Vital Annex: International Journal of Novel Research in Advanced Sciences
ISSN: 2751-756X
Volume 05 Number 04 (2026)
https://innosci.org/IJNRAS

Article

Reliability-Calibrated Surrogate Models for the Nonlinear Seismic Assessment of Existing Reinforced Concrete Frames

Shaymaa Abdelghany Mohamed[1]

1. College of Electrical Engineering, University of Technology – Iraq, Baghdad, Iraq, Civil Engineering
* Correspondence: Shaymaa.A.Eldawy@uotechnology.edu.iq

Abstract: Surrogate models are routinely substituted for nonlinear finite element models in the probabilistic seismic assessment of existing reinforced concrete buildings, and are accepted on the basis of global goodness-of-fit statistics such as the leave-one-out coefficient of determination. This paper shows that such statistics do not control the accuracy of the resulting reliability estimate, and derives the relation that does. A first-order expansion gives the error in the reliability index as the conditional mean surrogate error evaluated at the limit-state threshold, scaled by the ratio of the response density at that threshold to the standard normal density at the reliability index. The global error variance enters only at second order. The relation is tested on a four-storey, three-bay non-seismically detailed frame modelled with fibre beam-column elements, assessed at four drift-based limit states with reference reliability indices of -0.49, 0.71, 1.34 and 1.99, and with eleven random variables covering seismic intensity, materials, loads and model error. Across quadratic response surfaces, sparse polynomial chaos expansions, multilayer perceptrons, Gaussian process models and polynomial-chaos Kriging, the relation predicts the observed reliability-index error with a regression slope of 1.04 and a coefficient of determination of 0.93, whereas conventional fit statistics leave that error undetermined by up to two orders of magnitude. Variance decomposition attributes about 94% of the response variance to seismic intensity and demand model error, and less than 1% to all material variables combined. Three practical consequences are demonstrated: logarithmic transformation of the drift response changes the reliability-index error by an order of magnitude; misclassification-directed enrichment reaches an error below 10^{-3} within about 100 finite element evaluations; and the epistemic uncertainty introduced by the surrogate can be propagated into bounds on the failure probability that contain the reference solution at every training budget.

Keywords: structural reliability; surrogate models; Gaussian process regression; polynomial chaos expansion; existing reinforced concrete buildings; seismic fragility; active learning.

Citation: Mohamed S. A. Reliability-Calibrated Surrogate Models for the Nonlinear Seismic Assessment of Existing Reinforced Concrete Frames. Vital Annex: International Journal of Novel Research in Advanced Sciences 2026, 5(4), 37-51.

Received: 10th Jul 2026
Revised: 11th Jul 2026
Accepted: 24th Aug 2026
Published: 22th Sep 2026

1. Introduction

A large part of the reinforced concrete (RC) building stock in seismic regions was designed to gravity loads alone or to early seismic codes without capacity design or closely spaced transverse reinforcement [1]. Assessment standards such as EN 1998-3 and ASCE/SEI 41-17 therefore require nonlinear analysis and express acceptance in terms of deformation demands checked against deformation capacities, an approach grounded in the mechanics of RC members [2]. The material properties, load levels and modelling assumptions entering such an analysis are all uncertain, and their probabilistic treatment is well established: seismic risk is decomposed into hazard, demand, damage and loss, fragility functions are fitted by maximum likelihood, and the influence of modelling uncertainty on drift demands and collapse risk has been quantified for RC frames in several studies [3], [4].

The obstacle is computational. A crude Monte Carlo estimate of a failure probability of order 10^{-2} to 10^{-3} at an acceptable coefficient of variation requires 10^4 to 10^6 model evaluations, while a single fibre-element analysis of a mid-rise frame costs seconds to minutes. Surrogate modelling is the standard remedy: a small designed set of finite element (FE) evaluations trains an inexpensive approximation of the structural response, and the reliability integral is evaluated on that approximation. Quadratic response surfaces, sparse polynomial chaos expansions [5], Gaussian process models [6], polynomial-chaos Kriging [7], support vector regression and neural networks [8] are all in use, together with metamodel-based importance sampling and active-learning schemes that place training points near the limit-state surface [9]. Two reviews consolidate this material [10].

How these surrogates are validated has received far less attention than how they are built. In most applications, and in nearly all earthquake-engineering applications, a surrogate is accepted on the strength of a global goodness-of-fit statistic, the leave-one-out coefficient of determination Q^2LOO on the training set or the coefficient of determination R^2 on a held-out population, and is then reused unchanged across every limit state of the assessment [11]. Such statistics are integrals of the squared surrogate error weighted by the input distribution, and are therefore dominated by the region of high probability mass. The failure probability is a functional of the response distribution at and beyond a threshold, which for the severe limit states lies where probability mass is small. There is no reason for a statistic dominated by the bulk of the distribution to control an estimate governed by its tail.

The active-learning literature implicitly recognises this, since every learning function proposed since Bichon et al. targets the neighbourhood of the limit-state surface. That recognition, however, remains operational: the schemes are justified by benchmark performance rather than by an analytical relation between the residual surrogate error and the residual error in the reliability index. Consequently, stopping rules are heuristic, no criterion states how much local error a given limit state can absorb, and there is no basis for allocating a fixed FE budget among the limit states of a multi-level assessment [12]. In parallel, studies that quantify FE modelling uncertainty treat the FE model as the reference, so that the surrogate standing in for it is not itself treated as a source of epistemic uncertainty with a formal error budget [13].

This paper addresses that gap. Section 2 derives a first-order relation between the surrogate error field and the error in the reliability index, and converts it into an acceptance criterion that is specific to each limit state. Sections 3 and 4 test the relation on a non-seismically detailed RC frame against a reference solution obtained from direct FE evaluations, quantify how weakly conventional fit statistics predict the reliability error, and examine the three modelling decisions that the relation governs: the response transformation, the enrichment strategy, and the reporting of surrogate-induced epistemic uncertainty [14].

RELIABILITY-INDEX ERROR INDUCED BY SURROGATE SUBSTITUTION

The reliability index $\beta = -\Phi^{-1}(P_f)$ is used throughout as the scalar carrier of safety (Hasofer & Lind, 1974; Melchers & Beck, 2018). Let $Y = g(X)$ denote the exact FE response and $\hat{Y} = \hat{g}(X)$ the surrogate response, with pointwise error $\varepsilon(X) = \hat{g}(X) - g(X)$. For a threshold $y*$, the exact and surrogate failure probabilities are

$$P_f = P(Y > y*) = 1 - F_Y(y*), \quad \hat{P}_f = P(\hat{Y} > y*) = P(Y + \varepsilon > y*). \quad (1)$$

Conditioning on the exact response and expanding to first order in the error gives

$$\hat{P}_f - P_f \approx f_Y(y*) \cdot E\,[\varepsilon \mid Y = y*], \quad (2)$$

where f_Y is the probability density of the exact response. The perturbation of the failure probability is therefore governed by the conditional mean error on the limit-state surface, scaled by the response density there. The global error variance $E[\varepsilon^2]$ does not appear at this order. A surrogate with large but locally unbiased error is harmless to first order, whereas a surrogate with small global error and a systematic local offset at the threshold is not [15].

Transferring to the reliability index through $\beta = -\Phi^{-1}(P_f)$, with $d\beta/dP_f = -1/\varphi(\beta)$ and φ the standard normal density,

$\Delta\beta = \tilde{\beta} - \beta \approx -\,[\, f_Y(y_*) / \varphi(\beta) \,] \cdot E[\varepsilon \mid Y = y_*]$. (3)

Equation (3) defines the dimensionless amplification factor

$A(y_*) = f_Y(y_*) / \varphi(\beta)$, (4)

which converts a local response bias into a bias in the reliability index. Since $f_Y(y_*)$ is a property of the structure and the hazard while $\varphi(\beta)$ is a property of the safety level, A varies between limit states, and a single acceptance threshold applied to all of them is not consistent with Equation (3).

Inverting Equation (3) gives an acceptance criterion for surrogate calibration [16]. For a required tolerance δ on the reliability index at limit state j,

$\mid E[\varepsilon \mid Y = y_{*j}] \mid \leq \delta \cdot \varphi(\beta_j) / f_Y(y_{*j}) = \delta / A(y_{*j})$. (5)

Criterion (5) is computable during the analysis. The conditional mean error at the threshold is estimated from the residuals of the training points nearest the limit-state surface, or from the Gaussian process posterior mean where one is available; $f_Y(y_*)$ is estimated from the surrogate response population; and $\varphi(\beta)$ follows from the current reliability estimate [19], [20]. The criterion is local, so it can be satisfied by a surrogate that is globally inaccurate, and it is specific to each limit state, so it prescribes how a fixed FE budget should be distributed across a multi-level assessment.

2. Materials and Methods

CASE STUDY AND NUMERICAL SETUP

Structure and finite element model

The case study is a planar four-storey, three-bay RC frame (Figure 1a). Bay widths are 5.00 m, the ground storey is 3.50 m high and the three upper storeys are 3.00 m. Interior columns are 400 × 400 mm with eight 16 mm longitudinal bars and 8 mm stirrups at 250 mm centres (Figure 1b); beams are 300 × 500 mm with three plus three 16 mm bars and 8 mm stirrups at 250 mm centres (Figure 1c). The stirrup spacing, of the order of the effective depth, provides negligible confinement and does not restrain bar buckling, placing the structure in the category of existing buildings addressed by EN 1998-3 (CEN, 2005). Column bases are fixed, and the lateral load pattern is proportional to the fundamental modal shape, consistent with the N2 procedure (Fajfar, 2000).

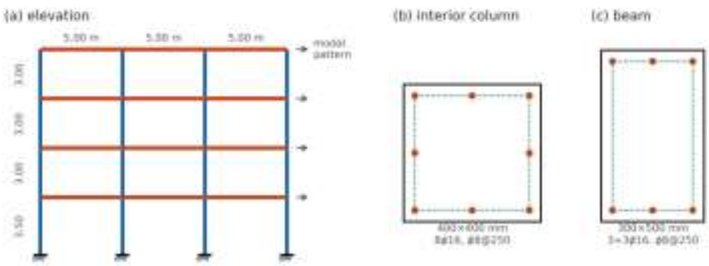

Figure 1. Case-study frame: (a) elevation with the modal lateral load pattern; (b) interior column section; (c) beam section. Storey heights and bay widths in metres.

Members are discretised with fibre beam-column elements, in which the section response is obtained by integrating uniaxial fibre constitutive laws over the cross-section and the element response by numerical integration along the length [17]. Concrete fibres follow a confined stress-strain law of the type proposed by Mander et al. with confinement effectiveness computed from the transverse reinforcement provided, and steel fibres follow a nonlinear hardening law. Because force-based fibre elements exhibit non-objective post-peak response under strain softening [18], plastic-hinge integration with a specified hinge length is used to regularise the section response.

Three response quantities are monitored: the base shear capacity Vmax, the equivalent single-degree-of-freedom period T∗, and the equivalent SDOF ultimate displacement Du∗, the latter two entering the N2 demand-capacity comparison. Two discretisation parameters control the fidelity of the model, and both were verified (Figure 2). For pushover displacement increments of 1, 2, 4 and 8 mm, deviations from the finest step remain below about 0.16% for Vmax, 0.45% for T∗ and 0.85% for Du∗, with Du∗ reaching its plateau at a 2 mm increment. For 6, 8, 12, 16 and 20 concrete fibres through the section depth, deviations from the finest mesh are about 3.4% for Du∗ and 0.6% to 1.1% for T∗ at six to eight fibres, and fall below 0.1% for all three quantities from twelve fibres upward. The production model adopts twelve fibres and the converged step size, so that the residual discretisation error is at least an order of magnitude smaller than the smallest surrogate-induced reliability errors reported below.

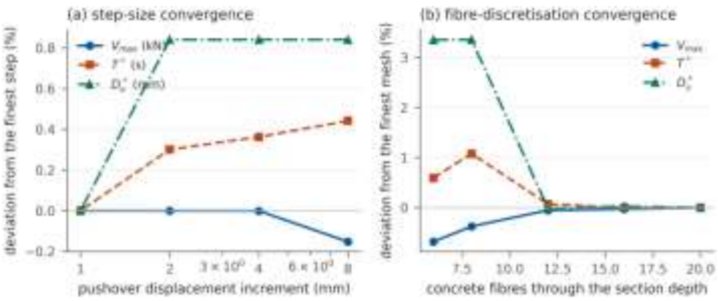

Figure 2. Numerical verification of the fibre model: (a) pushover step-size convergence; (b) fibre-discretisation convergence. Deviations are expressed relative to the finest discretisation.

Random variables, limit states and reference solution

Eleven random variables are propagated: the seismic intensity ag; the material properties fc, fy, Es and the hardening ratio b; the geometric and load variables c, kG, kQ and ψs; and two model-error terms, θD on the demand side and θC on the capacity side, in the sense of the probabilistic capacity-model formalism of Gardoni et al. Marginal distributions and their parameters follow the conventions of the JCSS Probabilistic Model Code and the fib Model Code 2010.

Performance is verified at four drift-based limit states (Table 1 and Figure 3c). The limit-state functions are gj(X) = y∗j - Y(X), where Y is the peak interstorey drift ratio (IDR) over the height and y∗j the threshold of limit state j, so that failure is the event Y > y∗j. Reference failure probabilities were obtained on a reference population of direct FE evaluations and correspond to reliability indices of -0.49, 0.71, 1.34 and 1.99. Global sensitivity is quantified by Sobol' indices, computed from the polynomial chaos representation where available and by pick-and-freeze estimators otherwise.

Table 1. Limit states, reference failure probabilities and reference reliability indices.

Limit state	Description (EN 1998-3)	Threshold IDR	P_f (50 yr)	β_{ref}
LS1	Damage limitation	0.5%	≈ 0.688	−0.49
LS2	Significant damage	1.5%	≈ 0.239	0.71
LS3	Near collapse	3.0%	≈ 0.090	1.34
LS4	Sideway collapse	6.0%	≈ 0.023	1.99

Note. Reference values obtained from direct finite element evaluations on the reference population; $Pf = \Phi(-\beta ref)$.

Surrogate families, enrichment and performance measures

Five surrogate families are compared under matched FE budgets: a full quadratic response surface fitted by ordinary least squares (QRS); a sparse polynomial chaos expansion in polynomials orthogonal to the input measure, with basis selection by least-angle regression (PCE); a multilayer perceptron (MLP); a Gaussian process with anisotropic stationary covariance and hyperparameters estimated by maximum likelihood (GP-MLE); and polynomial-chaos Kriging, which uses a sparse expansion as the trend of a Gaussian process (PCK). A sixth configuration, referred to below as the proposed scheme, combines the Gaussian process surrogate with the reliability-oriented enrichment and acceptance logic of Section 2: points are added where the misclassification criterion is largest, and the design is judged complete when Equation (5) is satisfied at the limit state considered. All surrogates are trained on identical experimental designs at each budget, so that differences in performance are attributable to the surrogate family rather than to the sample. Two response scales are used, the untransformed peak IDR and its natural logarithm, the latter reflecting the lognormal demand convention.

Enrichment is compared in three modes. Random enrichment draws points from the input distribution. U-MLE enriches at the point maximising the misclassification criterion of Echard et al. evaluated with the maximum-likelihood Gaussian process. U-LSW applies the same criterion with limit-state weighting, so that candidates are additionally weighted by their proximity to the limit-state surface in probability.

The primary performance measure is the absolute reliability-index error $|\Delta\beta| = |\hat{\beta} - \beta ref|$ at each limit state, together with the failure-probability bias ratio. The reliability index is used because code-calibrated targets are expressed in it (CEN, 2002) and because it compresses failure probabilities spanning orders of magnitude onto an interpretable scale. Two conventional statistics are reported for comparison, $1 - Q^2LOO$ on the training set and $1 - R^2$ on the independent reference population. Where a Gaussian process is used, its posterior predictive distribution induces a distribution on the failure probability, from which a plug-in estimate, a posterior mean estimate and epistemic bounds are obtained following the reasoning of Bect et al.

3. Results

Reference response and fragility

Figure 3a plots the peak IDR against the roof drift ratio for the 5th, 50th and 95th percentiles of the propagated population. The relation is close to linear, with a peak-to-roof drift amplification of approximately 3.3 to 3.4, so that a roof drift of 4% corresponds to a peak interstorey drift near 13.5%. This value is consistent with a ground-storey mechanism: if the entire lateral deformation concentrates in the 3.50 m ground storey of a structure 12.50 m tall, the amplification is 12.50/3.50 = 3.57, which bounds the observed value from above. The three percentile curves are almost coincident and the dispersion band is narrow, indicating that the drift concentration mechanism is insensitive to the material and load uncertainty and that the dispersion in the response originates elsewhere.

Figure 3b gives the complementary cumulative distribution of the peak IDR over the 50-year reference period. Exceedance probabilities at the four thresholds are approximately 0.688 at 0.5% drift, 0.239 at 1.5%, 0.090 at 3.0% and 0.023 at 6.0%, and the curve continues smoothly to probabilities of order 10^{-3} near 20% drift. Figure 3c presents the corresponding fragility functions conditioned on peak ground acceleration [21]. The median capacity increases and the curves flatten as the limit state becomes more severe: LS1 is almost certainly exceeded above about 0.3 g, whereas LS4 reaches an exceedance probability of about 0.95 only at 1.0 g. The four curves are well separated, so the limit states interrogate distinct regions of the response distribution.

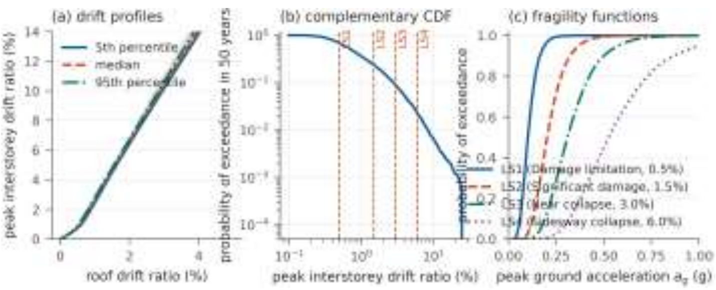

Figure 3. Reference structural response: (a) peak interstorey drift ratio versus roof drift ratio (5th, 50th and 95th percentiles); (b) complementary cumulative distribution of the peak interstorey drift ratio over the 50-year reference period, with the four limit-state thresholds; (c) fragility functions conditioned on peak ground acceleration.

Global sensitivity

Figure 4 reports first-order and total Sobol' indices for the eleven variables. Seismic intensity ag dominates, with S1 ≈ 0.61 and S^T ≈ 0.615, followed by the demand model error θD with S1 ≈ 0.33, and the capacity model error θC with S1 ≈ 0.05. The permanent-load factor kG and the concrete strength fc each contribute of the order of 0.003, and the remaining six variables are indistinguishable from zero on the scale of the figure [22].

The first-order indices sum to approximately unity and no total index exceeds its first-order counterpart by a resolvable amount, so the response is very nearly additive in the transformed input space, with negligible interaction. This structure is favourable to polynomial surrogates and accounts for the respectable performance of the sparse PCE in Figure 6. More consequentially, the material variables are effectively irrelevant to the reliability of this structure: concrete strength, yield strength, steel modulus, hardening ratio and cover together account for well under 1% of the response variance. Material variability is not mechanically unimportant, but once the hazard and the model-error terms are given realistic dispersions it is swamped by them. Reducing the coefficient of variation on fc through destructive testing will therefore not measurably improve the reliability estimate for a frame of this type, whereas reducing the demand model error will [24].

Because seismic intensity and the demand model error jointly carry about 94% of the response variance and act essentially as scaling factors, a surrogate can reproduce the bulk of the response distribution while distorting the conditional response near a high threshold. This is the mechanism by which a high value of R^2 can coexist with a poor reliability estimate.

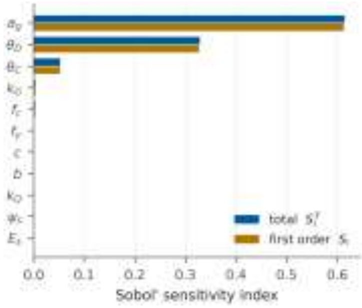

Figure 4. First-order and total Sobol' sensitivity indices for the eleven random variables.

Relation between fit statistics and reliability error

Figure 5 plots $|\Delta\beta|$ against a conventional fit statistic for each limit state, with four surrogate families distinguished; the upper row uses $1 - Q^2LOO$ on the training set and the lower row $1 - R^2$ on the independent reference population. A positive association is present in every panel, so the fit statistics are not uninformative, but the association is loose. At a fixed training-set error of $1 - Q^2LOO \approx 10^{-2}$, the observed reliability-index error ranges from below 3×10^{-3} to above 10^{-1}, that is, over one and a half to two orders of magnitude. The same dispersion appears when the statistic is computed on the independent population, which excludes overfitting of the leave-one-out estimate as an explanation [23]. The dotted line at $|\Delta\beta| = 0.1$ marks a tolerance that would be considered demanding in code calibration; points fall on both sides of it at every value of the fit statistic examined.

The scatter is structured. At LS3 and LS4 the multilayer perceptron produces the largest reliability errors at any given level of fit, with several configurations reaching $1 - R^2$ of order 10^{-2} while incurring $|\Delta\beta|$ of order 10^0. The Gaussian process occupies the opposite corner, attaining $|\Delta\beta|$ below 10^{-2} at fit levels where other families are an order of magnitude worse. Quadratic response surfaces and sparse PCE lie between, with the PCE showing a plateau in which further improvement in fit ceases to improve reliability accuracy, because the residual error of a truncated expansion is systematic near the tail rather than random [25]. A global fit statistic therefore constrains the reliability-index error only to within one and a half to two orders of magnitude, which is insufficient to certify an assessment.

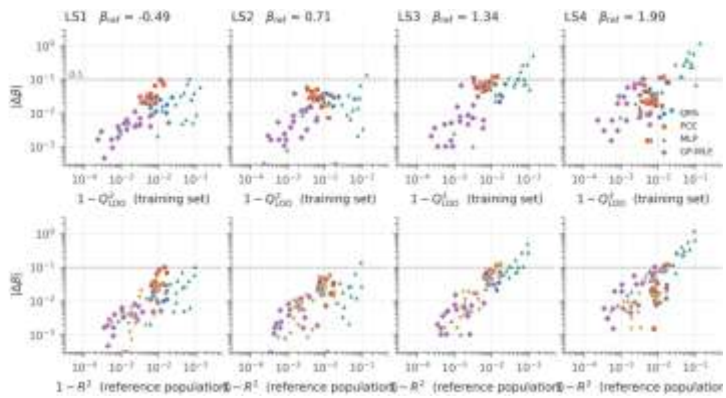

Figure 5. Reliability-index error against conventional goodness-of-fit statistics for four surrogate families at the four limit states. Upper row: leave-one-out error on the

training set. Lower row: error on the independent reference population. The dotted line marks an error of 0.1 in the reliability index.

Effect of the response transformation

Figure 7a reports the mean |Δβ| over budgets of at least 128 FE evaluations, by limit state, for each surrogate family fitted on the logarithm of the drift and on the untransformed drift. The logarithmic scale is superior for every family at LS1 to LS3. For the quadratic response surface the mean error falls from about 2.5×10^{-1} on the raw scale to about 2×10^{-2} on the log scale at LS1, and a comparable ratio persists at LS2 and LS3. The Gaussian process is the most accurate family on both scales but still gains from the transformation at LS1 and LS2 [26]. At LS4 the two scales converge, and for the MLP the raw-scale fit is marginally better. The transformation stabilises the variance of a response whose conditional dispersion grows with its mean (Box & Cox, 1964), which holds over most of the drift range; at the most severe limit state the contributing population lies in the far tail, where the conditional variance is more nearly constant and the transformation confers less benefit.

Figure 7b plots the failure-probability bias ratio at LS4 against R^2 recomputed on the untransformed response. Surrogates fitted on the raw scale attain values as low as -1×10^8 on this metric, that is, they are worse than the constant predictor by eight orders of magnitude, yet their bias ratios lie in the same band as those of surrogates whose R^2 on this metric is close to zero. The bias ratio is not ordered by this statistic [28]. The coefficient of determination evaluated on the untransformed response is therefore not merely a weak diagnostic but is decoupled from the reliability estimate without bound, because it is dominated by a small number of large-drift realisations carrying almost no probability mass. This follows directly from Equation (2), in which the reliability error depends on the conditional mean error at the threshold, a quantity that R^2 does not constrain.

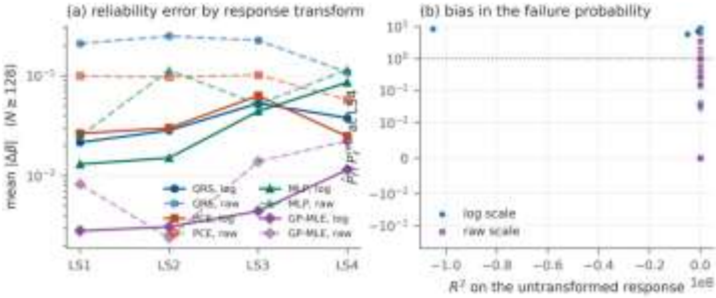

Figure 7. Effect of the response transformation: (a) mean reliability-index error for budgets of at least 128 finite element evaluations, by limit state, on the logarithmic and untransformed response scales; (b) failure-probability bias ratio at LS4 against the coefficient of determination recomputed on the untransformed response.

Convergence of the surrogate families

Figure 6 reports the mean |Δβ| against the number of FE evaluations, from 32 to about 1024, for the five surrogate families and the proposed scheme. The families separate consistently across limit states. The quadratic response surface converges rapidly at first and then plateaus at |Δβ| of about 2×10^{-2} to 6×10^{-2} without further improvement, the expected behaviour of a misspecified parametric model in which additional points reduce estimation variance but cannot remove structural bias [27]. The sparse PCE plateaus near 2.5×10^{-2} for the same reason. The MLP is erratic: at LS2 it descends below 10^{-3} by 1024 evaluations, but at LS3 and LS4 it remains at or above 10^{-2} over most of the budget range and is non-monotonic, reflecting the sensitivity of network training to initialisation in the small-sample regime characteristic of FE-based reliability problems.

The Gaussian process families dominate. GP-MLE and PCK reach $|\Delta\beta|$ of order 10^{-3} by 512 to 1024 evaluations at every limit state, one to one and a half orders of magnitude better than the parametric families at the same budget. The proposed scheme traces the lower envelope of the family curves at almost every budget and limit state, and retains its advantage in the small-budget regime: at 32 to 64 evaluations it already attains $|\Delta\beta|$ below 2×10^{-2} at LS1 and LS2, where the response surface is at 10^{-1} or worse. The ordering of families is stable across limit states, but the absolute error at fixed budget rises with the severity of the limit state for every family, from of order 10^{-2} at LS1 to 10^{-1} at LS4 for the weaker families at 128 evaluations. This is the behaviour predicted by Equation (4), since the amplification factor and the response density at the threshold both change with the limit state [29].

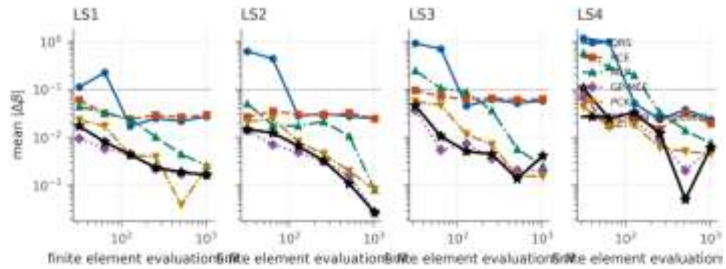

Figure 6. Convergence of the mean reliability-index error with the number of finite element evaluations, by surrogate family and limit state.

Enrichment strategy

Figure 8 compares the three enrichment strategies over budgets from about 40 to 250 evaluations. The two misclassification-directed strategies behave almost identically and both dominate random enrichment at every limit state. At LS2, LS3 and LS4 they drive $|\Delta\beta|$ below 10^{-3}, and in places below 10^{-4}, within approximately 80 to 150 FE evaluations, whereas random enrichment plateaus between 10^{-3} and 10^{-2} and does not reach 10^{-3} within the 250-evaluation budget [30]. At LS1 the advantage is smaller in the early phase, which is expected given a reference failure probability near 0.69, so that a large fraction of randomly drawn points already lie near the limit-state surface; beyond about 90 evaluations the directed strategies separate decisively.

The efficiency gain is therefore largest where the limit-state surface occupies a small fraction of the input probability mass, which is where random sampling rarely places points in the region that Equation (2) identifies as controlling. Attaining $|\Delta\beta| = 10^{-3}$ at LS3 requires roughly 100 to 130 evaluations with U-MLE and is not achieved by random enrichment within 250. The near-equivalence of U-MLE and U-LSW indicates that, once enrichment is directed at the limit-state surface, additional probability weighting adds little for this problem, consistent with the smooth unimodal response distribution of Figure 3b [33].

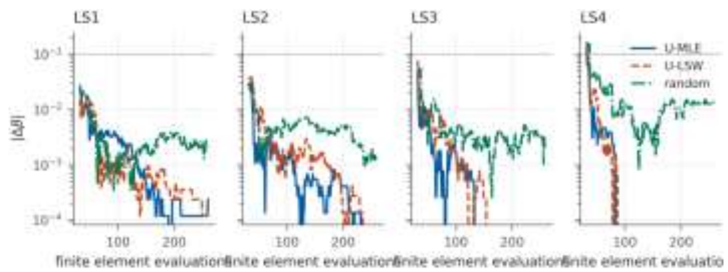

Figure 8. Active learning: reliability-index error against the number of finite element evaluations for two misclassification-directed enrichment strategies and for random enrichment.

Epistemic bounds on the failure probability

Figure 9 reports the failure probability at each limit state against the FE budget, with the plug-in estimate, the posterior mean estimate and the epistemic bounds derived from the Gaussian process posterior. The bounds contain the reference solution at every budget examined, so they express a statement about the reference value given the surrogate posterior rather than a heuristic error bar. They contract monotonically with the budget: at LS1 the interval at about 40 to 50 evaluations spans roughly 0.64 to 0.74 in failure probability, some ±7% of the reference value of 0.688, and has narrowed to a width negligible on the plotted scale by about 150 evaluations [31]. At LS4 the relative interval is initially wider, consistent with the smaller reference probability, and closes on the reference after roughly 100 evaluations.

The plug-in and posterior mean estimates are nearly indistinguishable beyond the first few tens of evaluations and both converge to the reference. Their small early separation at LS2 and LS4 reflects the nonlinearity of the indicator function under the posterior: when the surrogate is uncertain, averaging the indicator over the posterior differs from evaluating it at the posterior mean. The separation vanishes as the posterior concentrates [34]. Reported alongside the point estimate, these bounds quantify how much further FE evaluations could move the answer.

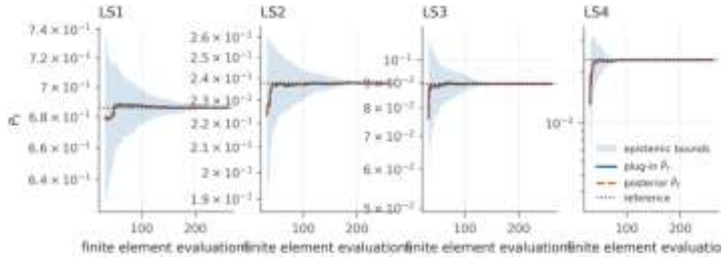

Figure 9. Failure probability against the finite element budget, with epistemic bounds derived from the Gaussian process posterior, the plug-in and posterior-mean estimates, and the reference solution.

Verification of the propagation law

Figure 10a plots the observed $\Delta\beta$ against the value predicted by Equation (3), pooling all four limit states and all surrogate configurations [32]. The points fall on the identity line, with a fitted slope of 1.04, within 4% of the unit slope that exact first-order agreement would give, and a coefficient of determination of 0.93 over a range of $\Delta\beta$ from about -0.08 to +0.12. The relation therefore accounts for 93% of the variance in the reliability-index error across surrogate families that differ in structure, in training budget and in the limit state to which they are applied [35]. Its predictor requires only the response density at the threshold, the standard normal density at the current reliability index, andthe conditional mean surrogate error at the threshold, none of which requires further FE evaluations, so Equation (3) can be evaluated during the analysis and used to decide whether to enrich the design of experiments.

Figure 10b tests the amplification factor of Equation (4). The predicted values are 1.38, 0.87, 0.94 and 1.00 at LS1 to LS4, taking a non-monotonic shape that is largest at the least severe limit state, falls to a minimum near LS2 and returns towards unity at LS4. The observed amplification, defined as the median of $|\Delta\beta| / |E[\varepsilon]|$ with the interquartile range shown, is 1.66, 1.21, 0.94 and 1.31. The predicted values lie within the observed interquartile range at all four limit states, and the agreement is essentially exact at LS3. The tendency of the observed medians to exceed the predicted values at LS1, LS2 and LS4 is consistent with the first-order nature of the derivation, since Equation (2) neglects a second-order term involving the conditional variance of the error at the threshold whose contribution to the magnitude of the perturbation is non-negative.

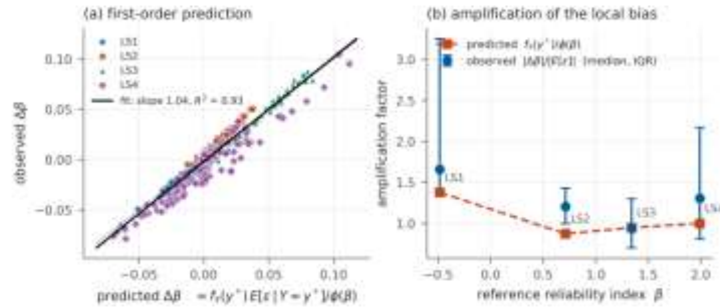

Figure 10. Verification of the first-order propagation law: (a) observed against predicted reliability-index error, pooled over limit states and surrogate configurations; (b) predicted and observed amplification factor against the reference reliability index (observed values: median and interquartile range).

4. Discussion

The reliability index responds to surrogate error through a narrow channel, the conditional mean error on the limit-state surface, whose width is set by the ratio of the response density at the threshold to the standard normal density at the reliability index; everything else about the error field enters at second order. The empirical findings then follow as predictions. Global fit statistics scatter against $|\Delta\beta|$ because they weight the bulk of the distribution rather than the threshold [36]. The coefficient of determination on the untransformed response can be extremely negative while the failure probability remains usable, because it is dominated by rare large-drift realisations of negligible probability. Threshold-directed enrichment outperforms random enrichment because it reduces the error where Equation (2) locates the sensitivity. Gaussian process surrogates dominate parametric ones not because they fit better on average, which at some budgets they do not, but because their local flexibility permits a small conditional bias near the threshold and their posterior supplies the local error estimate that Equation (5) requires. The sensitivity result is the mechanical counterpart: with two variables carrying about 94% of the variance and acting as scaling factors, a surrogate can reproduce the bulk of the distribution while distorting the conditional response near a high threshold.

The finding that surrogate accuracy must be assessed locally is consistent with the trajectory of the active-learning literature and with the benchmark of Moustapha et al.. What the present analysis adds is the justification that literature has lacked. The learning functions of Echard et al. Sun et al, Wang and Shafieezadeh and Kim and Song are, in the language of Equation (2), devices for driving the conditional error at the threshold towards zero; the derivation shows why that is the right target and, through Equation (5), how close is close enough [37], [38], [39]. Teixeira et al. call for principled stopping criteria in adaptive metamodel-based reliability analysis, and Equation (5) supplies one that is tied to a stated tolerance on the reported reliability index rather than to the stabilisation of an estimated failure probability.

Three implications extend beyond the case study. Since $A(y*)$ varies with the threshold, a surrogate validated once cannot be assumed adequate for every limit state of a multi-level assessment, and FE budget should be allocated in inverse proportion to the tolerance implied by Equation (5). The response transformation is part of the reliability model rather than a presentational choice, since it changes both the conditional bias at the threshold and the density fY there, and the two effects need not act in the same direction, which is why the benefit of the logarithmic scale largely disappears at LS4 [40]. Finally, the epistemic uncertainty contributed by the surrogate is reducible by computation and belongs in the reported result [41].

For practice the sequence is therefore: verify discretisation first, since surrogate error cannot otherwise be separated from discretisation error; fit the surrogate on the logarithm of the drift demand unless the governing limit state lies in the far tail; enrich the design with a misclassification-directed learning function; check Equation (5) at each limit state separately; and report the epistemic bounds with the failure probability [42]. The sensitivity result adds that investigative effort is better directed at reducing demand model error, through improved modelling of the storey mechanism or calibration of θD against experimental evidence, than at additional material testing.

The scope of the numerical evidence should be kept in view. It rests on one planar frame with one detailing configuration, assessed through an equivalent-SDOF pushover procedure in which higher-mode and record-to-record effects are carried by the lumped term θD rather than resolved explicitly as incremental dynamic analysis would resolve them (Vamvatsikos & Cornell, 2002), over reference failure probabilities between 0.688 and 0.023 [43], [44]. The derivation of Equations (2) to (5) makes no reference to the structural system, but the reported magnitudes, in particular amplification factors near unity and the ordering of surrogate families, are properties of this structure and this probabilistic model [45]. Extension to three-dimensional and irregular structures with dynamic demand estimation, derivation of the second-order term for the small failure probabilities relevant to new design, and embedding of Equation (5) as a stopping rule inside an active-learning loop are the natural next steps.

5. Conclusion

A first-order relation was derived showing that the error in the reliability index induced by surrogate substitution is governed by the conditional mean surrogate error at the limit-state threshold, amplified by the factor $A(y*) = fY(y*)/\varphi(\beta)$, with the global error variance entering only at second order. The relation was verified on a four-storey non-seismically detailed RC frame across five surrogate families, four limit states and FE budgets from 32 to 1024 evaluations, with a regression slope of 1.04 and a coefficient of determination of 0.93, and with predicted amplification factors falling inside the observed interquartile range at all four limit states.

Conventional goodness-of-fit statistics were found to leave the reliability-index error undetermined by one and a half to two orders of magnitude, and the coefficient of determination computed on the untransformed response was found to be decoupled from the reliability estimate without bound. Variance decomposition attributed approximately 61% of the response variance to seismic intensity and 33% to demand model error, with all material variables together contributing under 1% and interactions negligible. Logarithmic transformation of the response improved the reliability-index error by up to an order of magnitude at the less severe limit states, with the benefit vanishing at sideway collapse. Misclassification-directed enrichment attained $|\Delta\beta|$ below 10^{-3} within roughly 100 FE evaluations, where random enrichment did not reach that level within 250. Epistemic bounds derived from the Gaussian process posterior contained the reference failure probability at all budgets and contracted monotonically with the training set.

The practical consequence is that surrogate validation in reliability-oriented assessment should be conducted against the local criterion of Equation (5), separately at each limit state, rather than against a global fit statistic, and that the epistemic bounds induced by the surrogate should be reported with the failure probability. These recommendations are conditional on the scope stated in Section 5 and do not by themselves establish compliance with any assessment standard.

REFERENCES

[1] American Society of Civil Engineers, Seismic Evaluation and Retrofit of Existing Buildings, ASCE/SEI 41-17. Reston, VA, USA: ASCE, 2017.

[2] J. W. Baker, “Efficient analytical fragility function fitting using dynamic structural analysis,” Earthquake Spectra, vol. 31, no. 1, pp. 579–599, 2015, doi: 10.1193/021113EQS025M.

[3] J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez, “Sequential design of computer experiments for the estimation of a probability of failure,” Statistics and Computing, vol. 22, no. 3, pp. 773–793, 2012, doi: 10.1007/s11222-011-9241-4.

[4] B. J. Bichon, M. S. Eldred, L. P. Swiler, S. Mahadevan, and J. M. McFarland, “Efficient global reliability analysis for nonlinear implicit performance functions,” AIAA Journal, vol. 46, no. 10, pp. 2459–2468, 2008, doi: 10.2514/1.34321.

[5] D. Biskinis and M. N. Fardis, “Deformations at flexural yielding of members with continuous or lap-spliced bars,” Structural Concrete, vol. 11, no. 3, pp. 127–138, 2010, doi: 10.1680/stco.2010.11.3.127.

[6] G. Blatman and B. Sudret, “Adaptive sparse polynomial chaos expansion based on least angle regression,” Journal of Computational Physics, vol. 230, no. 6, pp. 2345–2367, 2011, doi: 10.1016/j.jcp.2010.12.021.

[7] J.-M. Bourinet, “Rare-event probability estimation with adaptive support vector regression surrogates,” Reliability Engineering & System Safety, vol. 150, pp. 210–221, 2016, doi: 10.1016/j.ress.2016.01.023.

[8] G. E. P. Box and D. R. Cox, “An analysis of transformations,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 26, no. 2, pp. 211–243, 1964, doi: 10.1111/j.2517-6161.1964.tb00553.x.

[9] D. Celarec and M. Dolšek, “The impact of modelling uncertainties on the seismic performance assessment of reinforced concrete frame buildings,” Engineering Structures, vol. 52, pp. 340–354, 2013, doi: 10.1016/j.engstruct.2013.02.036.

[10] European Committee for Standardization, Eurocode—Basis of Structural Design, EN 1990:2002. Brussels, Belgium: CEN, 2002.

[11] European Committee for Standardization, Eurocode 8: Design of Structures for Earthquake Resistance—Part 3: Assessment and Retrofitting of Buildings, EN 1998-3:2005. Brussels, Belgium: CEN, 2005.

[12] J. Coleman and E. Spacone, “Localization issues in force-based frame elements,” Journal of Structural Engineering, vol. 127, no. 11, pp. 1257–1265, 2001, doi: 10.1061/(ASCE)0733-9445(2001)127:11(1257).

[13] A. Der Kiureghian, “Analysis of structural reliability under parameter uncertainties,” Probabilistic Engineering Mechanics, vol. 23, no. 4, pp. 351–358, 2008, doi: 10.1016/j.probengmech.2007.10.011.

[14] A. Der Kiureghian and O. Ditlevsen, “Aleatory or epistemic? Does it matter?” Structural Safety, vol. 31, no. 2, pp. 105–112, 2009, doi: 10.1016/j.strusafe.2008.06.020.

[15] M. Dolšek, “Incremental dynamic analysis with consideration of modeling uncertainties,” Earthquake Engineering & Structural Dynamics, vol. 38, no. 6, pp. 805–825, 2009, doi: 10.1002/eqe.869.

[16] V. Dubourg, B. Sudret, and F. Deheeger, “Metamodel-based importance sampling for structural reliability analysis,” Probabilistic Engineering Mechanics, vol. 33, pp. 47–57, 2013, doi: 10.1016/j.probengmech.2013.02.002.

[17] B. Echard, N. Gayton, and M. Lemaire, “AK-MCS: An active learning reliability method combining Kriging and Monte Carlo simulation,” Structural Safety, vol. 33, no. 2, pp. 145–154, 2011, doi: 10.1016/j.strusafe.2011.01.002.

[18] P. Fajfar, “A nonlinear analysis method for performance-based seismic design,” Earthquake Spectra, vol. 16, no. 3, pp. 573–592, 2000, doi: 10.1193/1.1586128.

[19] M. N. Fardis, Seismic Design, Assessment and Retrofitting of Concrete Buildings: Based on EN-Eurocode 8. Dordrecht, The Netherlands: Springer, 2009, doi: 10.1007/978-1-4020-9842-0.

[20] Fédération Internationale du Béton, fib Model Code for Concrete Structures 2010. Berlin, Germany: Ernst & Sohn, 2013, doi: 10.1002/9783433604090.

[21] P. Gardoni, A. Der Kiureghian, and K. M. Mosalam, “Probabilistic capacity models and fragility estimates for reinforced concrete columns based on experimental observations,” Journal of Engineering Mechanics, vol. 128, no. 10, pp. 1024–1038, 2002, doi: 10.1061/(ASCE)0733-9399(2002)128:10(1024).

[22] J. Ghosh, J. E. Padgett, and L. Dueñas-Osorio, “Surrogate modeling and failure surface visualization for efficient seismic vulnerability assessment of highway bridges,” Probabilistic Engineering Mechanics, vol. 34, pp. 189–199, 2013, doi: 10.1016/j.probengmech.2013.09.003.

[23] B. U. Gokkaya, J. W. Baker, and G. G. Deierlein, “Quantifying the impacts of modeling uncertainties on the seismic drift demands and collapse risk of buildings with implications on seismic design checks,” Earthquake Engineering & Structural Dynamics, vol. 45, no. 10, pp. 1661–1683, 2016, doi: 10.1002/eqe.2740.

[24] A. M. Hasofer and N. C. Lind, “Exact and invariant second-moment code format,” Journal of the Engineering Mechanics Division, vol. 100, no. 1, pp. 111–121, 1974, doi: 10.1061/JMCEA3.0001848.

[25] F. Jalayer and C. A. Cornell, “Alternative non-linear demand estimation methods for probability-based seismic assessments,” Earthquake Engineering & Structural Dynamics, vol. 38, no. 8, pp. 951–972, 2009, doi: 10.1002/eqe.876.

[26] Joint Committee on Structural Safety, JCSS Probabilistic Model Code. JCSS, 2001.

[27] J. Kim and J. Song, “Probability-Adaptive Kriging in n-ball (PAK-Bn) for reliability analysis,” Structural Safety, vol. 85, Art. no. 101924, 2020, doi: 10.1016/j.strusafe.2020.101924.

[28] A. B. Liel, C. B. Haselton, G. G. Deierlein, and J. W. Baker, “Incorporating modeling uncertainties in the assessment of seismic collapse risk of buildings,” Structural Safety, vol. 31, no. 2, pp. 197–211, 2009, doi: 10.1016/j.strusafe.2008.06.002.

[29] J. B. Mander, M. J. N. Priestley, and R. Park, “Theoretical stress-strain model for confined concrete,” Journal of Structural Engineering, vol. 114, no. 8, pp. 1804–1826, 1988, doi: 10.1061/(ASCE)0733-9445(1988)114:8(1804).

[30] S. Mangalathu, G. Heo, and J.-S. Jeon, “Artificial neural network based multi-dimensional fragility development of skewed concrete bridge classes,” Engineering Structures, vol. 162, pp. 166–176, 2018, doi: 10.1016/j.engstruct.2018.01.053.

[31] S. Marelli and B. Sudret, “An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis,” Structural Safety, vol. 75, pp. 67–74, 2018, doi: 10.1016/j.strusafe.2018.06.003.

[32] R. E. Melchers and A. T. Beck, Structural Reliability Analysis and Prediction, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2018, doi: 10.1002/9781119266105.

[33] M. Moustapha, S. Marelli, and B. Sudret, “Active learning for structural reliability: Survey, general framework and benchmark,” Structural Safety, vol. 96, Art. no. 102174, 2022, doi: 10.1016/j.strusafe.2021.102174.

[34] T. B. Panagiotakos and M. N. Fardis, “Deformations of reinforced concrete members at yielding and ultimate,” ACI Structural Journal, vol. 98, no. 2, 2001, doi: 10.14359/10181.

[35] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. Cambridge, MA, USA: MIT Press, 2006.

[36] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola, “Variance based sensitivity analysis of model output: Design and estimator for the total sensitivity index,” Computer Physics Communications, vol. 181, no. 2, pp. 259–270, 2010, doi: 10.1016/j.cpc.2009.09.018.

[37] R. Schöbi, B. Sudret, and S. Marelli, “Rare event estimation using polynomial-chaos Kriging,” ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, vol. 3, no. 2, Art. no. D4016002, 2017, doi: 10.1061/AJRUA6.0000870.

[38] R. Schöbi, B. Sudret, and J. Wiart, “Polynomial-chaos-based Kriging,” International Journal for Uncertainty Quantification, vol. 5, no. 2, pp. 171–193, 2015, doi: 10.1615/Int.J.UncertaintyQuantification.2015012467.

[39] M. H. Scott and G. L. Fenves, “Plastic hinge integration methods for force-based beam-column elements,” Journal of Structural Engineering, vol. 132, no. 2, pp. 244–252, 2006, doi: 10.1061/(ASCE)0733-9445(2006)132:2(244).

[40] I. M. Sobol’, “Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates,” Mathematics and Computers in Simulation, vol. 55, nos. 1–3, pp. 271–280, 2001, doi: 10.1016/S0378-4754(00)00270-6.

[41] E. Spacone, F. C. Filippou, and F. F. Taucer, “Fibre beam-column model for non-linear analysis of R/C frames: Part I. Formulation,” Earthquake Engineering & Structural Dynamics, vol. 25, no. 7, pp. 711–725, 1996, doi: 10.1002/(SICI)1096-9845(199607)25:7<711::AID-EQE576>3.0.CO;2-9.

[42] B. Sudret, “Global sensitivity analysis using polynomial chaos expansions,” Reliability Engineering & System Safety, vol. 93, no. 7, pp. 964–979, 2008, doi: 10.1016/j.ress.2007.04.002.

[43] H. Sun, H. V. Burton, and H. Huang, “Machine learning applications for building structural design and performance assessment: State-of-the-art review,” Journal of Building Engineering, vol. 33, Art. no. 101816, 2021, doi: 10.1016/j.jobe.2020.101816.

[44] Z. Sun, J. Wang, R. Li, and C. Tong, “LIF: A new Kriging based learning function and its application to structural reliability analysis,” Reliability Engineering & System Safety, vol. 157, pp. 152–165, 2017, doi: 10.1016/j.ress.2016.09.003.

[45] R. Teixeira, M. Nogal, and A. O’Connor, “Adaptive approaches in metamodel-based reliability analysis: A review,” Structural Safety, vol. 89, Art. no. 102019, 2021, doi: 10.1016/j.strusafe.2020.102019.