Available online at www.CivileJournal.org

Civil Engineering Journal

(E-ISSN: 2476-3055; ISSN: 2676-6957)

Vol. 12, No. 04, April, 2026

Spatially-Adaptive Calibration for Reliable Uncertainty Quantification in Seismic Response Prediction of RC Frames

Donwoo Lee [1*], Seungjae Lee [1]

[1] School of Industrial Design & Architectural Engineering, Korea University of Technology & Education, Cheonan 31253, Republic of Korea.

Received 25 January 2026; Revised 23 March 2026; Accepted 28 March 2026; Published 01 April 2026

Abstract

Data-driven models offer the computational speed needed for rapid post-earthquake assessment, but their uncertainty estimates must be trustworthy to support safety decisions. This study reveals that Monte Carlo dropout uncertainty for RC frame seismic response prediction is severely miscalibrated: 95% prediction intervals capture only 46.6% of actual responses, meaning Immediate Occupancy assessments under ASCE 41-17 would be unconservative in over half of cases. We address this through post-hoc Temperature Scaling calibration. While a global scaling parameter ($T^* = 4.40$) reduces calibration error by 91.4%, we discover that the optimal calibration factor varies systematically across structural locations: T^* ranges from 1.94 at fixed-base nodes to 5.52 at mid-height floors—a 2.8-fold variation that single-parameter approaches cannot capture. This spatial variation reflects physical differences in prediction uncertainty: boundary-constrained nodes exhibit lower uncertainty requiring less scaling, while mid-height nodes dominated by higher-mode contributions show greater uncertainty underestimation. Building on this finding, we propose floor-adaptive calibration using location-specific scaling factors. Compared to global calibration, this approach reduces average calibration error by an additional 62%, with improvements of 61-70% at ground and top floors, where global calibration performs worst. The method requires no model retraining—only a lookup table mapping floor levels to optimal scaling factors. Validation across 12 RC frames (3-7 stories), 2,400 analysis cases, and 35,000+ node-level predictions confirms that spatially adaptive calibration provides more reliable uncertainty estimates across all structural locations, enabling trustworthy confidence intervals for performance-based post-earthquake assessment.

Keywords: Uncertainty Quantification; Post-hoc Calibration; Seismic Response Prediction; Graph Neural Network; Reinforced Concrete Frames; Post-Earthquake Assessment.

1. Introduction

Following an earthquake, engineers face urgent decisions. Is a building safe for continued occupancy? Can the structure serve emergency responders? Does it require immediate repairs? Immediate Occupancy (IO) and Serviceability Limit State (SLS) assessments, as defined in ASCE 41-17 [1], FEMA 273 [2], and Eurocode 8 [3], provide quantitative frameworks for these decisions but require reliable estimates of structural response. Detailed structural analysis within hours of an earthquake is typically impractical, creating demand for rapid assessment methods.

Data-driven prediction models offer the speed required for real-time assessment [4]. CNNs [5, 6], LSTMs [7], and other deep learning architectures [8] have been successfully applied to structural response prediction, with Graph Neural Networks (GNNs) showing particular promise because they can directly model the joint-member connectivity of building frames [9-11]. GNN-based damage detection [12] and response prediction [13] are active research areas, and

* Corresponding author: lov1004ely@koreatech.ac.kr

https://doi.org/10.28991/CEJ-2026-012-04-02

Graph Attention Networks (GAT) [14, 15] can learn to weight structurally important joints more heavily, providing physically interpretable predictions. PyTorch Geometric [16] provides efficient tools for these implementations. Recent studies further demonstrate the growing capability of these approaches: Liu et al. [17] proposed a graph network-based simulator for spatiotemporal structural response modeling in buildings, Chou et al. [18] developed a graph-based LSTM network for predicting nonlinear floor responses of steel buildings under bidirectional ground motions, and Shen et al. [19] introduced a CNN-LSTM architecture with attention mechanism and transfer learning for seismic response prediction. These advances have substantially improved prediction accuracy, yet they focus on point estimation without examining the reliability of associated uncertainty estimates.

But prediction accuracy alone is insufficient for safety decisions [20]. Engineers need not only point predictions but also information about how much those predictions can be trusted [21]. If a predicted inter-story drift is below the 1.0% IO threshold, but actual drift could substantially exceed that prediction, the prediction alone cannot support a safety decision. Monte Carlo (MC) Dropout provides a practical approach to uncertainty estimation in neural networks [22, 23]—by enabling dropout at inference and performing multiple forward passes, one obtains both mean predictions and variance estimates. However, prior work has shown that MC Dropout often produces miscalibrated uncertainty [24, 25]: intervals labeled as 95% may contain far fewer actual values. The importance of uncertainty quantification in seismic response analysis is increasingly recognized: Kim & Wang [26] proposed a dimensionality reduction-based stochastic simulator for seismic response UQ, and Choi et al. [27] developed a neural network-parameterized Gaussian process for regional seismic loss estimation that inherently provides uncertainty bounds. A recent comprehensive review [28] identifies uncertainty quantification as a critical frontier in deep learning for earthquake engineering. However, these studies do not examine whether predicted uncertainty estimates are well-calibrated—whether stated confidence levels correspond to actual coverage probabilities.

Why does this matter for safety assessment? Suppose an engineer declares a building safe based on a 95% prediction interval upper bound falling below the IO drift threshold. If that interval is properly calibrated, there is roughly a 2.5% chance the actual drift exceeds the upper bound. But if the interval is severely overconfident—capturing only 50% of actual responses—the safety judgment carries an entirely different risk profile, and the concept of a confidence interval loses its meaning.

Calibration [29, 30] is the statistical consistency between predicted confidence levels and observed frequencies. In a well-calibrated model, a 90% interval should contain approximately 90% of actual values. Without such statistical consistency, the core premise of performance-based assessment—probabilistic demand-capacity comparison—loses its validity. Post-hoc calibration techniques [31-33] adjust variance estimates from trained models to restore this consistency. Notably, single-parameter scaling approaches require no model retraining and add negligible inference-time cost (a single multiplication per prediction), making them preferable to Deep Ensembles [24] or Bayesian Neural Networks [34] for real-time applications. In the broader machine learning community, recent work has begun to address calibration challenges specific to graph architectures; Zhuang et al. [35] demonstrated that standard temperature scaling is insufficient for GNNs and proposed ensemble-based calibration for graph classification. However, calibration for regression-based GNN predictions in engineering applications—where prediction intervals directly inform safety decisions—remains unaddressed. Three specific gaps motivate this study: (1) no prior work has quantified the degree of miscalibration in data-driven structural response prediction; (2) whether optimal calibration factors vary spatially within a structure has not been investigated; and (3) no practical post-hoc calibration methods have been validated for structural engineering applications.

This study addresses two research questions: (RQ1) Can miscalibrated uncertainty estimates from data-driven models be corrected to provide statistically valid prediction intervals for post-earthquake safety assessment? (RQ2) Does the optimal calibration factor vary spatially within a structure, and if so, can location-specific calibration improve reliability? We investigate using 2,400 analysis cases from 12 RC frame structures (3-7 stories) subjected to 200 ground motions, comprising over 35,000 node-level predictions. Our contributions are: (1) quantifying severe miscalibration in GNN-based seismic response prediction—95% prediction intervals capture only 46.6% of actual responses, rendering IO assessments unconservative in over half of cases; (2) demonstrating that global post-hoc calibration ($T^* = 4.40$) reduces calibration error by 91.4%, restoring 95% coverage to 94.9%; (3) discovering systematic spatial variation in optimal calibration factors—T^* ranges from 1.94 at boundary-constrained nodes to 5.52 at mid-height floors, revealing that uncertainty characteristics differ across structural locations; (4) proposing floor-adaptive calibration that reduces calibration error by an additional 62% compared to global calibration, with 61-70% improvement at locations where global calibration performs worst; and (5) providing physical interpretation linking spatial T* variation to structural mechanics: boundary constraints, higher-mode contributions, and load transfer paths.

We use linear response, which aligns with the IO and SLS assessment context [1] and provides exact ground truth for validating the methodology. Since linear analysis introduces no inherent uncertainty, any calibration error observed can be attributed to the methodology, providing a clean testbed. Extension to nonlinear response, measured data, and inverse problems is left to future work. Section 2 describes the methodology, Section 3 the dataset, Section 4 the results, Section 5 discusses implications and limitations, and Section 6 concludes.

2. Methodology

2.1. Data-Driven Prediction Model

The prediction model uses a GAT [14] to process RC frame structures directly, as illustrated in Figure 1. Joints correspond to graph nodes and beam-column members to edges, so structural connectivity is naturally represented. The GAT attention mechanism [15] learns differential contributions from neighboring nodes, automatically assigning higher weights to connections that significantly influence structural response. For node i, the attention coefficient from neighbor j is computed using Equation 1, normalized via SoftMax in Equation 2, and the node representation is updated as shown in Equation 3:

$$e_{ij} = \text{LeakyReLU}(a^T[Wh_i||Wh_j||W_e f_{ij}]) \quad (1)$$

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) \quad (2)$$

$$h'_i = \sigma\left(\sum_{j\in N(i)} \alpha_{ij} W h_j\right) \quad (3)$$

where, W is the node transformation matrix, a is the attention vector, W_e is the edge feature transformation matrix, f_{ij} is the edge feature vector for edge (i, j), || denotes concatenation, N(i) is the neighborhood of i, and σ is the activation function (ELU). Multi-head attention with K heads, expressed in Equation 4, captures diverse structural relationships:

$$h'_i = ||_{k=1}^{K} \sigma\left(\sum_{j\in N(i)} \alpha_{ij} W h_j\right) \quad (4)$$

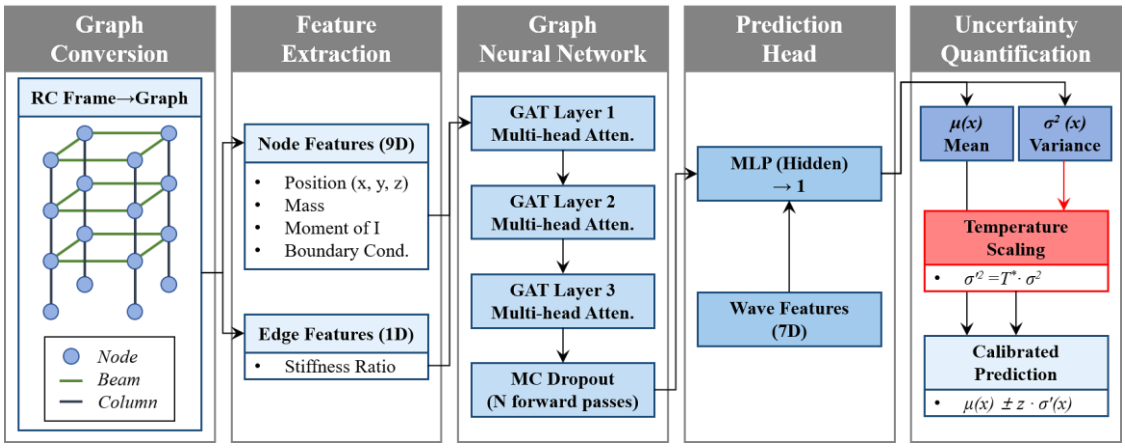

Figure 1. GAT model architecture for seismic response prediction

In structural terms, the attention mechanism learns which neighboring joints most influence each node's seismic response—for example, a node connected to a stiff column may receive higher attention than one connected to a flexible beam, reflecting the relative contribution of each member to load transfer. The model architecture consists of an input layer (9-dimensional node features, 1-dimensional edge features), three GAT layers (hidden dimensions 64/128/256, heads 8/4/2, dropout 0.1), and a final MLP combining GAT output with 7-dimensional ground motion features to predict per-node displacement. Implementation uses PyTorch Geometric [16].

2.2. Uncertainty Quantification

Monte Carlo Dropout [22, 23] estimates prediction distributions by enabling dropout at inference and performing N forward passes. For input x, the t-th forward pass yields $f_{\theta t}(x)$, where θ_t denotes the model parameters with the t-th dropout mask. The prediction mean and variance are computed using Equations 5 and 6, respectively:

$$\mu(x) = (1/\text{N}) \sum_{t=1}^{N} f_{\theta_t}(x) \quad (5)$$

$$\sigma^2(x) = (1/\text{N}) \sum_{t=1}^{N} [f_{\theta_t}(x) - \mu(x)]^2 \quad (6)$$

We use N=100, at which variance estimates converge [19]. Under Gaussian assumptions, the (1-α) confidence prediction interval is defined by Equation 7:

$$PI_{1-\alpha}(x) = [\mu(x) - z_{\alpha/2} \cdot \sigma(x), \mu(x) + z_{\alpha/2} \cdot \sigma(x)] \quad (7)$$

where $z_{\alpha/2}$ is the standard normal quantile (1.96 for 95%).

2.3. Post-hoc Calibration

Temperature Scaling [31] adjusts predicted variance by a single parameter T, as shown in Equation 8.

$$\sigma'^2(x) = T \cdot \sigma^2(x) \tag{8}$$

Calibrated standard deviation is $\sigma'(x) = \sqrt{T} \cdot \sigma(x)$. The optimal temperature T^* minimizes negative log-likelihood (NLL) on the validation set, as defined inequation 9, where y_i is the observed response value:

$$T^* = \text{argmin}_T \sum_i [(y_i - \mu(x_i))^2/(T \cdot \sigma^2(x_i)) + \log(T \cdot \sigma^2(x_i))] \tag{9}$$

This objective is convex in T, enabling efficient global optimization. Computational cost is minimal: calibration requires one-time optimization on the validation set (seconds), and inference adds only a single multiplication per prediction. This efficiency enables real-time application, contrasting with computationally intensive alternatives like Deep Ensembles [24] or Bayesian Neural Networks [34].

2.4. Floor-Adaptive Calibration

Standard Temperature Scaling assumes a single T^* applies uniformly across all predictions. However, structural mechanics suggests that prediction uncertainty may vary spatially—nodes at different floor levels experience different dynamic characteristics, boundary conditions, and load transfer paths. We therefore investigate whether optimal calibration factors vary by structural location.

For floor-adaptive calibration, we partition nodes by floor level and compute location-specific optimal temperatures. Let $F=\{f_1, f_2, \ldots, f_k\}$ denote floor groups (e.g., ground, lower floors, mid-height, upper floors). For each group f_j, we optimize:

$$T^*_{f_j} = \text{argmin}_T \sum_{i \in f_j} [(y_i - \mu(x_i))^2/(T \cdot \sigma^2(x_i)) + \log(T \cdot \sigma^2(x_i))] \tag{10}$$

At inference, calibrated variance for node i at floor level f_j is computed as:

$$\sigma'^2(x_i) = T_{f_j} \cdot \sigma^2(x_i) \tag{11}$$

This approach requires only a lookup table mapping floor levels to scaling factors, adding negligible computational overhead while potentially capturing spatial variation in uncertainty characteristics.

2.5. Evaluation Metrics

Coverage probability is the fraction of prediction intervals containing actual values, as expressed in Equation 12, where 1[·] is the indicator function:

$$\text{Coverage}_{1-\alpha} = (1/N) \sum_{i=1^N} 1[y_i \in PI_{1-\alpha}(x_i)] \tag{12}$$

For well-calibrated models, 95% intervals should contain approximately 95% of actual values. Average Calibration Error (ACE) is the mean absolute difference between expected and observed coverage across confidence levels, defined by Equation 13:

$$ACE = (1/|L|) \sum_{\alpha \in L} |\text{Coverage}_{1-\alpha} - (1-\alpha)| \tag{13}$$

where L = {0.50, 0.80, 0.90, 0.95, 0.99}. ACE = 0 indicates perfect calibration.

2.6. Training Configuration

Optimization uses Kingma & Ba (Adam) [36], learning rate 0.001, ReduceLROnPlateau scheduler, batch size 32, maximum 200 epochs, early stopping at 20 epochs. Temperature Scaling uses search range [0.1, 20] with L-BFGS optimization. Hardware: NVIDIA GeForce RTX 4070 Laptop GPU, Intel Core i9-13900H, 32GB RAM. Software: Python 3.10, PyTorch 2.8 [37], PyTorch Geometric 2.6 [16], OpenSeesPy 3.5 [38].

3. Dataset

3.1. RC Frame Structures

We analyze 12 RC frame structures representing a range of configurations common in seismic regions: 3, 5, and 7 stories; span layouts of 2×2, 3×2, 2×3, and 3×3 bays. $struct_{001\text{-}003}$ are 3-story (36-48 joints), $struct_{004\text{-}006}$ are 5-story (54-72 joints), $struct_{007\text{-}009}$ are 7-story (72-96 joints), $struct_{010}$ is 5-story 3×3 (90 joints), $struct_{011}$ is 7-story 3×3 (126 joints), $struct_{012}$ is 3-story 3×3 (64 joints). Total: 894 joints, 1,638 members, fundamental periods 0.42-1.02 s. Design parameters follow typical practice: concrete f'c = 30 MPa, Ec = 27,000 MPa, ν = 0.2; reinforcement fy = 400 MPa, Es = 200,000 MPa; column sections 400×400 to 600×600 mm; beam sections 300×500 to 400×700 mm; story height 3.2 m; bay width 6.0 m; floor mass 500 kg/m^2 as lumped mass; 5% Rayleigh damping based on modes 1 and 3.

Structural analysis uses OpenSeesPy [38] for linear time-history analysis (elasticBeamColumn elements, Newmark-β integration β=0.25/γ=0.5, 0.01 s time step). Each structure is modeled as a three-dimensional frame with six degrees of freedom per node, with fixed-base boundary conditions and rigid diaphragm constraints at each floor level. Ground motions are applied as uniform base excitation, and the analysis follows a three-stage procedure: static gravity analysis, eigenvalue analysis to extract natural periods, and transient dynamic analysis. Linear analysis is appropriate for IO and SLS assessment scenarios [1] motivating this work and provides exact ground truth for rigorous calibration evaluation. Figure 2 shows representative 3-story, 5-story, and 7-story structure geometries, and Figure 3 illustrates the distribution of 12 structures by story count, fundamental period, and bay configuration.

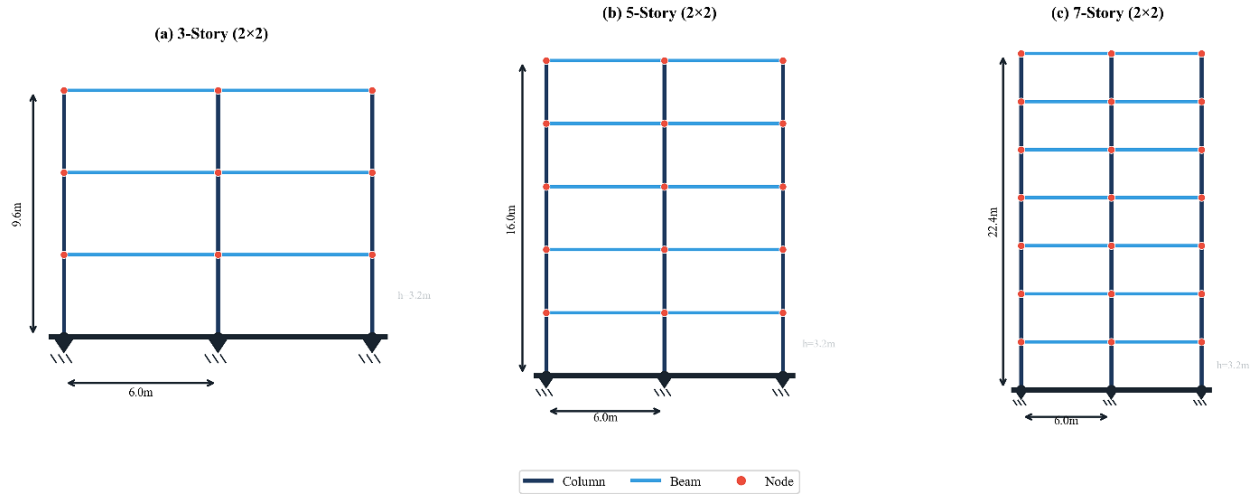

Figure 2. RC frame structure geometry: (a) 3-story (2×2), (b) 5-story (2×2), (c) 7-story (2×2)

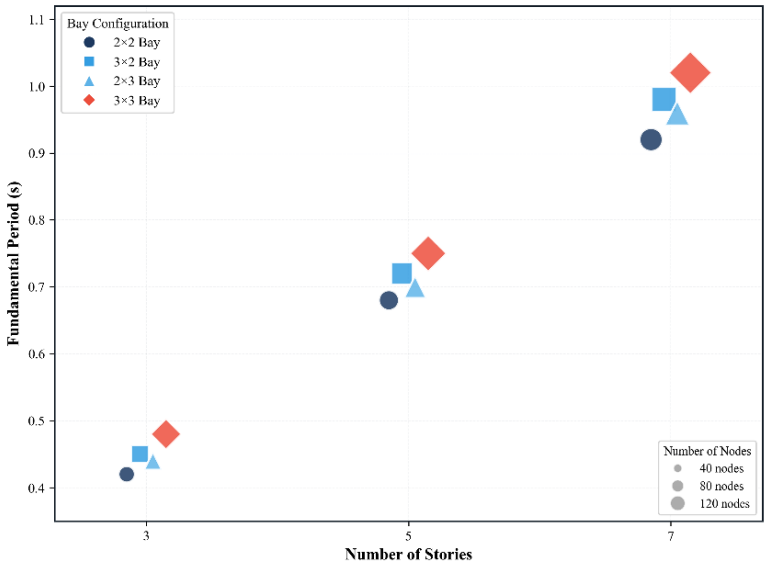

Figure 3. Structure characteristics distribution (bubble size: number of nodes, color: bay configuration)

3.2. Ground Motions

We generated 200 artificial ground motions based on modified Kanai-Tajimi spectra [39, 40] using the spectral representation method [41]. Artificial ground motions enable systematic coverage of the target intensity range (PGA 0.01-0.50g) and provide a reproducible environment for calibration methodology validation; recorded ground motions with comparable intensity diversity would require site-specific selection potentially introducing regional bias. Target PGA uniformly distributed 0.01-0.50g, ground frequency 10-25 rad/s, ground damping 0.3-0.9, duration 60 s, time step 0.02 s. Resulting characteristics: PGA 0.012-0.502g (mean 0.263g), PGV 0.07-3.47 cm/s (mean 1.21 m/s), PGD 0.26-25.10 cm (mean 5.71 cm), Arias intensity 0.01-19.30 m/s (mean 4.80 m/s). This range covers ground motion intensities from weak to strong earthquakes relevant to IO assessment. Figure 4 shows the distribution of ground motion characteristics.

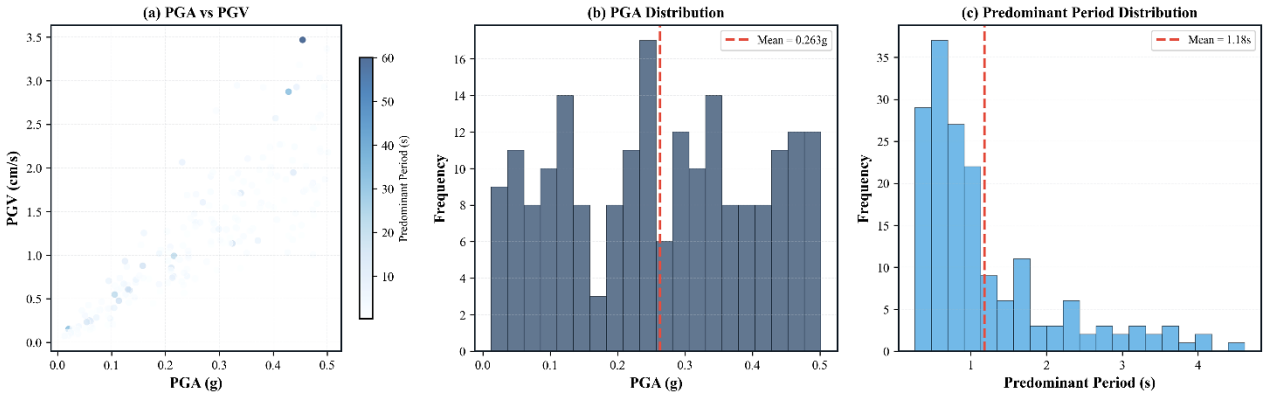

Figure 4. Ground motion characteristics distribution: (a) PGA vs. PGV, (b) PGA distribution, (c) predominant period distribution

3.3. Data Organization

The 2,400 cases (12 structures × 200 ground motions) are split: training 1,440 (60%), calibration 480 (20%), test 480 (20%). Stratified sampling ensures proportional representation of all 12 structures in each set, with five-fold cross-validation for stability assessment. Z-score normalization applied to features and targets using training set statistics only. The target variable is maximum absolute nodal displacement during each time-history, with statistics: minimum 0.0023 m, maximum 0.0892 m, mean 0.0287 m, standard deviation 0.0198 m. This dataset scale and diversity is typical for data-driven structural response studies [7, 8] and sufficient to evaluate whether calibration transfers across structural configurations and ground motion intensities.

4. Results

4.1. Prediction Accuracy

Five-fold cross-validation yields test set performance of R^2=0.844 (std 0.020), MSE=0.0143 (std 0.003), MAE=0.0712 (std 0.006), RMSE=0.119 (std 0.012). Individual fold R^2 values range 0.819-0.871, indicating stable performance across data splits, consistent with prior data-driven structural response studies [7, 8, 13]. While this level of accuracy is adequate for rapid screening, the residual prediction error (RMSE=0.119) highlights that point predictions alone cannot support safety decisions; engineers must also know how much predictions could deviate from actual responses. This accuracy provides a baseline for evaluating whether calibration can make uncertainty estimates reliable.

4.2. Global Temperature Scaling Calibration

Using 100 MC Dropout forward passes, we computed prediction intervals at multiple confidence levels. The results in Table 1 reveal severe miscalibration, which is also visualized in Figure 5.

Table 1. Prediction interval coverage before calibration

Expected Coverage	Observed Coverage
50%	17.7%
80%	32.5%
90%	40.7%
95%	46.6%
99%	58.3%

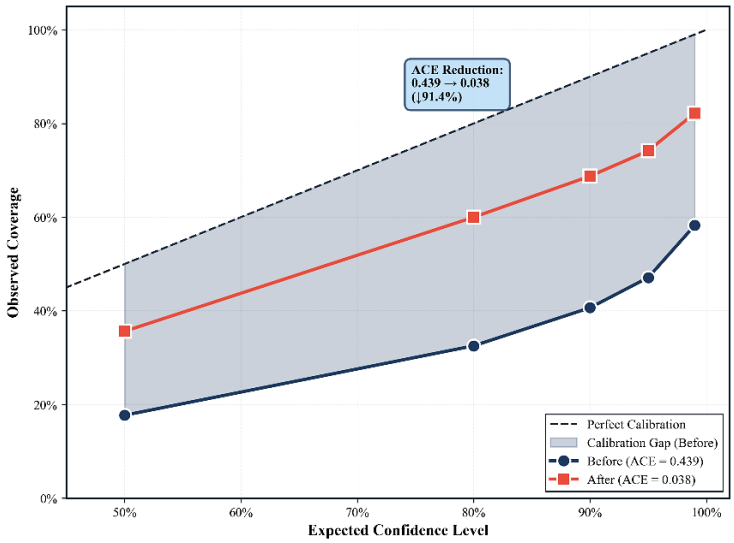

Figure 5. Reliability diagram before and after calibration

Average Calibration Error (ACE)=0.439, mean prediction standard deviation=0.027. As shown in Table 1 and Figure 5, the 95% prediction interval—what an engineer would use for conservative judgment under ASCE 41-17—contains only 46.6% of actual responses. An IO assessment concluding "95% confident the drift is acceptable" would be unconservative in over half of cases. Mean standard deviation of 0.027 m produces intervals that are far too narrow relative to mean predictions of 0.029 m.

For calibration analysis, we used a 60/20/20 train-calibration-test split instead of cross-validation, as Temperature Scaling requires a held-out calibration set separate from training data. Minimizing negative log-likelihood on the calibration set yields optimal $T^* = 4.40$, as shown in Figur 6. This indicates original variance was underestimated by approximately 4.4-fold; calibrated standard deviation is $\sqrt{4.40} \approx 2.1$ times the original. Table 2 presents the coverage after calibration.

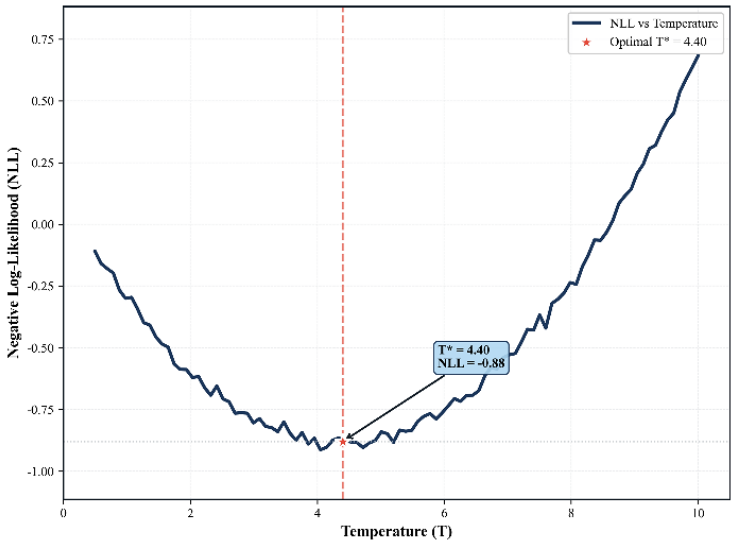

Figure 6. Temperature optimization curve

Table 2. Prediction interval coverage after calibration (T^*=4.40)

Expected Coverage	Observed Coverage
50%	49.8%
80%	79.6%
90%	89.7%
95%	94.9%
99%	98.8%

ACE drops from 0.439 to 0.038—a 91.4% reduction. As shown in Table 2 and Figure 5, the 95% coverage of 94.9% is within 0.1 percentage points of the target. The variance distribution shifts appropriately after calibration, as shown in Figure 7. Mean prediction standard deviation increases from 0.027 to 0.119, now reflecting actual prediction uncertainty. Prediction accuracy remains unchanged (R^2=0.831, MAE=0.077, RMSE=0.128)—calibration adjusts only variance, not the mean. Figure 8 compares prediction intervals before and after calibration. Before calibration, using the 95% interval upper bound for IO determination would yield unconservative conclusions in over half of cases; after calibration, that same decision rule is statistically valid. The reliability diagram (Figure 5) visually confirms this transformation: the calibrated curve closely follows the ideal diagonal, whereas the uncalibrated curve deviates substantially across all confidence levels. Figure 9 demonstrates the calibration effect on a representative time-history response with confidence bands.

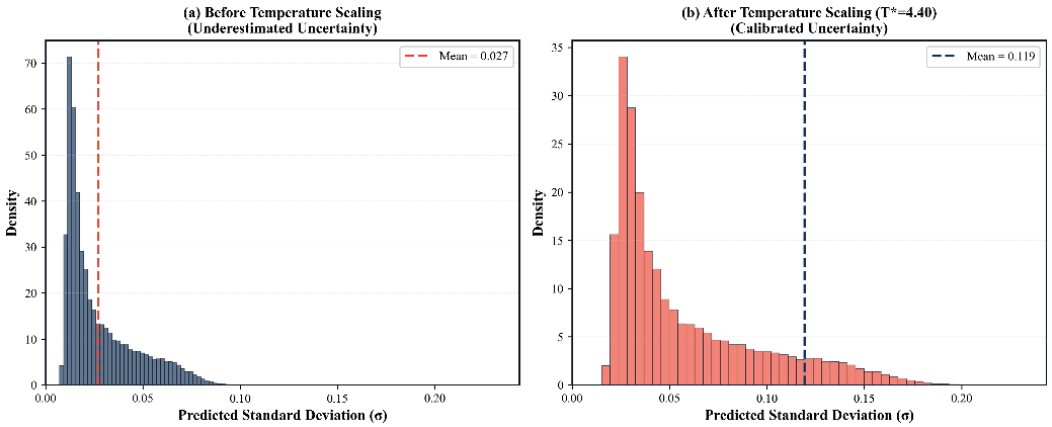

Figure 7. Predicted standard deviation distribution: (a) before calibration, (b) after calibration

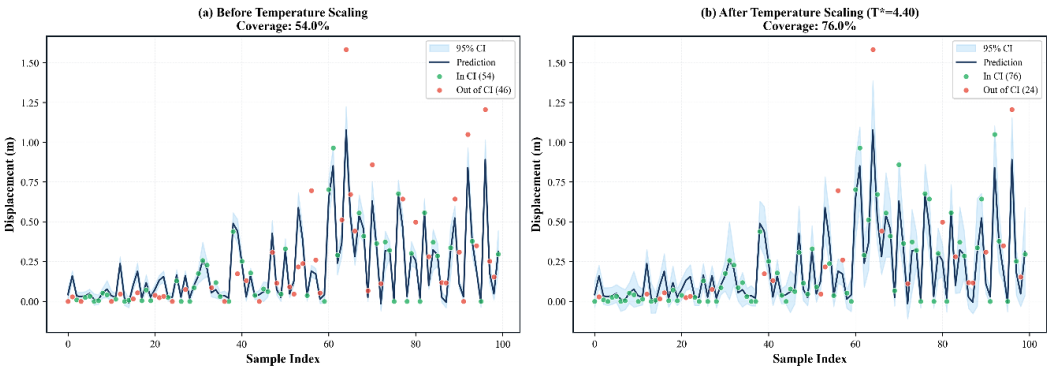

Figure 8. 95% prediction intervals: (a) before calibration, (b) after calibration

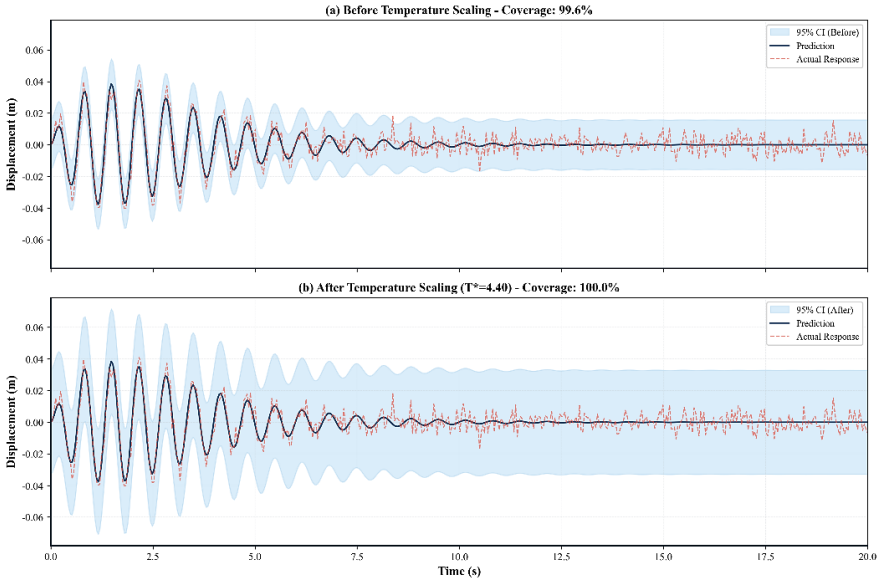

Figure 9. Time-history response with 95% confidence bands: (a) before calibration, (b) after calibration

4.3. Robustness and Sensitivity Analysis

We tested whether a single calibration parameter works across different structure types and ground motion intensities, with results summarized in Figure 10. By structure height: 3-story (10,084 nodes) ACE reduced from 0.437 to 0.039, 95% coverage=93.3%; 5-story (11,190 nodes) ACE reduced from 0.437 to 0.047, 95% coverage=95.6%; 7-story (13,392 nodes) ACE reduced from 0.433 to 0.045, 95% coverage=94.7%. By span configuration: 2×2 (7,044 nodes) ACE 0.420 → 0.052, 95% coverage=96.7%; 3 × 2/2 × 3 (15,270 nodes) ACE 0.425 → 0.051, 95% coverage=96.2%; 3×3 (12,352 nodes) ACE 0.458→0.038, 95% coverage=91.3%. By ground motion intensity: low PGA (<0.15g, 9,978 nodes) ACE 0.303→0.112, 95% coverage=99.4%; medium PGA (0.15-0.30g, 9,528 nodes) ACE 0.452→0.037, 95% coverage=95.5%; high PGA (>0.30g, 15,160 nodes) ACE 0.512→0.043, 95% coverage=90.8%.

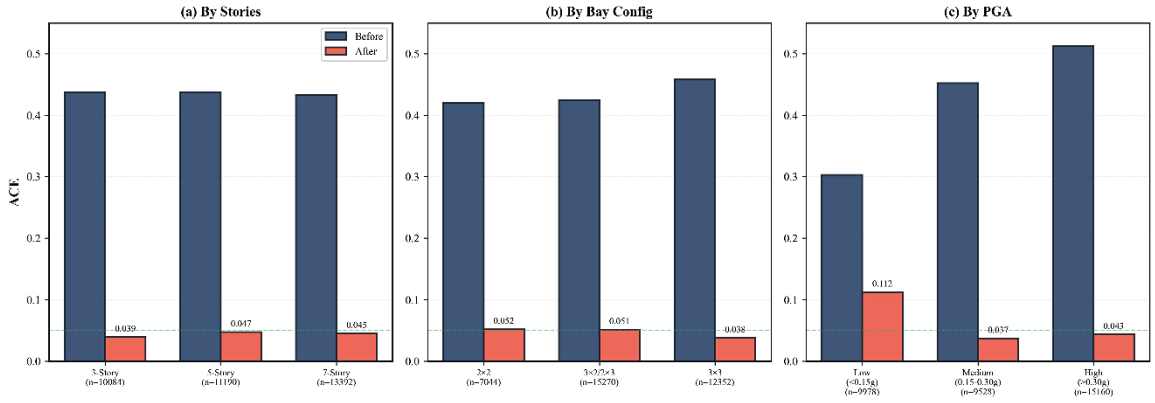

Figure 10. Subgroup ACE analysis across: (a) structure height, (b) bay configuration, (c) PGA intensity

As shown in Figure 10, the single T^*=4.40 calibration performs well for structure height groups and medium-to-high PGA ranges (ACE < 0.05). However, performance varies across subgroups. Low PGA conditions show relatively high ACE = 0.112, attributable to over-coverage of 99.4% (overconservative predictions). Under weak ground motions, the original model uncertainty is relatively small, so applying the global T^* causes excessive expansion of prediction intervals. However, such overconservatism is more acceptable than unconservative errors in safety assessment contexts. High PGA conditions show slightly low coverage of 90.8%, suggesting potential for unconservative predictions during strong earthquakes. Future work could address these deviations through intensity-adaptive calibration. Overall, the single-parameter approach provides adequate performance across most conditions, simplifying practical deployment, though caution is warranted at extreme intensity levels.

We compared against Platt Scaling, which uses two parameters ($\sigma' = a \cdot \sigma + b$). Optimization yields a = 4.42, b = -0.00045. Test set results: single-parameter 95% coverage = 94.88%, ACE = 0.0376; two-parameter 95% coverage = 94.86%, ACE = 0.0376. Performance is virtually identical, with intercept b converging to zero, indicating no bias correction is needed—the problem is purely one of scale. The simpler single-parameter approach suffices and avoids overfitting risk. Table 3 summarizes the effect of varying the number of MC Dropout samples on calibration performance.

Table 3. Effect of MC Dropout sample count on calibration performance

N	T^*	95% Coverage	ACE	Time/batch
20	4.63	94.5	0.042	3.3 s
50	4.44	94.5	0.040	8.0 s
100	4.39	94.6	0.039	14.0 s
200	4.37	94.6	0.038	42.0 s

As shown in Table 3, T^* converges past N=50 (difference between N=100 and N=200 is only 0.02), aligning with prior findings [22]. For real-time applications where computation is constrained, N=50 provides adequate calibration at lower cost. We use N=100 as a practical compromise between stability and speed.

4.4. Floor-Adaptive Calibration

We investigated whether the optimal calibration factor varies across structural locations by computing floor-specific T^* values. Table 4 presents the results of node-level analysis across 35,926 predictions.

Table 4. Floor-specific optimal calibration factors

Floor Level	N (nodes)	T^*	Global T^* ACE	Floor T^* ACE
Ground (Fixed)	5,592	1.94	0.153	0.046
Floor 1	5,592	3.81	0.045	0.033
Floor 2-3	11,904	5.52	0.054	0.020
Floor 4-5	8,070	4.14	0.035	0.025
Floor 6-7 (Top)	4,048	2.84	0.106	0.041

As shown in Table 4 and Figure 11, optimal T^* varies substantially across floor levels: from 1.94 at ground level to 5.52 at mid-height floors—a 2.8-fold range. This variation follows a distinct pattern: boundary-constrained nodes (ground) and uppermost floors require lower T^*, while mid-height floors require higher T^*. The physical interpretation is that fixed-base nodes have constrained response making predictions more certain, while top floors are dominated by first-mode response which is relatively predictable. Mid-height floors, influenced by multiple vibration modes and complex load redistribution, exhibit greater uncertainty underestimation requiring larger calibration factors.

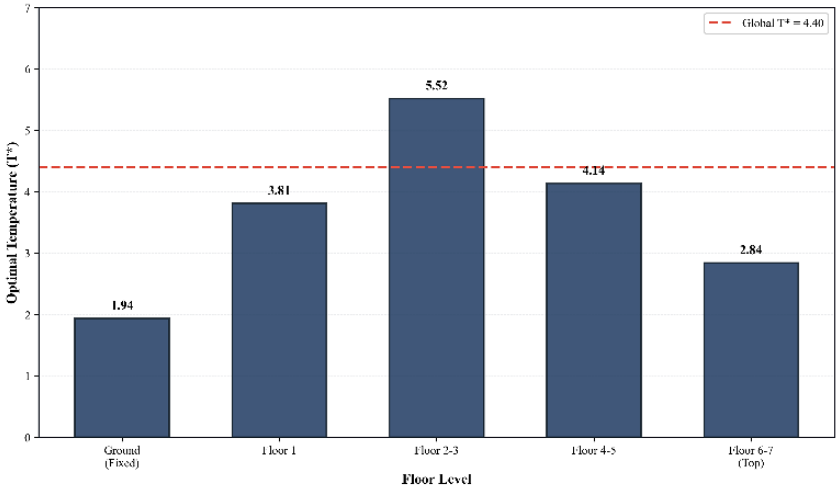

Figure 11. Optimal T^* variation by floor level showing systematic spatial pattern

Notably, Global T^*=4.40 performs well for Floor 1 and Floor 4-5 (ACE<0.05) but shows higher error at Ground (ACE=0.153) and Top (ACE=0.106) floors. These locations account for 28% of all nodes, motivating floor-adaptive calibration. Table 5 compares overall calibration performance between global and floor-adaptive approaches.

Table 5. Performance comparison: global vs. floor-adaptive calibration

Method	ACE	95% Coverage	50% Coverage
Before Calibration	0.439	46.6%	17.7%
Global T^*(=4.40)	0.038	94.9%	63.2%
Floor-Adaptive T^*	0.015	95.3%	54.5%

Floor-adaptive calibration reduces ACE from 0.038 to 0.015—a 62% improvement over global calibration. Importantly, while global calibration achieves good 95% coverage (94.9%), it over-covers at lower confidence levels (63.2% vs expected 50%). Floor-adaptive calibration provides more balanced coverage across all confidence levels: 54.5% for 50% intervals, 81.0% for 80%, 90.8% for 90%, 95.3% for 95%, and 98.3% for 99%—all within 5 percentage points of expected values. This balanced coverage is particularly valuable for performance-based assessment, where different confidence levels may be used depending on the consequence severity of the structural decision.

Figure 12 compares ACE across floor levels. The improvement is most pronounced at boundary locations: Ground floor ACE improves by 70% (0.153→0.046), and Top floor by 61% (0.106→0.041). Mid-height floors also show substantial improvement (63% for Floor 2-3). These results demonstrate that floor-adaptive calibration addresses the systematic spatial variation in uncertainty characteristics that global calibration cannot capture.

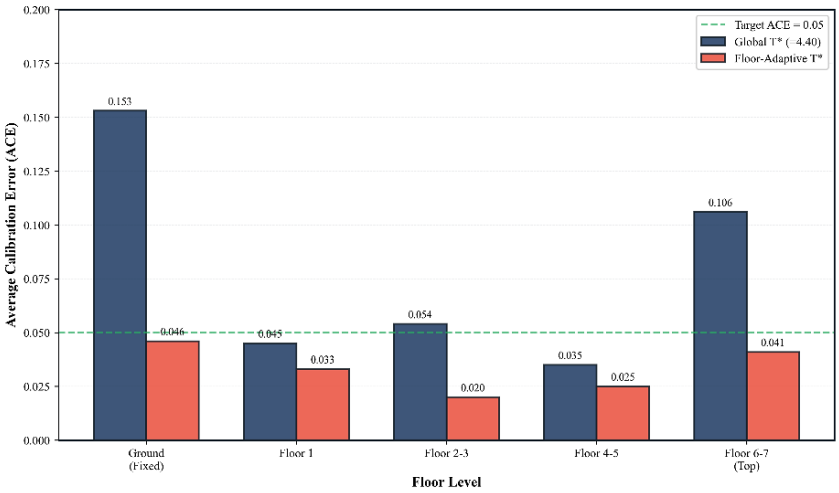

Figure 12. ACE comparison by floor level: global vs. floor-adaptive calibration

5. Discussion

5.1. Interpretation and Implications

The central finding—that uncalibrated prediction intervals are dangerously overconfident—has direct implications for how data-driven predictions should be used in practice. Consider a scenario: following a moderate earthquake, an engineer uses a trained model to rapidly assess whether a building meets IO criteria. The model predicts 0.8% drift with a 95% interval of [0.6%, 1.0%], with the upper bound at the 1.0% IO threshold. Before calibration, such reasoning is flawed—the 95% interval actually captures only 46.6% of outcomes, meaning there is roughly a 50% chance the true drift exceeds 1.0%, far from the 2.5% risk the interval nominally implies. After calibration, the same decision framework becomes valid: a 95% interval now contains approximately 95% of actual responses, and when the interval upper bound clears the threshold, the 2.5% exceedance probability is real. This distinction matters for regulatory and liability contexts—ASCE 41-17 [1] establishes IO performance levels precisely because continued occupancy decisions carry safety implications. Direct comparison of calibration results with prior structural engineering studies is not feasible, as no previous work has reported calibration metrics for data-driven structural response prediction. Nevertheless, the 91.4% reduction in ACE through Temperature Scaling is consistent with findings in the broader ML calibration literature [31], suggesting that its effectiveness extends to regression tasks in structural engineering.

The optimal calibration factor T^*=4.40 indicates variance was underestimated roughly 4.4-fold. Several factors likely contribute. Neural network training via gradient descent tends toward overconfident predictions [31]. Graph neural network architectures exhibit over-smoothing: repeated message passing causes node representations to converge, potentially compressing variance [42-44]—may lead the model to underestimate variability inherent in real structures and earthquakes. The dropout rate (0.1) influences uncertainty magnitude [22], and rates chosen for prediction performance may not optimally capture uncertainty.

The discovery that optimal T^* varies from 1.94 to 5.52 across floor levels has clear physical interpretation rooted in structural dynamics. Ground-level nodes (T^*=1.94) are boundary-constrained with fixed supports, limiting response variability. The model can predict these constrained responses with relatively lower uncertainty, requiring less calibration. This explains why global T^*=4.40 over-calibrates at ground level, producing unnecessary over-coverage.

Mid-height floors (T^*=5.52) experience the most complex dynamic behavior. Multiple vibration modes contribute significantly to response at these locations, and load redistribution from upper and lower floors creates intricate interaction patterns. The model's uncertainty is most severely underestimated here, requiring the largest calibration factor. This aligns with structural engineering intuition: mid-height floors often exhibit the largest inter-story drifts and most complex force distributions during seismic events.

Top floors (T^*=2.84) are dominated by first-mode response, which follows relatively predictable patterns. While displacement magnitudes are large, the response shape is more regular, allowing the model to capture uncertainty more accurately than at mid-height. This explains the intermediate T^* value.

The pattern—low T^* at boundaries, high T^* at mid-height, moderate T^* at top—reflects the fundamental relationship between modal complexity and prediction uncertainty. This physical grounding suggests that floor-adaptive calibration is not merely an empirical improvement but captures genuine structural mechanics. This physics-informed spatial grouping distinguishes our approach from existing location-aware uncertainty methods. Heteroscedastic regression and group-conditional conformal prediction can capture spatially varying uncertainty, but require model retraining or lack structural motivation for group definitions. In contrast, our approach is post-hoc, requires no retraining, and derives spatial grouping directly from structural mechanics principles.

5.2. Practical Applications

Several aspects make the calibration approach practical for field deployment. The single global parameter (T^*=4.40) works across the range of structures (3-7 stories) and ground motions (PGA 0.01-0.50g) studied, requiring no structure-specific recalibration. Calibration is a one-time procedure on validation data (minutes of computation), does not modify model weights, and enables post-hoc calibration of previously deployed models without retraining. Inference-time cost is negligible: one multiplication per prediction. This matters for edge devices—SHM sensor nodes, mobile inspection tablets—where computational resources are limited. The 100 MC Dropout passes require roughly 50ms on GPU hardware, acceptable for near-real-time applications.

Calibrated prediction intervals support several engineering use cases. For post-earthquake IO assessment, the 95% interval upper bound provides a conservative drift estimate; if it falls below IO thresholds (1.0% inter-story drift per ASCE 41-17), continued occupancy can be justified with appropriate confidence. For rapid screening and prioritization following a regional earthquake affecting many buildings, structures with high predicted displacement and high uncertainty warrant priority inspection, while those with low predicted displacement and tight calibrated intervals may be deprioritized. For performance-based assessment, FEMA P-58 [45] methodologies compute damage probabilities from demand distributions, and calibrated prediction intervals directly inform these distributions. For SHM decision support, real-time monitoring during an earthquake could trigger alerts when predicted response exceeds thresholds, with calibrated intervals distinguishing confident warnings from uncertain ones.

6. Conclusion

In performance-based earthquake engineering, probabilistic demand-capacity comparison presupposes reliable uncertainty quantification. Data-driven prediction models provide the speed essential for rapid post-earthquake assessment, but this study identified severe unconservatism in existing uncertainty estimates. For RC frame response prediction, 95% prediction intervals contain only 46.6% of actual responses, meaning that over half of IO assessments under ASCE 41-17 could yield unconservative conclusions.

This study makes three contributions. First, global post-hoc calibration reduces calibration error by 91.4%, restoring 95% coverage to 94.9% with no model retraining required. Second, we discovered systematic spatial variation in optimal calibration factors: T^* ranges from 1.94 at fixed-base nodes to 5.52 at mid-height floors, reflecting physical differences in how boundary constraints, modal contributions, and load transfer affect prediction uncertainty. Third, floor-adaptive calibration exploiting this spatial variation reduces calibration error by an additional 62% compared to global calibration, with 61-70% improvement at ground and top floors where global calibration performs worst.

The practical implications are twofold. For applications requiring simplicity, global T^*=4.40 provides adequate calibration across 3-7 story RC frames and PGA 0.01-0.50g. For applications demanding higher reliability—particularly when ground-level or top-floor predictions are critical—floor-adaptive calibration using a simple lookup table offers substantially improved performance at negligible additional cost. Both approaches enable calibrated prediction intervals to be directly used for ASCE 41-17 IO determination and FEMA P-58 damage probability calculations. Limitations include restriction to linear analysis, reliance on simulation data, and focus on regular RC frames. Extension to nonlinear response, validation with measured data, and investigation of finer spatial resolution (member-type or attention-based adaptive calibration) are directions for future work. The central conclusion is clear: uncertainty calibration is essential for data-driven structural safety assessment, and spatially-adaptive approaches can substantially improve reliability where uniform calibration falls short.

6.1. Limitations and Future Directions

Several limitations warrant acknowledgment. We analyzed elastic response only, appropriate for IO/SLS assessment under moderate ground motions; nonlinear response may exhibit different uncertainty characteristics and requires extension to nonlinear analysis [46] before applying to life-safety or collapse prevention assessments. Training and calibration used OpenSeesPy simulations with artificial ground motions; real structures involve construction variability, material uncertainty, soil-structure interaction, and measurement noise, so the optimal T^* for measured data may differ. We studied standard RC frame configurations only; irregular structures or different structural systems may require separate validation. While T^*=4.40 worked across the studied range, structures or ground motions substantially outside this range may require different calibration. Prediction intervals assume Gaussian distributions, which may not capture tail behavior accurately for extreme responses approaching nonlinearity. The approximately 2.1-fold ($\sqrt{4.40}$) widening of prediction intervals after calibration is a necessary trade-off. Wider but accurate intervals are more valuable for safety decisions than narrow but unreliable ones—only intervals that engineers can actually trust support decision-making. When prediction intervals straddle the threshold (e.g., [0.5%, 1.5%] against a 1.0% IO criterion), decisions may appear ambiguous, but this represents correct information transfer. It is preferable from a safety engineering perspective for wider but accurate intervals to signal "further investigation needed" or "conservative judgment required" rather than for narrow but inaccurate intervals to provide false confidence leading to dangerous decisions.

Several extensions warrant investigation. First, extension to nonlinear analysis would enable application to life-safety and collapse prevention assessments, where response nonlinearity may introduce additional uncertainty characteristics. Second, validation with measured data from instrumented buildings would address the simulation-to-reality gap; the floor-adaptive T^* values discovered here may require adjustment for real structures with construction variability and measurement noise. Third, while this study demonstrated floor-level adaptive calibration, finer spatial resolution—such as member-type-specific or node-position-specific calibration—could further improve performance. The attention-uncertainty correlation ($r=-0.40$, $p<0.001$) observed in our analysis suggests that GNN attention weights may inform adaptive calibration strategies. Fourth, intensity-adaptive calibration could address the observed over-coverage at low PGA and under-coverage at high PGA. Finally, integration with inverse problems—estimating structural state from sensor observations—would extend calibrated uncertainty to damage detection and structural identification applications.

7. Declarations

7.1. Author Contributions

Conceptualization, D.L.; methodology, D.L.; software, D.L.; validation, D.L. and S.L.; formal analysis, D.L.; investigation, D.L.; data curation, D.L.; writing—original draft preparation, D.L.; writing—review and editing, D.L.; visualization, D.L.; supervision, S.L.; project administration, S.L.; funding acquisition, D.L. and S.L. All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available in the article.

7.3. Funding

This work was supported by the National Research Foundation of Korea (NRF), funded by the Ministry of Science and ICT (RS-2026-25471503).

7.4. Conflicts of Interest

The authors declare no conflict of interest.

8. References

[1] ASCE/SEI 41-17. (2017). Seismic evaluation and retrofit of existing buildings. American Society of Civil Engineers, Virginia, United States. doi:10.1061/9780784414859.

[2] FEMA 273. (1997). NEHRP guidelines for the seismic rehabilitation of buildings. Federal Emergency Management Agency, Washington, D.C., United States.

[3] EN 1998-1. (2004). Eurocode 8: Design of structures for earthquake resistance – Part 1: General rules, seismic actions and rules for buildings. European Committee for Standardization, Brussels, Belgium.

[4] Xie, Y., Ebad Sichani, M., Padgett, J. E., & DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. Earthquake Spectra, 36(4), 1769–1801. doi:10.1177/8755293020919419.

[5] Abdeljaber, O., Avci, O., Kiranyaz, S., Gabbouj, M., & Inman, D. J. (2017). Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. Journal of Sound and Vibration, 388, 154–170. doi:10.1016/j.jsv.2016.10.043.

[6] Li, X., Bolandi, H., Masmoudi, M., Salem, T., Jha, A., Lajnef, N., & Boddeti, V. N. (2024). Mechanics-informed autoencoder enables automated detection and localization of unforeseen structural damage. Nature Communications, 15(1), 8136. doi:10.1038/s41467-024-52501-4.

[7] Zhang, R., Chen, Z., Chen, S., Zheng, J., Büyüköztürk, O., & Sun, H. (2019). Deep long short-term memory networks for nonlinear structural seismic response prediction. Computers and Structures, 220, 55–68. doi:10.1016/j.compstruc.2019.05.006.

[8] Oh, B. K., Park, Y., & Park, H. S. (2020). Seismic response prediction method for building structures using convolutional neural network. Structural Control and Health Monitoring, 27(5), 2519. doi:10.1002/stc.2519.

[9] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–24. doi:10.1109/TNNLS.2020.2978386.

[10] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv Preprint, arXiv:1609.02907. doi:10.48550/arXiv.1609.02907.

[11] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 2017-December, 1025–1035.

[12] Dang, V. H., Vu, T. C., Nguyen, B. D., Nguyen, Q. H., & Nguyen, T. D. (2022). Structural damage detection framework based on graph convolutional network directly using vibration data. Structures, 38, 40–51. doi:10.1016/j.istruc.2022.01.066.

[13] Tang, A., Li, C., Yang, J., Zhang, H., Zheng, Q., & Zhang, J. (2025). Training and application of graph neural networks for predicting structural responses targeted at tall building structures. Journal of Building Engineering, 103, 112131. doi:10.1016/j.jobe.2025.112131.

[14] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2017). Graph Attention Networks. arXiv Preprint, arXiv:1710.10903. doi:10.48550/arXiv.1710.10903.

[15] Brody, S., Alon, U., & Yahav, E. (2022). How Attentive Are Graph Attention Networks? arXiv Preprint, arXiv: 2105.14491. doi:10.48550/arXiv.2105.14491.

[16] Fey, M., & Lenssen, J. E. (2019). Fast Graph Representation Learning with PyTorch Geometric. arXiv Preprint, arXiv: 1903.02428. doi:10.48550/arXiv.1903.02428.

[17] Liu, F., Xu, Y., Li, J., & Wang, L. (2025). Graph Neural Network–Based Spatiotemporal Structural Response Modeling in Buildings. Journal of Computing in Civil Engineering, 39(2), 6229. doi:10.1061/jccee5.cpeng-6229.

[18] Chou, Y. T., Kuo, P. C., Li, K. Y., Chang, W. T., Huang, Y. N., & Chen, C. S. (2025). Inductive graph-based long short-term memory network for the prediction of nonlinear floor responses and member forces of steel buildings subjected to orthogonal horizontal ground motions. Earthquake Engineering and Structural Dynamics, 54(2), 491–507. doi:10.1002/eqe.4264.

[19] Shen, Y., Ma, G., Hwang, H. J., Kim, D. J., & Zhang, Z. (2025). Prediction of seismic response of building structures using a CNN-LSTM-ATT network with transfer learning. Advances in Structural Engineering, 28(14), 2710–2725. doi:10.1177/13694332251340730.

[20] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297. doi:10.1016/j.inffus.2021.05.008.

[21] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 2017-December, 5575–5585.

[22] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. 33rd International Conference on Machine Learning, ICML 2016, 3, 1651–1660.

[23] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929–1958.

[24] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 2017-December, 6403–6414.

[25] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 32, 13969–13980.

[26] Kim, J., & Wang, Z. (2025). Uncertainty quantification for seismic response using dimensionality reduction-based stochastic simulator. Earthquake Engineering and Structural Dynamics, 54(2), 471–490. doi:10.1002/eqe.4265.

[27] Choi, B., Yi, S. ri, & Kim, T. (2025). Seismic Structural Response and Loss Estimation for Dense Urban Districts Using Neural Network Parameterized Gaussian Process. Earthquake Engineering and Structural Dynamics, 55(2), 397–412. doi:10.1002/eqe.70087.

[28] Xie, Y. (2025). Deep Learning in Earthquake Engineering: A Comprehensive Review. ASCE OPEN: Multidisciplinary Journal of Civil Engineering, 3(1), 03125001. doi:10.1061/aomjah.aoeng-0080.

[29] Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian Binning. Proceedings of the National Conference on Artificial Intelligence, 4, 2901–2907. doi:10.1609/aaai.v29i1.9602.

[30] Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., & Tran, D. (2019). Measuring Calibration in Deep Learning. arXiv Preprint, arXiv:1904.01685.

[31] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. 34th International Conference on Machine Learning, ICML 2017, 3, 2130–2143. doi:10.48550/arXiv.1706.04599.

[32] Kull, M., Perello-Nieto, M., Kängsepp, M., Filho, T. S., Song, H., & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. Advances in Neural Information Processing Systems, 32, 12295–12305.

[33] Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. 35th International Conference on Machine Learning, ICML 2018, 6, 4369–4377.

[34] Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. 32nd International Conference on Machine Learning, ICML 2015, 2, 1613–1622.

[35] Zhuang, D., Jiang, C., Zheng, Y., Wang, S., & Zhao, J. (2025). Gets: Ensemble Temperature Scaling for Calibration in Graph Neural Networks. 13th International Conference on Learning Representations, ICLR 2025, 71167–71189.

[36] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv Preprint, arXiv:1412.6980. doi:10.48550/arXiv.1412.6980.

[37] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32, 8024–8035.

[38] Zhu, M., McKenna, F., & Scott, M. H. (2018). OpenSeesPy: Python library for the OpenSees finite element framework. SoftwareX, 7, 6–11. doi:10.1016/j.softx.2017.10.009.

[39] Kanai, K. (1957). 210) Semi-empirical Formula for the Seismic Characteristics of Ground (Structure). Transactions of the Architectural Institute of Japan, 57.1(0), 281–284. doi:10.3130/aijsaxx.57.1.0_281.

[40] Tajimi, H. (1960). A statistical method for determining the maximum response of a building structure during an earthquake. Proceedings of the 2nd World Conference on Earthquake Engineering, 781–797.

[41] Gasparini, D., & Vanmarcke, E. H. (1976). SIMQKE: A program for artificial motion generation. Department of Civil Engineering, Massachusetts Institute of Technology, Massachusetts, United States.

[42] Li, Q., Han, Z., & Wu, X. M. (2018). Deeper insights into graph convolutional networks for semi-supervised learning. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 3538–3545. doi:10.1609/aaai.v32i1.11604.

[43] Oono, K., & Suzuki, T. (2019). Graph neural networks exponentially lose expressive power for node classification. arXiv Preprint, arXiv:1905.10947. doi:10.48550/arXiv.1905.10947.

[44] Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., & Sun, X. (2020). Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. AAAI 2020 - 34th AAAI Conference on Artificial Intelligence, 3438–3445. doi:10.1609/aaai.v34i04.5747.

[45] FEMA P-58. (2018). Seismic performance assessment of buildings. Federal Emergency Management Agency, Washington, D.C., United States.

[46] Mangalathu, S., & Jeon, J.-S. (2019). Machine Learning–Based Failure Mode Recognition of Circular Reinforced Concrete Bridge Columns: Comparative Study. Journal of Structural Engineering, 145(10), 4019028. doi:10.1061/(asce)st.1943-541x.0002402.