4%)   (b) size32 × num4 (MP=17%)   (c) size64 × num8 (MP=70%)

Figure 9: Response spectrum of sample generations from EPSD: three representative missing gap scenarios with increasing missing ratio

For completeness, quantitative performance evaluation of the reconstructions in respect to various missing gap scenarios are tabulated in Table 2 (reported in terms of the power spectrum) and Table 3 (reported in terms of the response spectrum), in which all the metrics are computed and averaged over 10 runs to obtain representative results against randomness. The total missing percentage (MP) of various combinations of gap numbers and sizes are listed

Table 2: Performance comparison on power spectral density of reconstructions under various configurations of missing gaps (averaged over 10 runs)

PSD	gap size	number of gaps				
		2	4	6	8	10
e (e-3)	16	0.958	1.181	1.935	2.282	2.879
	32	1.703	2.389	3.202	3.846	4.336
	64	2.806	4.232	5.343	7.986	-
A_{LU}	16	0.524	0.630	0.848	1.006	1.205
	32	0.830	1.274	1.618	2.262	2.418
	64	1.707	2.920	3.528	5.301	-
P_{95} (%)	16	86.095	86.243	79.734	74.556	73.077
	32	83.876	83.432	76.479	78.107	80.030
	64	83.136	86.686	81.065	81.361	-

e denotes the mean absolute error; A_{LU} the area metric; P_{95} prediction interval coverage probability

Table 3: Performance comparison on response spectrum of reconstructions under various configurations of missing gaps (averaged over 10 runs)

PSA	gap size	number of gaps				
		2	4	6	8	10
e (e-2)	16	0.538	0.667	1.134	1.313	1.600
	32	0.925	1.328	1.785	2.039	2.229
	64	1.621	2.029	2.658	3.157	-
A_{LU}	16	0.013	0.015	0.020	0.023	0.026
	32	0.020	0.029	0.035	0.043	0.045
	64	0.037	0.049	0.060	0.070	-
P_{95} (%)	16	81.615	89.769	89.231	88.308	83.462
	32	80.385	84.000	86.077	82.923	88.077
	64	85.154	87.615	82.385	85.308	-

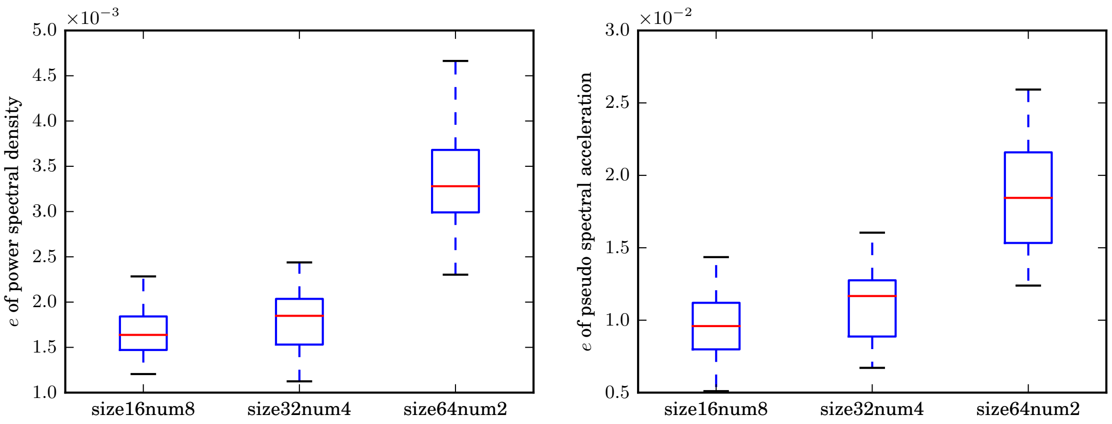
e denotes the mean absolute error; A_{LU} the area metric; P_{95} prediction interval coverage probability

as a reference in a color coded way in Table A.4. For both spectra, larger deviations and higher uncertainties are found as with the increase of missing percentage, which is intuitively understandable as a result of the iterative nature of the approach. Particularly, the error of PSD roughly increases by 60% when doubling the gap length (under the same gap numbers), which suggests the accumulation of errors propagated from the reconstructions. Generally, the estimated credible intervals covered both target spectrum quite well, with P_{95} higher than 80% for most scenarios. However, it should be noted that the high coverage probability of scenarios with missing percentage are at the cost of wider interval bounds, as suggested by A_{LU}. The detailed scenario case in Section 3.3, along with three more scenarios shown in Fig. 8 and Fig. 9, exemplify the scale of results and demonstrates the performance.

Note that, while included for completeness, the scenario with 10 gaps of size 64 is not compatible with our Bayesian updating setting, since too much of the empirical observations are missing (i.e. 87%), indicating that only very sparse samples of data are left. It is suggested by Eq. (2) that the partial chunks adopted for updating should be at least the size of p.

3.5. Impact of different data-loss scenarios

In addition to exploring the impacts of missing levels, this analysis further investigates more complicated patterns, since a certain missing data percentage could be associated with different scenarios, for example a 17.41% data loss in the strong motion phase may be attributed to three combinations: 8 gaps of size 16, 4 gaps of size 32, or 2 gaps of size 64. As a result, Fig. 10 shows the comparison of errors on both power spectral density and response spectral acceleration amplitudes, over 10 runs, in box plots. For power spectral estimates, under the same missing level, the first two scenarios (namely, 8 gaps of size 16 and 4 gaps of size 32) achieve comparable accuracy on average, though the second has slight higher error and slightly larger variability. But more significantly, the third scenario with the longest gap and least number of gaps (i.e. 2 gaps of size 64) has much higher error and much higher variability. For response spectral acceleration amplitudes, differences manifest a similar trend as the results in terms of power spectra. As with longer gaps, in spite of fewer gaps, the average error increases. Still, the third scenario (2 gaps of size 64) results in the worst performance, with largest error and variability. This may suggest that the performance is more sensitive to the gap length (especially quite long gaps) than the quantity of gaps.



(a) Mean absolute error on power spectral density

(b) Mean absolute error on response spectrum

Figure 10: Comparison of mean absolute error for investigating the effects of 3 different missing gap scenarios with same missing level

4. Conclusion

In this paper, a Bayesian framework to stochastically characterize ground motions in the presence of missing data is presented. This framework features the use of Bayesian neural networks that allow for epistemic uncertainty quantification, and a Bayesian model updating component that allows for the combination of seismological knowledge (*a priori* knowledge) with empirical observations (even incomplete) via Bayesian inference. The effect of missing gaps has been comprehensively studied via various missing scenarios, based on which the performance of the proposed method has been quantitatively demonstrated. Results show that the proposed method is highly effective even in serious cases of data-loss with about half of data missing in the strong motion phase, being capable of providing imputed waveforms, spectral estimates and stochastic synthetic generations that agree well with the target recording.

A host of representations of ground motion, consistent with an underlying stochastic process, are provided in a probabilistic manner, suggesting the versatility of the proposed approach as a general solution to dealing with missing data for various engineering and seismological applications, whether waveform-based or spectrum-based. The proposed approach helps in recovering the information conveyed from faulty or incomplete observations, for example, from low-cost temporary instruments deployed at scale. The Bayesian framework provides a building block on which it could be developed to enrich the database of ground motions in data scare areas (eg. near-field strong motions), facilitating stochastic dynamic analyses of engineering structures and boosting the understanding of earth structures. Of particular note is its mechanism that combines *a priori* information with empirical observations, remedying the causality dilemma concerning the dependence of observations and the extracted knowledge/information. Finally, we consider that, such Bayesian framework could serve as a complementary approach to current stochastic ground-motion models under the growing interests of PBEE (performance-based earthquake engineering), and ultimately a fundamental solution to the limited data problem in data scarce regions.

5. Acknowledgement

This work was supported by the EU Horizon 2020 - MSCA Actions project URBASIS [Project no. 813137];

Appendix A. Missing percentages for various scenarios

Table A.4: The total missing percentage (MP) for various missing scenarios

gap size	number of gaps				
	2	4	6	8	10
16	4.35	8.71	13.06	17.41	21.77
32	8.71	17.41	26.12	34.83	43.54
64	17.41	34.83	52.24	69.66	87.07

Appendix B. Monte Carlo estimator

Consider a general probabilistic objective function of the form:

$$\mathcal{F}(\boldsymbol{\theta}) = \int p(\mathbf{x};\boldsymbol{\theta}) f(\mathbf{x};\boldsymbol{\phi}) d\mathbf{x} = \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[f(\mathbf{x};\boldsymbol{\phi})] \tag{B.1}$$

where $f(\mathbf{x};\boldsymbol{\phi})$ denotes a general function of an input variable $\mathbf{x}$ with structural parameters $\boldsymbol{\phi}$; $p(\mathbf{x};\boldsymbol{\theta})$ represents a probability distribution parameterized by $\boldsymbol{\theta}$.

The usual Monte Carlo estimator for expectation is given by:

$$\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[f(\mathbf{x};\boldsymbol{\phi})] \backsimeq \frac{1}{N}\sum_{1}^{N} f(\hat{\mathbf{x}}^{(n)}), \text{ where } \hat{\mathbf{x}}^{(n)} \sim p(\mathbf{x};\boldsymbol{\theta}) \tag{B.2}$$

It suggests that a complex integral in B.1 can be numerically evaluated by drawing samples from the probability distribution $p(\mathbf{x};\boldsymbol{\theta})$ and then computing the average of the function evaluated at these samples. Furthermore, as many problems in Machine Learning generally focused on the computation of gradients, such as $\nabla_{\boldsymbol{\theta}}\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[f(\mathbf{x};\boldsymbol{\phi})]$. Several techniques exist to do further approximation, see additional details in (Mohamed et al., 2020). As an example, a Monte Carlo gradient estimator by the score function is given as:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[f(\mathbf{x};\boldsymbol{\phi})] &= \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})}[f(\mathbf{x};\boldsymbol{\phi})\nabla_{\boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta})] \\ &= \frac{1}{N}\sum_{1}^{N} f(\hat{\mathbf{x}}^{(n)})\nabla_{\boldsymbol{\theta}} \log p(\hat{\mathbf{x}}^{(n)};\boldsymbol{\theta})]\end{aligned}$$

where $\hat{\mathbf{x}}^{(n)} \sim p(\mathbf{x};\boldsymbol{\theta})$

Table C.5: Source and path parameters of the stochastic finite fault model (sourced from (Bindi and Kotha, 2020; Razafindrakoto et al., 2021))

Parameter	Description	Value
ρ_s	density of the medium	2.7
β	shear wave velocity	3.2
V	horizontal partition	$1/\sqrt{2}$
$R_{\theta\Phi}$	radiation pattern	0.55
F	free-surface factor	2
R_0	reference distance	10
Q	quality factor	$Q = 250.4f^{0.29}$

Appendix C. Seismological parameters of the finite-fault model

Appendix D. Spectral moments

The spectral moments are key statistical parameters in frequency domain analyses, which are of particular importance in evaluating survival probability or reliability assessment for structural systems. Consider stationary random processes, the jth spectral moment λ_j are given as (Lai, 1982; Zhang et al., 2017):

$$\lambda_j = \int_{-\infty}^{+\infty} \omega^j S(\omega) d\omega \tag{D.1}$$

where $S(\omega)$ denotes the two-sided power spectral density. Specifically, the zero spectral moment λ_0, which is also the variance of the excitation, is given as:

$$\lambda_0 = \int_{-\infty}^{+\infty} S(\omega) d\omega \tag{D.2}$$

then the central frequency ω_c, and the shape factor δ (also known as bandwidth measure) of the stochastic process can be computed from the first few spectra moments:

$$\begin{aligned} \omega_c &= [\lambda_1/\lambda_2]^{1/2} \\ \delta &= [1 - (\lambda_1^2/\lambda_0\lambda_0)]^{1/2} \end{aligned}$$

Appendix E. List of symbols

References

Atkinson, G.M., Boore, D.M., 2006. Earthquake ground-motion prediction equations for eastern north america. Bulletin of the seismological society of America 96, 2181–2205.

Babu, P., Stoica, P., 2010. Spectral analysis of nonuniformly sampled data–a review. Digital Signal Processing 20, 359–378.

Baker, J., Bradley, B., Stafford, P., 2021. Physics-Based Ground-Motion Characterization. Cambridge University Press. p. 196–246. doi:10.1017/9781108425056.007.

Beer, M., Spanos, P.D., 2009. A neural network approach for simulating stationary stochastic processes. Structural engineering and mechanics: An international journal 32, 71–94.

Bindi, D., Kotha, S., 2020. Spectral decomposition of the engineering strong motion (esm) flat file: regional attenuation, source scaling and arias stress drop. Bulletin of Earthquake Engineering 18, 2581–2606.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association 112, 859–877. doi:10.1080/01621459.2017.1285773, arXiv:1601.00670.

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network, in: International conference on machine learning, PMLR. pp. 1613–1622.

Boore, D.M., 2003. Simulation of ground motion using the stochastic method. Pure and applied geophysics 160, 635–676.

Bos, R., De Waele, S., Broersen, P.M., 2002. Autoregressive spectral estimation by application of the burg algorithm to irregularly sampled data. IEEE Transactions on Instrumentation and Measurement 51, 1289–1294.

Bottou, L., 2012. Stochastic gradient descent tricks, in: Neural networks: Tricks of the trade. Springer, pp. 421–436.

Symbols	Description
Θ_e	event metadata from the incomplete recording
Θ_g	region-specific seismological parameters
e	mean absolute error
A_{LU}	area between the lower and upper bounds of credible interval
P_{95}	prediction interval coverage probability
ρ_s	density of the medium
f_0	corner frequency
β	shear wave velocity
V	horizontal partition
$R_{\theta\Phi}$	radiation pattern
F	free-surface factor
R_0	reference distance
Q	quality factor
M_0	seismic moment
R	hypocentral distance
r	epicentral distance
d	depth
$\Delta\sigma$	stress drop
β	shear wave velocity
$Q(f)$	an inverse measure of anelastic attenuation
υ	site amplification factor in log units
κ_0	kappa value
b	geometric spreading coefficients
$\mathbf{w}$	collectively the weights and biases of a neural network model
$\boldsymbol{\theta}$	the parameters of the variational distribution
p	lagged window width in autoregressive modelling
η	learning rate
$m(t)$	sample simulation compatible with a given stochastic process
ϕ_n	the independent random phase angle
λ_j	jth spectral moment
ω_c	central frequency
δ	shape factor

Broersen, P.M., De Waele, S., Bos, R., 2004. Autoregressive spectral analysis when observations are missing. Automatica 40, 1495–1504.

Christmas, J., 2013. The effect of missing data on robust bayesian spectral analysis, in: 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE. pp. 1–6.

Church, E.D., Bartlett, A.H., Jourabchi, M.A., 2013. Raster-to-vector image analysis for fast digitization of historic seismograms. Seismological Research Letters 84, 489–494.

Comerford, L., Jensen, H., Mayorga, F., Beer, M., Kougioumtzoglou, I., 2017. Compressive sensing with an adaptive wavelet basis for structural system response and reliability analysis under missing data. Computers & Structures 182, 26–40.

Comerford, L., Kougioumtzoglou, I.A., Beer, M., 2015a. An artificial neural network approach for stochastic process power spectrum estimation subject to missing data. Structural Safety 52, 150–160.

Comerford, L., Kougioumtzoglou, I.A., Beer, M., 2015b. On quantifying the uncertainty of stochastic process power spectrum estimates subject to missing data. International Journal of Sustainable Materials and Structural Systems 2, 185–206.

Comerford, L., Kougioumtzoglou, I.A., Beer, M., 2016. Compressive sensing based stochastic process power spectrum estimation subject to missing data. Probabilistic Engineering Mechanics 44, 66–76.

Conte, J., Peng, B., 1997. Fully nonstationary analytical earthquake ground-motion model. Journal of Engineering Mechanics-Proceedings of the ASCE 123, 15–24.

Edwards, B., Fäh, D., 2013. A stochastic ground-motion model for switzerland. Bulletin of the Seismological Society of America 103, 78–98.

Edwards, B., Zurek, B., Van Dedem, E., Stafford, P., Oates, S., Van Elk, J., DeMartin, B., Bommer, J., 2019. Simulations for the development of a ground motion model for induced seismicity in the groningen gas field, the netherlands. Bulletin of Earthquake Engineering 17, 4441–4456.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR. pp. 1050–1059.

Gatti, F., Clouteau, D., 2020. Towards blending physics-based numerical simulations and seismic databases using generative adversarial network. Computer Methods in Applied Mechanics and Engineering 372, 113421.

Graves, A., 2011. Practical variational inference for neural networks. Advances in neural information processing systems 24.

Graves, A., 2013. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 .

Hernández-Lobato, J.M., Adams, R., 2015. Probabilistic backpropagation for scalable learning of bayesian neural networks, in: International conference on machine learning, PMLR. pp. 1861–1869.

Hung, J.C., 2008. A genetic algorithm approach to the spectral estimation of time series with noise and missed observations. Information Sciences 178, 4632–4643.

Jalayer, F., Beck, J., 2008. Effects of two alternative representations of ground-motion uncertainty on probabilistic seismic demand assessment of structures. Earthquake engineering & structural dynamics 37, 61–79.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 .

Kiureghian, A.D., Fujimura, K., 2009. Nonlinear stochastic dynamic analysis for performance-based earthquake engineering. Earthquake Engineering & Structural Dynamics 38, 719–738.

Kondrashov, D., Denton, R., Shprits, Y., Singer, H., 2014. Reconstruction of gaps in the past history of solar wind parameters. Geophysical Research Letters 41, 2702–2707.

Kondrashov, D., Ghil, M., 2006. Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Processes in Geophysics 13, 151–159.

Lai, S.S.P., 1982. Statistical characterization of strong ground motions using power spectral density function. Bulletin of the Seismological Society of America 72, 259–274.

Lanzano, G., Luzi, L., Cauzzi, C., Bienkowski, J., Bindi, D., Clinton, J., Cocco, M., D'Amico, M., Douglas, J., Faenza, L., et al., 2021. Accessing european strong-motion data: An update on orfeus coordinated services. Seismological Research Letters 92, 1642–1658.

Laurendeau, A., Cotton, F., Bonilla, L.F., 2012. Nonstationary stochastic simulation of strong ground-motion time histories: Application to the japanese database. arXiv preprint arXiv:1212.3938 .

Liang, J., Chaudhuri, S.R., Shinozuka, M., 2007. Simulation of nonstationary stochastic processes by spectral representation. Journal of Engineering Mechanics 133, 616–627.

Maranò, S., Edwards, B., Ferrari, G., Fäh, D., 2017. Fitting earthquake spectra: colored noise and incomplete data. Bulletin of the Seismological Society of America 107, 276–291.

McCallen, D., Petersson, A., Rodgers, A., Pitarka, A., Miah, M., Petrone, F., Sjogreen, B., Abrahamson, N., Tang, H., 2021a. Eqsim—a multidisciplinary framework for fault-to-structure earthquake simulations on exascale computers part i: Computational models and workflow. Earthquake Spectra 37, 707–735.

McCallen, D., Petrone, F., Miah, M., Pitarka, A., Rodgers, A., Abrahamson, N., 2021b. Eqsim—a multidisciplinary framework for fault-to-structure earthquake simulations on exascale computers, part ii: Regional simulations of building response. Earthquake Spectra 37, 736–761.

Mohamed, S., Rosca, M., Figurnov, M., Mnih, A., 2020. Monte carlo gradient estimation in machine learning. J. Mach. Learn. Res. 21, 1–62.

Musial, J.P., Verstraete, M.M., Gobron, N., 2011. Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series. Atmospheric chemistry and physics 11, 7905–7923.

Narayana Iyengar, R., Sundara Raja Iyengar, K., 1969. A nonstationary random process model for earthquake accelerograms. Bulletin of the Seismological Society of America 59, 1163–1188.

Palombo, B., Pino, N.A., 2013. On the recovery and analysis of historical seismograms. Annals of Geophysics .

Paolucci, R., Smerzini, C., Vanini, M., 2021. Bb-speedset: A validated dataset of broadband near-source earthquake ground motions from 3d physics-based numerical simulations. Bulletin of the Seismological Society of America 111, 2527–2545.

Pearce, T., Brintrup, A., Zaki, M., Neely, A., 2018. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach, in: International conference on machine learning, PMLR. pp. 4075–4084.

Razafindrakoto, H.N., Cotton, F., Bindi, D., Pilz, M., Graves, R.W., Bora, S., 2021. Regional calibration of hybrid ground-motion simulations in moderate seismicity areas: Application to the upper rhine graben. Bulletin of the Seismological Society of America 111, 1422–1444.

Rezaeian, S., Der Kiureghian, A., 2010. Simulation of synthetic ground motions for specified earthquake and site characteristics. Earthquake Engineering & Structural Dynamics 39, 1155–1180.

Rezaeian, S., Luco, N., 2012. Example applications of a stochastic ground motion simulation methodology in structural engineering, in: 15th World Conf. Earthquake Engineering,(WCEE).

Roberts, D.H., Lehár, J., Dreher, J.W., 1987. Time series analysis with clean-part one-derivation of a spectrum. The astronomical journal 93, 968.

Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T., 2020. Deepar: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting 36, 1181–1191.

Scargle, J.D., 1982. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal 263, 835–853.

Shinozuka, M., Deodatis, G., 1988. Stochastic process models for earthquake ground motion. Probabilistic engineering mechanics 3, 114–123.

Shinozuka, M., Deodatis, G., 1991. Simulation of stochastic processes by spectral representation .

Smith-Boughner, L., Constable, C., 2012. Spectral estimation for geophysical time-series with inconvenient gaps. Geophysical Journal International 190, 1404–1422.

Spanos, P., Kougioumtzoglou, I., 2012. Harmonic wavelets based statistical linearization for response evolutionary power spectrum determination. Probabilistic Engineering Mechanics 27, 57–68.

Spanos, P.D., Failla, G., 2004. Evolutionary spectra estimation using wavelets. Journal of Engineering Mechanics 130, 952–960.

Stoica, P., Larsson, E.G., Li, J., 2000. Adaptive filter-bank approach to restoration and spectral analysis of gapped data. The Astronomical Journal 120, 2163.

Tobar, F., 2018. Bayesian nonparametric spectral estimation, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/abd1c782880cc59759f4112fda0b8f98-Paper.pdf.

Tsioulou, A., Taflanidis, A.A., Galasso, C., 2018. Modification of stochastic ground motion models for matching target intensity measures. Earthquake Engineering & Structural Dynamics 47, 3–24.

Vannoli, P., Vannucci, G., Bernardi, F., Palombo, B., Ferrari, G., 2015. The Source of the 30 October 1930 Mw 5.8 Senigallia (Central Italy) Earthquake: A Convergent Solution from Instrumental, Macroseismic, and Geological Data. Bulletin of the Seismological Society of America 105, 1548–1561. URL: https://doi.org/10.1785/0120140263, doi:10.1785/0120140263, arXiv:https://pubs.geoscienceworld.org/ssa/bssa/article-pdf/105/3/1548/3656198/1548.pdf.

Vetter, C., Taflanidis, A.A., 2012. Global sensitivity analysis for stochastic ground motion modeling in seismic-risk assessment. Soil Dynamics and Earthquake Engineering 38, 128–143.

Vetter, C., Taflanidis, A.A., 2014. Comparison of alternative stochastic ground motion models for seismic risk characterization. Soil Dynamics and Earthquake Engineering 58, 48–65.

Vlachos, C., Papakonstantinou, K.G., Deodatis, G., 2018a. Predictive model for site specific simulation of ground motions based on earthquake scenarios. Earthquake Engineering & Structural Dynamics 47, 195–218.

Vlachos, C., Papakonstantinou, K.G., Deodatis, G., 2018b. Structural applications of a predictive stochastic ground motion model: Assessment and use. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 4, 04018006.

Wang, Y., Stoica, P., Li, J., Marzetta, T.L., 2005. Nonparametric spectral analysis with missing data via the em algorithm. Digital signal processing 15, 191–206.

Welch, P., 1967. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Transactions on audio and electroacoustics 15, 70–73.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian processes for machine learning. volume 2. MIT press Cambridge, MA.

Yang, W., Ben-Zion, Y., 2010. An algorithm for detecting clipped waveforms and suggested correction procedures. Seismological Research Letters 81, 53–62.

Zeng, Y., Anderson, J.G., Yu, G., 1994. A composite source model for computing realistic synthetic strong ground motions. Geophysical Research Letters 21, 725–728.

Zhang, J., Hao, J., Zhao, X., Wang, S., Zhao, L., Wang, W., Yao, Z., 2016. Restoration of clipped seismic waveforms using projection onto convex sets method. Scientific reports 6, 1–10.

Zhang, Y., Comerford, L., Kougioumtzoglou, I.A., Patelli, E., Beer, M., 2017. Uncertainty quantification of power spectrum and spectral moments estimates subject to missing data. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 3, 04017020.