

Comparison of Site- and Structure-Specific Surrogate Models for Estimating Bridge Structural Response in Regional Earthquake Scenario Studies

Meredith Lochhead¹, Sang-ri Yi², Kuanshi Zhong³, and Gregory G. Deierlein¹

¹Department of Civil and Environmental Engineering, Stanford University, mlochhead@stanford.edu

²Department of Civil and Environmental Engineering, Rice University

³Department of Civil and Architectural Engineering and Construction Management, University of Cincinnati

This manuscript is a preprint and has not been peer reviewed.

ABSTRACT

Surrogate models offer an efficient method for predicting site-specific and structure-specific performance within a regional seismic risk assessment. However, selecting a surrogate model for regional Performance-Based Earthquake Engineering applications requires weighing predictive accuracy against other considerations such as workflow setup, model sharing, treatment of uncertainty, correlation-handling, and regional scalability, among other factors. This study compares three specific implementations of surrogate modeling approaches—Probabilistic Learning on Manifolds (PLOM), Gaussian Process (GP), and multivariate linear regression—for predicting the seismic performance of concrete highway bridges under seismic loading, including both collapse and non-collapse behavior. The analysis includes 1) a quantitative comparison of the three models and 2) a discussion of the strengths and weaknesses of each model when applied at the regional scale. In this case study, GP provides the most consistent accuracy across collapse and non-collapse behavior and supports deterministic mapping from input realizations to output response, which is a critical feature for downstream transportation network analysis. Linear regression, though effective for collapse prediction, lacks the flexibility to model detailed engineering demand parameters, which range from elastic to highly nonlinear across earthquake intensities. By design, PLOM is well suited to multi-modal or non-Gaussian output distributions and cases where conditional correlation between output variables matters; however, its implementation here does not support deterministic input-to-output mapping, limiting its use for transportation network analysis.

1 INTRODUCTION

Regional seismic risk assessments provide critical insight into the expected seismic performance of communities and infrastructure systems, supporting the identification of mitigation priorities and improvements in community resilience. In the context of regional earthquake scenario studies, estimating structural response and damage due to earthquake ground motions is a critical step. At the regional scale, structural damage is typically quantified using fragility functions, which describe a broad typology of structures and map directly from an earthquake intensity measure to a probabilistic damage state [1]. While this approach is computationally efficient, typology-level fragilities can lack structural specificity and have limited ability to account for multiple earthquake intensity measures [2, 3, 4, 5].

In contrast, within the Performance-Based Earthquake Engineering (PBEE) framework [6], structural response is typically estimated using nonlinear response history analysis (NLRHA). NLRHA explicitly simulates the dynamic response history of a structure under a given set of ground motion records. NLRHA results, such as the peak and cumulative response quantities across the time history, are referred to as Engineering Demand Parameters (EDPs), which can be used to estimate the damage of structural and nonstructural components [7]. NLRHA is typically conducted for multiple input ground motions and/or structural parameters for a given structure at a given site to account for record-to-record variability in the ground motions and uncertainty in structural modeling and material properties, respectively [8, 7, 9]. While NLRHA is effective for understanding the detailed performance of a specific structure under a range of ground motions, there are several trade-offs when applied at the regional scale. Most prominently, these include excessive computational cost, as well as additional effort to 1) select appropriate site-specific ground motions and 2) document detailed structural design parameters for every structure across the region [10, 4].

Surrogate modeling techniques have emerged as a way to bridge this gap between generalized typology fragility functions and structure-specific, site-specific NLRHA using the PBEE methodology. In this study, the term *surrogate model* is used to describe a wide range of statistical functions, ranging from simple linear regression to complex machine learning models, which can replace computationally expensive analysis by mapping from input parameters to output parameters. In the context of PBEE, surrogate models often map from structural and/or ground motion parameters to structural responses, thus replacing the expensive NLRHA required to evaluate the structure [10, 11, 12, 5, 13].

There are many different types of surrogate models that have been used for PBEE applications. Samadian et al. [14] present a comprehensive literature review of data-driven surrogate models and metamodeling techniques used in the field of structural engineering, including a discussion of pure data-driven methods, multi-fidelity based surrogates, explainable machine learning methods, and projection-based methods. Surrogates can also be deployed at different scales for regional assessments. Rincon and Padgett [4] discuss fragility function development, including more traditional archetype fragilities, parameterized fragilities, and structure-specific fragilities, providing insight into varying levels of surrogate detail and the implications of alternative seismic fragility modeling practices.

The No Free Lunch Theorem [15] highlights that no single algorithm is universally superior across all problem domains; rather, algorithmic effectiveness is closely tied to the specific nature of the task. This underscores the critical need to integrate problem-specific knowledge when selecting, designing, and adapting surrogate models, as well as when formulating workflows that can effectively accommodate them. Thus, while existing literature discusses the use of surrogate models to replace NLRHA in regional assessments [16, 17, 18, 5, 19], few studies have explicitly compared site-specific, structure-specific surrogate modeling techniques across a suite of considerations relevant for regional surrogate use.

Through the example of a two-span, single-column concrete highway bridge, this study compares three surrogate models with distinct probabilistic inference mechanisms for predicting site- and structure-specific response in a regional seismic risk assessment: Probabilistic Learning on Manifolds (PLoM) and Gaussian Process (GP), both available through SimCenter's quoFEM tool [20], and multivariate linear regression, included as a baseline.

The main objective of this study is twofold: 1) a quantitative comparison of the models in predicting both collapse and non-collapse behavior, and 2) a discussion of broader suitability considerations for regional use, including workflow setup, surrogate types and usage, uncertainty, and correlations in the input and output variables. Through these comparisons, this study aims to offer a practical guide for surrogate selection, focusing on the PLoM and GP implementations available through SimCenter's quoFEM tool, together with a linear regression baseline.

To compare models quantitatively, this study adopts the case study, underlying datasets, and the PLoM surrogate model framework from Lee et al. [21]. The current study extends this work by systematically comparing three surrogate models, introducing a validation dataset, evaluating an alternate strategy for selecting PLoM hyperparameters, and assessing their applicability at the regional level. While the focus is on three specific model implementations for a single concrete bridge typology, the general processes and considerations aim to be relevant to other surrogate models used for structural performance prediction in regional scenario studies.

2 MODEL SELECTION

This study compares three surrogate models: a multivariate linear regression model, the widely used Gaussian Process (GP) model, and the Probabilistic Learning on Manifolds (PLoM) model, which has been recently applied in PBEE applications. A brief overview of each model and the rationale for selection is provided here.

First, a multivariate linear regression model is selected as a baseline for comparison. In the context of PBEE, lognormal distributions are often used to represent relationships between variables, such as fragility functions mapping from EDPs to damage states [6]. Furthermore, other studies have used and adapted forms of multivariate linear regression as surrogate models for structural response prediction [5]. Thus, linear regression is used here as a baseline model to evaluate the relative performance of the GP and PLoM models in predicting quantities relevant to earthquake engineering and structural assessment.

This study also considers an implementation of the GP model, also known as Kriging in the field of geostatistics. GP is currently one of the most extensively used surrogate models in the field [14, 22, 23, 24]. A GP model is a nonparametric model that defines an output distribution over functions of input variables, and mathematically, it is fully specified by a mean function and a covariance (kernel) function [24]. For any finite set of points in the input space, the resulting function values would follow a multivariate Gaussian distribution. GP models do not assume a fixed functional form, and thus can model complex patterns in the data and naturally quantify different uncertainties (i.e., both aleatory and epistemic) in predictions [24]. The implementation of the GP model used in this study is available through the quoFEM application [20] from SimCenter [25] that builds on the GPy Python package [26] for basic GP modeling and calibration. An extension is available to model flexible prediction variances across the parameter space (i.e., a heteroskedastic GP model, termed stochastic emulation [27]), as opposed to a homoskedastic GP model, which assumes constant variance. Recent findings suggest that using a heteroskedastic GP model can improve seismic response predictions [28]. Both the homoskedastic and heteroskedastic GP models are considered in this study. Overall, the GP model is selected for this study because it is widely used, there are existing tools for implementation [20, 29], and it naturally captures variation and uncertainty in output variables, which is important in

PBEE applications.

Finally, this study adopts the PLoM surrogate model, which is based on the nonparametric probabilistic learning method originally developed by Soize and Ghanem [30, 31]. This model has been adapted for several earthquake engineering contexts, including Zhong et al. [10], which used PLoM to estimate structural responses for systems with variations in design/modeling parameters or ground motion characteristics, and Lee et al. [21], which extended the PLoM algorithm to simultaneously consider both variations in design/modeling parameters and ground motion characteristics. The main functionality of the PLoM model is that it takes an input dataset and generates synthetic realizations of additional data that preserve the original data structure while complying with user-defined statistical constraints [10]. Several steps depend on model hyperparameters, which are user-defined settings that control the model's structure or learning process that are not learned directly from the data. First, the input of the model is scaled and normalized to regularize the range of each dimension. Principal component analysis is then performed and a certain number of principal components are used, as defined by a threshold hyperparameter, ϵ_{db} . The joint probability is estimated using kernel density estimation with a Gaussian kernel. Before sampling, the local concentration of the data is preserved by applying diffusion maps, dependent on a user-specified model hyperparameter, ϵ_k . In the final step of the PLoM procedure, new samples are generated via a Markov Chain Monte Carlo simulation. At each iteration, the method adjusts samples to simultaneously minimize the Kullback–Leibler divergence from the original dataset and satisfy user-specified constraints. Finally, the new samples are then transformed back to the original coordinate system. A more in-depth description of the PLoM algorithm can be found in Zhong et al. [10], and more details on how it is applied in the context of bridge earthquake response prediction can be found in Lee et al. [21]. The PLoM model is selected for the comparison study because it has been shown to have strong performance [10, 21], there are existing tools for implementation [20], and it captures distributions and correlations between variables.

Across these three models, PLoM behaves quite differently than the others. Models like GP and linear regression directly infer output values conditional on a given input point, i.e., they receive an input point and generate the corresponding output. The selected PLoM algorithm in this study is not a "traditional" surrogate, meaning it does not make specific output predictions for specific input points. Instead, based on the specified distribution of input data and user-defined constraints, PLoM generates a distribution of entirely new data points, none of which correspond directly to a specific single input point. This model setup, while it lends itself well to preserving data structures and correlations between different dimensions in the data, changes the way the PLoM model can be used and evaluated. Further discussion comparing the setup and use cases of these three models is included in Section 6.

3 REGIONAL HIGHWAY BRIDGE APPLICATION AND DATASET

Transportation networks are inherently distributed systems, and across California thousands of highway bridges are located in seismically active regions. These bridges exhibit substantial variability in structural design, while also being subjected to spatially varying ground motions due to site-specific seismicity. As a result, this infrastructure system naturally encompasses both variability in structural parameters across bridges and in ground motion characteristics across locations, making it a realistic and challenging test case for site- and structure-specific surrogate modeling. Thus, this study adopts the case study from Lee et al. [21], which focuses on pre-1971 two-span single-column concrete highway bridges in California subjected to seismic loading. Figure 1a shows the spatial distribution of highway bridges in the Bay Area from the National Bridge Inventory, while Figure 1b presents the bridge typology used for comparison of the three surrogate models considered in this study.

3.1 Structural Model and Selected Parameters

The surrogate models in this study aim to be generalizable across the selected bridge typology. To train such surrogates, a sufficiently comprehensive suite of structural models describing bridge designs in this era is required. To achieve this task, Lee et al. [21] develop a python package that can automatically sample bridge design parameters and generate corresponding OpenSees models. The modeler takes in three inputs: (1) a bridge information file, (2) a bridge design database file, and (3) the number of models to create. Based on the bridge era and other user-provided design constraints, the bridge modeler uses Monte Carlo sampling to generate realizations of bridge designs. Samples are drawn from underlying statistical distributions developed by Mangalathu [32] for reinforced concrete bridges in California, as provided in the bridge design database file.

The bridge modeler generates continuous-span bridge models consisting of detailed fiber-based reinforced concrete column elements and simplified girder and foundation elements. The girders are modeled using a series of elastic beam-column elements to better capture possible flexural and torsional response. Distributed translational and rotational masses are assigned to the girder elements. At the ends of the girders, the boundary conditions are simplified to roller supports. The bridge column model consists of one *Rigid* element representing the geometry of the girder,

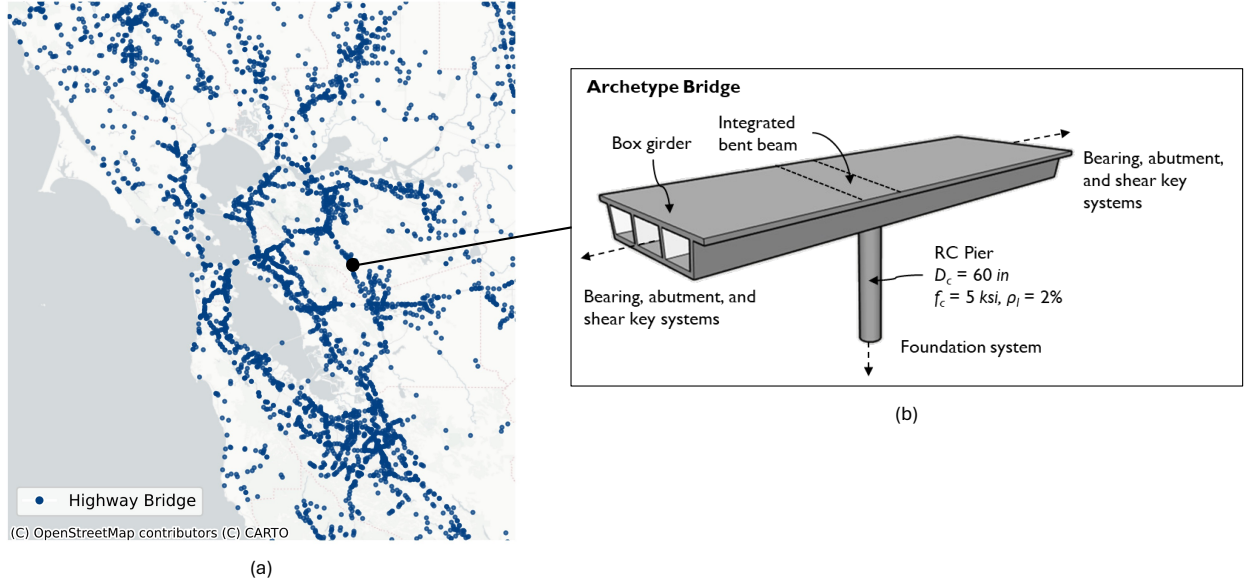


Fig. 1. (a) Highway bridge locations in the Bay Area from the National Bridge Inventory. (b) Schematic of the bridge typology used in this study.

two 3D bond-slip *Zero-Length Section* elements, and one fiber element with aggregated flexure and shear sections. Following the study by Zhao and Sritharan [33], the *Bond_SP01* material is used to simulate bond slip responses. The *DuctileFracture* material is used along with the *ReinforcingSteel* material to simulate stress-strain response and ductile fracture of steel reinforcement. The *Concrete02* material is used to simulate concrete stress-strain response where the confined concrete model is based on Mander et al. [34]. The concrete materials are regularized to eliminate mesh-dependency in the fiber element method [35]. Beneath the column element is a concentrated soil-foundation hinge whose responses are condensed in two translational springs and two rotational springs, and the torsional DOF is fully fixed. A schematic of the OpenSees model is shown in Figure 2a.

Using this package, Lee et al. [21] sample a suite of more than 600 two-span, single-column concrete bridge models representing the pre-1971 era. While there are 25 structural parameters that are defined in the models, only four are recorded for use as surrogate model inputs. The four selected parameters, slenderness ratio, axial load ratio, longitudinal reinforcement ratio, and longitudinal reinforcement bar yield strength are selected based on a sensitivity study using Sobol indices [21]. Histograms of the four selected structural parameters across the sampled bridge designs are shown in Figure 2b, which demonstrates that even within the single selected typology, there are significant differences across possible bridge designs. Several of the selected parameters are derived from other parameters; for example, slenderness ratio is column height over the diameter. Thus, while the underlying distributions of the bridge parameters come from Mangalathu [32], those of the computed parameters shown in Figure 2b are based on their components. All sampling in the concrete bridge modeler is conducted independently for each parameter because correlation information is not available. Thus, it is possible that some bridge designs represent unrealistic combinations of parameters. This study assumes these designs do not impede model accuracy, prioritizing complete input-space coverage.

3.2 Ground Motion Selection and Response History Analysis

Beyond bridge design variability, the surrogate models also aim to be site-specific, capturing the variability in seismicity across California. Recorded ground motions, which are described completely by their time histories, can be summarized using various intensity measures. Following previous work on surrogate models, this study adopts an existing gridded set of recorded ground motions, which aims to capture variability in frequency content and duration [5, 21]. More specifically, the ground motion set describes spectral shape using *SaRatio* [36], defined as the ratio of the geometric mean *Sa* over a period range to the *Sa* at a single conditioning period (here, 1.0 seconds), and ground motion duration using D_{S5-75} [37]. This study focuses on ground motions with recorded D_{S5-75} values of less than 40 seconds to be reflective of California. The gridded ground motions are later scaled to provide coverage of possible earthquake intensities.

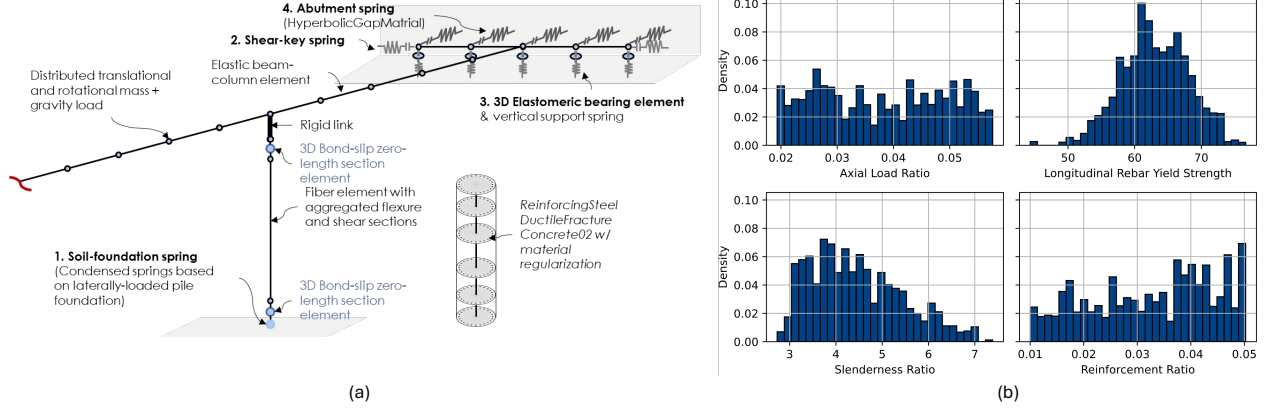


Fig. 2. (a) OpenSees bridge model developed using the concrete bridge modeler. (b) Histogram of values across sampled bridge designs for selected structural parameters (X_S).

Lee et al. [21] randomly pair the gridded ground motions with the sampled OpenSees models and perform NLRHA using the incremental dynamic analysis (IDA) procedure [38], where ground motions are incrementally scaled based on spectral acceleration at a fixed period of 1.0 seconds (referred to as S_a) to the point of structural collapse. This procedure produces a dataset of NLRHA results used to train and evaluate the surrogate models, consisting of peak responses and the intensity at bridge collapse ($S_{a_{collapse}}$) for each ground motion. In this study, the EDPs of interest include the maximum curvature (κ_{max}) and maximum fracture index (FI_{max}) at the base and tip of the column. While these are not the only metrics for measuring bridge performance, κ_{max} is selected because it is commonly used to represent structural damage and is a key input in determining curvature ductility, which is used in the Caltrans-defined damage state fragility functions [39, 32]. FI_{max} is also considered because, unlike κ_{max} , it is a cumulative-damage response parameter.

Figure 3a shows the gridded ground motions used for IDA and Figure 3b shows the IDA results as reported in Lee et al. [21]. All non-collapse data points represent scaled intensities at which the paired bridge and ground motion did not collapse, and each collapse data point represents the intensity at which the curvature exceeds the collapse limit. Because IDA produces only a single collapse point for each bridge/ground motion pairing, the collapse dataset is much smaller than the non-collapse dataset.

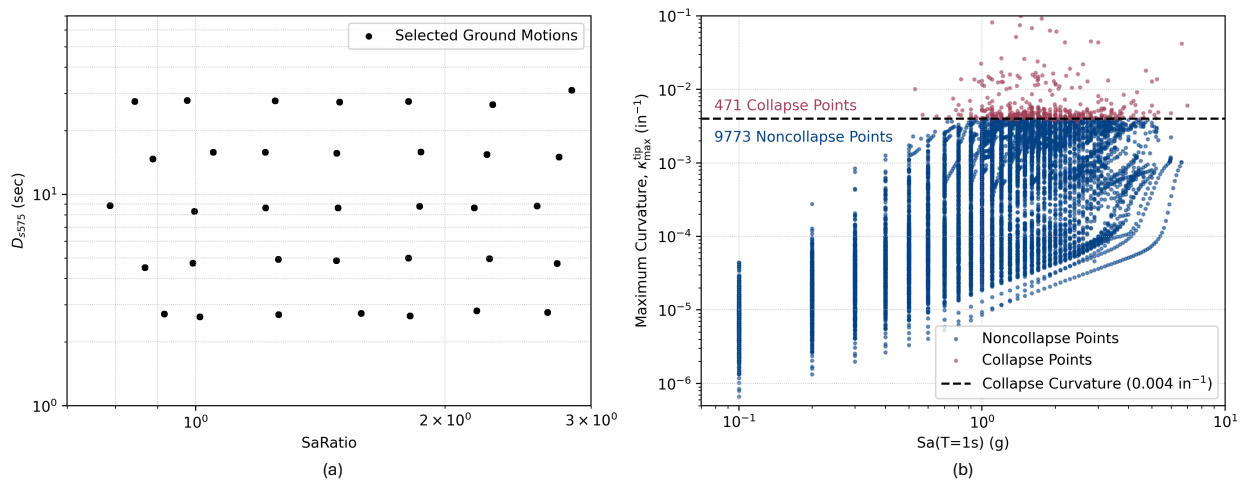


Fig. 3. (a) Gridded ground motions used for IDA. (b) IDA scaling of S_a and the resulting κ_{max}^{tip}

While the current study adopts the dataset from Lee et al. [21], there are other sampling methods by which to

generate data for surrogate model development, including random sampling, Latin Hypercube Sampling, which targets space-filling and can be more efficient than random sampling for high-dimensional problems, and adaptive sampling, which uses results from the model to iteratively guide sampling, among many other possible strategies [40]. The topic of sampling strategies for surrogate model development is not fully described here, and there is a substantial amount of work done on this topic.

3.3 Dataset Summary and Description

Following Lee et al. [21], the predictions of collapse and non-collapse EDP performance are separated into two surrogate models used in sequence to assess the full probabilistic performance of a bridge. Each surrogate has structural and ground motion input parameters characterizing the bridge and ground motion of interest, and appropriate output parameters, depending on either collapse or non-collapse EDP prediction. All input and output parameters are listed in Table 1. All data is transformed into log space in both datasets for use in the PLoM and GP models, and comparing log and linear space is a modeling choice for linear regression.

TABLE 1. Collapse and Non-Collapse EDP Surrogate Inputs and Outputs

	Collapse Surrogate Model	Non-Collapse EDP Surrogate Model
Structural Input Parameters	$X_S = [\text{Axial Load Ratio, Reinforcement Ratio, Slenderness Ratio, Longitudinal Reinforcement Yield Strength}]$	$X_S = [\text{Axial Load Ratio, Reinforcement Ratio, Slenderness Ratio, Longitudinal Reinforcement Yield Strength}]$
Ground Motion Input Parameters	$X_{GM}^{col} = [Sa_{Ratio}, D_{S5-75}]$	$X_{GM}^{EDP} = [Sa(T=1s), Sa_{Ratio}, D_{S5-75}]$
Output Parameters	$X_R^{col} = [Sa_{collapse}]$	$X_R^{EDP} = [\kappa_{max}^{tip}, \kappa_{max}^{base}, FI_{max}^{tip}, FI_{max}^{base}]$

Collapse prediction can be naturally formulated as a binary outcome (collapse or non-collapse). Although surrogate models exist for binary classification problems (e.g., logistic regression), the collapse models developed in this study are based on the same surrogates used for the non-collapse EDP case (PLoM, GP, linear regression). Leveraging the IDA curves, the collapse problem is reformulated from a binary to a continuous output variable by defining the collapse surrogate output as the intensity at which a bridge with the specified structural and ground motion characteristics collapses, $Sa_{collapse}$. This allows for the use of consistent surrogate types for both collapse and non-collapse.

Figure 4 shows scatter plots of several of the collapse input parameters against the output value, showcasing Sa_{Ratio} (Figure 4a), $Axial Load Ratio$ (Figure 4b), and $Reinforcement Ratio$ (Figure 4c), which are the three inputs with the highest magnitude pairwise linear correlation with $Sa_{collapse}$. Figure 4d shows a scatter of Sa and κ_{max}^{tip} . Among the input parameters, Sa has the highest pairwise linear correlation coefficient with the outputs, ranging from 0.21-0.39 across the four output parameters. The four output dimensions (X_R^{EDP}) are also highly correlated with one another, with values ranging from 0.35 to 0.8. Figure 4e shows that FI_{max}^{tip} has a strong cluster around 1.0. As defined in the numerical simulations, the bar fracture has a median fracture index value of 1.0; although fracture index can continue to increase beyond this point, load typically transfers away from fractured bars once fracture occurs, keeping many values close to 1.0. Figure 4f shows the relationship between two output parameters, κ_{max}^{tip} and FI_{max}^{tip} . At lower intensities, these values are highly correlated as the bridge is in the linear elastic range, while there is more scatter at higher intensities due to the highly nonlinear bridge response under large intensity earthquakes.

3.4 Surrogate Training, Validation, and Test Datasets

It is common practice in machine learning to separate available data into training, validation, and test sets [41]. *Training data* is typically the largest portion of the data and is used by the model to learn patterns from the dataset. *Validation data* is used to evaluate the model's performance during the training process; specifically, this data is commonly used to tune hyperparameters and optimize model performance. Finally, *test data* is used to evaluate the final model's performance on completely unseen data. The split between training, validation, and test data is important to prevent overfitting of the model, provide a final, unbiased evaluation of the model's performance, and appropriately tune any hyperparameters required within each model [41]. Hyperparameters can relate to different features of each model, and tuning them properly is crucial for good predictive performance [41]. This study follows the train/validation/test set split, and an overview of the data split is provided in Figure 5.

20% of the non-collapse IDA data is split into six non-overlapping test sets, which are reserved as unseen data for final model assessment (shown in blue in Figure 5). The remaining data is used to form six complementary sets of validation and training data, each containing 5% validation and 95% training data (shown in darker and lighter purple in Figure 5). Rather than sampling randomly, test and validation sets are constructed to be reflective of how the model

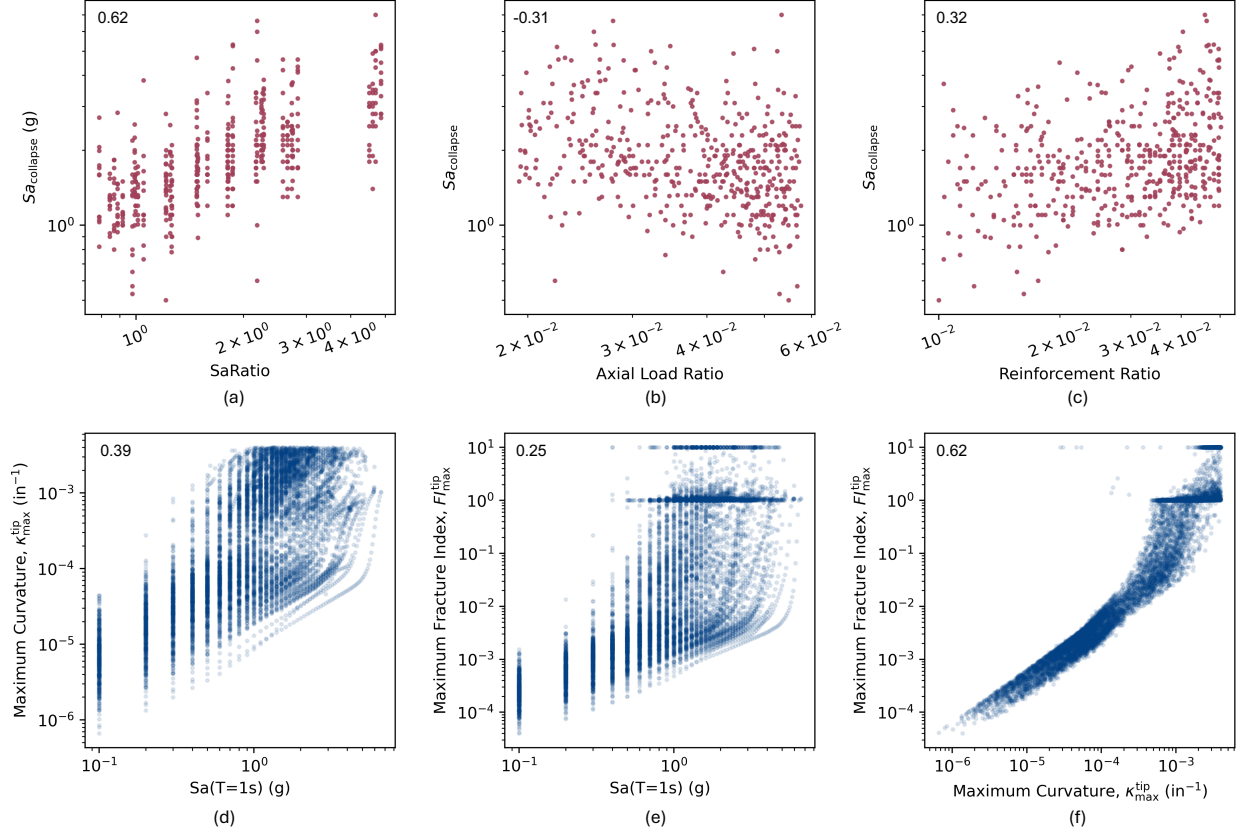


Fig. 4. Scatter plots and pairwise linear correlation coefficients for various dimensions of the collapse dataset (a-c) and non-collapse dataset (d-f).

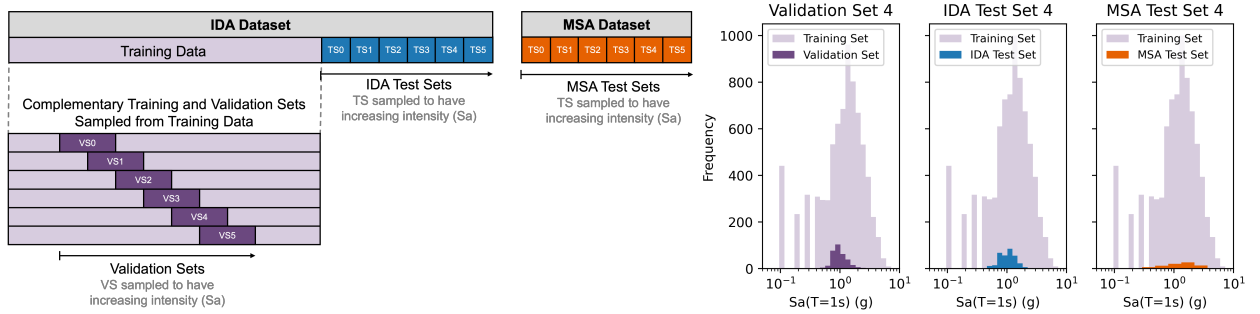


Fig. 5. Overview of training, validation, and test sets for non-collapse surrogate model development using IDA and MSA data.

would be used in a regional earthquake scenario study. In a true "use case," the model would be interrogated for a given earthquake scenario, corresponding to a narrow band of input intensity values ($Sa(T = 1s)$). Thus, both test and validation data are sampled to be lognormally distributed around a single specified value of earthquake intensity, which increases for each set. A similar sampling technique is introduced in Lee et al. [21] to examine the dependence between earthquake intensity (Sa) and the optimal PLoM hyperparameters. Histograms of a validation and test set from the IDA data are shown in Figure 5. Validation sets are referred to as VS0 through VS5, and test sets are TS0 through TS5.

Similar methods are used to develop training, validation, and test sets for the collapse data. IDA produces many more non-collapse points than collapse points, so 20% of the collapse data is reserved as only two test sets, with target median Sa values at the 40th and 60th quantiles of the collapse dataset, each with a dispersion of 0.3 (referred to as TS40 and TS60). The remaining data is used to sample 10 complementary validation and training sets: five at the 40th quantile and five at the 60th, each validation set containing 15% and each corresponding training set 85% of the remaining collapse data. These two quantiles were selected to be as far apart as possible, while maintaining a sufficient number of data points within each set, and multiple validation sets were generated at each level to reduce model sensitivity to small training samples [21]. Validation sets sampled at the 40th quantile are denoted VS40_0 through VS40_4, and those at the 60th quantile are denoted VS60_0 through VS60_4.

For both the collapse and non-collapse models, hyperparameters are tuned using the complementary training and validation sets. A final model for each of the three surrogate types is then trained on the full pool of IDA training data (i.e., all data not reserved for testing). The three models are then compared using the *IDA Test* datasets reserved for this purpose. The separation of training, validation, and test data provides a robust way to tune hyperparameters, avoid overfitting, and provide an unbiased evaluation of the model's performance relative to the IDA data [41].

However, in the context of PBEE, IDA is not the only format used to assess seismic performance. A robust methodology for evaluating performance is multiple stripe analysis (MSA), where site-specific ground motions are selected and scaled at each return period of interest and used to conduct NLRHA for a specific bridge. For a surrogate model that aims to pick up both site- and bridge-specific characteristics, performance would be considered strong if the model is able to replicate the results of an MSA. Thus, in addition to the IDA-based test sets, a set of test data is developed using the MSA procedure. This second set is an out-of-distribution evaluation, where the test dataset comes from a different distribution than the training dataset (MSA instead of IDA), which assesses model generalization. Distribution shifts between training and real model use are common in "in-the-wild" applications and can degrade model accuracy [42]. Thus, it is important to include MSA data to evaluate surrogate robustness.

The *MSA Test* data used in this study is reflective of a specific bridge design under site-specific conditions in Long Beach, CA (Structure 1, Site 3 from Lee et al. [21]). Six seismic hazard levels are adopted as six *MSA Test* sets for the surrogates, corresponding to return periods of 224, 475, 975, 2475, 4975, and 9950 years. At each hazard level, surrogates are compared to NLRHA results obtained from MSA using 100 ground motions. These ground motions are selected and scaled to match the target conditional mean spectrum and the median significant duration for the site and return period of interest. Each ground motion is paired with an OpenSees model representing a unique bridge design realization, with the ensemble of models sampled to capture both modeling and structural parameter uncertainties representative of a realistic bridge. Figure 6 provides an overview of the MSA data, including the proportion of collapse cases per stripe, and several sample scatter plots showing the relationship between various dimensions as stripe intensity increases.

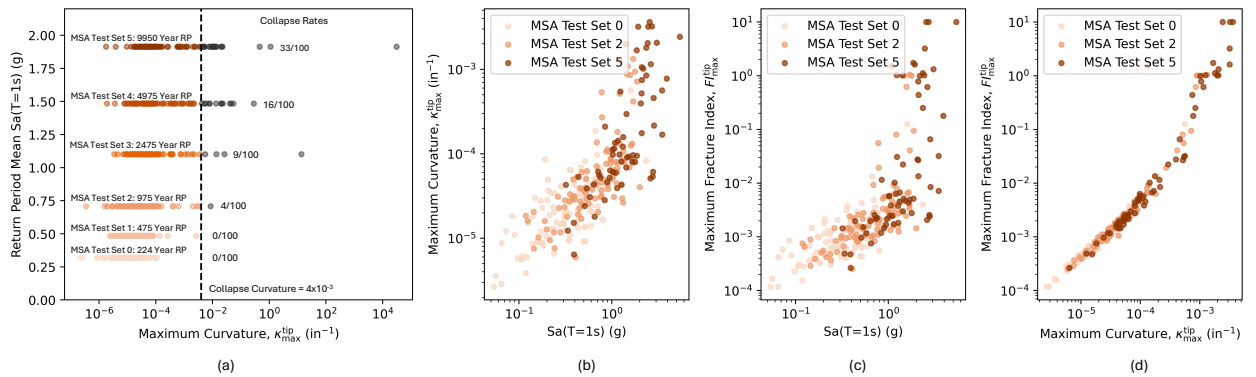


Fig. 6. MSA data used for out-of-distribution evaluation of surrogate models. (a) Proportion of collapse cases per test set. (b) $\kappa_{\max}^{\text{tip}}$ versus Sa for three sample test sets. (c) $FI_{\max}$ versus Sa for three sample test sets. (d) $\kappa_{\max}^{\text{tip}}$ versus $FI_{\max}$ for three sample test sets. Figure adapted from Lee et al. [21].

In summary, the three surrogates are evaluated and compared using the *IDA Test* sets (shown in blue in Figure 5) and the *MSA Test* sets (shown in orange in Figure 5). The test sets have three key differences: 1) the IDA data represents a broad set of bridge structures and ground motions, whereas the MSA data represents a specific bridge and

site-specific ground motions for a given return period, 2) the data in both sets are generated using NLRHA, but they are generated using different methodologies (IDA versus MSA), which may change its characteristics, and 3) the IDA data is sampled from the same dataset as the training and validation data, whereas the MSA data is generated separately.

3.5 Evaluation Metrics

Broadly in this study, surrogates are evaluated relative to the results of the NLRHA conducted using the IDA and MSA methodologies. Using OpenSees NLRHA results as a source of "ground truth" is not perfect, as OpenSees models have modeling uncertainty and potential inaccuracies. However, the goal of these surrogates is to replace this type of analysis for bridge structures at the regional scale, and thus, replicating results of a NLRHA is considered to be an appropriate way to measure performance.

Most surrogate models predict output values for a specified input point, enabling the use of point-specific evaluation metrics such as mean squared error (MSE). However, as discussed in Section 2, the PLoM model operates at the distribution level. As a result, traditional point-wise error metrics such as MSE are not suitable, and a distribution-to-distribution discrepancy metric is required to compare predicted and reference output distributions. Although GP and linear regression could be evaluated using point-wise error metrics, this study instead evaluates all three models by comparing their predicted output distributions against the reference distributions from the validation and test sets. This is done for two primary reasons. First, the PLoM model requires distribution-level assessment, meaning distribution-level metrics enable consistency and comparability across model types. Second, in a regional seismic risk assessment (the intended use case of the surrogates), uncertainty in the ground motion and structural parameters results in distributions in both inputs and outputs at a given site and bridge. Thus, capturing distribution-level performance indicates ability to capture site- and structure-specific behavior at the regional scale.

This work adopts the two-sample Kolmogorov–Smirnov (KS) statistic for this purpose (Equation 1), which measures the maximum absolute difference between two empirical cumulative distribution functions (CDFs). In this study, the KS statistic is used as a descriptive measure of distributional discrepancy between surrogate response predictions, denoted generically as X_R for the collapse and non-collapse surrogates, and corresponding OpenSees responses, rather than as a formal hypothesis test.

$$D_j = \max_x |F_{1,j}(x) - F_{2,j}(x)| \quad (1)$$

Where:

- D_j is the KS statistic computed for the j -th response variable in X_R
- $F_{1,j}(x)$ is the empirical CDF of the j -th response variable in X_R as predicted by the surrogate model
- $F_{2,j}(x)$ is the empirical CDF of the j -th response variable in X_R obtained from the OpenSees model

To obtain a single summary error metric for model comparison, D_j is averaged where applicable, yielding a mean KS statistic to use for comparison, $\bar{D}$. For the collapse model, since there is only a single output parameter in X_R^{col} , $\bar{D}$ is computed by averaging across validation or test sets, as appropriate. For the non-collapse model, $\bar{D}$ is computed by averaging D_j across the four output parameters in X_R^{EDP} and across validation or test sets, as appropriate.

The KS statistic provides a summary of discrepancy at the distribution level between surrogate predictions and OpenSees results. However, the KS statistic alone does not fully capture differences in central tendency, spread, or distributional shape. Furthermore, the KS statistic calculations are conducted separately for each response variable in X_R ; thus, achieving perfect fit would mean perfectly predicting the marginal distributions, not necessarily the joint distribution across output parameters. Due to these limitations, several secondary diagnostic measures are used to qualitatively compare performance, including the median and standard deviation of predicted versus true output distributions, as well as scatter plots and empirical copula plots to examine the joint (rather than marginal) distribution of output parameters. The empirical copula is obtained by transforming each marginal to a uniform distribution between [0, 1], isolating the dependence structure between variables independent of their marginal distributions.

4 CASE STUDY RESULTS: PREDICTION OF EDPS CONDITIONED ON NON-COLLAPSE

The development of the non-collapse EDP surrogate models involves selecting appropriate hyperparameters and other modeling decisions using complementary training and validation datasets, then assessing final model performance using the IDA and MSA Test Sets.

4.1 PLoM EDP Model Development

The primary task involved in training a PLoM surrogate model is to select appropriate hyperparameters (ϵ_k and ϵ_{db}). As demonstrated in Lee et al. [21], PLoM performance is highly sensitive to the choice of hyperparameters, underscoring the importance of appropriate selection.

A defining feature of the model in Lee et al. [21] is the use of Sa -dependent PLoM hyperparameters. This approach reflects how $X_R EDP$ evolves as the system transitions from elastic to highly nonlinear behavior; consequently, hyperparameter values that yield good performance at low intensities may be less appropriate at higher intensities. In contrast, the present study adopts a more conventional hyperparameter tuning approach; global values of ϵ_k and ϵ_{db} are explored via grid search across validation sets, similar to K-fold cross-validation.

There are tradeoffs in these two approaches for selecting PLoM model hyperparameters. The grid-based approach offers a transparent, repeatable, and fully automated selection criterion; however, using $\bar{D}$ as the hyperparameter selection metric deprioritizes PLoM's ability to capture the full correlation structure, which is not measured by the KS statistic. Conversely, the Sa -dependent functional hyperparameters may better capture correlation structure and performance across intensities, but this approach relies on semi-qualitative, manual evaluation of hyperparameter performance at each intensity level, which is time-intensive and requires user judgment that may not transfer easily to other use cases. Because a primary aim of this study is to evaluate the practical usability of surrogate models for regional seismic risk assessment, this study adopts the grid search method for selecting hyperparameters. We note that this may understate PLoM's best achievable performance relative to the approach used in Lee et al. [21].

Figure 7a shows the empirical CDFs of $\kappa_{\max}^{\text{tip}}$ and $FI_{\max}^{\text{tip}}$, as predicted given each hyperparameter pair included in the grid search. VS0 (the lowest intensity validation set) has less variability in the empirical CDFs than VS5, demonstrating that PLoM sensitivity to hyperparameters increases as the earthquake intensity (Sa) increases. Furthermore, the empirical CDF for VS5 $FI_{\max}^{\text{tip}}$ shows an example of how PLoM can capture multi-modal distributions given appropriate model hyperparameters. Figure 7b shows the median $\kappa_{\max}^{\text{tip}}$ and $FI_{\max}^{\text{tip}}$, indicating accurate median predictions for VS0–VS4 and underestimation at VS5, likely caused by the model's limited ability to fully reproduce the kinked shape of the $FI_{\max}$ distribution.

The selected hyperparameters that minimize $\bar{D}$ are $\epsilon_{db} = 0.01$ and $\epsilon_k = 5.57$, which are referred to as the "best pair" and are adopted for subsequent analyses. Additionally, a quantitative comparison shows that the PLoM model with the selected best pair of hyperparameters was comparable to the PLoM model using the Sa -dependent hyperparameters from Lee et al. [21] across validation sets, supporting the grid-search based approach adopted here.

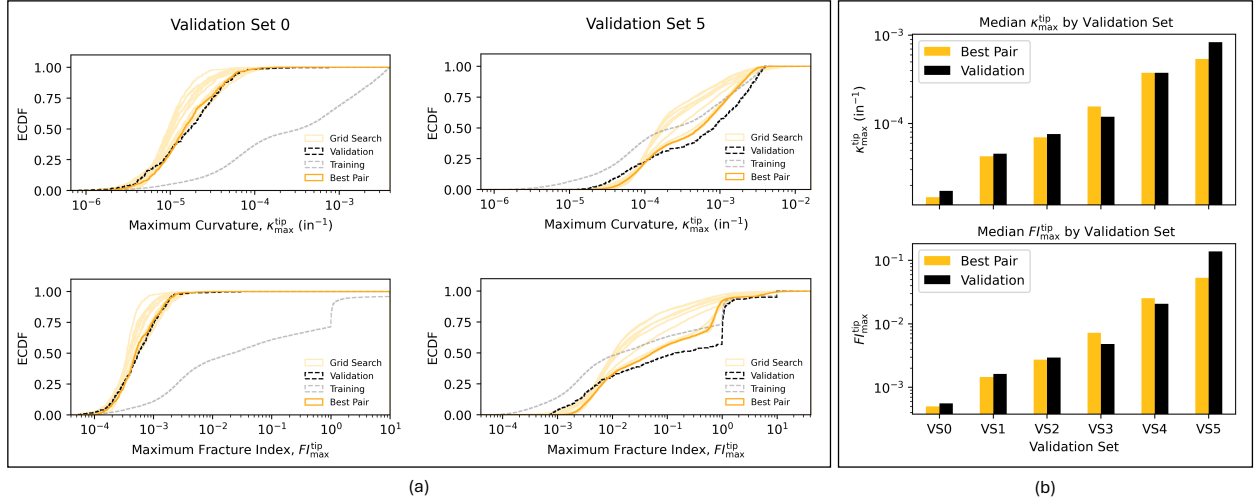


Fig. 7. Comparison of PLoM Models. (a) Empirical CDF plots for VS0 and VS5 comparing all PLoM predictions resulting from the hyperparameter grid search against the training and validation sets. (b) Median of $\kappa_{\max}^{\text{tip}}$ and $FI_{\max}^{\text{tip}}$ as predicted by the PLoM model with the selected best pair of hyperparameters against the validation data.

4.2 GP EDP Model Development

Several GP variations are considered in this study, including variation in the nugget, which is used to modify the covariance matrix to provide improved numerical stability and model noise in the observed data, and variation in the output sampling method. Two possible nuggets are evaluated, a homoskedastic nugget, which is a single scalar noise variance that is assumed constant across all training points, and a heteroskedastic nugget, which allows for noise to vary across the input training points. For a given input point, the GP outputs a median and variance value, and three possible sampling techniques are considered in this study: 1) to adopt the output median as the prediction for the specified input point, 2) to sample from the output distribution once as the prediction of the specified input point, and 3) to sample from the output distribution ten times and obtain multiple prediction samples for a single specified input point. For multi-output sampling, following a typical GP implementation, we assume that each output is independent at a given test point, and therefore, the sampling for 2) and 3) is conducted independently for each output. By pairing each nugget and sampling method, there are six possible GP model setups that are evaluated and compared.

Figure 8a shows that GP models perform similarly across nugget/sampling variants and validation sets for $\kappa_{\max}$ prediction. For $FI_{\max}$, patterns differ by intensity. For VS0, median-based sampling (regardless of nugget selection) outperforms random sampling, which overestimates standard deviation (Figure 8b). This could be partly explained by the fact that sampling introduces additional uncertainty into the outputs, as compared to the deterministic median selection. At higher intensities (VS5), the models converge, with less differentiation between output sampling methods, with median-based models slightly underestimating standard deviation (Figure 8b). Beyond these effects, all GP models struggle somewhat in the prediction of $FI_{\max}$ due to the shape of the underlying distribution, as shown in Figure 8a. GP models predict smooth responses, and thus these models are limited in their ability to describe the shape of the distribution accurately. However, all models capture median $FI_{\max}$ fairly well (see Figure 8b), even though they do not characterize the shape of the distribution well.

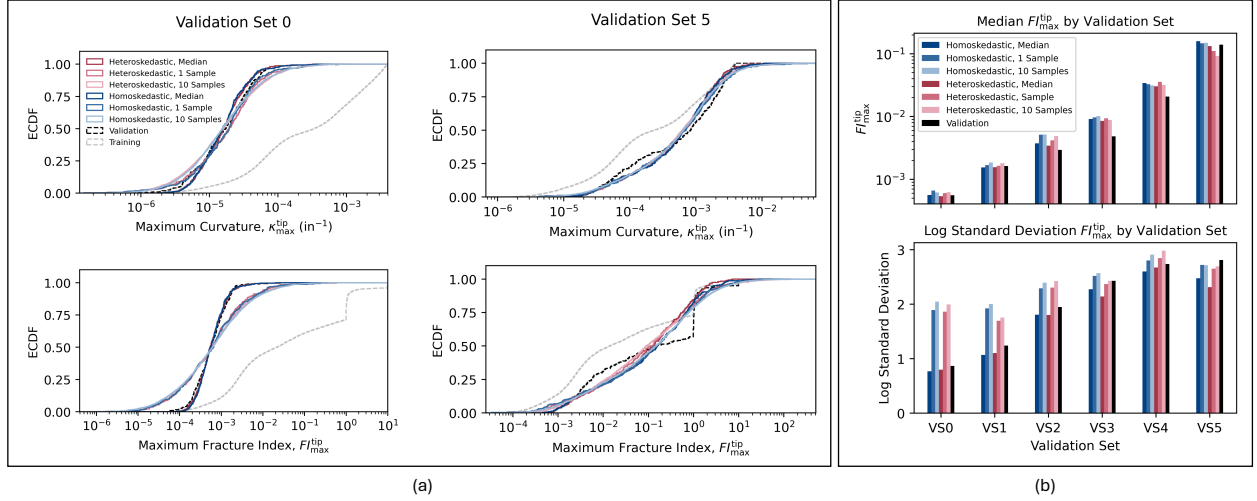


Fig. 8. Comparison of GP Models. (a) Empirical CDF example plots for VS0 and VS5 comparing GP predictions from six variants against the training and validation sets. (b) Median and standard deviation of $FI_{\max}^{\text{tip}}$ as predicted by the GP models against the validation data.

Figure 9, which shows the KS statistic, reflects the same patterns. The six models perform somewhat similarly when predicting $\kappa_{\max}$ (Figure 9a). For the prediction of $FI_{\max}$ at lower intensities (VS0-VS2), models that sample the output distribution rather than using the median have much higher KS statistics (higher error), as shown in Figure 9b. Meanwhile, the heteroskedastic GP outperforms the homoskedastic GP for $\kappa_{\max}$ at higher intensities (VS3-VS5), suggesting that the true data in these validation cases likely exhibit heteroskedasticity.

The GP model with the heteroskedastic nugget with the output sampled at the median had the lowest $\bar{D}$, and thus is selected for future use. The fact that sampling at the median performs best indicates that, given the optimal selection of IMs and bridge parameters, the nugget variance is small compared to the variability in input variables.

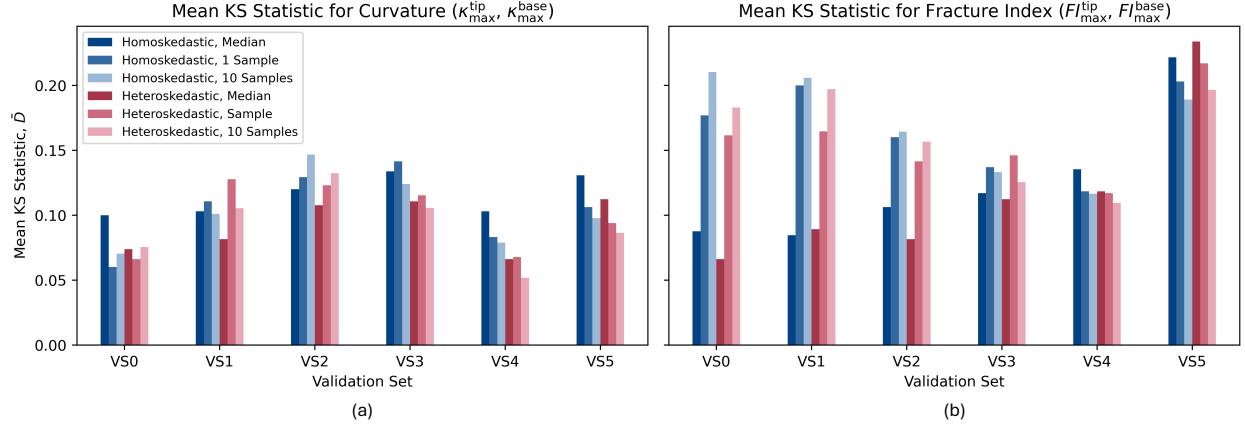


Fig. 9. Comparison of mean KS statistic for $\kappa_{\max}$ and $FI_{\max}$ across six GP variants.

4.3 EDP Linear Regression Model Development

Multivariate linear regression estimated using ordinary least squares (OLS) is adopted as a baseline point of comparison. Two formulations are compared: (1) a linear regression model in the original scale of the input and output variables and (2) a linear regression model in which both inputs and outputs are transformed to log space. For each formulation, six regression models are fit using the training sets and evaluated using the six complementary validation sets. OLS does not require tuning hyperparameters, so the validation set results are primarily used to select between the linear and log regression models.

Outputs demonstrate that log space is much more appropriate than linear space for this task. As shown in Figure 10, although the regression in log space cannot capture all of the nonlinearities in bridge performance, it far outperforms regression in linear space, and is thus selected for future use. This is not a surprising result because lognormal distributions have long been used in the field of PBEE to capture relationships between ground motion and structural parameters and damage states. For simplicity in terminology, the selected model will be referred to as ‘linear regression’ for the remainder of this paper.

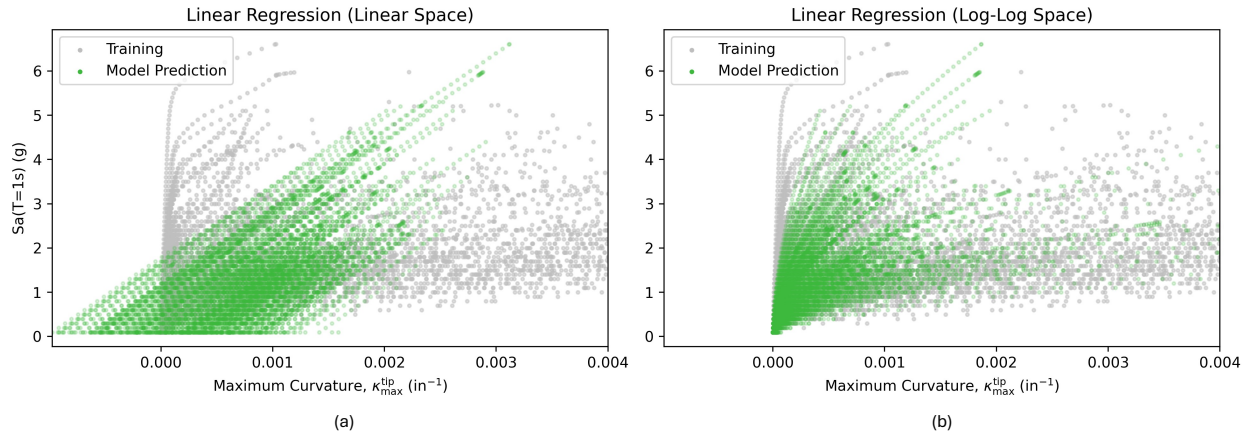


Fig. 10. Comparison of Linear Regression Models. (a) Example EDP Prediction using multivariate linear regression. (b) Example EDP Prediction using multivariate linear regression in log space.

4.4 Test Set Comparison Across Non-Collapse EDP Surrogate Models

The selected models of each type – PLoM (grid search hyperparameters), GP (heteroskedastic nugget, median sampling), and linear regression (log space) – are compared using the *IDA Test*, which is created from the same

database as the training and validation data, and the *MSA Test*, which is more reflective of a "true" use case.

Figure 11 shows the results of the *IDA Test*, with empirical CDFs shown for TS0 and TS5 in Figure 11a and the computed KS statistics in Figure 11b. Based on these plots, all three models perform better at lower intensities than at higher intensities; this is reflected in the ECDFs matching better in TS0 than in TS5 (11a), as well as the lower KS statistic values for all models for the lower intensity test sets across both $\kappa_{\max}$ and $FI_{\max}$ (11b). This may be because at low earthquake intensities, bridges behave more elastically, so data is more constrained, whereas at higher intensity earthquakes, bridges behave more nonlinearly and thus, behavior becomes harder to predict.

Figure 11 also shows that the GP and PLoM models outperform the linear regression model in predicting $\kappa_{\max}$, particularly at higher earthquake intensities. In the TS5 example test set, it is clear from the ECDFs that linear regression struggles to predict $\kappa_{\max}$ under high-intensity ground motions, and this is reflected in the large KS statistic for the linear regression model under TS4 and TS5. Based on Figure 11b, GP best predicts $\kappa_{\max}$ for this case study, followed closely by the PLoM model.

Similar patterns exist for the $FI_{\max}$ predictions. GP and PLoM both outperform linear regression in predicting $FI_{\max}$ across intensities (Figure 11b). Although the GP and PLoM models have similar KS statistics for $FI_{\max}$, they behave in different ways. As discussed previously, the GP model cannot replicate the large kink in the shape of the $FI_{\max}$ distribution, whereas PLoM is able to replicate the shape somewhat (see IDA Test Set 5, Figure 11a). In doing so, the PLoM model underpredicts the median, which the GP model predicts more accurately.

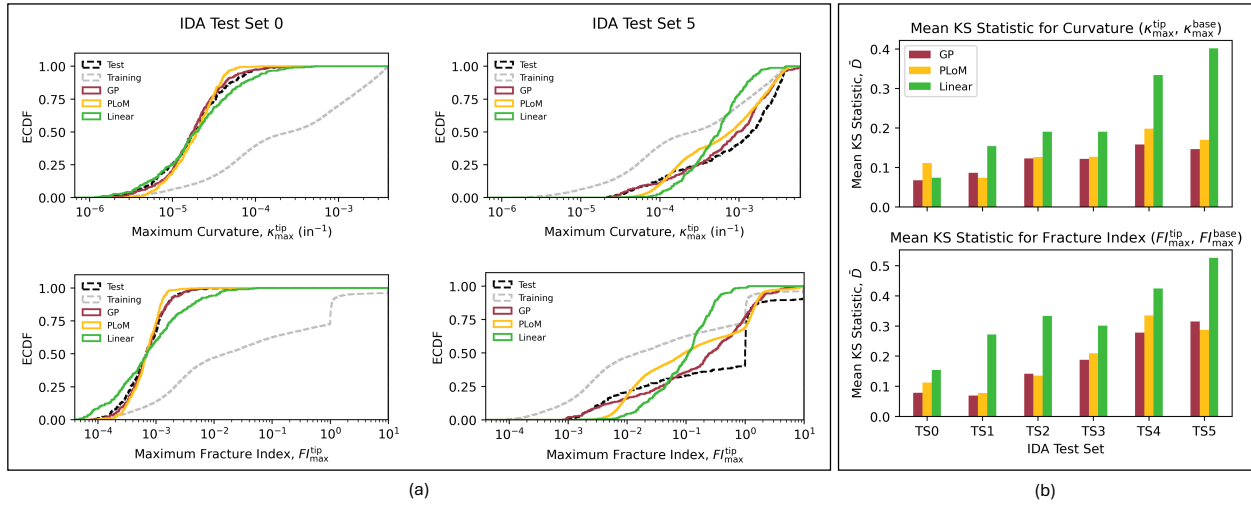


Fig. 11. Model comparison using *IDA Test*. (a) Empirical CDF example plots for TS0 and TS5 comparing predictions from the three surrogates against training and test sets. (b) Comparison of KS Statistic for each surrogate's prediction of $\kappa_{\max}$ and $FI_{\max}$ across test sets.

Moving beyond KS statistic, Figure 12a shows a scatter of the $FI_{\max}^{\text{tip}}$ and $\kappa_{\max}^{\text{tip}}$ and Figure 12b shows the corresponding empirical copula plots for the PLoM, GP, and linear regression predictions against the test set. Together, these plots demonstrate that PLoM more effectively captures the correlations and shape of the joint distribution in the output space of the test set. The GP and linear regression models pick up overall patterns, but correlation between output variables is not explicitly handled.

The results of the *MSA Test* are shown in Figure 13, which mirrors the *IDA Test* results shown in Figure 11. In contrast to the *IDA Test*, trends in model performance are less distinct in the results of the *MSA Test*. In particular, model performance is not necessarily best at lower intensities.

Linear regression exhibits similar behavior to that observed in the *IDA Test*. While it captures overall trends in the EDP response, it consistently struggles to reproduce the empirical CDFs as reflected by comparatively high KS statistics. The GP model demonstrates more consistent performance across earthquake intensities. Although it does not reproduce the full shape of output distributions, GP generally captures the median EDP response well. In contrast, the PLoM model shows the clearest difference in performance between the *IDA* and *MSA Tests*. Unlike in the *IDA Test*, PLoM struggles to predict EDPs at lower intensity levels in the *MSA Test* data. This behavior is evident from both the higher KS statistics for lower-intensity test sets and the empirical CDF comparisons for TS0 (Figure 13a), where

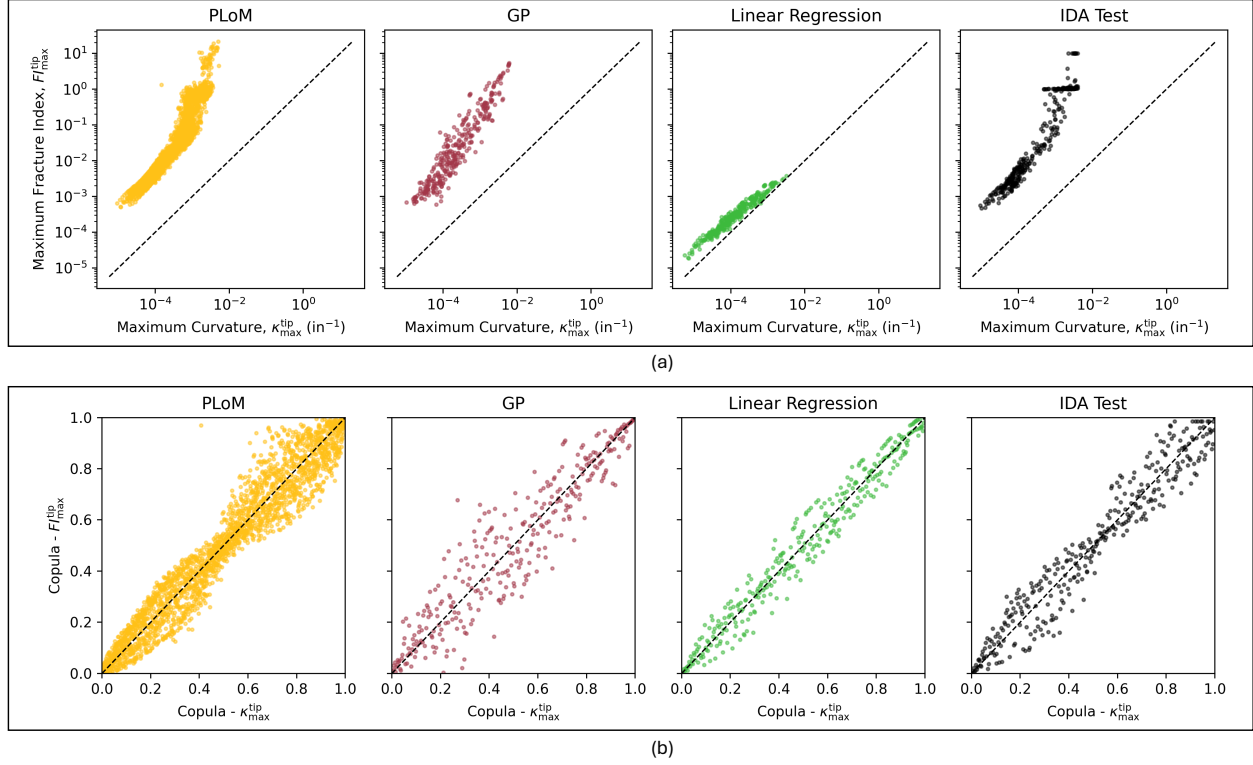


Fig. 12. Intercorrelation between variables for three surrogate predictions and test set data, shown for IDA TS3. (a) Scatter plot of κ_{max}^{tip} versus F_{max}^{tip} for each dataset. (b) Empirical copula plot of the κ_{max}^{tip} versus F_{max}^{tip} for each dataset.

noticeable discrepancies from the empirical distributions are observed.

This reduced performance of PLoM at low intensities is likely related to the way input parameters describing the bridge structure are defined in the *MSA Test*. By design, PLoM operates directly on distributions of input variables; thus, structure-specific predictions in PLoM are achieved by defining very narrow distributions around known parameter values, rather than by conditioning on a single deterministic input. When sufficient training data exist in the vicinity of these narrowly defined distributions, PLoM can approximate structure-specific behavior effectively; otherwise, predictive accuracy may degrade. Because PLoM relies directly on the density and structure of the training data, the narrow structural parameter definitions used in the *MSA Test* may exacerbate data sparsity at lower intensity levels, contributing to reduced accuracy. This limitation is less pronounced at higher intensities, where the underlying IDA data generation process produces a larger number of samples as bridges approach collapse. As a result, more training data is available in these regions, which may explain the improved PLoM performance at higher intensities. The hyperparameter selection method may also contribute to reduced performance of the PLoM model in the *MSA Test*. By selecting a single pair of hyperparameters for all earthquake intensities, it is possible that performance degrades when the task becomes more constrained (i.e., tighter structural parameter distributions), which was not observed as clearly in the PLoM model with functional hyperparameters proposed in Lee et al. [21].

In the context of this case study, the GP model performs most consistently across the *IDA* and *MSA Test* sets for non-collapse performance, achieving the lowest overall error based on KS statistic. Linear regression and PLoM each have contexts where model performance degrades: linear regression at high intensities and PLoM under tight structural parameter constraints. A more extended discussion of model tradeoffs and performance can be found in Section 6.

5 CASE STUDY RESULTS: PREDICTION OF GROUND MOTION INTENSITY AT COLLAPSE

Similar methods to the non-collapse surrogate are used to make modeling choices and select hyperparameters for the collapse surrogate.

For the PLoM model, the same grid-search procedure used for the EDP model is applied to the collapse dataset, using the ten validation datasets. The selected pair of hyperparameters based on the results of the grid search is ϵ_{db}

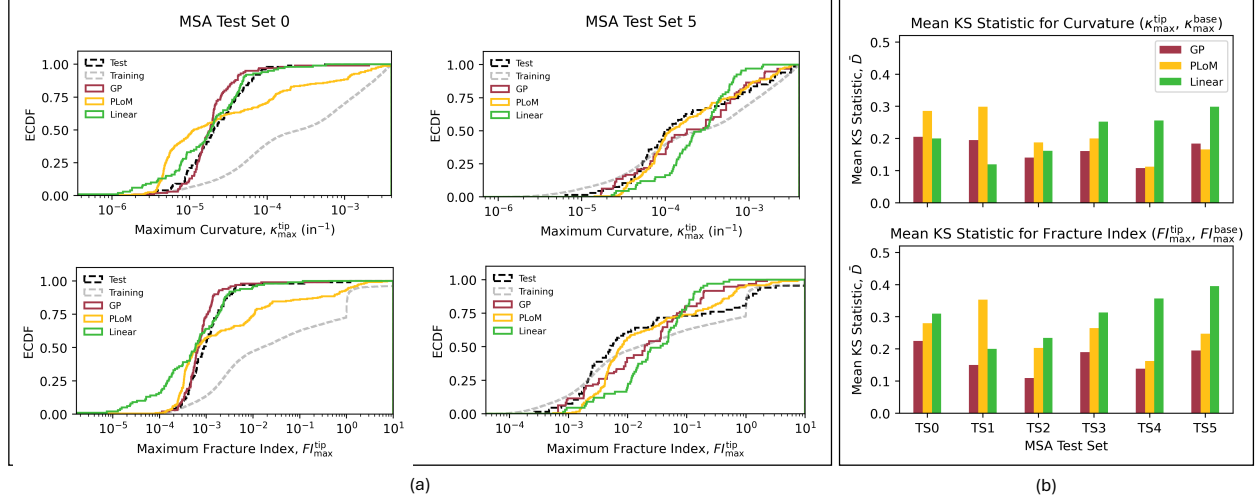


Fig. 13. Model comparison using *MSA Test*. (a) Empirical CDF example plots for TS0 and TS5 comparing predictions from the three surrogates against training and test sets. (b) Comparison of KS Statistic for each surrogate's prediction of $\kappa_{\max}$ and $FI_{\max}$ across test sets.

$= 0.001$ and $\epsilon_k = 1.0$. For the linear regression model, as in the EDP model, the log-space regression outperforms linear-space and is selected for future use.

The same six variations of the GP model are considered for the collapse prediction task. Based on the predicted empirical CDFs (14a) and the KS statistics (14b), it appears that sampling at the median, rather than sampling the output distribution randomly, improves performance (lower KS statistic) for almost all validation sets. Second, there does not seem to be a large difference between the homoskedastic and heteroskedastic nugget models, indicating the variability in the dataset is likely homoskedastic. Based on $\bar{D}$ across the ten validation sets, the GP model with the best performance is the homoskedastic nugget with the output sampled at the median, which is selected for use.

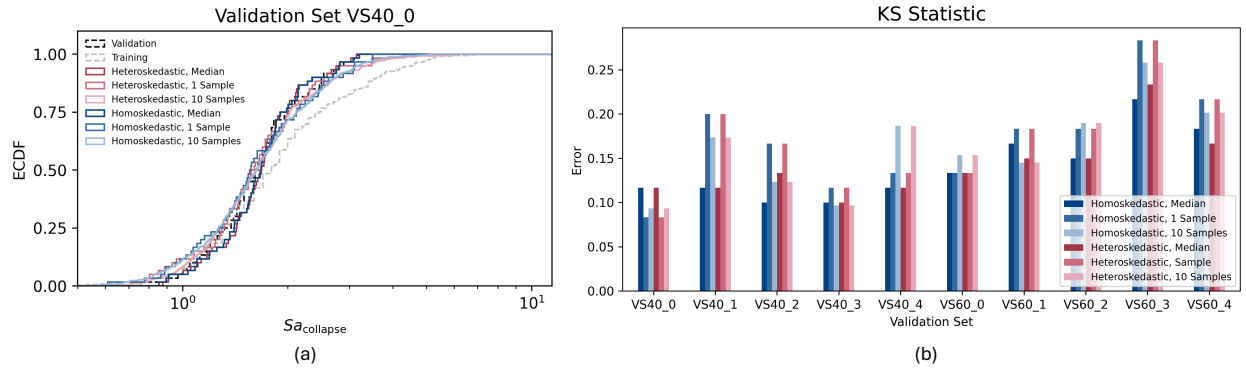


Fig. 14. Comparison of GP Collapse Models. (a) ECDF example plot for VS40_0. (b) Comparison of KS Statistic for six GP variants across validation sets.

5.1 Test Set Comparison Across Collapse Surrogate Models

The comparison across the selected version of each model (PLoM, GP, linear regression) follows the same procedure as the EDP model. Figures 15a and 15b show the predicted distributions of Sa_{collapse} against the results of the corresponding test set, demonstrating the strong performance of both the GP and linear regression models. The predicted PLoM distribution does not match the test distribution as well as the other two models. This pattern is again demonstrated by $\bar{D}$, shown in Figure 15c. While the GP and linear regression models have similar KS statistics, the

PLoM value is much higher. It is possible that the PLoM model is struggling due to limited training data. PLoM generates additional realizations (makes predictions) using the underlying training data available, and in the case of collapse, there is somewhat limited training data available due to the formulation of the datasets using IDA data.

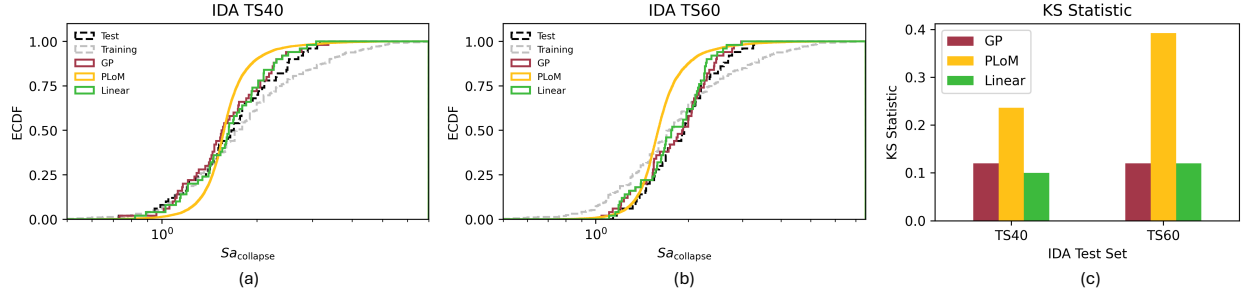


Fig. 15. Collapse Model comparison using *IDA Test*. Empirical CDF example plot for TS40 (a) and TS60 (b) comparing predictions from the three surrogates against training and test sets. (c) Comparison of KS Statistic for each surrogate prediction of $Sa_{collapse}$ across test sets.

While direct point-level error metrics cannot be computed for the PLoM model due to its distribution-based formulation, such metrics can be evaluated for the GP and linear regression models. These comparisons provide additional insight into model performance beyond distribution-level agreement. Figure 16 shows the $Sa_{collapse}$ values of a selected test set against the corresponding predictions from the GP and linear regression models. Although neither model appears clearly superior based on visual inspection alone, quantitative metrics (R-squared and MSE) indicate that the linear regression model performs slightly better than the GP model in predicting $Sa_{collapse}$.

Some of the variation in the true $Sa_{collapse}$ values is not explained by the six model input parameters and is therefore not captured by the surrogate models. We refer to this as aleatory uncertainty, which can degrade the accuracy of individual point predictions. The point-level inaccuracy, however, is masked when only comparing models at the distribution level. Both the GP and linear regression models appear to perform well when assessed as a distribution across a suite of input points, but this does not guarantee accurate individual point predictions. Since the PLoM model cannot be assessed in this way, this comparison is considered illustrative and is only relevant between the selected GP and linear regression models.

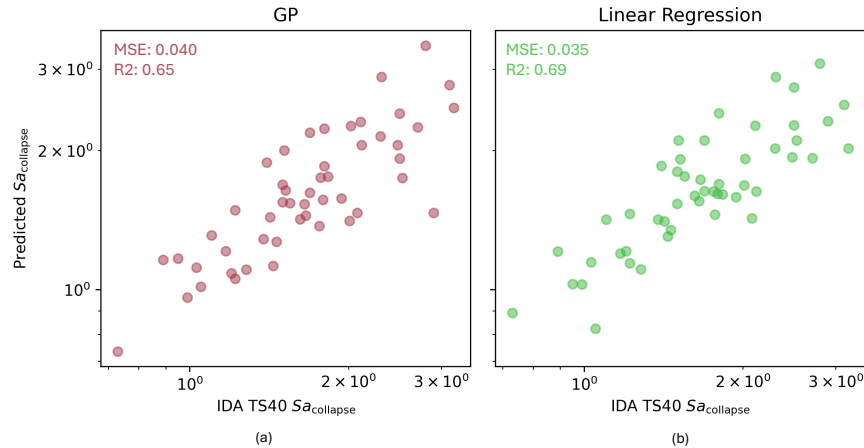


Fig. 16. Scatter plot of $Sa_{collapse}$ from the *IDA Test* set against predicted $Sa_{collapse}$ from the selected GP (a) and Linear Regression (b) models.

In addition to the *IDA Test*, the *MSA Test* serves as an out-of-distribution evaluation and is more representative of a practical use case for the surrogate models. Because the MSA data is formulated to reflect certain return periods (stripes), there is no corresponding "test" distribution of $Sa_{collapse}$ available for direct comparison with model

predictions at each stripe. Instead, the only available collapse information for each stripe is the number of ground motion records (out of 100) that result in collapse, which defines the empirical probability of collapse for a given bridge, site, and return period. Consequently, surrogate model performance is evaluated for the *MSA Test* using this scalar probability of collapse.

Figure 17a shows the probability density functions of S_a for the lowest (TS0) and highest (TS5) return periods included in the *MSA Test* alongside the predicted CDF of $S_{a_{collapse}}$ from the PLoM, GP, and linear regression models. For both example test sets, the GP and linear regression models predict very similar distributions of $S_{a_{collapse}}$, whereas the PLoM model is slightly different. By integrating the predicted collapse fragility with the distribution of S_a values from each stripe of the MSA data, a prediction of a single $P(collapse)$ value for each surrogate can be obtained. Figure 17b shows a comparison of predicted $P(collapse)$ values from each surrogate against the MSA data. Based on these values, linear regression is the most effective at predicting collapse behavior for this case study, closely followed by the GP model. While all models overpredict the probability of collapse, the PLoM model more severely overpredicts, particularly for the higher intensities.

It is also possible to use the predicted probability of collapse values alongside their associated mean S_a values to fit a collapse fragility for each surrogate, shown in Figure 17c. These collapse fragilities can then be integrated against the hazard curve for the original site used in the development of the MSA data to obtain the mean annual frequency (MAF) of collapse. The computed MAF of collapse values are shown for each of the three surrogates, as well as the *MSA test* data. All models provide estimates in the correct order of magnitude, but all models also overestimate collapse. These values confirm that linear regression slightly outperforms GP in collapse predictions, with PLoM more severely overestimating this value.

Predicting $S_{a_{collapse}}$ involves a scalar output rather than a full distribution of nonlinear EDPs, potentially explaining why the linear regression model, which is outperformed by more flexible models in the EDP task, is a stronger performer here. Section 6 contains additional discussion of model tradeoffs.

6 MODEL CONSIDERATIONS FOR REGIONAL ANALYSIS

Beyond a quantitative model performance comparison, it is also important to consider other aspects of surrogate models that make them more or less appropriate for use in regional seismic risk assessments. Considerations include how the workflow is set up (e.g., where and how the surrogates are implemented), the type and usage of the surrogate (e.g., dimensionality, shareability, and model use), and the treatment of uncertainty and correlations in the data.

6.1 Workflow Setup

In PBEE, each step of the assessment is treated probabilistically, allowing for uncertainties to be preserved and carried through the assessment [6]. When selecting a surrogate model, the first consideration is which portion of the framework will be "surrogated." While there are instances of surrogate models used in the ground motion modeling, this discussion focuses on surrogates pertaining to structural performance.

Because NLRHA is the most computationally expensive step in PBEE, studies often replace NLRHA with surrogate models, mapping from structural and/or ground motion parameters to a corresponding output. The choice of surrogate outputs, however, varies widely. In the most direct replacement of NLRHA, a surrogate model predicts EDPs. While conceptually straightforward, this approach presents several challenges. Multiple EDPs are often required, and these EDPs are typically correlated (e.g., peak floor accelerations at different locations). Furthermore, structural response under a given ground motion is inherently uncertain, requiring prediction of distributions rather than point estimates. Accurately capturing correlated, multi-output distributions requires specialized surrogate formulations, such as PLoM or multi-output stochastic emulators [13, 43].

Furthermore, outputting EDPs requires subsequent mapping from EDPs to damage, loss, or functionality metrics through fragility and consequence models. If these downstream models lack sufficient resolution, they may limit the benefits of a highly detailed EDP surrogate. Alternatively, surrogate models can be constructed to replace multiple sequential stages of the PBEE framework by directly predicting metrics like losses or downtime. These surrogates, which often use scalar outputs, can be simpler to develop but must be redeveloped for each output metric of interest. In contrast, EDP-based surrogates preserve flexibility, as predicted EDPs can be used to evaluate any downstream performance metric. In this study, the surrogate models map structural and ground motion parameters to EDPs, providing the most direct replacement of NLRHA.

Another key consideration for surrogates in the regional context is the typology resolution (level of structural specificity) they capture. Surrogates may be structure-specific, representing the response of an individual bridge, or typology-based, representing a class of similar structures. Rincon and Padgett [4] demonstrate that capturing structure-level variability can significantly influence regional loss estimates and help identify which structures drive

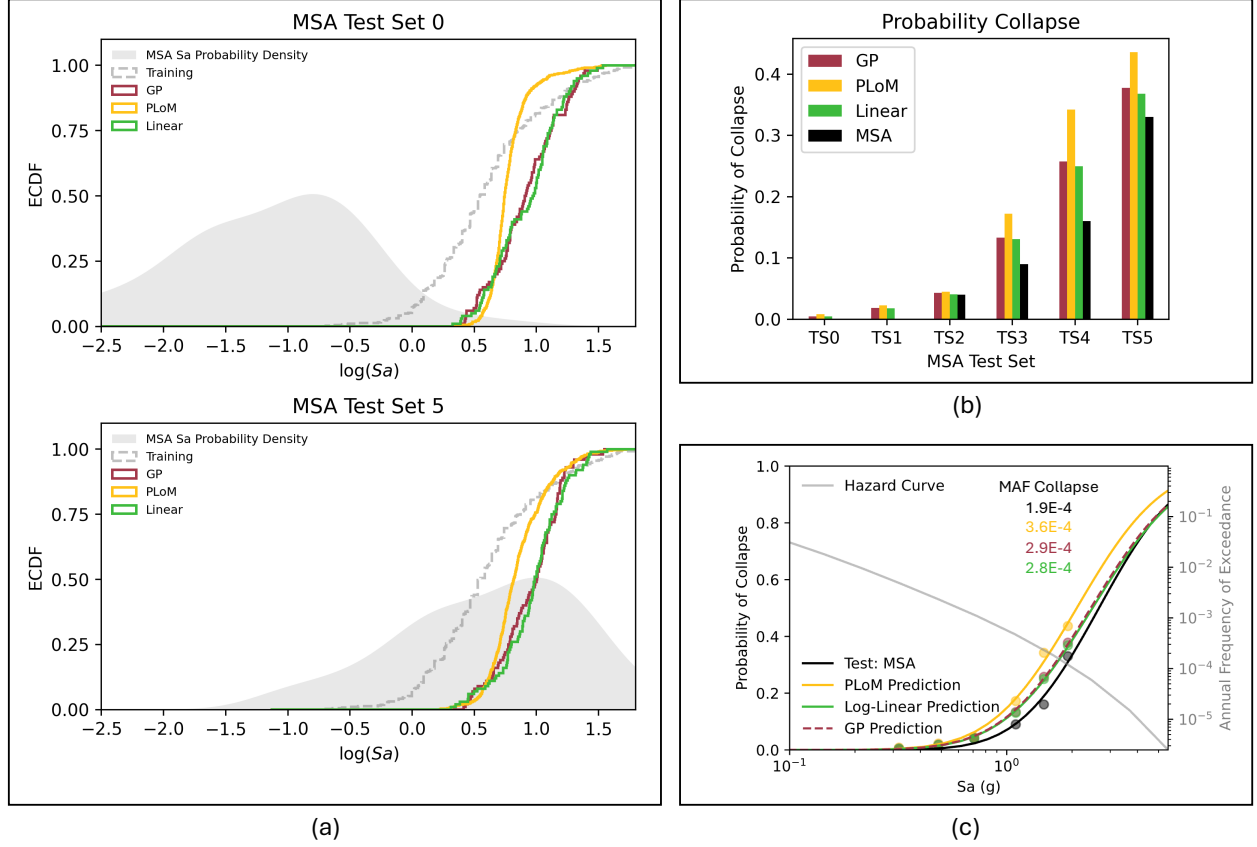


Fig. 17. Collapse Model comparison using *MSA Test*. (a) Empirical CDF of $Sa_{collapse}$ for each surrogate model plotted against the probability density function of the MSA Sa data for TS0 and TS5. (b) Predicted probability of collapse for each surrogate, computed by integrating the MSA Sa PDF with the predicted $Sa_{collapse}$ CDF, versus MSA data probability of collapse across *MSA Test* sets. (c) Lognormal fragility functions fit to MSA data and to each surrogate prediction plotted against the hazard curve originally used for ground motion selection. MAF values are displayed on the plot.

losses. While the current study opted for the use of structure-specific surrogate models, this selection should consider the available input inventory data. Structure-specific surrogates require detailed structural parameters (e.g., axial load ratio, slenderness ratio), whereas typology-based surrogates rely on more general information. When detailed design information is available, structure-specific surrogates can effectively leverage this data to identify risks; however, when this information is unavailable or highly uncertain, the advantages of structure-specific surrogates can be washed out.

6.2 Surrogate Types and Usage

Surrogate models have a range of types, and vary in the way they are deployed at regional scale. Several relevant characteristics are discussed here.

- **Parametric vs. Nonparametric:** Linear regression is an example of a parametric model, which assumes a predefined functional form with a fixed number of parameters to describe the relationship between inputs and outputs. The PLoM and GP models are examples of nonparametric models, which do not assume a fixed parametric functional form relating inputs and outputs. Parametric models can be quite accurate if the assumed functional form (e.g., linear regression) is close to the true functional form. However, they are much less accurate if that is not the case, for example predicting non-collapse EDPs using linear regression. In contrast, nonparametric approaches have greater flexibility to represent a wide range of functional relationships because they are not constrained to a specific form.

- **Dimensionality:** Dimensionality is also an important consideration when it comes to surrogate models. Dimensionality refers to the dimension of the input space (e.g., the non-collapse EDP model in this study has 7 input parameters). When this number is large, model performance can deteriorate due to the “curse of dimensionality,” where data points are dispersed across many dimensions, resulting in sparse coverage of the input space [41]. Nonparametric models are particularly sensitive to this effect because they rely on local information and data-driven flexibility to capture complex relationships. In contrast, parametric models, such as linear regression, are less affected by high dimensionality since they impose a fixed functional form on the input-output relationship. In general, GP models can struggle with high dimensionality or large training datasets because training requires inverting a covariance matrix, which scales near cubically with the number of training points. PLoM does not require GP-style covariance matrix inversion, but its computational cost also increases with the number of data points and the complexity of the manifold constraints, and it is incurred at each deployment.
- **Shareability:** Different model types also impact how surrogate models can be shared across users. Linear regression, which is a parametric model, can be transferred across users by simply sharing the functional form with the coefficients and relevant parameters. Nonparametric models are generally more memory demanding and complex to share, as their predictions depend not only on fitted hyperparameters, but also on the original training data. For example, sharing a GP model requires sharing the kernel selection, nugget selection, optimized hyperparameter values, and original training data. Sharing the PLoM model requires the selected hyperparameters, the underlying training data, among other factors. In this study, both the GP and PLoM models can be shared and evaluated using the quoFEM tool [20].
- **Model Use:** The three surrogates in this study also differ in the way they are trained and deployed in a regional seismic risk assessment. Both the GP and linear regression models are trained offline prior to deployment. While multiple candidate models may be evaluated during development, the final selected model is trained once and then queried many times during the regional analysis. As a result, most of the computational cost is incurred upfront during training, whereas model evaluation during deployment is computationally inexpensive. PLoM differs from these approaches in its deployment paradigm. Rather than producing a single pre-trained model that can be repeatedly queried, PLoM training consists of identifying appropriate hyperparameters. When deployed, the PLoM algorithm is executed separately for each site or bridge, resulting in a larger fraction of the computational cost for PLoM occurring during deployment. As a result, overall implementation time for PLoM in a regional analysis is higher than for the GP and linear regression models.

6.3 Uncertainty, Distributions, and Realizations

Regional seismic risk assessments involve uncertainty from several sources, including the intensity measure at a given site for a specified rupture, the structural input parameters, and the structural response under a given intensity measure. Developing surrogate models therefore requires considering how these uncertainties may affect both input and output parameters. Two types of uncertainty are relevant here, as described in Table 2.

TABLE 2. Forms of Uncertainty Captured in Surrogate Framework

	Input Uncertainty	Aleatory Uncertainty
Ground Motion Parameters [X_{GM}]	Variance in the ground motion model	Selected IMs do not fully capture record-to-record variability
Structural Parameters [X_S]	Material uncertainty and/or incomplete inventory description	The four selected structural parameters do not fully characterize the OpenSees model, which has 25 input parameters. <i>Note: there is additional uncertainty not fully captured by the OpenSees model at all (e.g., aggregate size). As this is not in the training data, this is also not captured in the surrogate models</i>

- **Input Uncertainty:** All three surrogate models accommodate input uncertainty, but in different ways. GP and linear regression predict at individual input points; thus, any distribution of input data can be propagated by sampling the input space and evaluating the surrogate at each sample. PLoM instead generates additional data realizations constrained to match target input distributions, using the standard deviation to filter nearby training data and the mean to guide realization generation (see Table 1 in Lee et al. [21]). Because only the mean and standard deviation are used as constraints, PLoM is not fully able to capture an arbitrary distribution.
- **Aleatory Uncertainty:** Selected input parameters do not comprehensively describe all relevant input informa-

tion, so aleatory uncertainty remains in the outputs associated with any single input point. Surrogates capture this in varying ways. GP naturally captures aleatory uncertainty by specifying output distributions for each input point. PLoM also predicts output distributions, but these correspond to an entire input distribution rather than a single input point. Linear regression can be sampled in the output space to introduce output variability, but does not capture this uncertainty as naturally as GP or PLoM. Despite this uncertainty, surrogates remain valuable at the regional scale. Current methods such as Hazus rely on simplified, typology-based fragility functions that do not capture time histories or detailed structural inputs. A surrogate that efficiently approximates NLRHA results, even imperfectly, can still offer substantially more resolution than existing approaches.

Allowing for distributions in the output space is important for capturing aleatory uncertainty; however, output distributions can also be limiting, specifically for PLoM. For example, ground motion uncertainty at a given site (e.g., sampled IM realizations), can be described as a distribution of input parameters. In regional analysis, though, single IM realizations are often used across the region to assess network performance, run traffic simulations, and similar tasks. Passing the input distribution into PLoM captures the site's ground motion variability, but at the cost of the one-to-one mapping between input and output points. As a result, the current PLoM implementation cannot translate specific IM realizations into corresponding realizations of bridge damage, limiting its use in regional network analysis. GP and linear regression preserve this one-to-one mapping and are better suited when realizations are important.

6.4 Correlations in Input/Output Data

When applied in the context of regional seismic risk assessment, it is important to consider the role of correlations. In this application, correlations can exist in both the input and output space, as well as either between different sites/structures or between parameters of a single site/structure.

- **Correlations in Model Inputs – Between Sites/Structures:** Previous research has shown that ground motion IMs are spatially correlated across sites for a given scenario earthquake [44]. Similarly, it is possible that bridges located along the same transit corridor and/or constructed at the same time may have similar material properties and structural designs; thus, there may also be correlations in the structural parameters between different bridges.
- **Correlations in Model Inputs – Within a Site/Structure:** Two structural design parameters may be correlated due to bridge design practices. Typically, these correlations are not prevalent in surrogate modeling because input parameters are often selected to have low correlation; a second highly correlated input parameter does not provide as much descriptive information on the outputs beyond the first parameter for a surrogate model.

The correlations in input parameters (both between different sites/bridges and between parameters of a single site/bridge) can be described in the input data and thus can be handled separately from the surrogate model itself. If inputs are specified with the correct correlations in place, the surrogates will process that correlated data, as long as there is comprehensive coverage of the input space.

- **Correlations in Model Outputs – Between Sites/Structures:** In this study, the only correlation that is captured in the outputs between different sites/structures comes from having similar input parameters. For example, if two bridges are close together and thus have highly correlated IMs, the results will reflect that. However, no further correlation is captured, as each site/structure is assessed by the surrogate independently, similarly to how NLRHA and Hazus fragilities are applied independently at each site. Additional steps can be taken to introduce this correlation, if desired, via postprocessing of results [45].
- **Correlations in Model Outputs – Within a Site/Structure:** There are two forms of correlations in output variables for a given site/bridge. First, there are correlations in the output parameters due to having similar inputs (e.g., if the input earthquake intensity is very large, both curvature and fracture index will likely be large). These correlations are captured by all three surrogates in this study. Second, there can also be *conditional correlation*, which is correlation conditional on the inputs, meaning correlations in output variables that are not fully described just by having similar input points. The GP and linear regression model do not explicitly correlate between different EDP outputs and thus do not capture this second form of correlation. It is possible that these models could be extended to do so by introducing latent variables for the joint outputs or additionally learning joint statistics [43]. PLoM explicitly and naturally handles conditional correlations because EDPs are not predicted independently; instead, PLoM generates an output dataset in multi-dimensional space. The importance of capturing correlations between outputs depends on the application. For example, in this case,

the main EDP used for downstream assessment is curvature ductility, which is computed using $\kappa_{\max}$ [32]. Thus, capturing correlations between EDPs is not as important in this application as it is for a multi-story building, where the relationship between EDPs on different floors is important for overall building damage and loss. Thus, PLoM’s ability to more fully capture correlations may not be fully used in this case study.

A summary of the main points in this discussion is provided in Table 3.

TABLE 3. Summary of comparison across surrogate model implementations for selected case study

Surrogate Model	PLoM	GP	Linear Regression
Model Type	Nonparametric	Nonparametric	Parametric
Training / Use in Region	Tune hyperparameters in advance; re-run algorithm at each site/structure in region	Train model once; use same model repeatedly at every site/structure in region	Train model once; use same model repeatedly at every site/structure in region
EDP Accuracy for Selected Case Study	Strong for predicting EDPs overall. More variable performance when given tight structural constraints	Strong performance overall. Struggles to predict multimodal output distributions (FI)	Insufficient flexibility to predict EDP performance, particularly under high intensity ground motions
Collapse Accuracy for Selected Case Study	Less accurate than GP and linear regression in both IDA and MSA tests; possible factors include the small collapse dataset and hyperparameter selection	Comparatively strong collapse prediction performance	Comparatively strong collapse prediction performance
Input Data Type	Distribution of input dataset used to define filtering and constraints in algorithm	Input data point sampled explicitly by model. Multiple points can be sampled to represent an input distribution	Input data point sampled explicitly by model. Multiple points can be sampled to represent an input distribution
Output Data Type	Generates additional realizations of data. The output prediction is taken as the distribution of the additional realizations along a given dimension	Generates an output distribution for corresponding input point, which can be sampled once at the median, once randomly, or multiple times	Generates an output point prediction for corresponding input point
Correlations due to Similar Inputs	Handles correlations	Handles correlations	Handles correlations
Conditional Correlations	Handles correlations	Does not handle correlations	Does not handle correlations
Deterministic mapping from input point to output (required for network analysis)	Does not support direct mapping; can only make predictions at the distribution level	Supports direct mapping	Supports direct mapping

7 CONCLUSIONS

This study compares three surrogate models for predicting site- and structure-specific performance of two-span, single-column highway bridges under seismic loading. The goals of this comparison are twofold: first, a quantitative comparison of the models in predicting earthquake intensity at collapse and EDPs conditioned on non-collapse; second, a discussion of considerations for using each surrogate for regional seismic risk assessments, including workflow setup, model types and sharing, input/output data types, and correlations.

This study does not propose a new surrogate modeling method, but instead compares three existing techniques: PLoM, which generates additional realizations of data subject to constraints and has recently been applied to bridge seismic response [21]; GP, which defines an output distribution over functions of the inputs and is widely used across earthquake engineering; and multivariate linear regression, included as a baseline. Each model is trained using IDA data from Lee et al. [21] and evaluated using two tests: the *IDA Test*, drawn from the same dataset as training, and the *MSA Test*, an out-of-distribution evaluation representative of a practical use case. A limitation of the analysis is that the comparisons are only made for a single application: the seismic performance of two-span one-column concrete highway bridges. It is possible that attempting to surrogate more complex structures would reveal certain benefits and/or limitations in the three surrogates. Further comparisons could be done to assess these models on different structure types, including more complex bridge models (e.g., multi-span, multi-column bridges).

Based on the quantitative results for the specific case study in this paper, the EDP surrogate must represent behavior ranging from elastic response at low intensities to highly nonlinear behavior near collapse, so the more flexible GP and PLoM models outperform linear regression. In contrast, the collapse dataset consists only of IDA collapse points, giving it a more uniform structure that the linear regression model captures more accurately. Beyond quantitative accuracy, the suitability of surrogate models for regional seismic risk assessment depends on broader considerations such as workflow setup, surrogate types and usage, input/output distributions, and correlations. While all three models can serve as replacements for NLRHA, each offers distinct advantages and limitations.

Linear regression is simple, efficient, and highly shareable, but its fixed functional form limits its ability to capture the nonlinear relationships inherent in EDP prediction. It performs well for $Sa_{collapse}$, where the underlying data structure is more uniform, but does not naturally represent output uncertainty or intercorrelation beyond what is due to shared input variables.

In this case study, GP offered the most consistent performance across EDP and collapse prediction. Furthermore, it can be trained once and deployed efficiently in a region with many bridges, and it supports deterministic mapping from specific ground motion realizations to outputs. The GP model provides flexible, probabilistic predictions that naturally represent output uncertainty, though they are computationally intensive to train and become less practical with high-dimensional or very large datasets. Like linear regression, it does not capture intercorrelation among outputs and it cannot accurately predict multi-modal output distributions like fracture index; however, these limitations matter less for the selected bridge typology than they might in other contexts (e.g., multi-story buildings).

The implemented PLoM model follows a different prediction paradigm than the other surrogates in this study. As discussed in Section 6, the algorithm must be rerun at every site/bridge, which may become computationally expensive at the regional scale. Furthermore, rather than predicting outputs for a single input point, it generates additional realizations, producing a distribution of outputs from a distribution of inputs. This is beneficial for preserving correlations between variables; however, in the context of two-span concrete highway bridges, this advantage may be underutilized as compared to other use cases (e.g., multi-story buildings). On the other hand, the PLoM formulation is limiting when individual IM realizations are needed, such as for downstream transportation network analysis. Recent work has explored PLoM applications that support this type of analysis (e.g., using the conditional expectation), and future implementations may resolve this limitation [46]. Finally, of the three models, PLoM is best able to capture multi-modal or non-Gaussian output distributions (such as fracture index), though this sometimes comes at the cost of accuracy of the median.

The broader contribution of this study is the set of considerations to help guide which surrogates are appropriate for a given regional application. For example, if a use case involves highly non-Gaussian distributions or complex conditional correlations between output variables, PLoM may be beneficial despite its added computational and workflow demands. If deterministic input-to-output mapping is critical, as in the case of transportation network analysis, the GP model is preferable among the models considered here. Ultimately, selecting a surrogate model for regional PBEE should be guided by application-specific demands and regional-scale use rather than by accuracy metrics alone.

Software and Data Availability

The GP models, linear regression models, and data are available on GitHub at <https://github.com/mloch/head/bridge-surrogate-comparison>. The PLoM model is available through the DesignSafe Data Depot under project PRJ-4170 (<https://doi.org/10.17603/ds2-ytpz-gj12>). Furthermore, the PLoM and GP models can be run through the NHERI-SimCenter's quoFEM application (<https://simcenter.designsafe-ci.org/research-tools/quoFEM-application/>).

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants CMMI-1612843 and CMMI-2131111 and the Graduate Research Fellowship Program under Grant DGE-1656518, and by the Pacific Earthquake Engineering Research (PEER) Center. 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 or PEER Center. The authors would like to thank Carmen Andrade von Hillebrandt for her thoughtful review and comments.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude to improve the language and clarity of selected sections of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

- [1] FEMA. Hazus earthquake technical manual. <https://www.fema.gov/flood-maps/products-tools/hazus>, 2020.
- [2] Sujith Mangalathu, Jong-Su Jeon, Jamie E Padgett, and Reginald DesRoches. Performance-based grouping methods of bridge classes for regional seismic risk assessment: Application of ANOVA, ANCOVA, and non-parametric approaches. *Earthquake Engineering & Structural Dynamics*, 46(14):2587–2602, 2017.
- [3] Pablo Heresi and Eduardo Miranda. Structure-to-structure damage correlation for scenario-based regional seismic risk assessment. *Structural Safety*, 95:102155, 2022.
- [4] Raul Rincon and Jamie Ellen Padgett. Fragility modeling practices and their implications on risk and resilience analysis: From the structure to the network scale. *Earthquake Spectra*, 40(1):647–673, 2024.
- [5] Kuanshi Zhong, Reagan Chandramohan, Jack W Baker, and Gregory G Deierlein. Site-specific adjustment framework for incremental dynamic analysis (SAF-IDA). *Earthquake Spectra*, 38(3):1893–1917, 2022.
- [6] Jack Moehle and Gregory G Deierlein. A framework methodology for performance-based earthquake engineering. In *13th World Conference on Earthquake Engineering*, volume 679, page 12. WCEE Vancouver, 2004.
- [7] Applied Technology Council. Seismic performance assessment of buildings, volume 1 – methodology, second edition. Technical Report FEMA P-58-1, Federal Emergency Management Agency, Washington, D.C., 2018.
- [8] Applied Technology Council. Quantification of building seismic performance factors. Technical Report FEMA 695, Federal Emergency Management Agency, Washington, D.C., 2009.
- [9] Applied Technology Council. Seismic performance assessment of buildings, volume 2 – implementation guide. Technical Report FEMA P-58-2, Federal Emergency Management Agency, Washington, D.C., 2018.
- [10] Kuanshi Zhong, Javier G Navarro, Sanjay Govindjee, and Gregory G Deierlein. Surrogate modeling of structural seismic response using probabilistic learning on manifolds. *Earthquake Engineering & Structural Dynamics*, 52(8):2407–2428, 2023.
- [11] Sujith Mangalathu and Jong-Su Jeon. Stripe-based fragility analysis of multispan concrete bridge classes using machine learning techniques. *Earthquake Engineering & Structural Dynamics*, 48(11):1238–1255, 2019.
- [12] Vamshi Krishna Gudipati and Eun Jeong Cha. Surrogate modeling for structural response prediction of a building class. *Structural Safety*, 89:102041, 2021.
- [13] Jungho Kim and Ziqi Wang. Uncertainty quantification for seismic response using dimensionality reduction-based stochastic simulator. *Earthquake Engineering & Structural Dynamics*, 54(2):471–490, 2025.
- [14] Delbaz Samadian, Imrose B Muhit, and Nashwan Dawood. Application of data-driven surrogate models in structural engineering: a literature review. *Archives of Computational Methods in Engineering*, pages 1–50, 2024.
- [15] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.
- [16] Sujith Mangalathu, Gwanghee Heo, and Jong-Su Jeon. Artificial neural network based multi-dimensional fragility development of skewed concrete bridge classes. *Engineering Structures*, 162:166–176, 2018.
- [17] Taeyong Kim, Junho Song, and Oh-Sung Kwon. Pre-and post-earthquake regional loss assessment using deep learning. *Earthquake Engineering & Structural Dynamics*, 49(7):657–678, 2020.
- [18] J Zou, DP Welch, A Zsarnoczay, A Taflanidis, and GG Deierlein. Surrogate modeling for the seismic response estimation of residential wood frame structures. In *Proceedings of the 17th world conference on earthquake engineering, Japan*, 2020.

- [19] Youjun Chen, Zeyang Sun, Ruiyang Zhang, Liuzhen Yao, and Gang Wu. Attention mechanism based neural networks for structural post-earthquake damage state prediction and rapid fragility analysis. *Computers & Structures*, 281:107038, 2023.
- [20] Frank McKenna, Sang-ri Yi, Aakash Bangalore Satish, Adam Zsarnoczay, Kuanshi Zhong, Michael Gardner, and Wael Elhaddad. NHERI-SimCenter/quoFEM: Version 3.5.0, 2023. URL <https://doi.org/10.5281/zenodo.10443180>.
- [21] Jeonghyun Lee, Meredith Lochhead, Kuanshi Zhong, and Gregory G Deierlein. Systematic training and validation of parameterized probabilistic learning on manifolds surrogate model for seismic performance assessment of highway bridges. *Earthquake Engineering & Structural Dynamics*, 54(15):3726–3745, 2025.
- [22] Daniel G Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.
- [23] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.
- [24] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT Press, Cambridge, MA, 2006.
- [25] Adam Zsarnoczay, Gregory G Deierlein, Frank McKenna, Matthew Schoettler, Sang-ri Yi, Barbaros Cetiner, Aakash B Satish, Jinyan Zhao, Justin Bonus, Abiy F Melaku, et al. An open-source simulation platform to support and foster research collaboration in natural hazards engineering. *Frontiers in Built Environment*, 11: 1590479, 2025.
- [26] GPy. GPy: A Gaussian process framework in Python. <http://github.com/SheffieldML/GPy>, since 2012. Accessed: 2026-09-30.
- [27] Sang-ri Yi and Alexandros A Taflanidis. Stochastic emulation with enhanced partial- and no-replication strategies for seismic response distribution estimation. *Earthquake Engineering & Structural Dynamics*, 53(7):2354–2381, 2024.
- [28] Aikaterini P Kyprioti and Alexandros A Taflanidis. Addressing the different sources of excitation variability in seismic response distribution estimation using kriging metamodeling. *Earthquake Engineering & Structural Dynamics*, 51(10):2466–2495, 2022.
- [29] G. G. Deierlein, F. McKenna, A. Zsarnóczay, T. Kijewski-Correa, A. Kareem, W. Elhaddad, L. Lowes, M. J. Schoettler, and S. Govindjee. A cloud-enabled application framework for simulating regional-scale impacts of natural hazards on the built environment. *Frontiers in Built Environment*, 6:558706, 2020. doi: 10.3389/fbuilt.2020.558706.
- [30] Christian Soize and Roger Ghanem. Data-driven probability concentration and sampling on manifold. *Journal of Computational Physics*, 321:242–258, 2016.
- [31] Christian Soize and Roger Ghanem. Physics-constrained non-Gaussian probabilistic learning on manifolds. *International Journal for Numerical Methods in Engineering*, 121(1):110–145, 2020.
- [32] Sujith Mangalathu. *Performance Based Grouping and Fragility Analysis of Box-girder Bridges in California*. Phd thesis, Georgia Institute of Technology, Atlanta, GA, 2017.
- [33] Jian Zhao and Sri Sritharan. Modeling of strain penetration effects in fiber-based analysis of reinforced concrete structures. *ACI structural journal*, 104(2):133, 2007.
- [34] John B Mander, Michael JN Priestley, and Robert Park. Theoretical stress-strain model for confined concrete. *Journal of structural engineering*, 114(8):1804–1826, 1988.
- [35] Joshua S Pugh, Laura N Lowes, and Dawn E Lehman. Nonlinear line-element modeling of flexural reinforced concrete walls. *Engineering Structures*, 104:174–192, 2015.

- [36] Laura Eads, Eduardo Miranda, and Dimitrios G Lignos. Average spectral acceleration as an intensity measure for collapse risk assessment. *Earthquake Engineering & Structural Dynamics*, 44(12):2057–2073, 2015.
- [37] Julian J Bommer and Alejandro Martínez-Pereira. The effective duration of earthquake strong motion. *Journal of earthquake engineering*, 3(02):127–172, 1999.
- [38] Dimitrios Vamvatsikos and C Allin Cornell. Incremental dynamic analysis. *Earthquake Engineering & Structural Dynamics*, 31(3):491–514, 2002.
- [39] Karthik Narayan Ramanathan. *Next Generation Seismic Fragility Curves for California Bridges Incorporating the Evolution in Seismic Design Philosophy*. Phd thesis, Georgia Institute of Technology, Atlanta, GA, 2012.
- [40] Henry Burton, Hongquan Xu, and Zhengxiang Yi. Design of computer experiments for developing seismic surrogate models. *Earthquake Spectra*, 38(1):384–406, 2022.
- [41] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. *An Introduction to Statistical Learning with Applications in Python*. Springer, 2023.
- [42] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.
- [43] Sang-ri Yi and Alexandros A Taflanidis. Multi-output stochastic emulation with applications to seismic response correlation estimation. *Structural Safety*, 115:102578, 2025.
- [44] Jack W Baker and Nirmal Jayaram. Correlation of spectral acceleration values from NGA ground motion models. *Earthquake Spectra*, 24(1):299–317, 2008.
- [45] Carmen Andrade, Meredith Lochhead, and Gregory Deierlein. Surrogate modeling of highway bridges for regional earthquake simulation of transportation networks. In *Proceedings of the 13th National Conference on Earthquake Engineering*, Portland, Oregon, 2026.
- [46] Xiaoshu Zeng, Bora Gencturk, and Olivier Ezvan. Probabilistic learning on manifolds (PLoM) for cross-scale diagnostics in structural dynamics. *Computer Methods in Applied Mechanics and Engineering*, 435:117635, 2025.