A Hybrid Probabilistic Framework for Model Validation with Application to Structural Dynamics Modeling

Subhayan De[a,*], Patrick T. Brewick[a,1], Erik A. Johnson[a], Steven F. Wojtkiewicz[b]

[a]*Sonny Astani Department of Civil and Environmental Engineering, University of Southern California, Los Angeles, California, USA 90089*
[b]*Department of Civil and Environmental Engineering, Clarkson University, Potsdam, New York, USA 13699*

Abstract

Identifying useful mathematical models of physical systems is an essential part of computational modeling and simulation. Once appropriate models are identified, they can be used for applications such as response prediction, structural control, monitoring structural integrity, lifetime prognosis, etc. The number of models and model classes available to the modeler to represent a physical phenomenon, however, can be very large. Retaining all available models throughout a study can be computationally burdensome, so the modeler has the significant problem of identifying the valid models to be used in further studies. To address this challenge, a probabilistic framework is proposed herein for validating models by intertwining the concepts of model falsification and Bayesian model selection. Model falsification, based

*Corresponding author

Email addresses: Subhayan.De@usc.edu (Subhayan De), patrick.brewick@gmail.com (Patrick T. Brewick), JohnsonE@usc.edu (Erik A. Johnson), swojtkie@clarkson.edu (Steven F. Wojtkiewicz)

[1]Present address: United States Naval Research Laboratory, Washington, DC, USA 20375.

on the philosophy that measurements can only be used to falsify models, is used in this framework in both pre- and postprocessing steps to eliminate models and model classes, respectively, that cannot explain the measurements. This is the first study to propose a framework to integrate these two paradigms. A likelihood-bound model falsification, previously introduced by the authors, determines the validity of the initial candidate model classes, using the false discovery rate (FDR), and removes most of the incorrect ones without incurring any significant additional computational burden. Next, Bayesian model selection, which assigns posterior model class probabilities based on Bayes' theorem, is applied to the remaining model classes to identify the the model(s) and model class(es) that provide predictions that probabilistically best fit the data. Finally, a postprocessing likelihood-bound falsification checks the validity of the final model class(es). The proposed framework is first illustrated through two nonlinear structural dynamics examples that show the efficacy of the proposed framework in identifying models for these structures as well as reducing the computational burden relative to Bayesian model selection applied alone. Finally, a third example uses measurement data from experiments performed on a full-scale four-story base-isolated building at the world's largest shake table in Japan's "E-Defense" laboratory.

Keywords: Model validation; model falsification; false discovery rate (FDR); Bayesian model selection.

---

1. Introduction

Even as computational capabilities continue to expand, facilitating increasingly intricate and complex simulations, the results obtained from those simulations will only be as accurate as the models used to represent real-world phenomena. In diverse areas of science and engineering, from dynamics to biomechanics, to fully harness advancing simulation power, significant gains must be similarly achieved in modeling. When attempting to model a complex dynamical system, the modeler (whether engineer or scientist, researcher or practitioner) is often presented with several types, or classes, of models that may be appropriate, but constraints on time and cost limit ensuing simulations and analysis to a single model class, or a select few, and sometimes only to a limited set of models within those model class(es). In structural dynamics, these models are used to design damping and control systems to mitigate the effects of natural hazards based on the outcome of different simulated scenarios [1]. A hysteretic element can be modeled using various linear and nonlinear models, but a modeler must choose a single model before designing an isolation layer using such devices. Changes in model parameters can be monitored to indicate when critical maintenance must be initiated, but these decisions first require identifying the model and/or model class of the system. In the area of fluid mechanics, particularly aerodynamics, modeling turbulence is a critical issue [2, 3] since the predictive performance of different models may vary based on problem specifics (*e.g.*, geometry, boundary conditions, etc.) that will ultimately affect the aerodynamic design. Climate models use different dynamic, thermodynamic, and biological process equations to quantify the interactions between the environ-

ment and its surroundings; the many models, of varying complexity, available for climate applications require a validation process be used before predicting future climates using these models.

Model selection refers to problems in which model(s) must be selected from a larger set, which may or (more commonly) may not include the *true* model [4]. The use of *Bayesian inference* to evaluate the plausibility of different models is known as *Bayesian model selection*, which has been applied across diverse fields [5–12]. A common form of Bayesian model selection uses the Bayes factor, which is a ratio of the marginal likelihood, or evidence, of two models [13, 14]. *Bayesian model class selection* [15] is Bayesian inference applied to quantify the likelihood of entire families of models, *i.e.*, model classes. This tool has been applied to structural models using dynamic response measurement data by Beck *et al.* [8–12]. The *Occam's razor* principle, which suggests that models with lesser complexity should be favored among models of comparable accuracy, is also embedded in Bayesian model selection, as discussed in Beck and Yuen [8], MacKay [16, 17], and Gull [18].

Another approach to model selection is *model falsification*, which is based on Popper's assertion first made in the 1930's that scientific models cannot be fully validated by data and can only be falsified [19]. In control theory, the concept has been used to design and select robust controllers that satisfy some performance criteria using measurement data by *model unfalsification* [20–22]. *Error domain model falsification* [23] was developed by Smith and Goulet and colleagues through a series of advances in structural modeling and monitoring [24, 25], with early applications quantifying the uncertainty and, subsequently, identifiability of models for civil structures, such as bridges and

pipe networks, based on characteristics of the measurement and monitoring systems, and linear dynamical systems using their natural frequencies [26–29]; the strength of this technique has been compared [30] to other methods, such as residual minimization and Bayesian inference. Beven and his colleagues [31–35] have proposed a different approach to model falsification by using the likelihood values of the measurement data in hydrological examples while defining the bounds in a subjective manner; however, they have been criticized for using empirical likelihood functions that do not represent a probability density for the residual errors [36]. De *et al.* [37] compared several approaches for model falsification and also proposed new ones that facilitate falsification for systems with many measurements distributed over multiple spatial and/or temporal domains.

The merits of both Bayesian model selection and model falsification are obvious, yet each method alone has its own inherent weaknesses. When all model classes in an initial set are inadequate representations of the system (an entirely plausible scenario, particularly for systems with high complexity or unknown features), Bayesian model selection will always choose a model class but without clear warning or indication to the modeler of its inadequacy. Consequently, future predictions produced by the inadequate model class may be highly inaccurate and, therefore, assumptions or decisions based on these predictions may have catastrophic consequences. While model falsification possesses the ability to eliminate incorrect models and inform the modeler if none of the options are valid, further judgment on the *usefulness* of a particular model class and parameter values are not possible in its current form [37].

Thus, this study proposes, and subsequently evaluates in numerical examples, a framework that selects one or more model(s) or model class(es) by integrating the principle of model falsification into Bayesian model selection to mitigate the weaknesses of these different identification schemes. The framework utilizes model falsification in a preprocessing step to eliminate models and model classes that do not fit the data to a statistical significance, and in a postprocessing step to ensure that the model classes evaluated via the central step of Bayesian model class selection are, on average, validated by the data as well. Exploiting model falsification's ability to significantly shrink the valid model class set will avoid numerous expensive computations required to evaluate the posterior parameter distribution for Bayesian model selection; these savings in computational cost will directly grow with increasing numbers of measurements, degrees of freedom (DOF), or space/time resolution. As Bayesian model selection already includes the effect of Occam's razor in evidence, no further extraneous steps are needed to penalize model classes with more parameters. Hence, the proposed framework not only identifies the most plausible model class and parameter estimates, it also requires fewer model simulations than other validation methods and provides checks on the suitability of the resulting models and model classes. Finally, the proposed framework extends Tarantola's comments [38] that the candidate models are from a prior distribution and the unfalsified models are from a *posterior distribution*; indeed, an alternate interpretation of the preprocessing falsification is as if one used a truncated density function for the prediction errors in a staged Bayesian selection. While the proposed framework remains true to his comments during the preprocessing step, the unfalsified models are further

processed to produce a posterior judgment between them as well as across different unfalsified model classes.

2. Hybrid Framework for Model Validation

Since Bayesian model selection leaves as most plausible a model that may be wrong, and model falsification can fail to falsify any model class, leaving a large number of unfalsified models, and provides no relative confidence in them, this section proposes the falsification of models in a Bayesian framework. Further, this falsification approach [37] is designed (unlike other falsification approaches) to accommodate many measurements such as those over multiple spatial and temporal dimensions from dynamical systems, both linear and nonlinear. The model validation framework proposed herein incorporates model falsification as both pre- and postprocessing steps before and after Bayesian model selection to combine the usefulness of both approaches while overcoming their individual shortcomings. These first and last steps of the framework use a form of model falsification developed by the authors [37]. A flowchart of the steps of the proposed validation procedure is given in Figure 1. The flowchart shows that the preprocessing falsification step first eliminates the model classes that do not reproduce reasonably well the responses of the physical system, thereby shrinking the set of candidate model classes. The next step is to implement Bayesian model selection with the remaining model classes by re-using computational results from the preprocessing step. Finally, a postprocessing step checks the validity, on average, of the final model class(es).

Clearly, just as with all other modeling and system identification tools,

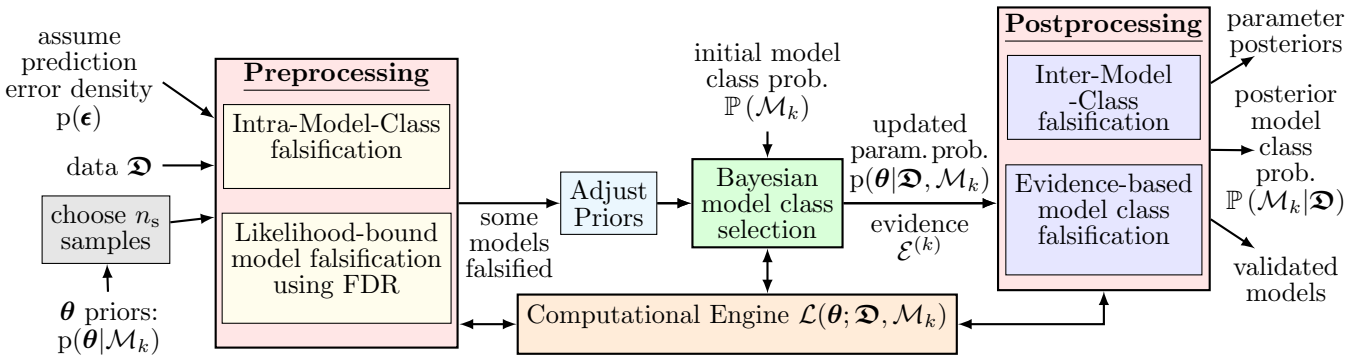


Figure 1: Proposed synergistic framework of model falsification and Bayesian model class selection.

this framework requires the modeler to ensure that the measurement data is sufficiently rich to exercise the ranges of system behavior for which the resulting validated models will be used. For example, if the models of a nonlinear system are validated with the proposed framework (or calibrated with any similar approach) using data that insufficiently exercises the nonlinear behavior, linear models will be preferred; if the resulting models will only be used for predictions of the system response when it is behaving essentially linearly, then those models are sufficient; on the other hand, if the model will be used to study and predict response over the entire breadth of the system behaviors, then the modeler is responsible for ensuring that the validation measurement data spans those behaviors.

Some other researchers have proposed different types of model validation frameworks. For example, a validation approach proposed by Babuška *et al.* [39] employs a model rejection step using a validation dataset, but only after fitting probability distributions to all model class parameters using a separate calibration dataset. Recently, Farrell *et al.* [40–42] proposed a framework for model validation that first selects a subset of candidate model

classes and calculates the calibrated posterior model parameter distribution, the evidence values and, finally, the validated posterior parameter distribution (and probability distribution for a quantity of interest) using Monte Carlo approaches; however, that framework is unable to find model classes with higher evidence values but with more parameters because it starts with the model class subsets that have the fewest parameters, stopping when it finds a valid model class, and does not investigate further.

2.1. Intra-Model-Class Falsification: Framework's Preprocessing Step

A model from a model class $\mathcal{M}_k$, which is in the set $\mathscr{M}$ of all candidate model classes, is specified by an $n_{\boldsymbol{\theta}} \times 1$ parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. (Technically, $\boldsymbol{\theta}$ should be written $\boldsymbol{\theta}^{(k)}$ since its size may be different for different model classes; however, the superscript (k) is omitted for notational simplicity.) Herein, $\boldsymbol{\theta}$ will be called a *model* as its value defines one model within the corresponding model class. The difference between the N_o outputs $\mathbf{h}(\boldsymbol{\theta})$ of the model and their corresponding measurements $\mathbf{d}$ is known as the residual error $\boldsymbol{\epsilon}$ [37]. These residuals are modeled as continuous random variables, herein characterized by the probability density function $\mathrm{p}_{\mathbf{E}}(\mathbf{e}|\boldsymbol{\theta})$, where $\mathbf{E}$ is a random vector, and $\mathbf{e}$ is a possible value of random vector $\mathbf{E}$, whereas $\boldsymbol{\epsilon}$ is the actual residual error. (The random vector $\mathbf{E}$ is henceforth omitted for brevity but is implied by context.)

Conventional error domain model falsification sets bounds on each error residual $\epsilon_i = h_i(\boldsymbol{\theta}) - d_i$; the bounds are chosen so that each residual has a given probability, defined by the model's assumptions on $\mathrm{p}(\mathbf{e}|\boldsymbol{\theta})$, of remaining within the bounds. This approach ignores significant information that may be available about the residuals, their uncertainty distributions and their

correlations. Instead, following De *et al.* [37], the proposed framework's first step is to falsify models based on the *likelihood* of their residual; in the case of a single measurement, the likelihood (omitting the model class $\mathcal{M}_k$) is

$$\mathcal{L}(\boldsymbol{\theta};\mathfrak{D}) = \mathrm{p}(\mathfrak{D}|\boldsymbol{\theta}) = \mathrm{p}(\epsilon|\boldsymbol{\theta}) = \mathrm{p}\Big([h(\boldsymbol{\theta}) - d]\Big|\boldsymbol{\theta}\Big) \tag{1}$$

The model would then be *unfalsified* (*i.e.*, accepted) if the likelihood is larger than some likelihood threshold, denoted $\underline{\mathcal{L}}$ and computed as described in a subsequent section, and *falsified* if below the threshold:

$$\mathcal{L}(\boldsymbol{\theta};\mathfrak{D}) < \underline{\mathcal{L}} \quad \Rightarrow \quad \text{falsify model } \boldsymbol{\theta} \tag{2}$$

For an arbitrary number of measurements, including time history responses of dynamical systems, the predicted outputs of model $\mathcal{M}_k$ are given again by $\mathbf{h}(\boldsymbol{\theta})$ but in a stacked form (*i.e.*, for predicted response time history vector $\mathbf{y}(t;\boldsymbol{\theta})$, sampled at times $\{t_0, t_1, \cdots\}$, let $\mathbf{h}(\boldsymbol{\theta}) = [\mathbf{y}^{\mathrm{T}}(t_0;\boldsymbol{\theta})\ \ \mathbf{y}^{\mathrm{T}}(t_1;\boldsymbol{\theta})\ \ \cdots]^{\mathrm{T}}$ or a similar arrangement). The $N_{\mathrm{o}} \times 1$ residual vector is $\boldsymbol{\epsilon} = \mathbf{h}(\boldsymbol{\theta}) - \mathbf{d}$. If the resulting residuals are zero-mean Gaussian distributed with covariance matrix $\boldsymbol{\Sigma}$, then the likelihood can be written

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta};\mathfrak{D}) = \mathrm{p}\Big(\mathbf{h}(\boldsymbol{\theta}) - \mathbf{d}\Big|\boldsymbol{\theta}\Big) &= \frac{\exp\left(-\frac{1}{2}[\mathbf{h}(\boldsymbol{\theta}) - \mathbf{d}]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}[\mathbf{h}(\boldsymbol{\theta}) - \mathbf{d}]\right)}{(2\pi)^{\frac{N_{\mathrm{o}}}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \\ &= \frac{\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}\right)}{(2\pi)^{N_{\mathrm{o}}/2}|\boldsymbol{\Sigma}|^{1/2}} \end{aligned} \tag{3}$$

This approach directly frames model uncertainty in terms of the dynamic response measurements with two advantages: (i) avoiding the intermediate

use of modal parameters helps limit error introduced via system identification methods and modal analysis; and (*ii*) the likelihood calculated for models in a class can be used to subsequently compute the model class evidence, an integral step in Bayesian model selection.

Figure 2 show the typical lower and upper bounds ($\underline{\epsilon}_i$ and $\bar{\epsilon}_i$) used by model falsification for a two-data-point case; for a symmetric, unimodal probability density of the residual error such as that shown, these bounds have the same likelihood $\underline{\mathcal{L}}$, though this is not the case for the more general residual densities, such as in Figure 3, where a likelihood threshold (chosen as discussed in subsequent paragraphs) would shift the residual error bounds to the area(s) of highest likelihood. Further, a likelihood threshold criteria can accommodate more general residual densities (skewed or multimodal), and allow the modeler to incorporate the correlation structure of multiple

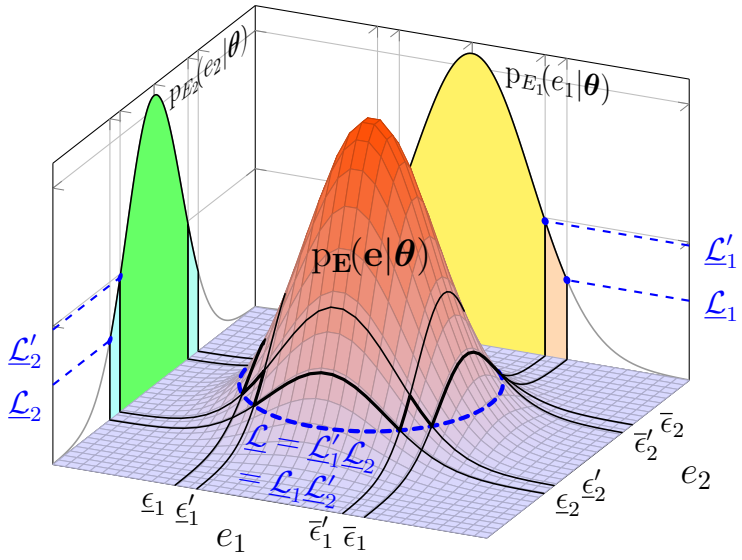


Figure 2: Likelihood bound for the two-measurement case [37]. Given a target identification probability $\phi = 1 - \alpha$, the significance levels $\bar{\alpha}_1 = \alpha/2$ and $\bar{\alpha}_2 = \alpha$ are used to compute $\underline{\epsilon}_j$, $\bar{\epsilon}_j$ and $\underline{\epsilon}'_j$, $\bar{\epsilon}'_j$, respectively. See Appendix A for details of this procedure.

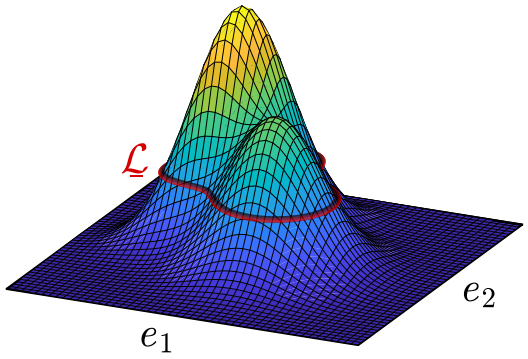


Figure 3: A likelihood-bound $\underline{\mathcal{L}}$ defined for a multidimensional non-Gaussian probability density of residual error.

measurement residuals.

Methods to calculate bounds on likelihood values

De *et al.* [37] discussed various model falsification approaches to determine the likelihood lower bound $\underline{\mathcal{L}}$ based on the false discovery rate (FDR) correction with the Benjamini-Hotchberg (BH) procedure [37, 43]. In multiple comparison problems, FDR control ensures that, on average, the number of false positives among all rejections will be below a specified value, and provides a better statistical power while allowing some false positive results [44]. (The statistical power, defined as the probability of rejecting a model when it is invalid, is related to probability of making a type II error, which is the error introduced by incorrectly accepting invalid models.) Herein, the BH procedure is used, as summarized in Appendix A, to compute the residual error bounds $[\underline{\epsilon}_i, \bar{\epsilon}_i]$ from the target identification probability ϕ (generally assumed as 0.95 or 0.90). The significance level $\alpha = 1 - \phi$ of the overall falsification step is the probability of rejecting a model when it is, in fact, valid. In short, the BH procedure first ranks the residual errors according to

their p-values (the probability of exceeding the residual error in magnitude) and compares them against their corresponding per-measurement significance levels $\bar{\alpha}_i$ (which would $\bar{\alpha}_1 = \alpha/2$ and $\bar{\alpha}_2 = \alpha$ for a two-measurement case such as depicted in Figure 2). Then, residual errors falling inside the limits are accepted; for the two-measurement case, the limits are shown as the solid thick lines on the joint probability density of residual errors in Figure 2 ($[\underline{\epsilon}_1, \overline{\epsilon}_1] \times [\underline{\epsilon}'_2, \overline{\epsilon}'_2]$ if the residual error ϵ_1 is less likely than the second ϵ_2, and $[\underline{\epsilon}'_1, \overline{\epsilon}'_1] \times [\underline{\epsilon}_2, \overline{\epsilon}_2]$ otherwise). These bounds are used next to compute $\underline{\mathcal{L}}$ using

$$\underline{\mathcal{L}} = \prod_{i=1}^{N_o} \min_{\underline{\epsilon}_i \le e_i \le \overline{\epsilon}_i} \mathrm{p}(e_i|\boldsymbol{\theta}) \tag{4}$$

(As discussed in a prior study [37], there may be other ways to use the bounds $[\underline{\epsilon}_i, \overline{\epsilon}_i]$ or other information to decide the lower bound $\underline{\mathcal{L}}$.)

The modeler must exercise judgement in choosing the number of candidate models from each model class. This number, which is problem dependent, should be sufficiently large to explore the parameter space and its prior, but also require reasonable computational resources. One approach for evaluating whether the number of candidate models is sufficient is to gradually (linearly or exponentially) increase the number of models for several iterations, stopping when the fraction of unfalsified models from the model class remains stable. Here, the initial number of models can be decided following the guidelines in Goulet and Smith [28].

When all models evaluated from a model class are falsified, the entire model class is considered falsified. $\mathscr{M}^1$, the set of model classes that passes this preprocessing step, is used as the possible model class set for the subse-

quent step of Bayesian model selection. (If $\mathscr{M}^1$ is empty, then the selection of the initial set of model classes must be reconsidered.) By shrinking the set of candidate model classes, this preprocessing step can achieve significant computational savings in the Bayesian model selection, which requires complete exploration of the high likelihood region, a computationally expensive task (discussed in the next subsection), whereas a satisfactory coverage of the high likelihood region is sufficient for model falsification. (It should also be noted that it is possible for a model class with all models falsified to occur because of an insufficient number of models sampled from that class — *e.g.*, any sampling-based approach is challenged when exploring a high-dimensional parameter space — or if the priors are not chosen appropriately. The modeler may choose to revisit and verify assumptions when a model class is eliminated.)

2.2. Bayesian model class selection

With the remaining model classes in $\mathscr{M}^1$, the *posterior* model class probabilities (*i.e.*, the probability of a model class conditioned on measurement data $\mathfrak{D}$) are given by Bayes' theorem:

$$\mathbb{P}(\mathcal{M}_k|\mathfrak{D}) = \frac{\mathrm{p}(\mathfrak{D}|\mathcal{M}_k)\mathbb{P}(\mathcal{M}_k)}{\mathrm{p}(\mathfrak{D})}, \quad \mathcal{M}_k \in \mathscr{M}^1 \tag{5}$$

where the probability of an event is denoted by $\mathbb{P}(\cdot)$; $\mathbb{P}(\mathcal{M}_k)$ is an *a priori* measure of model plausibility assigned by the modeler based on past experience, normalized so $\sum_{\mathcal{M}_k \in \mathscr{M}^1} \mathbb{P}(\mathcal{M}_k) = 1$; and denominator $\mathrm{p}(\mathfrak{D}) = \sum_{\mathcal{M}_k \in \mathscr{M}^1} \mathrm{p}(\mathfrak{D}|\mathcal{M}_k)\mathbb{P}(\mathcal{M}_k)$ using the theorem of total probability. For a par-

ticular model class $\mathcal{M}_k$, the model evidence $\mathcal{E}^{(k)} = \mathrm{p}(\mathfrak{D}|\mathcal{M}_k)$ is

$$\begin{aligned} \mathcal{E}^{(k)} &= \int_{\boldsymbol{\Theta}} \mathrm{p}(\mathfrak{D}|\boldsymbol{\theta}, \mathcal{M}_k)\mathrm{p}(\boldsymbol{\theta}|\mathcal{M}_k)\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta}; \mathfrak{D}, \mathcal{M}_k)\mathrm{p}(\boldsymbol{\theta}|\mathcal{M}_k)\mathrm{d}\boldsymbol{\theta} \end{aligned} \tag{6}$$

where $\mathrm{p}(\boldsymbol{\theta}|\mathcal{M}_k)$ is the prior probability, based on modeler judgment, of parameter vector $\boldsymbol{\theta}$ for model class $\mathcal{M}_k$; and $\mathcal{L}(\boldsymbol{\theta}; \mathfrak{D}, \mathcal{M}_k) = \mathrm{p}(\mathfrak{D}|\boldsymbol{\theta}, \mathcal{M}_k)$ is the likelihood function already computed in Section 2.1. These evidence values $\mathcal{E}^{(k)}$ are required to evaluate the posterior model class probabilities $\mathbb{P}(\mathcal{M}_k|\mathfrak{D})$, which can then be used to select a model class, or to be used as weights for multiple validated model classes. At the end of this step, based on the posterior model class probabilities, a smaller model class set $\mathscr{M}^2 \subseteq \mathscr{M}^1$, typically only the one model class with the highest $\mathbb{P}(\mathcal{M}_k|\mathfrak{D})$, is retained and will be subjected to the final postprocessing falsification step.

The assumption that the true model class is among the candidate model classes may not always be true. In such scenarios, the application of Bayesian model class selection alone may lead to erroneous conclusions; the framework's preprocessing falsification step avoids this error and ensures a meaningful result from the Bayesian model selection. Note: The unfalsified models' parameters can be used to formulate informative priors for the model selection, though the question of using the data twice can be raised. In that case, the dataset is divided in two datasets $\mathfrak{D}^{\mathrm{I}}$ and $\mathfrak{D}^{\mathrm{II}}$, where the falsification is done with $\mathfrak{D}^{\mathrm{I}}$. The prior of the model parameters is then formed using the unfalsified parameters for the model selection, which is performed with $\mathfrak{D}^{\mathrm{II}}$.

Sampling Algorithms

For efficient evaluation of the evidence, a notable computational challenge for applications of Bayesian model selection, several methods have been proposed, including the posterior harmonic mean estimator [45], importance sampling [13], nested sampling technique [46], annealed sampling [47], the Power Posterior method [48], the Transitional Markov Chain Monte Carlo method [49], the Monte Carlo splitting and subset methods [50–53], stochastic collocation [54], and polynomial chaos approaches [55]. For the numerical examples herein, a nested sampling algorithm is used to evaluate the evidence values. A brief description of this nested sampling algorithm is provided in Appendix B. After several iterations of the nested sampling algorithm, sampling from high likelihood region becomes difficult, especially if the high likelihood region is concentrated within a very small region. In such a case, Skilling [46] suggested using the Markov Chain Monte Carlo (MCMC) algorithm for generating samples from the prior constrained to high likelihood region. For the examples herein, the nested sampling uses conventional random sampling with rejection but switches, when the average acceptance rate drops below 5%, to an MCMC augmented by the modified Metropolis algorithm, introduced in Au and Beck [56] (see Appendix C), which will provide a high acceptance rate to efficiently sample from the high likelihood region.

2.3. Inter-Model-Class Falsification: Framework's Postprocessing Step

A postprocessing falsification procedure, following the Bayesian model class selection, is proposed here to provide a robust framework that is capable of indicating whether the model class(es) in the current pool $\mathscr{M}^2$ are valid in terms of their evidence value(s).

The evidence $\mathcal{E}^{(k)}$ for model class $\mathcal{M}_k$, which is the likelihood (6) of the data given a model class, can be written as the expected likelihood of the residuals; *i.e.*, $\mathcal{E}^{(k)} = \mathrm{p}(\boldsymbol{\epsilon}|\mathcal{M}_k) = \mathbb{E}_{\boldsymbol{\Theta}}\left[\mathcal{L}(\boldsymbol{\theta}; \mathfrak{D}, \mathcal{M}_k)\right]$. Similar to the likelihood threshold $\underline{\mathcal{L}}$ for a given model in the preprocessing falsification, an entire model class can be falsified if the model class evidence $\mathcal{E}^{(k)}$ is below a threshold $\underline{\mathcal{E}}$:

$$\mathcal{E}^{(k)} < \underline{\mathcal{E}} \quad \Rightarrow \quad \text{falsify model class } \mathcal{M}_k \tag{7}$$

As the evidence $\mathcal{E}^{(k)}$ has already been computed in the Bayesian model class selection, this proposed postprocessing falsification can be performed with no extra calculation. While different choices can be made for evidence lower bound $\underline{\mathcal{E}}$, it is taken herein to be equal to the likelihood threshold $\underline{\mathcal{L}}$ used in the preprocessing step because the evidence $\mathcal{E}$ is nothing but the expected value of the likelihood function $\mathcal{L}$; since the likelihood threshold $\underline{\mathcal{L}}$ has already been computed, the choice $\underline{\mathcal{E}} = \underline{\mathcal{L}}$ has the further convenience of not requiring additional computation. Another evidence threshold that could be used without further computation is $\underline{\mathcal{E}} = \gamma \left[\max_k \max_i \mathcal{L}(\boldsymbol{\theta}_i; \mathfrak{D}, \mathcal{M}_k)\right]$ with $\gamma \in (0, 1)$ depending on the modeler's expectation about the average likelihood value in a model class.

This postprocessing falsification is particularly beneficial in identifying the cases in which some few models remain from an entire model class after the preprocessing falsification but, on average, the model class is incorrect as well. The result, then, is a set $\mathcal{M}^3 = \{\mathcal{E}^{(k)} < \underline{\mathcal{E}} : \mathcal{M}_k \in \mathscr{M}^2\}$ of final, validated model classes.

2.4. Computational Advantage of the Synergistic Framework

The proposed approach provides significant computational savings when some competing model classes are excluded using the preprocessing step, thereby eliminating the most computationally expensive step of calculating evidence values for those model classes, as this step generally requires many more model evaluations. The second reduction in computational effort is obtained since the model likelihoods computed in the preprocessing falsification can be used as initial starting values for (nested) sampling. Third, the postprocessing model class falsification relies on evidence already computed in the Bayesian model class evaluation, so does not incur additional cost. A metric C for the computational savings can be introduced as

$$C = \frac{\sum_{k \notin \mathscr{M}^1} (N_{2,k} - N_{1,k})}{\sum_{k \in \mathscr{M}} N_{2,k}} \times 100\%$$

where $N_{1,k}$ and $N_{2,k}$ are number of model evaluations in the preprocessing falsification and Bayesian model selection steps, respectively, for the k^{th} model class. Generally $N_{2,k} \gg N_{1,k}$ because, in evidence estimation, a complete exploration of the high likelihood region must be performed (as shown in the two numerical illustrations given in the next section), leading to significant computational savings. Also, if some of the model classes are nonlinear and fail to pass the preprocessing step, additional savings in computation time can be expected, compared to falsification of some linear model classes, due to the additional computational complexity associated with the response calculation of a nonlinear model.

3. Numerical Illustrations

The proposed framework is illustrated using three numerical examples from structural dynamics. In each of these examples, several model classes are used to describe the system's nonlinear behaviors. The proposed framework is shown to systematically eliminate the incorrect candidate model classes to choose and validate the model class that probabilistically best fits the measurement data.

3.1. Example I: 3DOF Model with Nonlinear Stiffnesses

Consider the three DOF model shown in Figure 4 subjected to support acceleration $\ddot{x}_{\mathrm{s}}$. The equations of motion of the structure are given by

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{C}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} + \mathbf{L}\mathbf{g}(\mathbf{x}) = -\mathbf{M}\mathbf{1}_{3\times 1}\ddot{x}_{\mathrm{g}} \tag{8}$$

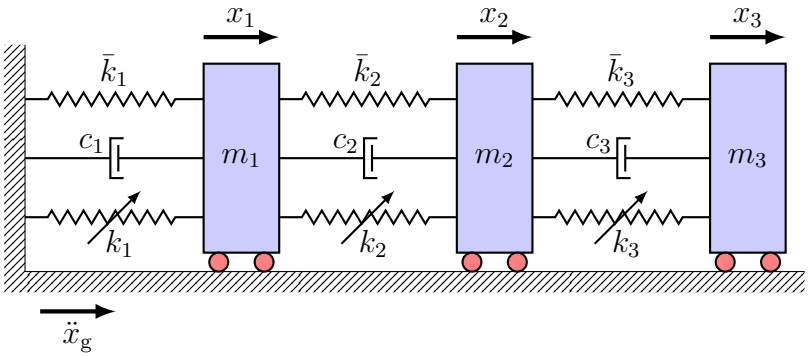


Figure 4: 3 DOF model with nonlinear stiffnesses.

where the mass matrix is

$$\mathbf{M} = \begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \tag{9}$$

with $m_1 = m_2 = m_3 = 300$ Mg; the nominal stiffness matrix is

$$\mathbf{K} = \begin{bmatrix} \bar{k}_1 + \bar{k}_2 & -\bar{k}_2 & 0 \\ -\bar{k}_2 & \bar{k}_2 + \bar{k}_3 & -\bar{k}_3 \\ 0 & -\bar{k}_3 & \bar{k}_3 \end{bmatrix} \tag{10}$$

with $\bar{k}_1 = \bar{k}_2 = \bar{k}_3 = 8$ MN/m; $\mathbf{g}(\mathbf{x})$ is a (possibly nonlinear) restoring force function with its influence matrix

$$\mathbf{L} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \tag{11}$$

$\mathbf{1}_{3\times 1}$ is a column vector of ones; and the mass displacement vector relative to the support is $\mathbf{x} = [x_1 \quad x_2 \quad x_3]^{\mathrm{T}}$. Rayleigh damping, *i.e.*, $\mathbf{C} = \beta_1\mathbf{M} + \beta_2\mathbf{K}$, is assumed with 3% damping for each of the first two modes, where β_1 and β_2 are constants evaluated from the damping ratios and the nominal natural frequencies.

3.1.1. Candidate Model Classes

Three possibilities for each element of $\mathbf{g}(\mathbf{x})$ are assumed — linear, quadratic and cubic stiffnesses — for a total of $3 \times 3 \times 3 = 27$ different model classes:

Table 1: Means and standard deviations of lognormal prior distributions for model class stiffness coefficients (units are MN/m^{p_i}).

Story	Stiffness	linear ($p_i = 1$)		quadratic ($p_i = 2$)		cubic ($p_i = 3$)	
i	coeff.	mean	σ	mean	σ	mean	σ
1	k_1	25	2.5	250	25	2500	250
2	k_2	0.18	0.025	1.8	0.25	18	2.5
3	k_3	1.8	0.25	18	2.5	180	25

i.e., $g_i = k_i \, |x_i - x_{i-1}|^{p_i} \operatorname{sgn}(x_i - x_{i-1})$, where $x_0 \equiv 0$ and $p_i \in \{1, 2, 3\}$ for $i = 1, 2, 3$. These model classes are denoted by concatenating the exponent values $p_1 - p_2 - p_3$; *e.g.*, combination $1 - 2 - 3$ has the vector $\mathbf{g}(\mathbf{x}) = [k_1 x_1 \quad k_2 |x_2 - x_1|^2 \operatorname{sgn}(x_2 - x_1) \quad k_3 (x_3 - x_2)^3]^{\mathrm{T}}$. The prior distributions for parameters k_1, k_2, and k_3 are assumed to be lognormal with means and standard deviations as shown in Table 1. These choices of model parameters for the different exponents are chosen so that the force levels are similar with different exponents given the responses of this system (*i.e.*, the interstory drifts are on the order of 0.1 m, so the coefficients in the linear, quadratic and cubic stiffness models have ratios of $1 : 10 : 100$).

3.1.2. True Model Class and the Measurement Data

The measurement data $\mathfrak{D}$, containing the time history $x_1(t)$ of the first mass displacement relative to the support, is generated using (8) for model class $1-3-2$ with $\mathbf{g}(\mathbf{x}) = [k_1 x_1, \; k_2 (x_2 - x_1)^3, \; k_3 |x_3 - x_2|^2 \operatorname{sgn}(x_3 - x_2)]^{\mathrm{T}}$ — *i.e.*, $p_1 = 1$, $p_2 = 3$ and $p_3 = 2$ — with stiffness coefficients $k_1 = 22.5\,\mathrm{MN/m}$, $k_2 = 20.0\,\mathrm{MN/m^3}$ and $k_3 = 20.0\,\mathrm{MN/m^2}$. The measurements are sampled at 20 Hz and include additive Gaussian pulse process sensor noise, in which the sensor noises at every time step are independent and identically-distributed zero-mean Gaussian random variables with a standard deviation that is 20%

of the noise-free RMS response. The support excitation $\ddot{x}_g(t)$ is the first 30 s of the N-S component of the 18 May 1940 Imperial Valley earthquake recorded at the Imperial Valley Irrigation District substation in El Centro, California, sampled at 50 Hz, with peak acceleration $3.42\,\mathrm{m/s^2}$.

3.1.3. Preprocessing: Intra-Model-Class Falsification

Model falsification is performed with probability density $\mathrm{p}(\epsilon_i)$ of each residual error assumed to be independent zero-mean Gaussian with a standard deviation σ that is 20% of the RMS of the measured first-mass displacement and with the target identification probability set at $\phi = 0.95$ resulting in an FDR-based likelihood bound of $\underline{\mathcal{L}} = 266.4$, computed using (4) and Appendix A, that will be used to falsify models. 1000 models are randomly generated for each model class using the prior parameter Gaussian distribution statistics listed in Table 1. Out of the 27 possible combinations of different nonlinear stiffnesses of the three stories, only nine remain after the preprocessing step of intra-model class falsification: $\mathscr{M}^1 = \{1 - i - j \mid i \in \{1,2,3\}, j \in \{1,2,3\}\}$, which means that the nonlinear models of the first spring are clearly unsuitable. The fractions of models unfalsified in each of the 27 model classes are shown in Table 2. The result of this step also shows the limitation of using model falsification alone as all model classes in $\mathscr{M}^1$ have almost the same fraction of unfalsified models.

3.1.4. Bayesian Model Selection

Bayesian model selection is performed next for only the nine model classes that passed the preprocessing falsification step, with prior model class probabilities set as $\mathbb{P}(\mathcal{M}_k) = 1/9$. The posterior model class probabilities are given

Table 2: Unfalsified models using proposed intra-model class falsification. (Bold means unfalsified model classes.)

Model class ($\mathcal{M}_k$)	Unfalsified (%)	Model class ($\mathcal{M}_k$)	Unfalsified (%)	Model class ($\mathcal{M}_k$)	Unfalsified (%)
$\mathbf{1-1-1}$	**54.9**	$2-1-1$	0.0	$3-1-1$	0.0
$\mathbf{1-1-2}$	**63.9**	$2-1-2$	0.0	$3-1-2$	0.0
$\mathbf{1-1-3}$	**65.5**	$2-1-3$	0.0	$3-1-3$	0.0
$\mathbf{1-2-1}$	**53.8**	$2-2-1$	0.0	$3-2-1$	0.0
$\mathbf{1-2-2}$	**63.4**	$2-2-2$	0.0	$3-2-2$	0.0
$\mathbf{1-2-3}$	**64.7**	$2-2-3$	0.0	$3-2-3$	0.0
$\mathbf{1-3-1}$	**50.5**	$2-3-1$	0.0	$3-3-1$	0.0
$\mathbf{1-3-2}$	**61.4**	$2-3-2$	0.0	$3-3-2$	0.0
$\mathbf{1-3-3}$	**63.4**	$2-3-3$	0.0	$3-3-3$	0.0

Table 3: Posterior model class probabilities after most of the model classes are rejected using a preprocessing step of intra-model-class falsification. (Relative log-evidence is with respect to the model class with the largest log-evidence.)

Model class ($\mathcal{M}_k$)	log(Evidence)	rel. log(Evidence)	$\mathbb{P}(\mathcal{M}_k\|\mathfrak{D})$
$1-1-1$	2290.7	-72.1	≈ 0.0
$1-1-2$	2315.5	-47.3	≈ 0.0
$1-1-3$	2309.7	-53.1	≈ 0.0
$1-2-1$	2317.6	-45.2	≈ 0.0
$1-2-2$	2351.0	-11.8	≈ 0.0
$1-2-3$	2334.3	-28.5	≈ 0.0
$1-3-1$	2338.0	-24.8	≈ 0.0
$\mathbf{1-3-2}$	2362.8	0.0	1.0
$1-3-3$	2334.2	-28.6	≈ 0.0

in Table 3, which shows that the model class $1-3-2$ has the maximum posterior model class probability of essentially 1.0; *i.e.*, $\mathscr{M}^2 = \{1-3-2\}$ and $\mathbb{P}(1-3-2|\mathfrak{D}) \approx 1.0$. The means and the standard deviations of the posterior distribution of the stiffness parameters are obtained from the nested sampling algorithm and are shown in Table 4 along with their true values.

Table 4: Posterior mean and standard deviation of model parameters for model class $1-3-2$, and the true values.

Stiffness coeff.	True value	Posterior	
		mean	Std. dev.
k_1 [MN/m]	22.5	22.5224	0.0630
k_2 [MN/m^3]	20.0	20.7529	0.9928
k_3 [MN/m^2]	20.0	19.3070	0.5485

In this numerical example, due to the intra-model class falsification step, 18 out of 27 model classes are entirely falsified, requiring no evidence calculation for them. As $N_{2,k} \gg N_{1,k}$ for all model classes, the computational savings here is given by $C \approx 66.67\%$.

3.1.5. Postprocessing: Inter-Model-Class Falsification

Using the likelihood-bound falsification approach with the FDR/BH procedure, the lower limit for the model class evidence for validation is again $\log \underline{\mathcal{E}} = \log \underline{\mathcal{L}} = 266.4$. For model class $1-3-2$, $\log \mathcal{E}^{(k)} = 2362.8 > \log \underline{\mathcal{E}} = 266.4$, *i.e.*, on average this model class also passes the postprocessing falsification step. Note that, with the current choice of $\underline{\mathcal{E}}$, all of the model classes that pass through the preprocessing step will also pass through the postprocessing step. However, if other model classes, *e.g.*, $2-1-1$ or $2-3-2$ somehow pass through the preprocessing, then this choice of $\underline{\mathcal{E}}$ would stop them since $\log \mathcal{E}^{(2-1-1)} = -2148.48 < \log \underline{\mathcal{E}}$ and $\log \mathcal{E}^{(2-3-2)} = -2899.43 < \log \underline{\mathcal{E}}$.

3.1.6. Variations on the Method

One way to improve the validity of the model classes in $\mathscr{M}^1$, is to adjust their prior parameter distributions based on those retained after preprocessing falsification, using insights gained by applying the likelihood-bound falsification. This optional step, shown after the preprocessing block in Fig-

ure 1, is implemented here; the measurement data is divided in to two sets: $\mathfrak{D}^{\mathrm{I}}$ consisting of every odd numbered measurement of the response and $\mathfrak{D}^{\mathrm{II}}$ consisting every even numbered measurement of the response. (Other simple divisions of the data in the time domain can be used but, for a response to a historical earthquake record that is nonstationary, this division of data is chosen to ensure that the nonlinear behavior of the system is pronounced in both datasets; a detailed discussion on dividing $\mathfrak{D}$ into $\mathfrak{D}^{\mathrm{I}}$ and $\mathfrak{D}^{\mathrm{II}}$ is beyond the scope of this study.) The results from the preprocessing falsification step using $\mathfrak{D}^{\mathrm{I}}$ again give the same nine unfalsified model classes as before. For each of these nine model classes, the priors (which were originally lognormal) are adjusted by fitting new lognormal distributions to the unfalsified models' parameters using maximum likelihood estimation, and are then used as priors for the next step of Bayesian model selection. For example, for the model class $1-2-2$, the prior for k_1 is changed to lognormal with mean $23.59\,\mathrm{MN/m}$ and standard deviation $1.47\,\mathrm{MN/m}$; *i.e.*, the mean is closer to the *true* value and the standard deviation is about half that of the original prior. The Bayesian model selection is performed with $\mathfrak{D}^{\mathrm{II}}$ for the nine model classes that pass the preprocessing step with adjusted priors for the model parameters. The results for two best model classes are shown in Table 5, which shows that the proposed framework is able to again find the correct model class. Table 5 also shows that using $\mathfrak{D}^{\mathrm{I}}$ to inform the priors makes model class $1-2-2$ more competitive with true model class $1-3-2$, as evidenced their relative log-evidence values reducing from a difference of 11.3 to only 3.9, though the posterior probability of $1-3-2$ is still near unity. Hence, this modification in the proposed framework is very useful in making

Table 5: Posterior model class probabilities of the two best model classes after some model classes are rejected using preprocessing intra-model-class falsification and an adjustment of prior parameter distribution for model selection.

Model class ($\mathcal{M}_k$)	log(Evidence)	rel. log(Evidence)	$\mathbb{P}(\mathcal{M}_k \mid \mathfrak{D})$
$1-2-2$	1169.3	-4.6	0.01
$\mathbf{1-3-2}$	1173.9	0.0	0.99

the candidate model classes competitive when they are very close to each other in their respective behavior.

A second variation is to adjust the choice of the target identification probability ϕ, changed here from 0.95 to 0.90, which gives fewer unfalsified models but still nine unfalsified model classes after the preprocessing intra-model-class falsification step, as shown in Table 6. Using these nine unfalsified model classes, the Bayesian model class selection and postprocessing again chooses the $1-3-2$ model class as the valid model class. This shows that adjusting ϕ within conventional ranges does not significantly affect the result (the effect of other values of ϕ is beyond the scope of this paper).

Table 6: Unfalsified models using proposed intra-model class falsification with target identification probability $\phi = 0.90$.

Model class ($\mathcal{M}_k$)	Unfalsified (%)
$1-1-1$	44.9
$1-1-2$	55.3
$1-1-3$	55.5
$1-2-1$	44.5
$1-2-2$	55.2
$1-2-3$	54.8
$1-3-1$	40.4
$1-3-2$	53.3
$1-3-3$	53.3

3.2. Example II: 4DOF model with Hysteretic Isolation layer

Consider the base-isolated building or mechanical equipment models shown in Figure 5. The isolation often exhibits hysteretic behavior, introducing nonlinearity into an otherwise linear dynamical system. To simulate the system response to various classes of inputs, the behavior of the hysteretic elements in the isolation layer must be accurately modeled. The true isolation model here is an elastoplastic element (*e.g.*, a lead-rubber bearing in building isolation) with modest linear viscous damping. This four DOF system is subjected to base excitation $\ddot{x}_g$, a stationary filtered white noise generated using

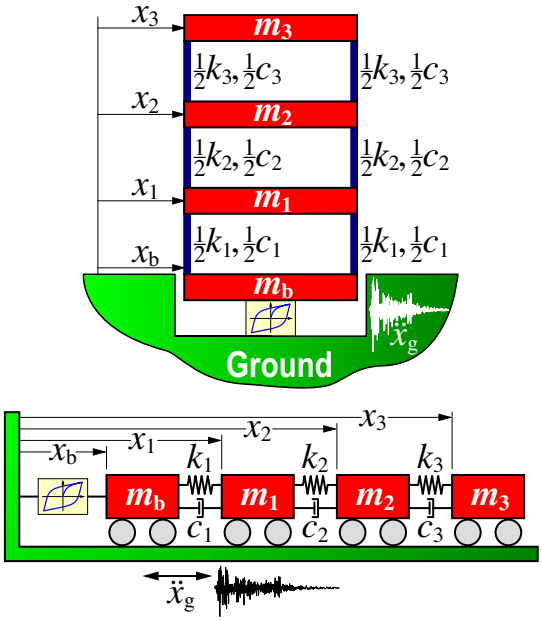


Figure 5: 4DOF models

a Kanai-Tajimi filter [57] with spectral density

$$S_{\ddot{x}_g \ddot{x}_g}(\omega) = \frac{S_0 \left(4\zeta_g^2 \omega_g^2 \omega^2 + \omega_g^4\right)}{\left(\omega^2 - \omega_g^2\right)^2 + 4\zeta_g^2 \omega_g^2 \omega^2} \tag{12}$$

where $\omega_g = 17\,\mathrm{rad/s}$ and $\zeta_g = 0.3$ are assumed following Ramallo *et al.* [58], and the spectral intensity S_0 is calculated using

$$S_0 = \sigma_w^2 \frac{0.03\zeta_g}{\pi\omega_g \left(4\zeta_g^2 + 1\right)} g^2 \tag{13}$$

where g is the gravitational acceleration. The constant $\sigma_w = 2$ is selected such that the nonlinearity in the system response is pronounced but not so large that the isolation layer is always beyond its yield point. Figure 6 shows a representative time history realization of $\ddot{x}_g(t)$.

The equations of motion of the superstructure, if it were fixed base, are

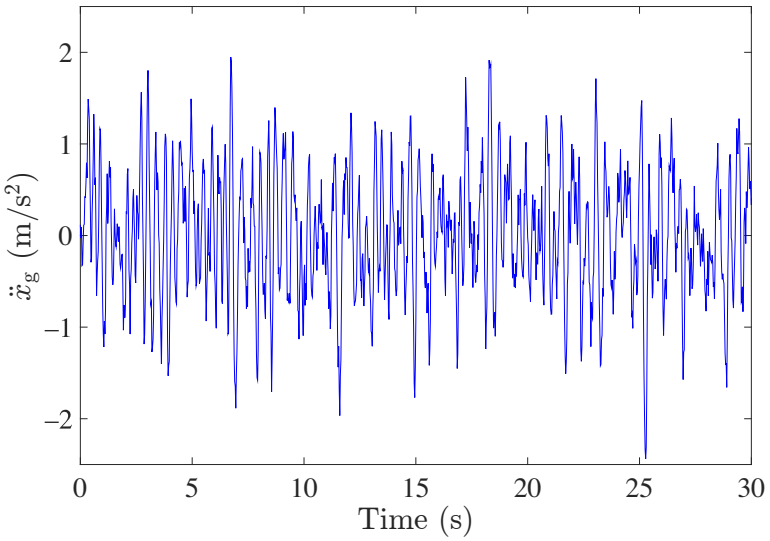


Figure 6: Representative time history realization of $\ddot{x}_g(t)$.

given by

$$\mathbf{M}_\mathrm{s}\ddot{\mathbf{X}}_\mathrm{s} + \mathbf{C}_\mathrm{s}\dot{\mathbf{X}}_\mathrm{s} + \mathbf{K}_\mathrm{s}\mathbf{X}_\mathrm{s} = -\mathbf{M}_\mathrm{s}\mathbf{1}\ddot{x}_\mathrm{g} \tag{14}$$

where $\mathbf{M}_\mathrm{s}$ is the 3×3 mass matrix as in (9) with $m_1 = m_2 = m_3 = 300\,\mathrm{Mg}$; $\mathbf{K}_\mathrm{s}$ is the 3×3 stiffness matrix similar to (10) with $\bar{k}_1 = \bar{k}_2 = \bar{k}_3 = 40\,\mathrm{MN/m}$; $\mathbf{1}_{3\times1}$ is a column vector of ones; and $\mathbf{X}_\mathrm{s} = [x_1 \quad x_2 \quad x_3]^\mathrm{T}$ is the vector of floor displacements relative to the ground. Again, Rayleigh damping, *i.e.*, $\mathbf{C}_\mathrm{s} = \beta_1\mathbf{M}_\mathrm{s} + \beta_2\mathbf{K}_\mathrm{s}$, is assumed with 3% damping for the first two modes, where β_1 and β_2 are constants evaluated from the damping ratios and the superstructure natural frequencies. Combining the isolation layer and the superstructure equations of motion, the full system can be described by

$$\begin{aligned}
&\mathbf{M}_\mathrm{s}\ddot{\mathbf{X}}_\mathrm{s} + \mathbf{C}_\mathrm{s}\dot{\mathbf{X}}_\mathrm{s} + \mathbf{K}_\mathrm{s}\mathbf{X}_\mathrm{s} = -\mathbf{M}_\mathrm{s}\mathbf{1}\ddot{x}_\mathrm{g} + \mathbf{C}_\mathrm{s}\mathbf{1}\dot{x}_\mathrm{b} + \mathbf{K}_\mathrm{s}\mathbf{1}x_\mathrm{b} \\
&m_\mathrm{b}\ddot{x}_\mathrm{b} + \mathbf{1}^\mathrm{T}\mathbf{C}_\mathrm{s}\mathbf{1}\dot{x}_\mathrm{b} + \mathbf{1}^\mathrm{T}\mathbf{K}_\mathrm{s}\mathbf{1}x_\mathrm{b} + f_\mathrm{b} = -m_\mathrm{b}\ddot{x}_\mathrm{g} + \mathbf{1}^\mathrm{T}\mathbf{C}_\mathrm{s}\dot{\mathbf{X}}_\mathrm{s} + \mathbf{1}^\mathrm{T}\mathbf{K}_\mathrm{s}\mathbf{X}_\mathrm{s}
\end{aligned} \tag{15}$$

where $m_\mathrm{b} = 500\,\mathrm{Mg}$ is the base mass; model classes for f_b, the sum of the isolation-layer damping and restoring forces, are discussed in the following paragraphs.

3.2.1. Candidate Model classes

A total of six model classes — four linear and two nonlinear (Bouc-Wen and bilinear) — are considered and described as follows.

Nonlinear Models for Hysteretic Damping

The two nonlinear model classes considered here (Figure 7) are a bilinear model and a Bouc-Wen hysteresis model [59], which is smoother and more

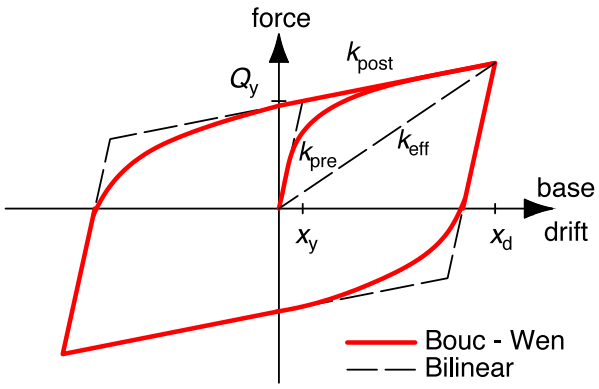


Figure 7: Model classes for hysteresis.

realistic [60]. In these nonlinear models, k_{pre}, k_{post}, and Q_{y} are the pre-yield and post-yield stiffnesses, and yield force, respectively. The Bouc-Wen model uses $q_{\mathrm{y}}z$ as the non-elastic force, where $q_{\mathrm{y}} = Q_{\mathrm{y}}\,(1 - 1/r_{\mathrm{k}})$, $r_{\mathrm{k}} = k_{\mathrm{post}}/k_{\mathrm{pre}}$ is the hardening ratio, and z is an evolutionary variable given by

$$\dot{z} = A\dot{u}_{\mathrm{b}} - \beta\dot{u}_{\mathrm{b}}|z|^{n_{\mathrm{pow}}} - \gamma z|\dot{u}_{\mathrm{b}}||z|^{n_{\mathrm{pow}}-1} \tag{16}$$

where $A = 2\beta = 2\gamma = k_{\mathrm{pre}}/Q_{\mathrm{y}}$ makes z stay in $[-1, 1]$ and consistent loading and unloading stiffnesses. For the Bouc-Wen model, $n_{\mathrm{pow}} = 1$ is assumed. As $n \to \infty$, the model approaches a bilinear hysteresis model; herein, $n_{\mathrm{pow}} = 100$ is used to represent the bilinear model. The effect of degradation can be incorporated by replacing (16) with the Baber-Wen model [61, 62]

$$\dot{z} = \frac{1}{\eta}\left[A\dot{u}_{\mathrm{b}} - \nu\left(\beta\dot{u}_{\mathrm{b}}|z|^{n_{\mathrm{pow}}} + \gamma z|\dot{u}_{\mathrm{b}}||z|^{n_{\mathrm{pow}}-1}\right)\right] \tag{17}$$

where the degradation shape functions are $\nu(e) = 1 + \delta_\nu e$ and $\eta(e) = 1 + \delta_\eta e$, with strength degradation parameter δ_ν, stiffness degradation parameter δ_η,

and $e = \int_0^t z(\tau)\dot{u}_{\rm b}(\tau)\mathrm{d}\tau$ is a measure of response duration and severity. The total force exerted by the isolation layer, the sum of the viscous damping and restoring forces, is given by

$$f_{\rm b} = c_{\rm b}\dot{x}_{\rm b} + k_{\rm post}u_{\rm b} + q_{\rm y}z \tag{18}$$

Linear Models for Hysteretic Damping

The linear model classes are approximately equivalent to a bilinear model of hysteresis in terms of energy dissipation in each cycle. The linear model classes represent the total isolator force as

$$\begin{aligned} f_{\rm b} &= \left[c_{\rm b} + c_{\rm eq}\right]\dot{x}_{\rm b} + k_{\rm eq}u_{\rm b} \\ &= \left[c_{\rm b} + 2\zeta_{\rm eq}\sqrt{k_{\rm eq}(m_{\rm b}+m_{\rm s})}\right]\dot{x}_{\rm b} + k_{\rm eq}u_{\rm b}. \end{aligned} \tag{19}$$

where $m_{\rm s}$ = mass of the superstructure. AASHTO (American Association of State Highway and Transportation Officials) and JPWRI (Japanese Public Works Research Institute) specified equivalent linear model classes [63, 64], where the equivalent damping ratio $\zeta_{\rm eq}$ and equivalent stiffness $k_{\rm eq}$ to approximate the energy dissipation of a hysteretic component are given by

$$\begin{aligned} \zeta_{\rm eq} &= \frac{2(1-r_{\rm k})(1-\rho^{-1})}{\pi[1+r_{\rm k}(\rho-1)]} \\ k_{\rm eq} &= \frac{k_{\rm pre}}{\rho}[1+r_{\rm k}(\rho-1)] \end{aligned} \tag{20}$$

The parameter $\rho = r_{\rm d}$ for the AASHTO model and $\rho = 0.7r_{\rm d}$ for the JPWRI model, where $r_{\rm d} = x_{\rm d}/x_{\rm y}$ is the shear ductility ratio of the design displacement $x_{\rm d}$ to the yield displacement $x_{\rm y}$. Hwang and Chiou [64] proposed a

correction to the AASHTO model using two correction factors, $r_{\mathrm{d}}^{0.58}/4.5$ and $[1-0.737(r_{\mathrm{d}}-1)/r_{\mathrm{d}}^2]^{-2}$ for k_{eq} and ζ_{eq}, respectively. The AASHTO linear model with the correction factor is used here as the third linear model class. In the CALTRANS model [64], the equivalent damping ratio and stiffness are given by

$$\begin{aligned} \zeta_{\mathrm{eq}} &= 0.0587(r_{\mathrm{d}}-1)^{0.371} \\ k_{\mathrm{eq}} &= k_{\mathrm{pre}}\{1+\ln[1+0.13(r_{\mathrm{d}}-1)^{1.137}]\}^{-2} \end{aligned} \tag{21}$$

To assess the effectiveness of the proposed model validation framework for a nonlinear system, the Baber-Wen model with $\delta_\nu = 0.04$ and $\delta_\eta = 0.02$, to introduce degradation within 30 s [62], is used to generate a set of nonlinear dynamic response data for base acceleration $\ddot{x}_{\mathrm{b}}$ at a sampling rate of 20 Hz for 30 s, giving $N_{\mathrm{o}} = 601$, to which 20% Gaussian noise is added (*i.e.*, the standard deviation of the Gaussian pulse process sensor noise is 0.2 times the noise-free RMS response).

3.2.2. Preprocessing: Intra-Model-Class Falsification

For each of the six model classes, 1000 models are drawn from independent distributions of the constitutive parameters using their prior distributions given in Table 7. The model falsification is applied with a zero-mean Gaussian likelihood function with standard deviation σ assumed to be 25% of the noisy RMS base acceleration measurements; *i.e.*, to 25% of the standard deviation of the data in $\mathfrak{D}$. This assumed residual standard deviation is close to the actual noise present in the measurement data. The target probability level is set at $\phi = 0.95$ for the falsification method. The results in Table 8

Table 7: Priors for model parameters as applicable to each model class.

Parameter	True value	Distribution	Mean	Std. dev.
k_{post}	4.0 MN/m	Lognormal	4.5 MN/m	0.25 MN/m
c_{b}	20 kN·s/m	Lognormal	18 kN·s/m	4 kN·s/m
r_{k}	0.1667	Uniform	0.1600	0.0058
r_{d}	n/a†	Uniform	2.5	0.2887
Q_{y} (% of W)*	5.00	Uniform	4.75	0.2887

* W = weight of the structure = $(m_{\text{s}} + m_{\text{b}})g$.

† Measurement data generated from Baber-Wen model does not require this parameter.

Table 8: Unfalsified models using proposed intra-model class falsification.

Model class ($\mathcal{M}_k$)	Unfalsified (%)
AASHTO	0
JPWRI	0
CALTRANS	0
mod. AASHTO	0
Bouc-Wen	**82.1**
Bilinear	**5.2**

demonstrate that the FDR likelihood bounds successfully falsify all linear model classes. Further, the results indicate that both Bouc-Wen and bilinear models may be valid representations and, therefore, deserving candidates for Bayesian model selection.

For this four DOF example, combined computational savings become $C \approx 66.67\%$. (The actual savings will be slightly smaller than this because the ease of simulation of the linear systems compared to a nonlinear system, but this will not be a significant factor since the structure in this example has only four degrees of freedom.)

3.2.3. Bayesian Model Selection

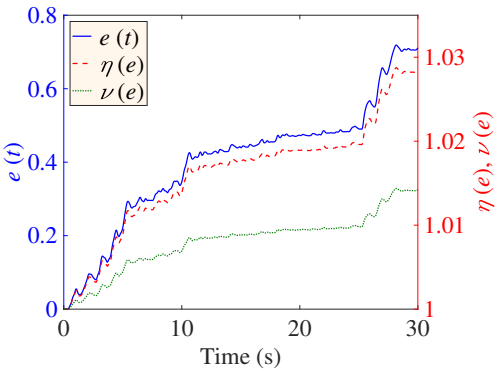
By falsifying all linear model classes with FDR, the subsequent Bayesian model selection is significantly reduced in scale, from six plausible model classes to two, saving computation time by avoiding unnecessary evidence calculations. Assuming equal priors $\mathbb{P}(\mathcal{M}_k) = 1/2$ for the nonlinear model classes, a subsequent Bayesian model class analysis assigns the Bouc-Wen model a posterior model class probability of essentially unity, leaving nearly zero probability of the bilinear model class (see Table 9).

3.2.4. Postprocessing: Inter-Model-Class Falsification

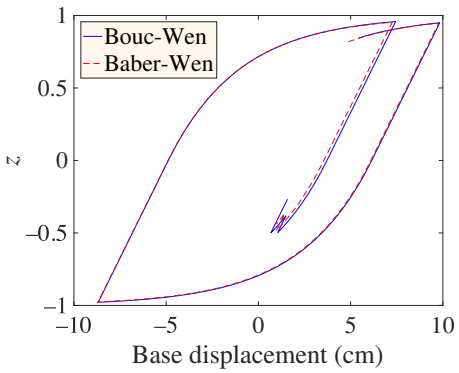
The Bouc-Wen model class also passes the postprocessing step of inter-model-class falsification step as the log of evidence value for this model class $\log \mathcal{E}^{(\text{Bouc-Wen})} = 490.3 > \log \underline{\mathcal{E}} = -986.5 = \log \underline{\mathcal{L}}$. (The evidence is the expected value of likelihood of the measurement data for a model class; hence, a wide range of values are expected.) This result is expected since the Bouc-Wen model is very similar to the *true* Baber-Wen model class. For example, the evolution of degradation parameters and hysteresis loops of the Bouc-Wen and Baber-Wen models for this building model are shown in Figure 8. The current choice of $\underline{\mathcal{E}} = \underline{\mathcal{L}}$ accepts the bilinear model class as well; it would also have rejected all of the linear model classes if they had passed

Table 9: Posterior model class probabilities after some model classes are rejected using a preprocessing step of intra-model-class falsification.

Model class ($\mathcal{M}_k$)	log(Evidence)	rel. log(Evidence)	$\mathbb{P}(\mathcal{M}_k\|\mathfrak{D})$
Bouc-Wen	490.3	0.0	≈ 1.0
Bilinear	27.2	-463.1	≈ 0.0



(a) Measure of response duration and severity e and degradation shape functions $\eta(e)$and $\nu(e)$.



(b) Hysteresis loops of Bouc-Wen and Baber-Wen models

Figure 8: Degradation parameters and hysteresis loops of the Bouc-Wen and Baber-Wen models using the *true* values of the parameters.

through the preprocessing. A stricter $\underline{\mathcal{E}}$ can be chosen, for example $\underline{\mathcal{E}} = 0.5\left[\max_k \max_i \mathcal{L}(\boldsymbol{\theta}_i; \mathfrak{D}, \mathcal{M}_k)\right]$, which will reject the bilinear model class even if it passes the preprocessing step and only accepts the Bouc-Wen model class. However, if the Baber-Wen model parameters were chosen so that it were to degrade faster, then the Bouc-Wen model might not remain a valid model class.

This framework identifies the Bouc-Wen model class as the valid one, and does so efficiently, drawing from a smaller candidate pool in which invalid model classes were systematically removed. The model validation framework here eliminates other model classes but validates the Bouc-Wen model class. However, the hysteresis loops for both Bouc-Wen and Baber-Wen models with the same parameter values are very similar, which justifies the result that the framework validates the Bouc-Wen model class. The posterior parameter means and standard deviations for the validated Bouc-Wen model class, shown in Table 10, are very close to the true values (as expected).

Table 10: Posterior model parameters and their true values for Bouc-Wen model class.

Parameter	True value	Mean	Std. dev.
k_{post} [MN/m]	4.0 MN/m	3.9082 MN/m	0.0764 MN/m
c_{b}	20 kN·s/m	18.0486 kN·s/m	1.1822 kN·s/m
r_{k}	0.1667	0.1628	0.0024
Q_{y} (% of W)	5.00	5.0017	0.0919

Table 11: Unfalsified models using the proposed intra-model class falsification with $\phi = 0.90$.

Model class ($\mathcal{M}_k$)	Unfalsified (%)
AASHTO	0
JPWRI	0
CALTRANS	0
mod. AASHTO	0
Bouc-Wen	**69.7**
Bilinear	**2.0**

3.2.5. Variations on the Method

To evaluate the framework with a different choice of target probability ϕ to assess its effects on this framework's performance, the preprocessing intra-model-class falsification step is applied using another conventional value $\phi = 0.90$, resulting in the fractions of remaining models listed in Table 11, which shows that two model classes, Bouc-Wen and bilinear, remain unfalsified. The Bayesian model selection and postprocessing inter-model-class falsification are applied next, resulting in validating the Bouc-Wen model class as before. This shows that the effect of this user-chosen parameter, the target identification probability ϕ, (if altered modestly) does not change the number of unfalsified model classes after the preprocessing step.

3.3. Example III: Full-scale Four-story Building Experiment

In 2013, a full-scale four-story base-isolated building was tested [65] on the world's largest six degree-of-freedom shake table at Japan's "E-Defense" laboratory, formally called the Hyogo Earthquake Engineering Research Center, part of Japan's National Research Institute for Earth Science and Disaster Resilience (NIED). The 676.6-ton asymmetric moment frame has dimensions $14\,\mathrm{m}\times10\,\mathrm{m}\times15\,\mathrm{m}$; its isolation layer, on the testing day 8 August 2013 concerned herein, consisted of two rubber bearings, two elastic sliding bearings, and two pairs of passive U-shaped steel yielding dampers. Random excitations along different table axes (which are the experimental data used herein) and scaled versions of historical and synthetic earthquake ground motions were used during that day of testing.

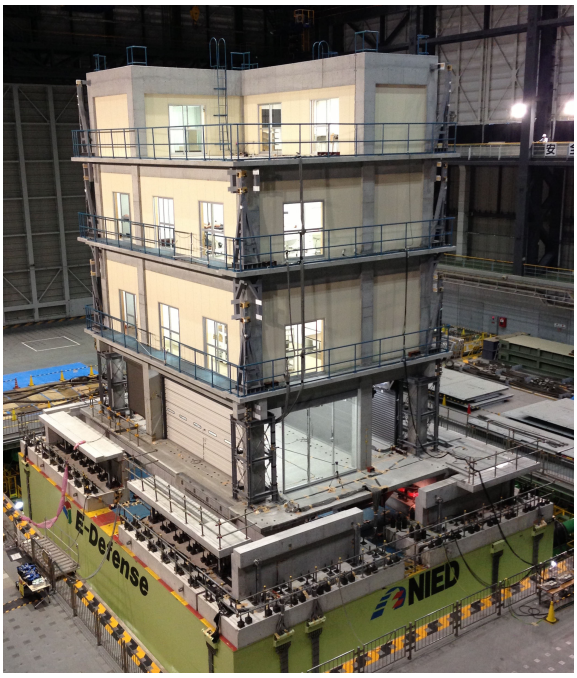


Figure 9: The experimental set-up.

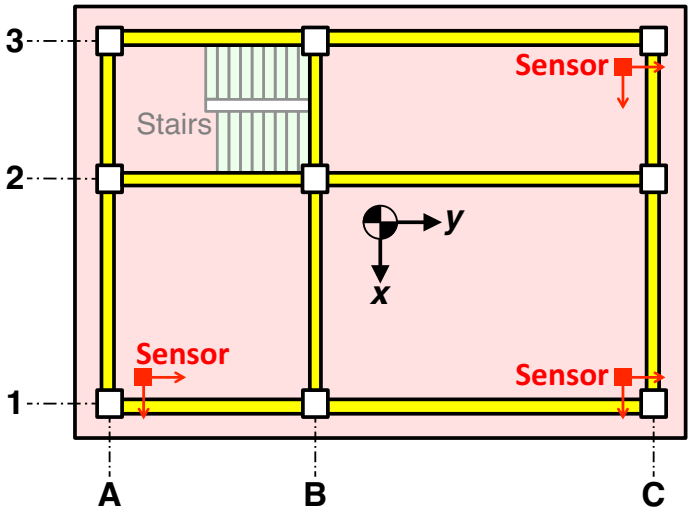


Figure 10: Typical accelerometer placements for the bottom three floors.

Tridirectional accelerometers recorded the responses at three corners of each floor (only two on the roof due to the structure set-back); four similar accelerometers were located on the shake table. Displacement transducers measured the displacements across the isolation layer and force transducers measured the forces induced by the isolation-layer devices. Measurements were recorded at 1 kHz sampling rate with a low-pass filtering using a 35 Hz cut-off frequency.

3.3.1. Candidate Model Classes

A finite element model with about 80,000 degrees of freedom was developed from design drawings. Beams, columns, and shear walls were modeled using solid elements, the steel reinforced bars using truss elements, and floor slabs and nonstructural walls using shell elements. For this example, the isolation-layer devices are modeled using bi-directional springs as the test responses used here showed a clear linear force-displacement relationship for

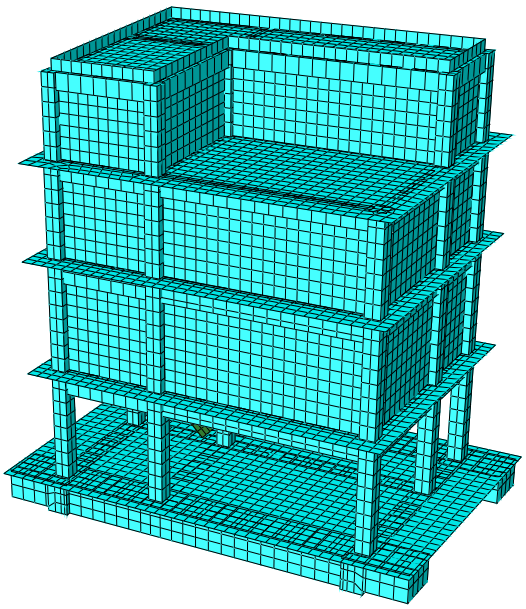


Figure 11: The finite element model with about 80,000 degrees of freedom.

the random excitation tests. The equations of motion of this model can be written as

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{C}(\boldsymbol{\theta})\dot{\mathbf{u}} + \mathbf{K}(\boldsymbol{\theta})\mathbf{u} = -\mathbf{M}\mathbf{r}\ddot{\mathbf{u}}_{\mathrm{t}} \tag{22}$$

where $\mathbf{M}$, $\mathbf{C}$, and $\mathbf{K}$ are model's mass, damping, and stiffness matrices; $\boldsymbol{\theta}$ is the vector of uncertain parameters; $\mathbf{u}$ is the displacement vector of the building with respect to the shake table; $\mathbf{r}$ consists of zeros and ones based on the table motion's influence on that degree-of-freedom; and $\ddot{\mathbf{u}}_{\mathrm{t}}$ is the table acceleration.

The candidate model classes are constructed by choosing a combination of different elastic moduli for different components of the building model. As shown in Table 12, model class $\mathcal{M}_1$ assumes all building components have same elastic modulus. Different elastic moduli are assumed for the floor

Table 12: Model classes with the list of each class's parameters (subscripts "Beam" and "Col" represent beams and columns in the building, respectively; subscript numbers represent floor/story numbers) and the means and standard deviations of their prior truncated Gaussian distributions.

Model Class $\mathcal{M}_k$	Parameter	Mean	Standard deviation
$\mathcal{M}_1$	$E_{\text{Beam,Col}}$	27 GPa	2.5 GPa
$\mathcal{M}_2$	E_{Beam}	27 GPa	2.5 GPa
	E_{Col}	23 GPa	2.5 GPa
$\mathcal{M}_3$	$E_{\text{Beam,1}}$	27 GPa	2.5 GPa
	$E^x_{\text{Beam,2,3,4}}$	27 GPa	2.5 GPa
	$E^y_{\text{Beam,2,3,4}}$	23 GPa	2.5 GPa
	E_{Col}	23 GPa	2.5 GPa
$\mathcal{M}_4$	$E_{\text{Beam,1}}$	27 GPa	2.5 GPa
	$E^x_{\text{Beam,2,3,4}}$	27 GPa	2.5 GPa
	$E^y_{\text{Beam,2,3,4}}$	23 GPa	2.5 GPa
	$E_{\text{Col,1,2,3}}$	23 GPa	2.5 GPa
	$E_{\text{Col,4}}$	24 GPa	2.5 GPa

beams and vertical columns in Model class $\mathcal{M}_2$. Similarly, model classes $\mathcal{M}_3$ and $\mathcal{M}_4$ differentiate between beams in different floors and different directions and columns in different stories. Prior distributions for these parameters, assumed as independent truncated Gaussians (truncated at zero to avoid negative stiffnesses), are shown in Table 12.

3.3.2. Preprocessing: Intra-Model-Class Falsification

The N4SID subspace identification algorithm [66] was used to identify the natural frequencies (see Table 13) and the mode shapes from Tests 010–012 that used low-intensity random excitations in the three table coordinate directions [67]. The *measurements* in this example are the first six natural frequency estimates and the 36 entries in the 6×6 MAC (modal assurance

Table 13: First six natural frequencies identified using the 42 responses and 12 recorded table acceleration inputs [67].

Mode	Natural frequency (Hz)
1^{st}	0.6853
2^{nd}	0.6975
3^{rd}	0.7095
4^{th}	4.7812
5^{th}	5.1749
6^{th}	6.1199

criterion) matrix; thus, $N_{\text{o}} = 42$. The probability density of residual error is assumed to be Gaussian for all entries, with a unit mean for the diagonal entries of the MAC matrix (as the ideal MAC matrix is an identity matrix) and zero-mean for all other MAC values and frequency residual errors. The assumed standard deviation of the residual errors are 0.02 Hz for the natural frequencies and 0.25 for MAC values based on a preliminary error analysis of the N4SID algorithm. With $\phi = 0.90$, the falsification results are shown in Table 14. By reducing the candidate model class set by half, the computational burden of evaluating such large models multiple times is reduced by approximately 50%.

Table 14: Results of preprocessing step for Example III.

Model Class	% Unfalsified
$\mathcal{M}_1$	0.0
$\mathcal{M}_2$	0.0
$\mathcal{M}_3$	3.7
$\mathcal{M}_4$	4.0

Table 15: Posterior model class probabilities after two candidate model classes are rejected using a preprocessing step of intra-model-class falsification.

Model class ($\mathcal{M}_k$)	log(Evidence)	rel. log(Evidence)	$\mathbb{P}(\mathcal{M}_k\|\mathfrak{D})$
$\mathcal{M}_3$	1852.70	0.00	≈ 0.60
$\mathcal{M}_4$	1852.30	-0.40	≈ 0.40

3.3.3. Bayesian Model Selection and Postprocessing

Next, Bayesian model selection is applied to the remaining two model classes assuming equal prior $\mathbb{P}(\mathcal{M}_k) = 1/2$ for each of them, resulting in posterior model class probabilities of 0.60 and 0.40, as shown in Table 15. Although model class $\mathcal{M}_4$ has more parameters, it is not assigned the highest posterior probability, likely due to the Occam's razor principle that is inherent in Bayesian model class selection. Finally, the postprocessing step using $\underline{\mathcal{L}}$ confirms the validity of both of these model classes. Hence, the proposed framework validates two of the four initial model classes.

4. Conclusions

The proposed hybrid probabilistic framework unites the philosophical ideas of model falsification and model selection into a single integrated model validation methodology. This framework applies falsification in a likelihood domain with false discovery rate control as pre- and postprocessing steps. This framework is shown to not only identify the correct models, but to also overcome the shortcomings of each of these methods applied alone, by efficiently and systematically eliminating incorrect model classes. The numerical examples demonstrate the efficacy of this framework and its enhanced computational efficiency for dynamical system modeling. The first example shows

how the proposed framework reduces the available model classes by 66.67%, giving large computational savings in the next step of Bayesian model selection, which requires the calculation of evidence — a costly enterprise. The second example uses a candidate model class set that does not contain the true model class used to generate the measurements, but the validated model class is very close to the true one. Further, this framework provides computational savings of approximately 66.67% by eliminating most of the incorrect model classes. (Note that the elimination of two-thirds of the model classes in the preprocessing is particular to these examples, and it is coincidence that they eliminated the same fraction of model classes.) The third example uses high-fidelity finite element model, with about 80,000 degrees of freedom, of a full-scale structure tested at E-Defense. Again, the proposed framework eliminates half of the candidate model classes during the preprocessing step. For more complex models, with more measurements and a higher-dimensional parameter space $\boldsymbol{\Theta}$, much greater savings are expected. Finally, computational efficiency can be further enhanced by exploring other sampling algorithms and exploiting the localized nature of uncertainties and/or nonlinearities in dynamical systems. In future studies, a feedback structure to this framework will be added where some of the previous measurements will be used to intelligently select the probability density of residual error and the parameter priors at the next time step for its on-line implementation.

Acknowledgment

The authors gratefully acknowledge the partial support of this work by the National Science Foundation through awards CMMI 14-36018/14-36058

and 16-63667/16-62992. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Mr. De and Dr. Brewick acknowledge the support of a Viterbi Ph.D. Fellowship and a Viterbi Postdoctoral Fellowship, respectively, from University of Southern California. The authors thank Dr. Eiji Sato (NIED) and Dr. Tomohiro Sasaki (Obayashi Corp.) for their extensive efforts on the base isolation experiments and their collaboration studying the resulting data that is also used in the third example, and Mr. Tianhao Yu for kindly providing the finite element model for that specimen.

Appendix A. Use of False Discovery Rate (FDR) to Compute the Residual Error Bounds [37]

The false discovery rate introduced in 1995 by Benjamini and Hotchberg [43] is defined as the expected value of the ratio of the number of times the model is falsely rejected (N_{vr}) to total number of times the model is rejected (N_{r}), where this ratio ($N_{\mathrm{vr}}/N_{\mathrm{r}}$) is assumed to be zero when $N_{\mathrm{r}} = 0$, *i.e.*, $\mathrm{FDR} = \mathbb{E}\left[\frac{N_{\mathrm{vr}}}{N_{\mathrm{r}}}\middle| N_{\mathrm{r}} > 0\right]\mathbb{P}(N_{\mathrm{r}} > 0)$. An algorithm, known as the Benjamini-Hotchberg (BH) procedure, is used here for keeping FDR at a prechosen level

α. Here, the residual errors are first sorted according to their p-values[2]

$$0 \leq p_1 \leq p_2 \leq \cdots \leq p_{N_\mathrm{o}} \leq 1 \tag{A.1}$$

and the significance level for residual error ϵ_i is chosen to be

$$\bar{\alpha}_i = \frac{i}{N_\mathrm{o}}\alpha, \qquad i = 1, \ldots, N_\mathrm{o} \tag{A.2}$$

where the target identification probability is $\phi = 1 - \alpha$, which comes from hypothesis testing (typically $\phi = 0.95$). The residual error bounds $\underline{\epsilon}_i$ and $\bar{\epsilon}_i$ are next computed from $\bar{\alpha}_i$ using

$$\begin{aligned} \frac{1}{2}\bar{\alpha}_i &= \mathbb{P}(E_i \leq \underline{\epsilon}_i|\boldsymbol{\theta}) = \mathbb{P}(E_i \geq \bar{\epsilon}_i|\boldsymbol{\theta}) \\ &= \int_{-\infty}^{\underline{\epsilon}_i} \mathrm{p}(e_i|\boldsymbol{\theta})\mathrm{d}e_i = \int_{\bar{\epsilon}_i}^{\infty} \mathrm{p}(e_i|\boldsymbol{\theta})\mathrm{d}e_i, \quad i = 1, \ldots, N_\mathrm{o} \end{aligned} \tag{A.3}$$

Appendix B. Nested Sampling Algorithm [46, 68]

Evidence integral (6) for model class $\mathcal{M}_k$ is rewritten using a probability integral transformation [46, 68]

$$\mathcal{E}^{(k)} = \int_0^1 \varphi(\chi(\lambda))\mathrm{d}\chi(\lambda) \approx \sum_i \varphi(\chi(\lambda_i))\Delta\chi(\lambda_i) \tag{B.1}$$

[2]The p-values for two-sided distributions can be defined as:

$$\begin{aligned} p_i &= 2\,\min\left\{\mathbb{P}\left(E_i \leq \epsilon_i|\boldsymbol{\theta}\right), \mathbb{P}\left(E_i \geq \epsilon_i|\boldsymbol{\theta}\right)\right\}, \qquad i = 1, \ldots, N_\mathrm{o} \\ &= 2\,\min\left\{\int_{-\infty}^{\epsilon_i} \mathrm{p}(e_i|\boldsymbol{\theta})\mathrm{d}e_i, \int_{\epsilon_i}^{\infty} \mathrm{p}(e_i|\boldsymbol{\theta})\mathrm{d}e_i\right\} \end{aligned}$$

where $\chi(\lambda)$ is the probability mass enclosed in the parameter space subset where likelihoods $\mathcal{L}(\boldsymbol{\theta})$ exceed λ and $\varphi(\chi)$ denotes its inverse, *i.e.*, $\varphi\left(\chi(\lambda)\right) \equiv \lambda$. This one-dimensional integral is then evaluated using Algorithm 1. A *stopping criterion* of $\frac{1}{\mathcal{E}_i}\max\{\mathcal{L}(\boldsymbol{\theta}_i)\}_{i=1}^{n_\mathrm{s}} < 0.01$, suggested in Skilling [46], is used for the numerical examples in this paper. The outputs of the nested sampling algorithm can be further used to estimate the moments of the posterior distribution of the parameters of a model class, as shown in Algorithm 1.

1 **Initialization:** Set $\chi_0 = 1$, $\mathcal{E}_0 = 0$, $\mathbf{m}_{\boldsymbol{\theta}} = \mathbf{0}_{n_{\boldsymbol{\theta}}\times 1}$, and $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{0}_{n_{\boldsymbol{\theta}}\times 1}$;
2 Generate n_s samples $\boldsymbol{\theta}_i, \quad i = 1, \ldots, n_\mathrm{s}$, from prior $\mathrm{p}(\boldsymbol{\theta})$ with corresponding likelihoods $\mathcal{L}(\boldsymbol{\theta}_i)$;
3 Start the sample counter: $i = 1$;
4 **while** *stopping criteria* = **FALSE do**
5 Find $j = \arg\min_{k=1,\ldots,n_\mathrm{s}} \mathcal{L}(\boldsymbol{\theta}_k)$;
6 Assign $\chi_i = [n_\mathrm{s}/(n_\mathrm{s}+1)]^i$;
7 Assign $\Delta\chi_i = \chi_{i-1} - \chi_i$;
8 Update evidence estimate by $\Delta\mathcal{E}_i = \Delta\chi_i\mathcal{L}(\boldsymbol{\theta}_j)$ and $\mathcal{E}_i = \mathcal{E}_{i-1} + \Delta\mathcal{E}_i$;
9 Update statistics $\mathbf{m}_{\boldsymbol{\theta}} = \mathbf{m}_{\boldsymbol{\theta}} + \Delta\mathcal{E}_i\boldsymbol{\theta}_j$ and $\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{s}_{\boldsymbol{\theta}} + \Delta\mathcal{E}_i\boldsymbol{\theta}_j^2$;
10 Replace $\boldsymbol{\theta}_j$ with a new sample $\boldsymbol{\theta}_\mathrm{new}$ that satisfies $\mathcal{L}(\boldsymbol{\theta}_\mathrm{new}) > \mathcal{L}(\boldsymbol{\theta}_j)$;
11 $i = i + 1$;
12 **end**
13 $\mathcal{E}_\mathrm{last} = \mathcal{E}_{i-1} + \frac{1}{n_\mathrm{s}}\sum_{k=1}^{n_\mathrm{s}} \mathcal{L}(\boldsymbol{\theta}_k)\chi_i$,
14 $\mathbb{E}[\boldsymbol{\theta}] = \mathbf{m}_{\boldsymbol{\theta}}/\mathcal{E}_\mathrm{last}$,
15 $\mathrm{Var}[\boldsymbol{\theta}] = (\mathbf{s}_{\boldsymbol{\theta}}/\mathcal{E}_\mathrm{last}) - (\mathbb{E}[\boldsymbol{\theta}])^2$;
Result: Evidence $\mathcal{E} = \mathcal{E}_\mathrm{last}$, posterior parameter mean $\mathbb{E}[\boldsymbol{\theta}]$, and posterior parameter variance $\mathrm{Var}[\boldsymbol{\theta}]$.

Algorithm 1: Evidence calculation using nested sampling. The exponents $(\cdot)^2$ in lines 9 and 15 are element-by-element, and $\mathrm{Var}[\boldsymbol{\theta}]$ is a vector of the variances of the elements of $\boldsymbol{\theta}$ (not the covariance matrix of $\boldsymbol{\theta}$).

Appendix C. Modified Metropolis-Hastings Algorithm [56]

The modified Metropolis-Hastings algorithm proposed in Au and Beck [56] is used here to generate $\boldsymbol{\theta}_{\text{new}}$ with a high acceptance rate in Algorithm 1. The joint prior distribution of parameters is assumed to be written as $\mathrm{p}(\boldsymbol{\theta}) = \prod_{i=1}^{n_{\boldsymbol{\theta}}} \mathrm{p}_i(\theta_i)$, at least using some approximate transformation, where $\boldsymbol{\theta} \in \mathbb{R}^{n_{\boldsymbol{\theta}} \times 1}$. At any iteration of the nested sampling, the chain starts from $\boldsymbol{\theta}_k$, which is chosen from one of the remaining samples. The proposal density to generate candidates is assumed as $\mathrm{q}(\boldsymbol{\theta}^{\mathrm{c}}|\boldsymbol{\theta}) = \prod_{i=1}^{n_{\boldsymbol{\theta}}} \mathrm{q}_i(\theta_i^{\mathrm{c}}|\theta_i)$. A sequence of $\boldsymbol{\theta}^{(l)}$ can then be generated such that $\mathcal{L}(\boldsymbol{\theta}^{(l)}) > \mathcal{L}(\boldsymbol{\theta}_j)$ using the steps shown in Algorithm 2. Please note that no burn-ins are required in this case since the new samples generated will always have $\mathcal{L}_{\text{new}} > \mathcal{L}$.

Initialization: Set $l = 0$ and $\boldsymbol{\theta}^{(l)} = \boldsymbol{\theta}_k$;
`/* Select candidate` $\boldsymbol{\theta}^{\mathrm{c}}$ `*/`
for $i = 1, \ldots, n_{\boldsymbol{\theta}}$ **do**
  Generate $\varphi \sim \mathrm{q}_i\left(\varphi|\theta_i^{(l)}\right)$;
  Evaluate acceptance probability $\mathrm{p_a} = \min\left(1, \frac{\mathrm{p}_i(\varphi)\mathrm{q}_i\left(\theta_i^{(l)}|\varphi\right)}{\mathrm{p}_i\left(\theta_i^{(l)}\right)\mathrm{q}_i\left(\varphi|\theta_i^{(l)}\right)}\right)$;
  Generate $u \sim \mathcal{U}(0,1)$ (*i.e.*, from a uniform distribution in $(0,1)$);
  if $u < \mathrm{p}_a$ **then**
    Set $\theta_i^{\mathrm{c}} = \varphi$;
  else
    Set $\theta_i^{\mathrm{c}} = \theta_i^{(l)}$;
  end
end
`/* Accept or reject` $\boldsymbol{\theta}^{\mathrm{c}}$ `*/`
if $\mathcal{L}(\boldsymbol{\theta}^c) > \mathcal{L}(\boldsymbol{\theta}_j)$ **then**
  Set $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{\mathrm{c}}$;
else
  Set $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)}$;
end
Set $l = l + 1$;
Repeat ;

Algorithm 2: Modified Metropolis-Hastings algorithm.

References

[1] B. F. Spencer, Jr., S. Nagarajaiah, State of the art of structural control, Journal of Structural Engineering 129 (7) (2003) 845–856.

[2] F. R. Menter, Two-equation eddy-viscosity turbulence models for engineering applications, AIAA Journal 32 (8) (1994) 1598–1605.

[3] S. Wallin, A. V. Johansson, An explicit algebraic Reynolds stress model for incompressible and compressible turbulent flows, Journal of Fluid Mechanics 403 (25) (2000) 89–132.

[4] C. Rao, Y. Wu, S. Konishi, R. Mukerjee, On model selection, Lecture Notes-Monograph Series 38 (2001) 1–64.

[5] H. Chipman, E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, R. A. Stine, The practical implementation of Bayesian model selection, Lecture Notes-Monograph Series 38 (2001) 65–134.

[6] K. M. Cremers, Stock return predictability: A Bayesian model selection perspective, Review of Financial Studies 15 (4) (2002) 1223–1249.

[7] C. Andrieu, P. Djurić, A. Doucet, Model selection by MCMC computation, Signal Processing 81 (1) (2001) 19–37.

[8] J. L. Beck, K.-V. Yuen, Model selection using response measurements: Bayesian probabilistic approach, Journal of Engineering Mechanics 130 (2) (2004) 192–203.

[9] J. L. Beck, Bayesian system identification based on probability logic, Structural Control and Health Monitoring 17 (7) (2010) 825–847.

[10] S. H. Cheung, J. L. Beck, Calculation of posterior probabilities for Bayesian model class assessment and averaging from posterior samples based on dynamic system data, Computer-Aided Civil and Infrastructure Engineering 25 (5) (2010) 304–321.

[11] M. Muto, J. L. Beck, Bayesian updating and model class selection for hysteretic structural models using stochastic simulation, Journal of Vibration and Control 14 (1-2) (2008) 7–34.

[12] L. Mthembu, T. Marwala, M. I. Friswell, S. Adhikari, Model selection in finite element model updating using the Bayesian evidence statistic, Mechanical Systems and Signal Processing 25 (7) (2011) 2399–2412.

[13] R. E. Kass, A. E. Raftery, Bayes factors, Journal of the American Statistical Association 90 (430) (1995) 773–795.

[14] S. N. Goodman, Toward evidence-based medical statistics. 2: The Bayes factor, Annals of Internal Medicine 130 (12) (1999) 1005–1013.

[15] E. T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, U.K., 2003.

[16] D. J. MacKay, Bayesian interpolation, Neural Computation 4 (3) (1992) 415–447.

[17] D. J. C. MacKay, Bayesian methods for adaptive models, Ph.D. thesis, California Institute of Technology (1992).

[18] S. F. Gull, Bayesian inductive inference and maximum entropy, in:

Maximum-entropy and Bayesian methods in science and engineering, Springer, 1988, pp. 53–74.

[19] K. R. Popper, The Logic of Scientific Discovery, Routledge, New York, 2002, translation of *Logik der Forschung*, first published in 1934 by Verlag von Julius Springer, Vienna, Austria.

[20] M. G. Safonov, T. Tsao, The unfalsified control concept and learning, IEEE Transactions on Automatic Control 42 (1997) 843–847.

[21] P. B. Brugarolas, M. G. Safonov, Learning about dynamical systems via unfalsification of hypotheses, International Journal of Robust and Nonlinear Control 14 (11) (2004) 933–943.

[22] H. Hjalmarsson, From experiment design to closed-loop control, Automatica 41 (3) (2005) 393–438.

[23] J.-A. Goulet, P. Kripakaran, I. Smith, Multimodel structural performance monitoring, Journal of Structural Engineering 136 (2010) 1309–1318.

[24] B. Raphael, I. F. C. Smith, Finding the right model for bridge diagnosis, Artificial Intelligence in Structural Engineering, Computer Science, Lecture Notes in Artificial Intelligence 1454 (1998) 308–319.

[25] I. F. C. Smith, S. Saitta, Improving knowledge of structural system behavior through multiple models, Journal of Structural Engineering 134 (2008) 553–561.

[26] J.-A. Goulet, I. F. C. Smith, Predicting the usefulness of monitoring for identifying the behavior of structures, Journal of Structural Engineering 139 (2013) 1716–1727.

[27] J.-A. Goulet, S. Coutu, I. F. C. Smith, Model falsification diagnosis and sensor placement for leak detection in pressurized pipe networks, Advanced Engineering Informatics 27 (2013) 261–269.

[28] J.-A. Goulet, I. F. Smith, Performance-driven measurement system design for structural identification, Journal of Computing in Civil Engineering 27 (4) (2013) 427–436.

[29] J.-A. Goulet, C. Michel, I. F. C. Smith, Hybrid probabilities and error-domain structural identification using ambient vibration monitoring, Mechanical Systems and Signal Processing 37 (2013) 199–212.

[30] J.-A. Goulet, I. F. C. Smith, Structural identification with systematic errors and unkown uncertainty dependencies, Computers and Structures 128 (2013) 251–258.

[31] K. Beven, A. Binley, The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes 6 (1992) 279–298.

[32] K. Beven, Prophecy, reality and uncertainty in distributed hydrological modelling, Advances in Water Resources 16 (1) (1993) 41–51.

[33] K. Beven, J. Freer, Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology, Journal of Hydrology 249 (1) (2001) 11–29.

[34] K. Beven, A manifesto for the equifinality thesis, Journal of Hydrology 320 (1) (2006) 18–36.

[35] K. J. Beven, Rainfall-runoff modelling: the primer, John Wiley & Sons, 2011.

[36] P. Mantovan, E. Todini, Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, Journal of Hydrology 330 (1) (2006) 368–381.

[37] S. De, P. T. Brewick, E. A. Johnson, S. F. Wojtkiewicz, Investigation of model falsification using error and likelihood bounds with application to a structural system, Journal of Engineering Mechanics 144 (9) (2018) 04018078.

[38] A. Tarantola, Popper, Bayes and the inverse problem, Nature Physics 2 (2006) 492–494.

[39] I. Babuška, F. Nobile, R. Tempone, A systematic approach to model validation based on Bayesian updates and prediction related rejection criteria, Computer Methods in Applied Mechanics and Engineering 197 (29) (2008) 2517–2539.

[40] K. Farrell, J. T. Oden, Calibration and validation of coarse-grained models of atomic systems: application to semiconductor manufacturing, Computational Mechanics 54 (1) (2014) 3–19.

[41] K. A. Farrell, Selection, calibration, and validation of coarse-grained models of atomistic systems, Ph.D. thesis, University of Texas at Austin (2015).

[42] K. Farrell, J. T. Oden, D. Faghihi, A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems, Journal of Computational Physics 295 (2015) 189–208.

[43] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, Journal of the Royal Statistical Society. Series B (Methodological) 57 (1995) 289–300.

[44] M. Bouaziz, M. Jeanmougin, M. Guedj, Multiple testing in large-scale genetic studies, in: Data Production and Analysis in Population Genomics, Springer, 2012, pp. 213–233.

[45] M. A. Newton, A. E. Raftery, Approximate Bayesian inference with the weighted likelihood bootstrap, Journal of the Royal Statistical Society. Series B (Methodological) 56 (1) (1994) 3–48.

[46] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Analysis 1 (4) (2006) 833–859.

[47] R. M. Neal, Annealed importance sampling, Statistics and Computing 11 (2) (2001) 125–139.

[48] N. Friel, A. N. Pettitt, Marginal likelihood estimation via power posteriors, Journal of the Royal Statistical Society. Series B (Methodological) 70 (3) (2008) 589–607.

[49] J. Ching, Y.-C. Chen, Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging, Journal of Engineering Mechanics 133 (7) (2007) 816–832.

[50] Z. I. Botev, D. P. Kroese, Efficient Monte Carlo simulation via the generalized splitting method, Statistics and Computing 22 (1) (2012) 1–16.

[51] M. Chiachio, J. L. Beck, J. Chiachio, G. Rus, Approximate bayesian computation by subset simulation, SIAM Journal on Scientific Computing 36 (3) (2014) A1339–A1358.

[52] F. A. DiazDelaO, A. Garbuno-Inigo, S. K. Au, I. Yoshida, Bayesian updating and model class selection with subset simulation, Computer Methods in Applied Mechanics and Engineering 317 (2017) 1102 – 1121.

[53] M. K. Vakilzadeh, Y. Huang, J. L. Beck, T. Abrahamsson, Approximate Bayesian computation by subset simulation using hierarchical state-space models, Mechanical Systems and Signal Processing 84 (2017) 2 – 20, recent advances in nonlinear system identification.

[54] Y. M. Marzouk, D. Xiu, A stochastic collocation approach to Bayesian inference in inverse problems, Communications in Computational Physics 6 (1) (2009) 826–847.

[55] Y. M. Marzouk, H. N. Najm, L. A. Rahn, Stochastic spectral methods for efficient Bayesian solution of inverse problems, Journal of Computational Physics 224 (2) (2007) 560–586.

[56] S.-K. Au, J. L. Beck, Estimation of small failure probabilities in high dimensions by subset simulation, Probabilistic Engineering Mechanics 16 (4) (2001) 263–277.

[57] B. F. Spencer, S. J. Dyke, H. S. Deoskar, Benchmark problems in structural control: part I – active mass driver system, Earthquake Engineering & Structural Dynamics 27 (11) (1998) 1127–1140.

[58] J. C. Ramallo, E. A. Johnson, B. F. Spencer, Jr., "Smart" base isolation systems, Journal of Engineering Mechanics 128 (10) (2002) 1088–1099.

[59] Y.-K. Wen, Method for random vibration of hysteretic systems, Journal of the Engineering Mechanics Division 102 (2) (1976) 249–263.

[60] S. Nagarajaiah, X. Sun, Response of base-isolated USC hospital building in Northridge earthquake, Journal of Structural Engineering 126 (10) (2000) 1177–1186.

[61] T. T. Baber, Y.-K. Wen, Random vibration hysteretic, degrading systems, Journal of the Engineering Mechanics Division 107 (6) (1981) 1069–1087.

[62] F. Ma, A. Bockstedte, G. C. Foliente, P. Paevere, H. Zhang, Parameter analysis of the differential model of hysteresis, Journal of Applied Mechanics 71 (3) (2004) 342–349.

[63] K. Kawashima, K. Hasegawa, H. Nagashima, Manual for Menshin design of highway bridges, in: 2nd US-Japan Workshop on Earthquake Protective Systems for Bridges, Public Works Research Institute (PWRI), Tsukuba City, Japan, 1992.

[64] J. S. Hwang, J. M. Chiou, An equivalent linear model of lead-rubber seismic isolation bearings, Engineering Structures 18 (7) (1996) 528–536.

[65] E. Sato, T. Sasaki, K. Fukuyama, K. Tahara, K. Kajiwara, Development of innovative base-isolation system based on E-Defense full-scale shake table experiments, Part I: outline of project research, in: AIJ Annual Meeting, Hokkaido, Japan, 2013, pp. 751–752, (In Japanese.).

[66] P. van Overschee, B. De Moor, N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems, Automatica 30 (1) (1994) 75–93.

[67] P. T. Brewick, E. A. Johnson, E. Sato, T. Sasaki, Constructing and evaluating generalized models for a base-isolated structure, Structural Control and Health Monitoring (in press) (2018) e2243.

[68] S. De, E. A. Johnson, S. F. Wojtkiewicz, P. T. Brewick, Computationally-efficient Bayesian model selection for locally nonlinear structural dynamical systems, Journal of Engineering Mechanics 144 (5) (2018) 04018022.