ARTICLE

Limitations of Fourier Neural Operators for Nonlinear Structural Response Prediction under Earthquake Excitation

Kexun Li | **Luis Ceferino**

[1]Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, CA, USA

Correspondence
Kexun Li, Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, USA.
Email: kexunli2@berkeley.edu

Abstract

Neural operators provide a promising framework for mapping earthquake ground motions directly to structural response histories, but their accuracy for nonlinear dynamics and out-of-distribution conditions remains poorly understood. We systematically evaluate a fixed Fourier Neural Operator architecture using single- and five-degree-of-freedom systems with linear-elastic, cubic-hardening, and bilinear-hysteretic restoring forces. Our numerical tests isolate the effects of restoring-force nonlinearity, structural period, excitation frequency content, and ground-motion amplitude. Within the training regime, the mean per-record relative L_2 error increases from 0.02 for the linear system to 0.28 and 0.38 for the cubic-hardening and bilinear-hysteretic systems, respectively; corresponding five-degree-of-freedom tests confirm the deterioration for hysteretic response. Generalization degrades more severely under distribution shifts. When structural periods fall outside the 0.5–1.0s training interval, relative L_2 errors exceed 1 for both single- and multi-degree-of-freedom systems. Broader excitation frequency content further reduces nonlinear prediction accuracy, even within otherwise interpolative conditions. In addition, when models trained on records with peak ground accelerations up to about 0.19 g are evaluated at amplitudes up to about 1.0 g, they underpredict the root-mean-square bilinear-hysteretic response by approximately 17%. Error decomposition further shows that residual mean offsets account for approximately 22% of the bilinear-hysteretic error, compared with 3% for the cubic-hardening system, whereas amplitude-scaling errors remain small. These results demonstrate that strong interpolation performance does not ensure accurate prediction of nonlinear structural response under changes in structural or excitation regimes.

KEYWORDS
Fourier Neural Operator; Tensorized FNO; nonlinear structural dynamics; bilinear hysteresis; cubic hardening; ground-motion response prediction

1 | INTRODUCTION

Time-series surrogates can rapidly predict structural response time histories under earthquake ground motions [1,2]. These models support regional risk assessment [3,4,5,6], structural health monitoring and post-earthquake assessment [7,8,9], and seismic digital twins, where repeated or near-real-time response predictions are needed [10,11,12]. In these settings, surrogates reduce the computational cost of many-query structural response prediction, which would otherwise require repeatedly solving the structural dynamic equilibrium equations over many time steps across ground motions and structural configurations [13,14].

Several deep-learning surrogates have been developed for seismic response prediction. Recurrent sequence models, including long short-term memory (LSTM) architectures, can predict nonlinear response time histories from ground-motion inputs [1,2]. Transformer-based models have also recently been used to capture longer-range temporal dependencies in structural seismic

Abbreviations: FNO, Fourier Neural Operator; TFNO, Tensorized Fourier Neural Operator; SDOF, single-degree-of-freedom; 5DOF, five-degree-of-freedom; LSTM, long short-term memory; PGA, peak ground acceleration; RMS, root mean square; GroupNorm, Group Normalization; GELU, Gaussian Error Linear Unit; MLP, multilayer perceptron.

response[15]. Other architectures incorporate structural or hysteretic response features to estimate engineering demand parameters[16]. These studies show that data-driven surrogates can predict seismic response, but each trained model is tied to a particular input-output format. Sequence models typically predict response histories on a fixed time grid with a fixed number of output channels[1,15]. Feature-based networks instead return one response quantity, such as the peak displacement of a hysteretic single-degree-of-freedom (SDOF) system[16].

Neural operators provide a natural way to generalize this task by learning mappings between functions rather than fixed-dimensional vectors[17,18]. In the seismic response setting, this task corresponds to learning a ground-motion-to-response operator: a map from an input ground-motion function to an output structural response function[19,20,21]. Different neural operator architectures construct this operator in different ways[17,18]. For example, DeepONet combines a branch network that encodes the input function and a trunk network that encodes the output coordinates[19]. This study focuses on the Fourier-based family, which gives a spectral construction of the ground-motion-to-response map[22]. Such a construction is useful for seismic response prediction because dominant frequency components in earthquake records strongly influence the resulting structural response histories.

Fourier Neural Operators (FNO) and their tensorized variants (TFNO) provide a spectral operator-learning approach for this task. FNOs learn the ground-motion-to-response relationship in the Fourier domain, allowing each spectral layer to capture global patterns that extend across the full time history[22,17]. Tensorized and factorized variants preserve this spectral structure while reducing the number of parameters[23,24]. This representation is especially appealing for linear structural dynamics, where each response quantity can be expressed as the convolution of the ground motion with an impulse-response function[25]. Under moderate or large earthquakes, however, this relationship is no longer fixed. Nonlinear restoring forces, hysteretic energy dissipation, and residual deformation make the excitation-to-response relationship depend on both amplitude and loading history[26]. These mechanisms can challenge a spectral surrogate in ways that are not apparent from linear response alone.

Prior work has identified limitations of FNO-type models that are relevant to earthquake response prediction. Deep networks tend to learn low-frequency content before high-frequency content[27], and FNOs can compound this tendency through Fourier-mode truncation[28,29]. As a result, FNO predictions may become over-smoothed, a limitation reported in turbulence modeling[30] and seismic wavefield simulations[31]. Other studies show that FNO accuracy can also degrade for complex geometries and non-smooth outputs[18], and for resolution changes and unseen frequency content[32,33,34]; generalization analyses further tie FNO error bounds to the number of retained Fourier modes[35]. Recent work on nonlinear dynamical systems also identifies energy dissipation and frequency content distortion as sources of accuracy degradation[36]. Related physics-informed learning studies do not by themselves remove these failure modes. Soft physical constraints can make optimization harder[37], and band-limited or physics-informed corrections do not guarantee reliable zero-shot generalization[34].

Neural operator surrogates have been applied to nonlinear structural response problems, including multi-story buildings under stochastic ground motion[38], hysteretic and path-dependent seismic response[39], wavelet-domain nonlinear seismic response models[40], and floating offshore structures under irregular waves[21]. These studies focus on demonstrating when neural operator surrogates work: they report accurate predictions on test inputs drawn from within the sampled training regime, often refining architectures or adding physics information to improve that accuracy. They do not systematically identify where a surrogate ceases to be reliable for earthquake-induced structural dynamics. A deployed surrogate can leave its training regime in two principal ways: the structure can differ from the training systems in its restoring-force nonlinearity or its structural period, and the ground motion can exceed the training records in excitation amplitude or differ from them in input frequency content. In particular, FNO limitations have not been characterized across both structure-related and excitation-related sources of difficulty.

We therefore isolate these four factors using a single TFNO configuration: two structure-related factors—restoring-force nonlinearity and structural period, and two excitation-related factors—input frequency content and excitation amplitude. The architecture is held constant while each factor is varied. These factors are concurrent and partially coupled rather than independent.

Our goal is to identify applicability limitations of this spectral surrogate: the conditions under which it predicts nonlinear earthquake response well, and the conditions under which it loses accuracy. We test one fixed architecture across two structural scales, asking whether behavior seen in an SDOF system still holds in a five-degree-of-freedom (5DOF) system. We also measure two parts of each prediction error: a residual offset, in which the prediction oscillates about a wrong mean level, and an amplitude-scale error, in which the prediction has a wrong overall amplitude. We use this decomposition to identify which part differs between the cubic-hardening and bilinear-hysteretic cases.

The remainder of this paper is organized as follows. Section 2 defines the response operator and the three restoring-force laws. Section 3 describes the TFNO surrogate, training procedure, and evaluation metrics. Section 4 presents the SDOF and

5DOF experiments. Section 5 reports the four-factor results, the SDOF-to-5DOF comparison, the error decomposition, and the candidate error mechanisms. Section 6 states the main conclusions and the outlook.

2 | RESPONSE OPERATOR AND RESTORING-FORCE MODELS

This section defines the response operator approximated by the surrogate and the restoring-force laws used to evaluate it. The input is an earthquake ground-motion record, and the output is the corresponding structural displacement time history.

Let $a_g(t)$ denote the base ground acceleration time history, expressed as a dimensionless multiple of the gravitational acceleration g. For an SDOF system with mass m and damping coefficient c, the displacement response $x(t)$ satisfies

$$m\ddot{x} + c\dot{x} + f_r(x, h(t)) = -m\, g\, a_g(t),$$

where f_r is the restoring force, $h(t)$ denotes internal history variables required by path-dependent restoring laws[41,42], and an overdot denotes differentiation with respect to time. The structural response operator is the induced map from the ground-motion record to the displacement response time history[22],

$$\mathcal{G} : a_g \mapsto x. \tag{1}$$

The surrogate approximates this map with a parameterized operator,

$$\mathcal{G}_\theta : \mathcal{A} \to \mathcal{U}, \tag{2}$$

where θ denotes the trainable parameters, $\mathcal{A}$ is the input acceleration function space, and $\mathcal{U}$ is the displacement response function space[43,17,18].

Three restoring-force laws are considered: linear elastic, cubic hardening, and bilinear hysteretic. These laws define the response cases analyzed throughout the paper (Fig. 1). They form a controlled progression in response complexity: the linear law provides the baseline case most compatible with a spectral representation; the cubic law introduces smooth amplitude-dependent hardening; and the bilinear-hysteretic law introduces yielding with reduced post-yield stiffness, path dependence, energy dissipation, and residual deformation.

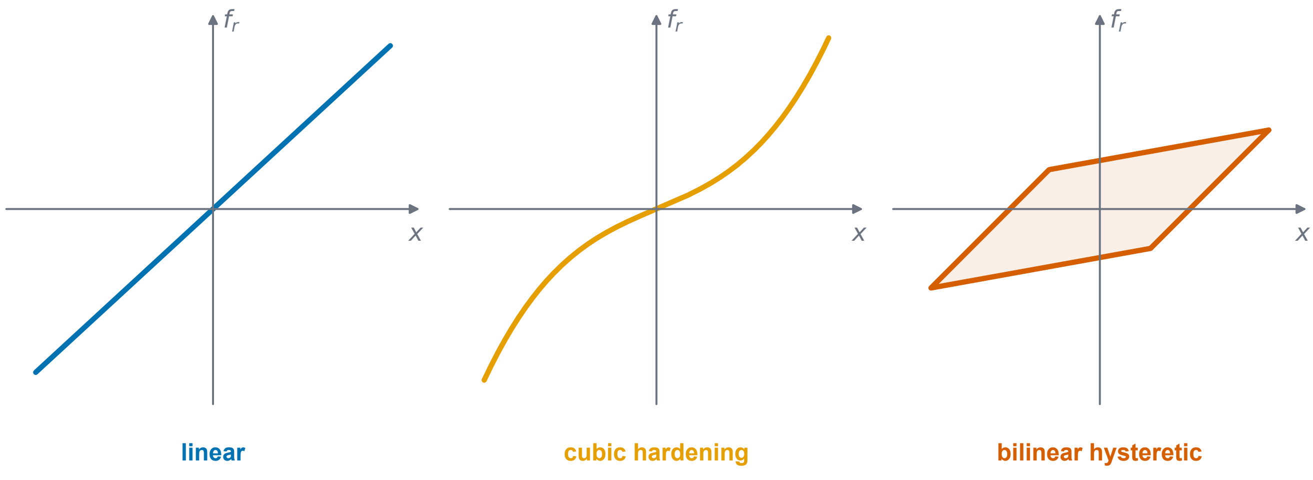


FIGURE 1 Force–displacement relations for the linear, cubic-hardening, and bilinear-hysteretic restoring-force laws. The horizontal axis is the displacement x and the vertical axis the restoring force f_r; for the 5DOF system, the law acts on each story.

- *Linear elastic*. For the SDOF system, the restoring force is

$$f_r(x) = kx, \tag{3}$$

where k is the constant elastic stiffness. The stiffness does not change with response amplitude, and the restoring force has no path dependence. Under zero initial conditions, the input-to-response map is linear and time invariant.

- *Cubic hardening*. For SDOF, the restoring force is

$$f_r(x) = k\,x + k_3\,x^3. \tag{4}$$

For the 5DOF system, the same law acts on each story. The cubic term increases the effective stiffness with displacement amplitude, so the effective dynamic properties depend on the response amplitude. However, the law remains memoryless: the restoring force depends on the current displacement, not on the loading history. The cubic stiffness is set using the dimensionless ratio $\eta = k_3\,x_{\text{ref}}^2/k = 0.40$ at $x_{\text{ref}} = 10\,\text{mm}$, the same reference displacement adopted for each story in the 5DOF system.

- *Bilinear hysteretic*. The bilinear-hysteretic law is modeled as an elastic-plastic spring with yield displacement $x_y = 5\,\text{mm}$, yield force $F_y = kx_y$, and post-yield stiffness ratio $\alpha = 0.10$. Below yield, the response is linear elastic. After yielding, the restoring force depends on the loading and unloading history through an internal plastic offset state $x_p(t)$. Plastic deformation can accumulate over loading cycles, the force–displacement loop dissipates energy, and residual displacement can remain after the ground motion ends[42,41,44]. Thus, the cubic-hardening and bilinear-hysteretic laws are both nonlinear, but only the bilinear-hysteretic law is path dependent.

Connection between restoring-force behavior and FNO accuracy. The three laws above define a progression in the complexity of the ground-motion-to-response operator. In the linear-elastic case, the structure keeps the same stiffness and fundamental period throughout the response. Therefore, changes in the displacement time history come from the changing ground motion, not from changes in the structural properties. In frequency-domain terms, the structure amplifies or attenuates each component of the ground motion according to the same period and damping ratio over the full record; this relation is described by a fixed transfer function[25,45,46,47]. This fixed frequency-domain relationship is naturally aligned with an FNO, which learns mappings through spectral kernels, although only over the finite set of retained Fourier modes[22,17,28].

The cubic-hardening law departs from this setting because the effective stiffness changes with response amplitude. At small displacements, the response is close to the linear-elastic case; at larger displacements, the tangent stiffness increases and the apparent dynamic properties of the oscillator change. Consequently, scaling a ground motion to a higher intensity does not simply scale the displacement response. It can also change the timing, dominant response frequencies, and relative contributions of different portions of the input record[26,48]. The surrogate must therefore learn a response operator whose effective dynamics vary with response amplitude, rather than one governed by a fixed transfer function.

The bilinear-hysteretic law introduces path dependence in addition to nonlinearity. After yielding, the restoring force depends on the loading, unloading, and reloading history through the accumulated plastic offset. Thus, the same displacement can correspond to different restoring forces depending on the current hysteretic branch. The surrogate must therefore infer an internal history-dependent state from the observed input-output histories. Hysteretic energy dissipation changes the oscillatory part of the displacement response, while residual displacement shifts the response away from oscillations centered around zero.

Together, the three laws provide a controlled progression from fixed linear behavior, to amplitude-dependent nonlinearity, to history-dependent hysteresis. This progression lets us interpret FNO errors in structural-dynamics terms, including changes in effective stiffness, hysteretic state evolution, and their interaction with period, input frequency content, excitation amplitude, and system scale.

3 | NEURAL OPERATOR ARCHITECTURE AND EVALUATION PROTOCOL

The model maps an input excitation time series to structural displacement responses, with one output channel for the SDOF systems and five floor-displacement output channels for the 5DOF systems (Fig. 2). Each ground-motion record $a_g(t)$ is represented on a uniform time grid with spacing $\Delta t_{\text{out}} = 0.02\,\text{s}$ and padded to a common length of $N_t = 14014$ points, corresponding to approximately 280 s, so each input record is a vector $\mathbf{a}_g \in \mathbb{R}^{N_t}$. In discrete form, a batch of n_s records has input shape

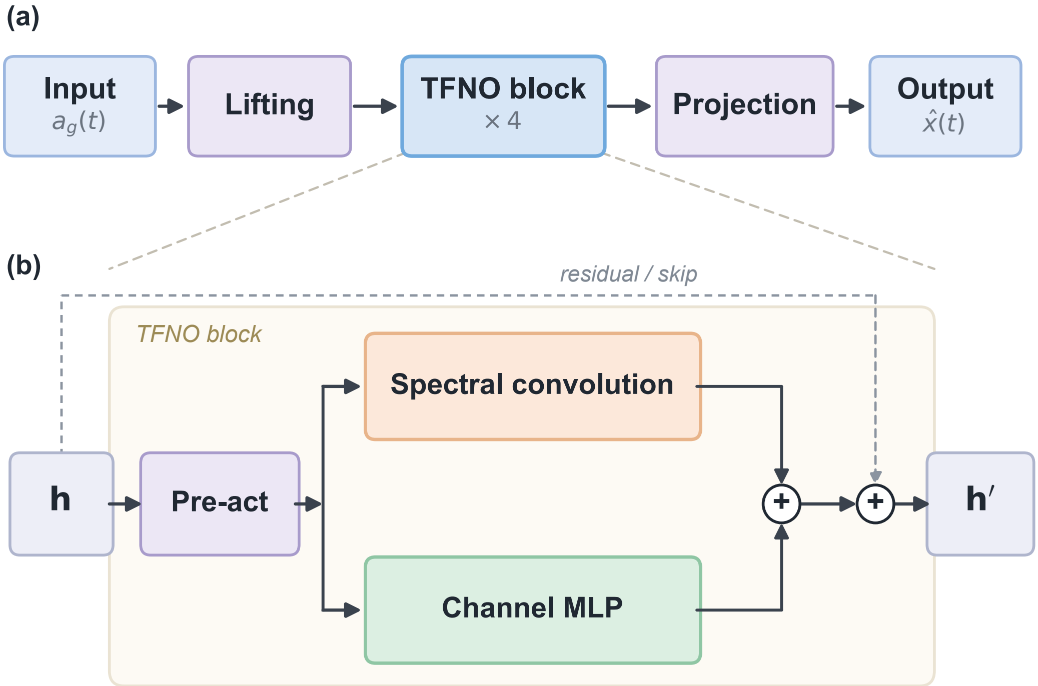


FIGURE 2 TFNO surrogate architecture. (a) The input excitation sequence is lifted to a hidden representation, processed by four TFNO blocks, and projected to one SDOF displacement channel or five 5DOF floor-displacement channels. (b) Each TFNO block combines a spectral-convolution branch on the retained Fourier modes, a pointwise channel-MLP branch, and residual addition; the block input and output shown correspond to v_ℓ and $v_{\ell+1}$ in the text.

$[n_s, N_t, 1]$ and output shape $[n_s, N_t, d_{\text{out}}]$, where $d_{\text{out}} \in \{1, 5\}$ is the number of output channels. The same operator-learning formulation is used at both scales; only the dimension of the response vector changes. The architecture has one input channel (two in the period test), 128 hidden channels, four spectral layers, 1024 retained Fourier modes per layer, and Tucker rank 32 in the spectral kernels[23]. The model has approximately 4.5 million trainable parameters.

Throughout this paper, TFNO denotes a one-dimensional Fourier Neural Operator in which the Fourier-domain convolution kernels are represented with a Tucker factorization[23,49]. This factorization replaces the full spectral-convolution weight tensor with a lower-rank tensor representation, reducing the number of spectral-convolution parameters while retaining frequency-dependent channel mixing[24,50].

The implementation follows the standard FNO design[22,17]. The input sequence is lifted to a hidden-channel representation, processed by four TFNO blocks, and projected back to the response channels (Fig. 2a). Figure 2b expands one block. A pre-activation step, Group Normalization (GroupNorm) followed by the Gaussian Error Linear Unit (GELU), feeds two parallel branches: a Tucker-factorized spectral convolution $\mathcal{S}_\ell$ and a pointwise channel-wise multilayer perceptron (MLP) $\mathcal{M}_\ell$. Their outputs are summed and added back to the block input through a residual connection. In compact form, one TFNO block is

$$h_\ell(t) = \text{GELU}(\text{GroupNorm}(v_\ell(t)))\,, \qquad v_{\ell+1}(t) = v_\ell(t) + \mathcal{S}_\ell(h_\ell)(t) + \mathcal{M}_\ell(h_\ell)(t),$$

where $v_\ell(t)$ is the latent representation at layer ℓ. The two branches act on complementary scales. The spectral branch $\mathcal{S}_\ell$ operates on the retained Fourier modes. These modes couple the entire record, so the branch acts over long temporal ranges. The channel-MLP branch $\mathcal{M}_\ell$ acts pointwise in time and refines each step independently. The residual keeps each block close to the identity, so each layer adds a small correction. The dropout rate is 0.2, the channel-MLP expansion ratio is 2.0, and the projection-head width is 128.

Both SDOF and 5DOF training optimize a mask-aware mean-squared error in normalized output space. Let $y_{n,t,c}$ and $\hat{y}_{n,t,c}$ denote the normalized target and predicted responses for sample n, time step t, and output channel c (with $d_{\text{out}} \in \{1,5\}$

channels), and let the mask $m_{n,t} \in \{0, 1\}$ mark the valid (unpadded) time steps of each record. The training loss is

$$\mathcal{L}_{\text{train}} = \frac{\sum_{n,t,c} m_{n,t} \left(\hat{y}_{n,t,c} - y_{n,t,c}\right)^2}{d_{\text{out}} \sum_{n,t} m_{n,t}}.$$

Per-channel normalization statistics for $\hat{y}$ and y are computed once from the training dataset using only valid time steps. Padded entries are kept at zero after normalization, and all reported test metrics are computed after de-normalization to physical units with the same masking procedure.

3.1 | Surrogate performance metrics

We evaluate surrogate performance using three record-level metrics. The relative L_2 error measures the overall mismatch between the predicted and true displacement histories. The root-mean-square (RMS) response amplitude ratio indicates whether the surrogate systematically underpredicts or overpredicts the typical displacement amplitude. The peak-response error compares the predicted and true peak responses.

Because ground-motion records have different durations, all metrics are computed only over the original, unpadded portion of each response. For a true response x and prediction $\hat{x}$, the relative L_2 error is

$$\varepsilon = \frac{\|\hat{x} - x\|_2}{\|x\|_2}.$$

Here and throughout this section, the norm is evaluated only over the unpadded portion of the record. This metric is computed separately for each record and then averaged over the test set. It captures the total time-history mismatch, including differences in response amplitude, timing of peaks and cycles, oscillation pattern, and mean level.

The RMS response amplitude ratio is

$$\frac{\text{RMS}(\hat{x})}{\text{RMS}(x)},$$

where $\text{RMS}(x)$ is the root-mean-square value of x over the unpadded samples of the record. The ratio is computed separately for each record. Values below one indicate amplitude underprediction, while values above one indicate amplitude overprediction. This metric does not measure time-history accuracy by itself; instead, it helps identify whether large relative L_2 errors are associated with systematic amplitude bias.

We also compare peak responses. For each record, the peak-response error is the relative error of the peak absolute response, $\left|\max_t |\hat{x}(t)| - \max_t |x(t)|\right| / \max_t |x(t)|$. In the excitation-amplitude tests, we plot the predicted peak $\max_t |\hat{x}(t)|$ against the true peak.

The relative L_2 error is reported for every test. The RMS response amplitude ratio is reported where amplitude bias is examined. The peak-response error is reported where peak accuracy is examined.

For 5DOF runs, metrics are first computed per record and output channel over valid samples. Reported aggregate values are means over the evaluated record-channel set unless explicitly labeled as roof or floor quantities.

4 | SURROGATE EVALUATION DESIGN

The evaluation tests whether the fixed TFNO surrogate remains accurate as the response changes across four factors: restoring-force nonlinearity, structural period, input frequency content, and excitation amplitude. We use real ground-motion records to evaluate performance under realistic earthquake inputs and to test amplitude extrapolation, and controlled synthetic excitations to isolate the effect of input frequency content. These tests are repeated for SDOF and 5DOF systems to compare error behavior between a single displacement response and a distributed multi-story response. Table 1 summarizes how each factor is tested.

We hold the surrogate architecture, loss definition, normalization, and evaluation metrics fixed across studies, with one design change in the period test: the period enters as a second input channel. A separate TFNO model is trained for each restoring-force case and for each study, but all models follow the same evaluation protocol. For the 5DOF system, the output changes from one displacement channel to five displacement channels, while the same comparison logic is retained. The same real ground-motion library is reused across the restoring-force and excitation-amplitude tests at both scales; the amplitude test divides this library by peak ground acceleration (PGA) instead of at random. Differences in error can therefore be interpreted with respect to the factor being tested.

TABLE 1 Evaluation factors used to test surrogate accuracy across structural and ground-motion regimes.

Group	Factor	Implemented through	Purpose
Structure	Restoring-force law	Linear, cubic-hardening, and bilinear-hysteretic cases on the shared ground-motion library	Test increasing nonlinearity and path dependence
Structure	Structural period	Train on intermediate periods and test on separate shorter and longer periods using fixed ground-motion inputs	Test extrapolation beyond the trained range of structural periods
Ground motion	Frequency content	Synthetic multi-frequency sinusoidal inputs with progressively broader frequency ranges	Test sensitivity to excitation bandwidth
Ground motion	Excitation amplitude	Train on lower-PGA real records and test on separate higher-PGA real records	Test extrapolation to stronger shaking

All runs at both scales share the same random seed. The restoring-force and excitation-amplitude runs on the real records use a 300-epoch budget with validation-based early stopping: a run ends once its validation loss has not improved for 40 consecutive epochs, and the checkpoint with the lowest validation loss is retained. The period-test runs use a 600-epoch budget with the same rule at a patience of 40 epochs, monitored on the validation periods; every run stopped before this cap. On the real records, the linear models used the full budget at both scales, whereas the nonlinear models stopped earlier: the 5DOF cubic run retained its best checkpoint at epoch 110 and stopped at epoch 150, and the 5DOF bilinear-hysteretic run retained epoch 70 and stopped at epoch 110. In these runs the validation loss had stopped improving while the training loss was still decreasing, consistent with the onset of overfitting rather than incomplete training. The synthetic input-frequency runs use the same 300-epoch budget, with the stopping rule applied to the training loss.

4.1 | Baseline systems and restoring-force tests

This test evaluates the effect of the restoring-force law while keeping the baseline structural properties fixed. The SDOF and 5DOF systems implement the same three restoring-force laws: linear, cubic-hardening, and bilinear-hysteretic (Fig. 3). The SDOF system has mass $m = 2.0 \times 10^5$ kg, stiffness $k = 4.0 \times 10^7$ N/m, damping ratio $\zeta = 0.05$, and natural period $T_n = 0.444$ s (frequency $f_n = 2.25$ Hz). The cubic-hardening law uses $k_3 = 1.6 \times 10^{11}$ N/m^3. The bilinear-hysteretic law uses a yield displacement $x_y = 5$ mm, a yield force $F_y = 200$ kN, and a post-yield ratio $\alpha = 0.10$.

The 5DOF system has five floor masses of 2.7, 2.7, 1.8, 1.8, and 1.8 $\times 10^5$ kg and five inter-story stiffnesses of 2.45, 1.95, 0.98, 0.98, and 0.98 $\times 10^8$ N/m, both listed from the base to the top. Its first mode period is $T_1 = 0.77$ s, and Rayleigh damping is anchored to $\zeta = 0.05$ at modes 1 and 2. The story yield displacement and post-yield ratio match the SDOF nondimensional settings; the cubic coefficient is set using the same dimensionless ratio $\eta = k_3 x_{\mathrm{ref}}^2 / k = 0.40$, with k taken as the mean story stiffness, giving a single cubic coefficient shared by all stories.

All restoring-force cases at both scales are trained and tested on real ground motions: a library of 1095 records from the PEER NGA-West2 database[51], selected by two-stage stratified sampling to balance moment magnitude and spectral acceleration. The library spans M_w from 5.6 to 7.9, rupture distances from 10 to 100 km, and peak ground accelerations from about 0.003 to 1.0 g. Fig. 4a shows the joint magnitude–spectral-acceleration coverage produced by this sampling, and Fig. 4b the PGA distribution with its median and the 0.19 g threshold of the amplitude-extrapolation dataset. For the restoring-force comparison, a random 80/10/10 train/validation/test partition gives 876, 109, and 110 records and evaluates interpolation within the sampled ground-motion distribution. The selection procedure and record processing are described in Supplementary Note 1 and Supplementary Figure 1.

4.2 | Structural-period tests

This test evaluates period extrapolation by training the surrogate over a range of structural periods and testing it on periods excluded from training. Each system uses a fixed bank of 100 real ground-motion records: 91 records from the 1095-record library and 9 additional near-fault pulse records from its 2946-record candidate pool (Supplementary Note 1). The bank is stratified to include near-fault pulse records with pulse periods between 0.5 and 6 s, strong-motion records with PGA above

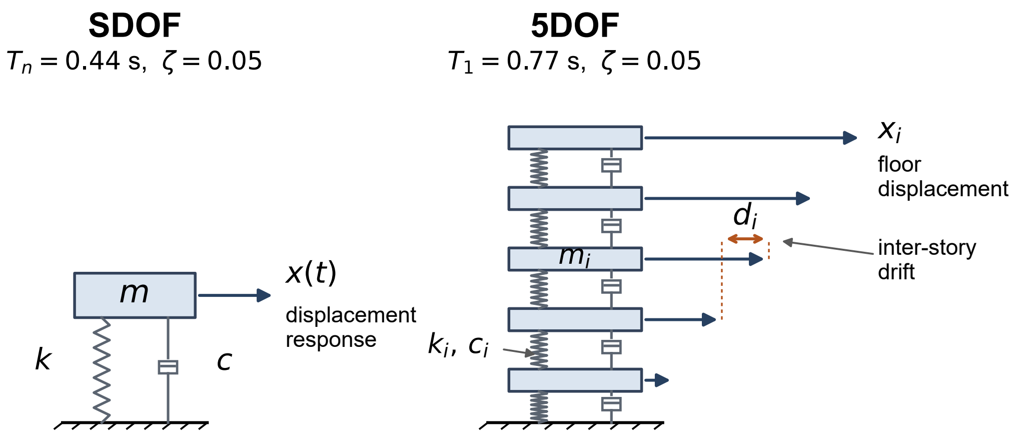


FIGURE 3 The two test structures. The SDOF system (left) is a single mass m connected to a spring k and damper c, with displacement response $x(t)$. The 5DOF system (right) stacks five floor masses m_i on inter-story springs k_i and dampers c_i. Its responses are the floor displacements x_i, from which the inter-story drifts $d_i = x_i - x_{i-1}$ are computed. Both systems use the same nondimensional restoring-force settings, allowing the 5DOF case to test the same nonlinear response behavior in a multi-story setting.

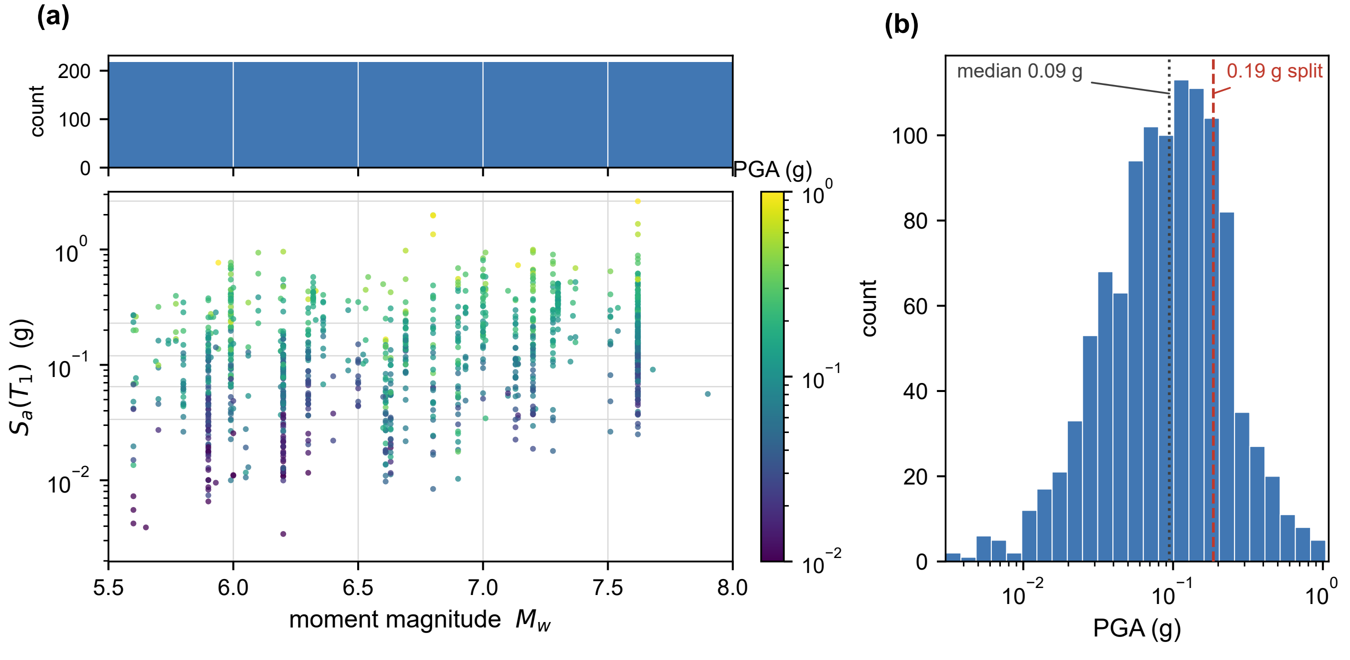


FIGURE 4 Distribution of the 1095-record real ground-motion library. (a) Joint coverage of moment magnitude M_w and major-component spectral acceleration $S_a(T_1)$ on a log scale, with the marginal magnitude histogram showing the five magnitude bins of 219 records each; points are colored by peak ground acceleration (PGA). (b) Major-component PGA, from 0.003 to 1.0 g (median 0.09 g); the dotted line marks the median and the dashed line the 0.19 g threshold of the amplitude-extrapolation dataset.

0.3 g, and records with strong long-period spectral content. The same bank is used at every period and for every restoring-force case, so the ground-motion content is held fixed and only the structural period varies.

The model receives the structural period through a second input channel, $(\log_{10} T)_{\mathrm{norm}}$, obtained by shifting and scaling $\log_{10} T$ to zero mean and unit variance over the training periods. This formulation allows a single trained model to be evaluated at any prescribed period without retraining. We use a logarithmic period input because, for a linear oscillator at fixed damping ratio, the dynamic amplification of each frequency component of the input depends on the ratio of the excitation frequency to the structural natural frequency. Equal period ratios map to equal increments in $\log_{10} T$, so equal input steps scale this frequency ratio by equal factors across the training band; the training periods are spaced uniformly on the same scale. The target displacement is normalized using the overall mean and standard deviation computed from the training data, while all reported errors are evaluated after converting the predicted and target displacements back to their original physical units.

At SDOF, the training periods cover 0.50 to 1.00 s. This band follows the first-mode estimate $T \approx 0.1N$, where N is the number of stories, and represents roughly five- to ten-story buildings. The grid holds 54 training periods plus 17 validation and 17 held-out test periods; the validation periods drive early stopping and checkpoint selection, and the in-range errors are reported at the held-out test periods. Period extrapolation is evaluated on two windows of equal width in period: a lower window $[0.10, 0.50)$ s and an upper window $(1.00, 1.40]$ s, each 0.40 s wide and sampled at 41 periods that the model never sees during training.

At 5DOF, the same design applies to the first-mode period T_1, which is varied by scaling all story stiffnesses together. This scaling keeps the mode shapes fixed and moves every modal period with T_1. The lower window cannot extend to 0.10 s at this scale: the fifth-mode frequency reaches the 25-Hz Nyquist limit of the 50-Hz time grid near $T_1 = 0.23$ s. The lower window is therefore set to $[0.357, 0.50)$ s, which keeps the fifth mode near 16 Hz, and the upper window is narrowed to $(1.00, 1.143]$ s to keep equal width.

4.3 | Input-frequency tests

Real records constrain how we can probe the input-frequency factor. Their energy concentrates below 3 Hz, near 1 to 2 Hz, and their spectral shape is set by magnitude, distance, and site [51]. Frequency content and amplitude vary together across records, so the real library cannot move the input spectrum on its own. We add synthetic two-frequency sinusoidal inputs to vary the input spectrum directly while holding the amplitude fixed. This isolates the input-frequency factor as a controlled variable.

Each synthetic input follows a two-frequency sinusoidal form,

$$a_g(t) = A_1 \sin(2\pi f_1 t + \varphi_1) + A_2 \cos(2\pi f_2 t + \varphi_2),$$

generated at 50 Hz over 40 s. The amplitudes A_1 and A_2 are drawn independently per sample from $\mathcal{U}(0.05, 0.30)$ g, the same medium-amplitude range in every regime. This amplitude range matches the moderate portion of the real library's PGA distribution (median 0.09 g; Fig. 4), so the synthetic inputs probe frequency content at realistic amplitudes. Two is the smallest number of frequency components whose spacing controls the width of the input spectrum: a single component concentrates all input energy at one frequency, whereas the separation between two components sets the excited band directly.

TABLE 2 Synthetic two-frequency forcing regimes for the SDOF system ($f_n = 2.25$ Hz). The four regimes broaden the input spectrum from a narrow band at f_n to a wide random band. Frequencies in $\mathcal{U}(\cdot,\cdot)$ are drawn uniformly per sample.

Regime	f_1, f_2 (Hz)	Spectral content
narrowband	$\mathcal{U}(2.0, 2.5)$	0.5 Hz band at f_n
medium-band	$\mathcal{U}(1.25, 3.25)$	2.0 Hz band at f_n
wide-band	$\mathcal{U}(0.25, 4.25)$	4.0 Hz band at f_n
broadband	$\mathcal{U}(0.3, 5.0)$, $\mathcal{U}(0.5, 8.0)$	0.3–8 Hz, spans all modes

Table 2 lists the four regimes. The first three are centered on the structural frequency f_n and widen the excited band in steps, from 0.5 to 2.0 to 4.0 Hz. The broadband regime is not centered on f_n: it draws its two frequencies from 0.3 to 5.0 and from 0.5 to 8.0 Hz, spanning all modal frequencies of the test systems. The narrowband regime sits inside the real-record energy band near f_n, and the broadband regime extends well beyond it. The real records are contrasted with these synthetic regimes in Figs. 5 and 6: the former shows the input acceleration time histories, while the latter compares their 5%-damped response spectra. We run each of the four regimes under the three restoring-force cases, giving twelve synthetic datasets. Each holds 4096 samples, split evenly into training and test.

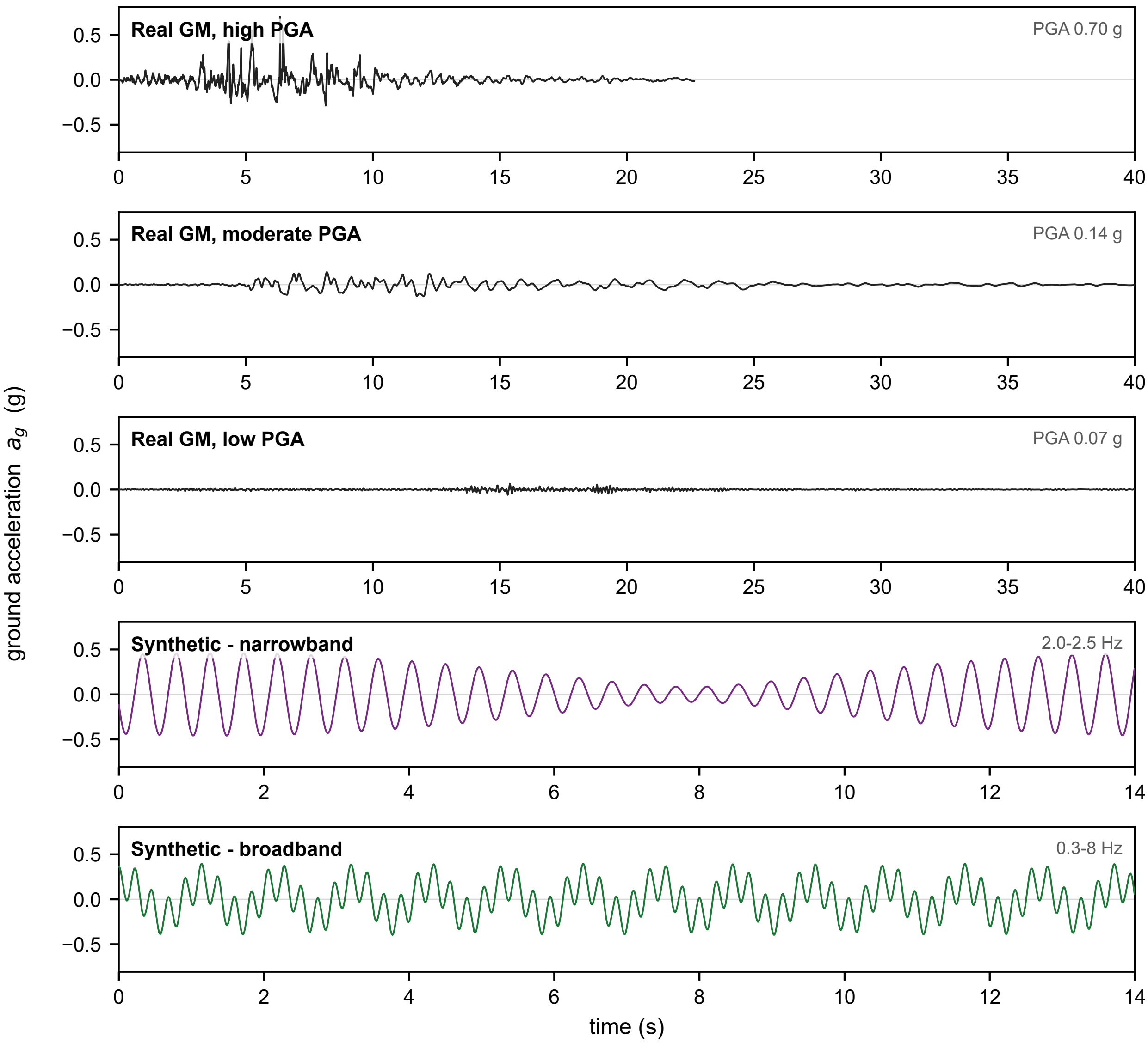


FIGURE 5 Input ground-motion time histories. The real records span the library PGA range (high, moderate, low), and the synthetic examples show the narrowband and broadband forcing regimes. All panels share the acceleration scale.

At 5DOF we repeat the comparison for the narrowband and broadband regimes. The five modal frequencies run from f_1 = 1.29 to f_5 = 7.49 Hz. The narrowband input draws both frequencies from 1.04 to 1.54 Hz at the first mode. The broadband input reuses the SDOF broadband frequency ranges and excites all five modes at once. The synthetic amplitude range only sets the overall excitation level. The excitation-amplitude factor is tested on the real records through the PGA-based division of the library.

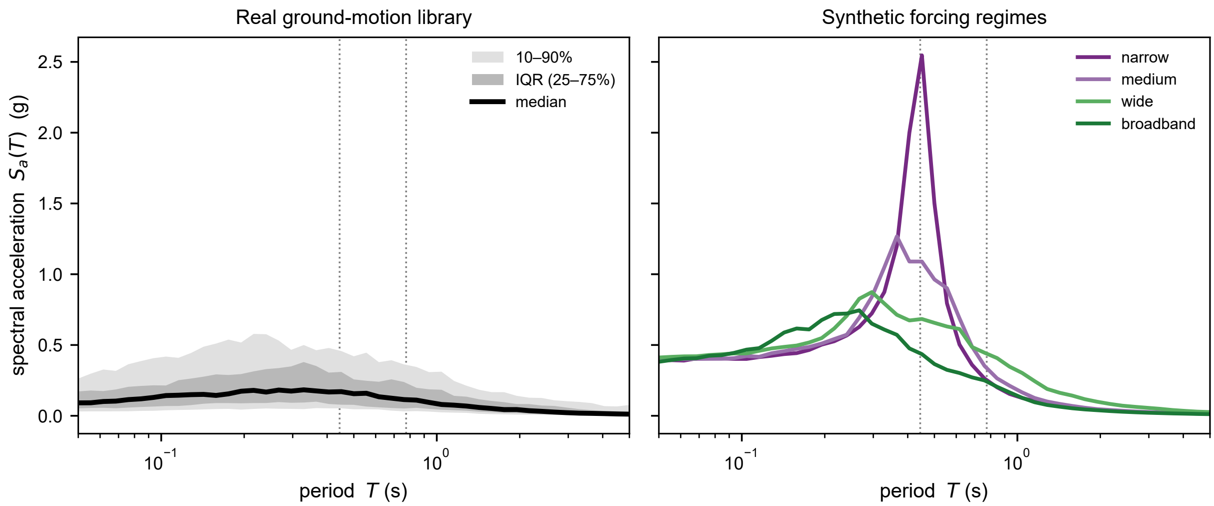


FIGURE 6 Spectral acceleration of the inputs. The real-library panel shows the median 5%-damped $S_a(T)$ with the interquartile (25–75%) and 10–90% spread bands; the synthetic-regime panel shows the median $S_a(T)$ for the four synthetic bands. Dotted lines mark the SDOF (T_n) and 5DOF (T_1) fundamental periods.

4.4 | Excitation-amplitude tests

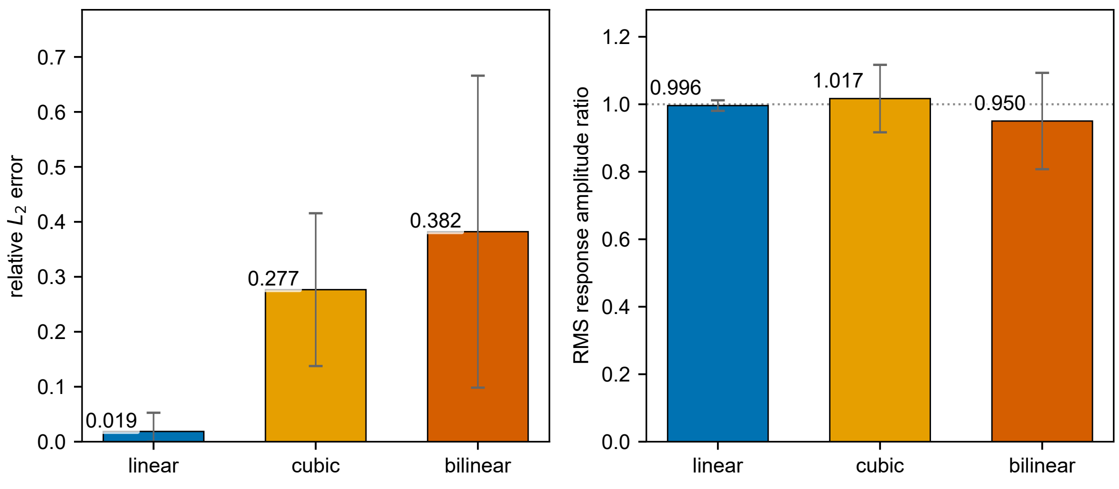
This test evaluates extrapolation to shaking stronger than the training dataset contains, the regime in which a deployed surrogate is most exposed. We split the real ground-motion library by peak ground acceleration (PGA). For each restoring-force case, a single model is trained on lower-PGA records, using 670 records for training and 96 for validation, and is evaluated on two disjoint held-out sets. The extrapolation test uses 219 higher-PGA records above approximately 0.19 g, reaching up to about 1.0 g; a separate set of 110 held-out lower-PGA records provides the interpolation reference for the same model. Because both held-out sets are evaluated with the same trained model, their contrast isolates the effect of moving from lower- to higher-amplitude shaking. Stronger records also drive the bilinear-hysteretic system deeper into yielding and the cubic-hardening system further into hardening, so this factor tests amplitude extrapolation together with stronger nonlinear response.

5 | EVALUATION ACROSS SDOF AND 5DOF SYSTEMS

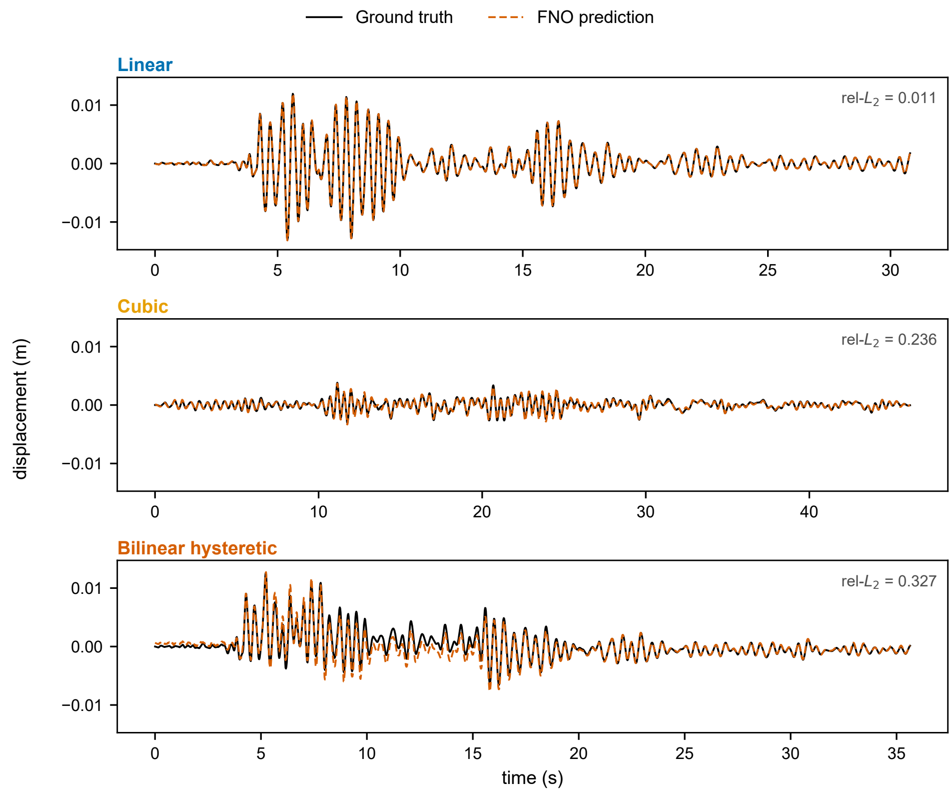
The results first use the SDOF system as a controlled baseline for identifying when TFNO predictions remain accurate and when they degrade. We evaluate four factors: restoring-force law, structural period, input frequency content, and excitation amplitude, each of which changes a distinct aspect of the response. Together, these tests separate in-distribution accuracy from extrapolation across structural and ground-motion regimes. We then repeat the same evaluation logic in the 5DOF system to test whether the SDOF error patterns persist in a higher-dimensional response with coupled floor motions. Finally, we examine the bilinear-hysteretic case in more detail because it shows the largest errors on the real-record tests and a consistent, systematic error pattern.

5.1 | Effect of restoring-force law

The surrogate loses accuracy as the restoring-force law moves from linear elasticity through cubic hardening to bilinear hysteresis. The random 80/10/10 split draws on the 1095 real ground-motion records, and the same 110 held-out records are used for every case. On this test set, the per-record relative L_2 error ε rises from 0.019 ± 0.034 for the linear case to 0.277 ± 0.139 for the cubic case and 0.382 ± 0.284 for the bilinear-hysteretic case (mean $\pm$ standard deviation; Fig. 7a). The linear prediction is nearly exact. The two nonlinear errors are roughly 15 and 20 times the linear error.



(a)



(b)

FIGURE 7 SDOF restoring-force comparison on the random ground-motion partition. (a) Per-case relative L_2 error ε (left) and RMS response amplitude ratio (right); error bars show mean $\pm$ one standard deviation over the 110 test records. (b) Representative displacement histories for the three cases on a shared displacement scale, with the per-record relative L_2 error annotated: the linear prediction overlays the true response (ε = 0.01), and both nonlinear predictions accumulate cycle-level differences (ε = 0.24 cubic, 0.33 bilinear-hysteretic).

The relative L_2 error and the RMS response amplitude ratio measure different things. The error ε counts every difference between the predicted and true histories, whether in amplitude, in timing, in oscillation pattern, or in mean level. The RMS ratio compares only the overall response amplitude: values below one mean the predicted motion is smaller than the true one. Only the bilinear-hysteretic case has a mean RMS response amplitude ratio below one (0.95 ± 0.14); the linear and cubic ratios stay near one (1.00 ± 0.02 and 1.02 ± 0.10). The bilinear-hysteretic ratio also varies most across records, so its amplitude bias is uneven: some records are reproduced at nearly the correct amplitude and others are underpredicted more strongly. The peak-response error stays small for all three cases: about 0.6% for the linear case, 8% for the cubic case, and 6% for the bilinear-hysteretic case.

Together, these numbers show that amplitude error explains only a small part of the nonlinear errors. On average, the predicted peak and RMS response amplitudes stay within about 8% of their true values, while the mean error ε is 0.28 to 0.38. The representative histories in Fig. 7b show this directly: the linear prediction overlays the true response, while both nonlinear predictions accumulate differences cycle by cycle, larger for the bilinear-hysteretic record shown. On this test set, the two nonlinear errors are of the same order. The bilinear-hysteretic case is the only one that slightly underpredicts the RMS response amplitude, and this bias appears even within the training distribution.

5.2 | Effect of structural period

We test how the surrogate performs across structural period and how it extrapolates outside the trained period range. One model per restoring-force case is trained on structures with periods between 0.50 and 1.00 s and is then evaluated on the two equal-width period windows outside this range: $[0.10, 0.50)$ s below it and $(1.00, 1.40]$ s above it, each 0.40 s wide (Fig. 8).

Within the training range, all three cases interpolate accurately. At the held-out in-range periods, the mean relative L_2 error is 0.03 for the linear case, 0.13 for the cubic case, and 0.06 for the bilinear-hysteretic case; the mean peak-response error is about 1.0% for the linear case, 2.9% for the cubic case, and 1.4% for the bilinear-hysteretic case, and the RMS response amplitude ratios stay near one. This in-range accuracy is not explained by weak nonlinearity: about 66 to 78% of the record–period pairs drive the bilinear-hysteretic system past yield.

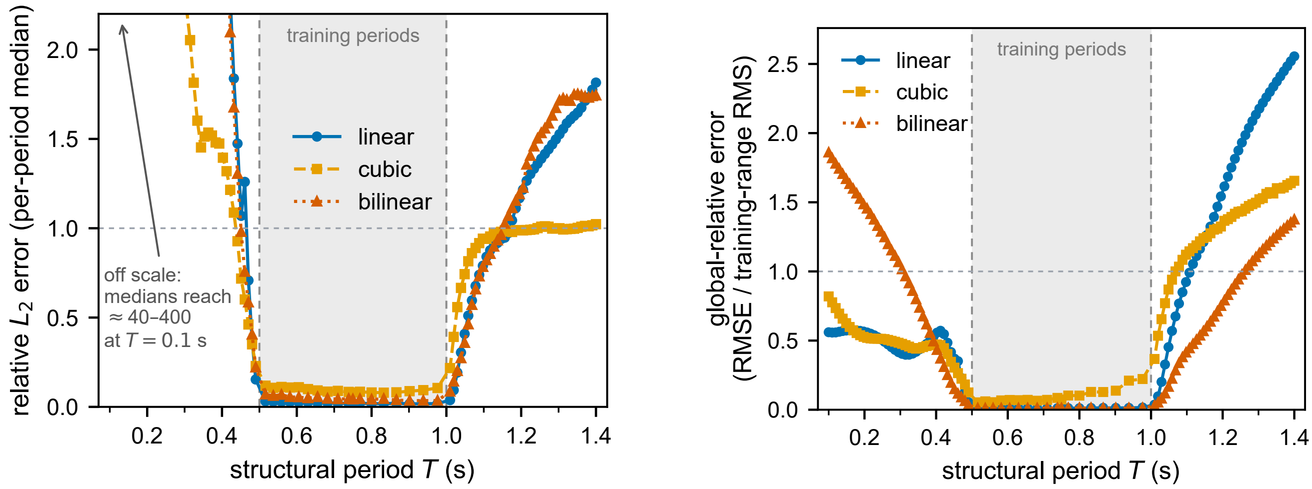
Outside the training range, prediction error increases for every restoring-force law in both windows (Fig. 8a). Relative to each case's interpolation error, the mean extrapolation error grows by a factor of 9 to 78 in the upper window and by two to three orders of magnitude in the lower window. Each case has a single training run, so we read these deterioration factors as orders of magnitude rather than precise values. The loss of accuracy outside the trained period range is shared across the three restoring-force laws.

The very large relative errors in the lower extrapolation range are partly due to the small denominator values: short-period responses are small, which inflates a per-record relative metric. The global-relative error in Fig. 8b removes this effect by dividing each record's root-mean-square error (RMSE) by the RMS response amplitude of the training-range data. Under this metric, the error still rises well above its in-range level on both sides of the training range, by up to two orders of magnitude, so the accuracy deterioration is not an artifact of small response amplitudes. Accuracy outside the trained period range is not assured for any restoring-force law, and a surrogate trained over one period range should be re-validated before it is applied outside that range.

5.3 | Effect of input frequency content

We test how the surrogate responds to broader input frequency content while keeping the structure and excitation amplitude fixed. The SDOF system is driven by synthetic two-frequency waves grouped into four bands, from a narrow band near the structural frequency to a wide random band (Table 2). For each band we train a separate model with the same architecture and training setup, so accuracy differences can be attributed to the input band. Figure 9 reports the result: panel (a) places the four synthetic bands against the central frequency band of the real records, and panel (b) reports the mean ε for the three cases across the four bands.

The linear case stays accurate in every band: its mean ε never exceeds 0.005. The two nonlinear cases degrade as the band widens. The cubic ε rises from 0.023 in the narrowband regime to 0.33 in the wide-band regime and stays near that level (0.29) under broadband input; the bilinear-hysteretic ε rises monotonically from 0.006 to 0.19. This result is consistent with the cubic and bilinear-hysteretic responses becoming harder to predict as the input frequency content broadens.



(a)

(b)

FIGURE 8 SDOF structural-period test with equal-width extrapolation windows. (a) Per-period median relative L_2 error versus structural period T. The shaded band marks the training periods ([0.50, 1.00] s); the windows below ([0.10, 0.50) s) and above ((1.00, 1.40] s) have equal width. In the lower window, the curves extend beyond the plotted range: the per-period medians reach about 40 to 400 at $T = 0.10$ s. (b) Per-period mean global-relative error (per-record RMSE divided by the RMS response amplitude of the training-range data) versus T.

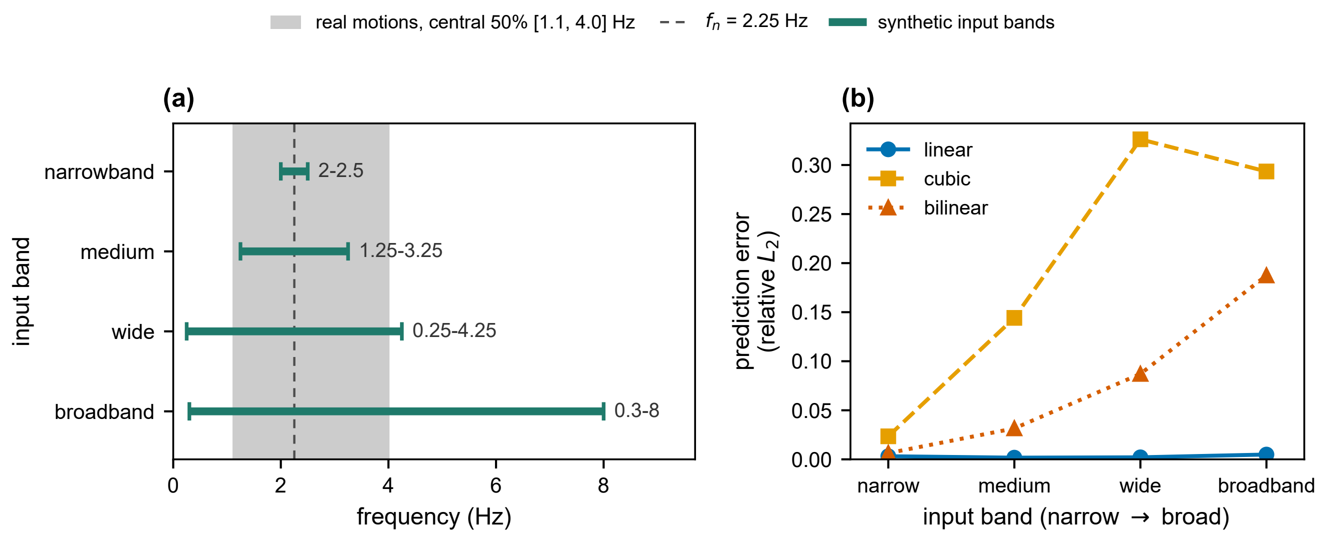


FIGURE 9 SDOF input-frequency test. (a) Synthetic input bands against the central 50% frequency band of the real ground-motion records. (b) Mean relative L_2 error for the three restoring-force cases across the four input bands. The linear error remains low across the bands, while the cubic and bilinear-hysteretic errors rise up to the wide band, and under broadband input the cubic error stays near its wide-band level.

5.4 | Effect of excitation amplitude

We test how the surrogate extrapolates to stronger shaking. The real records are split by PGA: a single model is trained only on the lower-PGA records (up to about 0.19 g) and then evaluated on two held-out sets: an interpolation set of 110 records with in-range PGA, and an extrapolation set of 219 records with higher PGA (up to about 1.0 g). Because both test errors come from the same model, the comparison isolates the effect of amplitude extrapolation.

Extrapolation increases error ε for all three cases: the interpolation errors are 0.018, 0.087, and 0.411, and the extrapolation errors are 0.039, 0.539, and 0.672 for the linear, cubic, and bilinear-hysteretic cases, respectively. The cubic interpolation error of 0.087 sits well below the 0.277 measured in the restoring-force test because the lower-PGA interpolation records activate the cubic nonlinearity only weakly. Figure 10 compares the predicted and true peak responses record by record. The linear points stay close to the 1:1 line in both sets, and the cubic extrapolation points scatter about the line, with several rising above it, without a systematic downward bias. The bilinear-hysteretic extrapolation points fall below the line, and the gap widens as the true response grows, so the model underpredicts the largest responses most. The RMS response amplitude ratio summarizes the same bias: for the bilinear-hysteretic case it falls from 0.88 at interpolation to 0.83 at extrapolation, while the linear and cubic ratios remain near one (0.99 and 1.04). The interpolation-versus-extrapolation bars and the per-record RMS-ratio scatter are in Supplementary Figures 2 and 3. The displacement histories in Fig. 11 show these patterns for representative records: the linear prediction tracks the response in both sets, the cubic prediction keeps the amplitude near the true level under extrapolation but misses individual cycles, and the bilinear-hysteretic prediction underpredicts the large post-yield displacements. Two further record pairs are in Supplementary Figures 4 and 5.

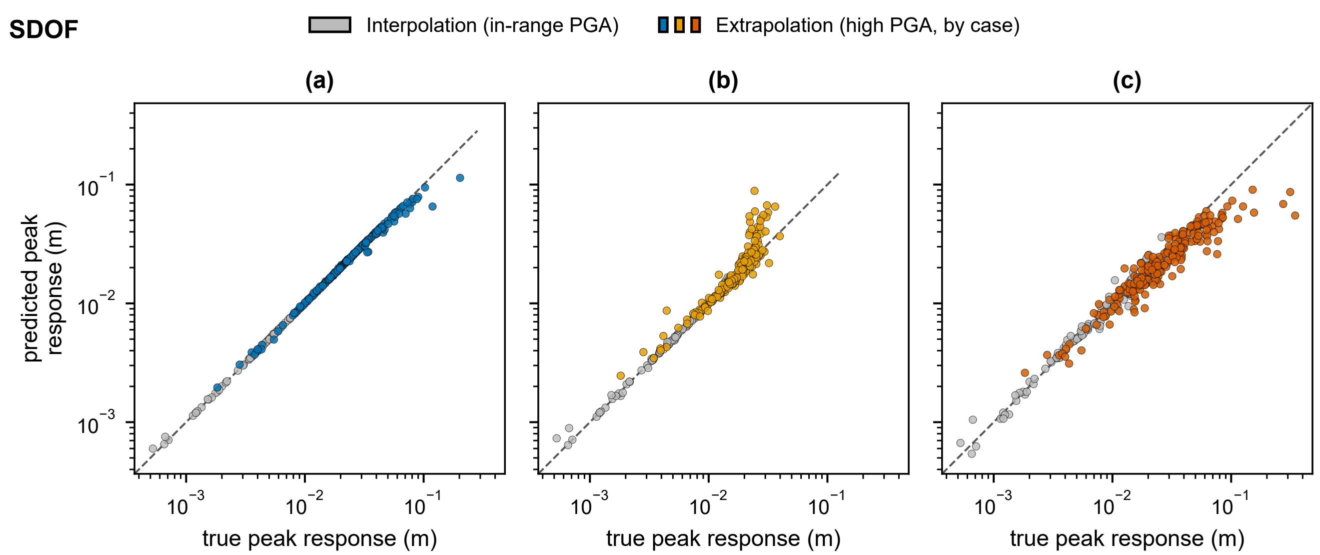


FIGURE 10 SDOF excitation-amplitude test: predicted versus true peak response for each record, with the dashed 1:1 line, for the (a) linear, (b) cubic, and (c) bilinear-hysteretic cases. Gray points are interpolation (in-range PGA, n = 110); colored points are extrapolation (high PGA, n = 219). For the bilinear-hysteretic case (c), the extrapolation points fall increasingly below the 1:1 line as the true response grows.

5.5 | Cross-scale replication at 5DOF

At 5DOF, we repeat the SDOF evaluation with the same ground-motion records, data splits, training protocol, and random seed. This comparison tests whether the SDOF conclusions carry over to a 5DOF system. In the restoring-force test, the per-record ε is 0.046, 0.153, and 0.454 for the linear, cubic, and bilinear-hysteretic cases (Fig. 12), keeping the SDOF ordering: the linear case is easy, the bilinear-hysteretic case is hard. The RMS response amplitude ratio repeats the SDOF pattern: only the bilinear-hysteretic ratio falls below one (about 0.96).

Yielding and prediction error are not concentrated at the first story, although it carries the largest story shear. The same restoring-force law applies to every story, and in a yielding check on a stratified subset of 30 records, every story yielded in 30–67% of them, most often story 3, where the story stiffness drops most sharply. The per-floor ε for the bilinear-hysteretic case

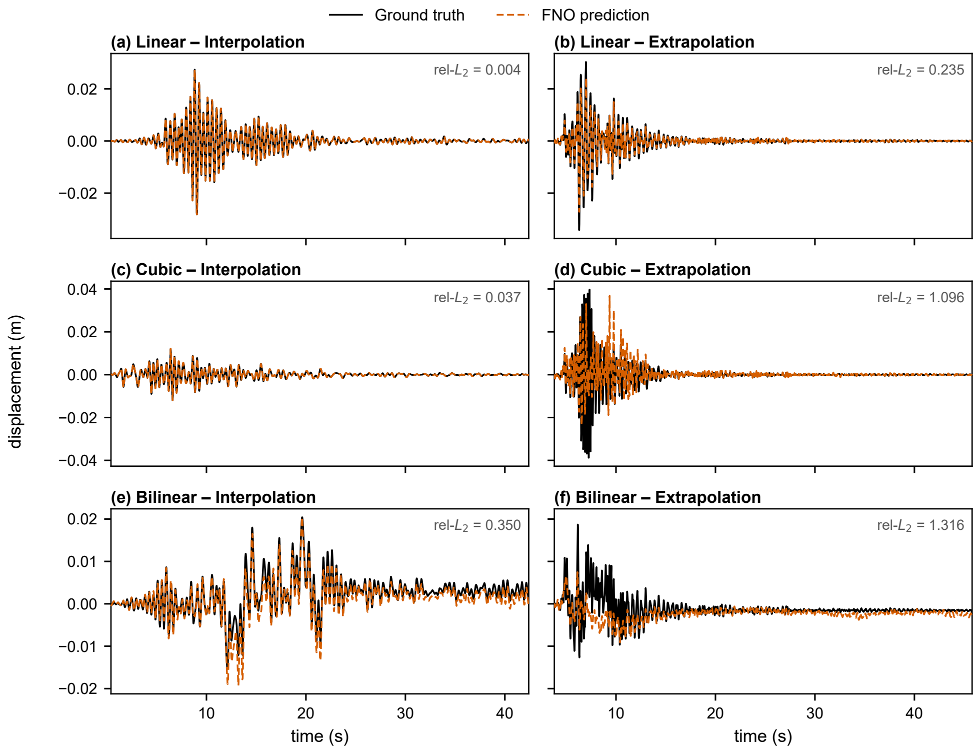


FIGURE 11 SDOF displacement histories under amplitude interpolation (left column) and extrapolation (right column) for representative records of each case, with the per-record relative L_2 error ε annotated. The linear prediction tracks the response (the extrapolation record shown is a comparatively hard one, $\varepsilon = 0.24$, against a case mean of 0.04). The cubic prediction keeps the response amplitude under extrapolation and misses individual cycles ($\varepsilon = 1.10$). The bilinear-hysteretic prediction underpredicts the large post-yield displacement excursions under extrapolation ($\varepsilon = 1.32$).

is likewise spread out, ranging from 0.38 to 0.52 across the five floors and peaking at floor 3 ($n = 5$ floors, single training run). Note that ε is computed on floor displacements. Because floor displacement is the cumulative sum of the story deformations below that floor, a floor-level error cannot be assigned to that floor's story alone. Taken together, the yielding check and the per-floor error profile indicate that the bilinear-hysteretic difficulty is not a base-story-only effect; it appears across the height of this 5DOF system.

The period and input-frequency difficulties also appear at 5DOF. The period test applies the SDOF protocol to the first-mode period T_1; the lower window is limited to $[0.357, 0.50)$ s by the Nyquist limit of the output time grid, and the upper window is narrowed to $(1.00, 1.143]$ s to keep equal width. Inside the training range, the mean error ε is 0.04, 0.09, and 0.08 for the linear, cubic, and bilinear-hysteretic cases. In both extrapolation windows, the error grows for every restoring-force law, by factors of 8 to 68 relative to each case's own in-range level. The same qualitative deterioration outside the trained period range thus appears at both scales; the 5DOF windows are narrower than the SDOF windows, so the two scales are not compared on extrapolation severity. As the input spectrum broadens from narrowband to broadband, the cubic error rises from 0.07 to 0.44 and the bilinear-hysteretic error from 0.06 to 0.43, while the linear error stays below 0.04. Supplementary Note 3 reports both tests (Supplementary Figures 6 and 7).

Amplitude extrapolation produces the most systematic underprediction observed in these tests, and it is specific to the bilinear-hysteretic case. Under PGA extrapolation, ε grows for all three cases. The linear and cubic RMS ratios stay near one (about 1.00 and 0.96), while the bilinear-hysteretic ratio falls from 0.98 at interpolation to 0.83 at extrapolation. For the bilinear-hysteretic case, the predicted-to-true RMS response amplitude ratio decreases as the true peak response grows (Supplementary Figure 12). As in the SDOF tests, the largest responses are underpredicted most strongly (Fig. 13). Under extrapolation, the surrogate increasingly underpredicts the bilinear-hysteretic response and shifts its mean level, while the linear and cubic predictions track the true response. Supplementary Figures 8–10 illustrate this with the roof displacement, which

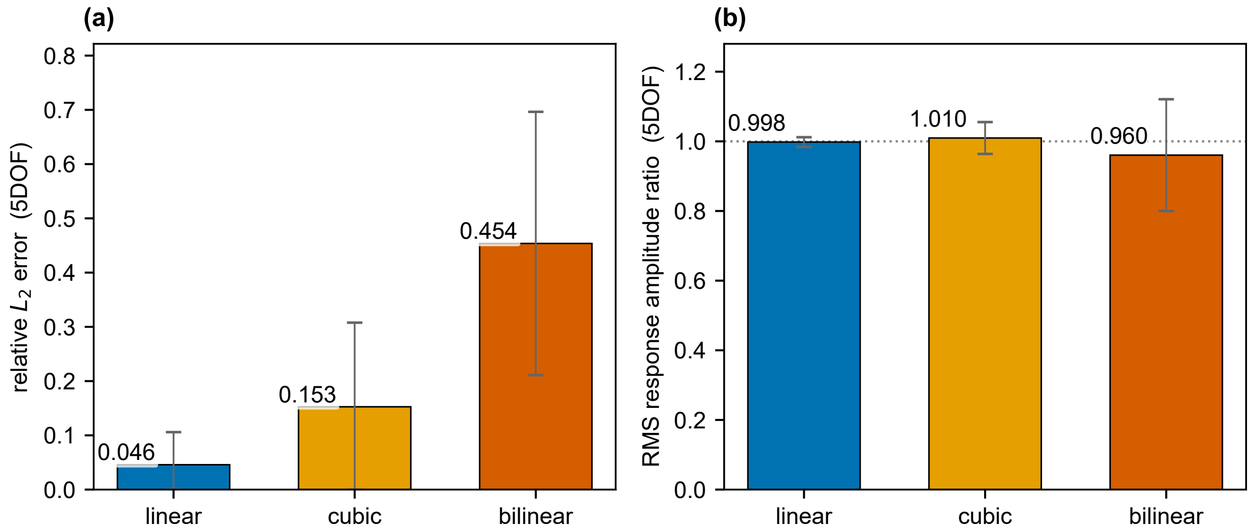


FIGURE 12 5DOF restoring-force case comparison on the random ground-motion partition: (a) per-case relative L_2 error; (b) RMS response amplitude ratio. Error bars show mean $\pm$ one standard deviation over the test records.

accumulates the drifts of all stories and serves as a representative response quantity. The interpolation-versus-extrapolation bars and the per-record RMS-ratio scatter are in Supplementary Figures 11 and 12.

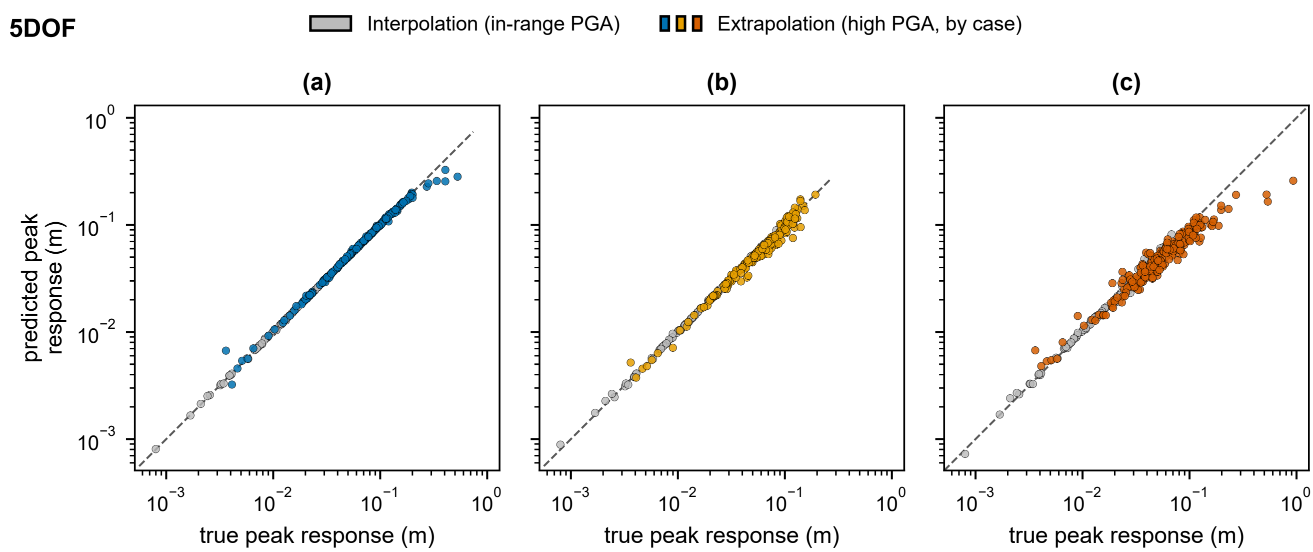


FIGURE 13 5DOF excitation-amplitude test: predicted versus true peak response for each record, with the dashed 1:1 line, for the (a) linear, (b) cubic, and (c) bilinear-hysteretic cases. Gray points are interpolation (in-range PGA); colored points are extrapolation (high PGA). For the bilinear-hysteretic case (c), the extrapolation points fall increasingly below the 1:1 line as the true response grows.

5.6 | Why the bilinear-hysteretic case is hardest on the real-record tests

To identify which part of the prediction error makes the bilinear-hysteretic case the hardest on the real-record tests, we measure two parts of each record's error (Fig. 14a,b; Supplementary Note 5 gives the formal definition): a residual offset, in which the prediction oscillates about a wrong mean level, and an amplitude-scale error, in which the prediction has a wrong overall amplitude. Residual displacement shifts the response of a yielding structure away from oscillations centered around zero, so a prediction that misses the residual displacement carries a residual offset. Each part is quantified by how much the error drops when that part is removed from both the prediction and the true response; Fig. 14c illustrates the mismatch that remains after both removals.

The decomposition shows that an appreciable residual-offset component is specific to the bilinear-hysteretic case. On the SDOF, removing the mean level reduces the bilinear-hysteretic error ε from 0.382 to 0.299, a 21.7% reduction, whereas the same operation reduces the cubic error by only about 3% (Fig. 15). Removing the amplitude scale changes the error little for either case. Of the two measured parts, only the residual offset separates the bilinear-hysteretic case from the cubic case, and it amounts to about a fifth of the bilinear-hysteretic error.

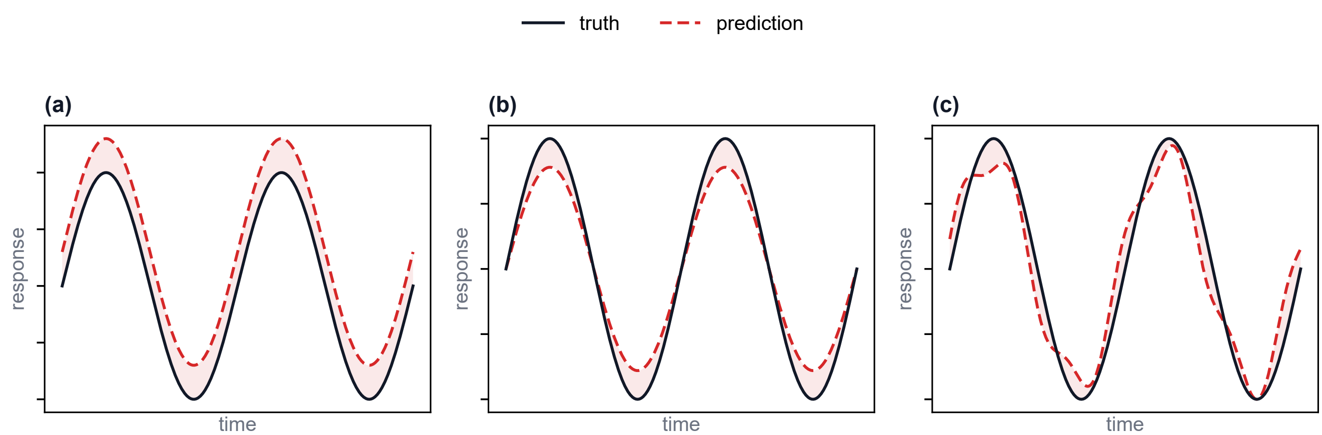


FIGURE 14 Error modes in the decomposition of the SDOF prediction error: the two measured parts, (a) residual offset and (b) amplitude scaling, and (c) the mismatch that remains after both parts are removed.

The same residual offset appears at 5DOF. The true responses carry residual displacement: at the first story, the residual deformation exceeds 1 mm in about 18% of the test records, reaching at most 9.7 mm, about twice the yield displacement. For the bilinear-hysteretic case, removing the mean level reduces the floor-displacement error by 19.8% (median over records and channels), while the cubic reduction stays near zero.

Table 3 collects the measured contrasts between the bilinear-hysteretic case and the cubic reference, and lists candidate mechanisms consistent with them. The contrast points to the restoring-force laws themselves: the cubic force is an instantaneous function of the current state, while the bilinear-hysteretic force depends on the accumulated plastic state, which the operator does not receive as an input.

6 | CONCLUSIONS

This paper identified where a fixed FNO surrogate predicts earthquake structural response accurately and where it loses accuracy. We tested one TFNO configuration on SDOF and 5DOF systems along four factors: the restoring-force law, the structural period, the input frequency content, and the excitation amplitude. These four factors are concurrent and partially coupled. The surrogate is most accurate for linear response, narrowband inputs, and responses within the trained period and amplitude ranges. Within these trained ranges, nonlinear response is the main source of degradation, and the bilinear-hysteretic case shows the clearest practical limitation. On the real-record tests, it shows the largest relative L_2 error of the three cases. It underpredicts

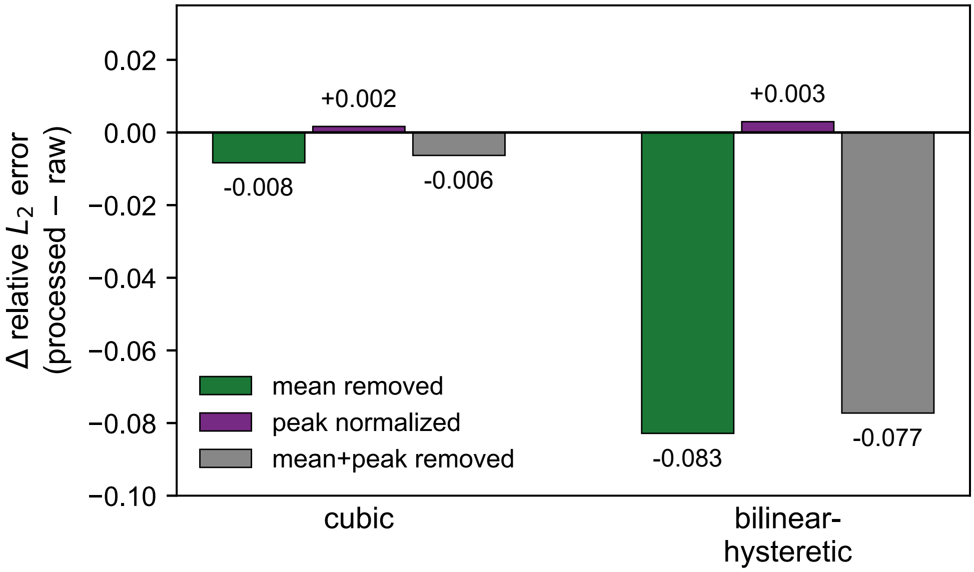


FIGURE 15 Change in relative L_2 error (processed – raw) after each processing step for the cubic and bilinear-hysteretic cases on the random-split interpolation set. Negative bars indicate error reduction.

TABLE 3 Cubic versus bilinear-hysteretic error patterns at SDOF and 5DOF. The first three rows report the measured metrics for each case. The last row lists candidate mechanisms consistent with these measurements.

Metric	Cubic	Bilinear-hysteretic
Whole-record relative L_2 error	Intermediate: above linear, below bilinear-hysteretic[a]	Largest of the three[a]
PGA-extrapolation RMS ratio[b]	Near one (1.04 SDOF, 0.96 5DOF)	Falls to 0.83 (SDOF and 5DOF); underprediction grows with the true response amplitude
Residual-offset share (percent of raw error)	About 3% at SDOF and near zero at 5DOF[c]	19.8–21.7%, the only case with an appreciable residual-offset component[d]
Candidate mechanism	Amplitude-dependent stiffness variation, frequency mixing, spectral-shape mismatch	Path-dependent memory, hidden-state sensitivity

[a] On the random ground-motion partition, at both SDOF and 5DOF.
[b] Predicted-to-true RMS response amplitude ratio; values below one indicate underprediction of the response amplitude.
[c] Cubic residual-offset share: 3.0% at SDOF on the random-split interpolation set; near zero at 5DOF.
[d] Bilinear-hysteretic residual-offset share: 21.7% (SDOF, random-partition interpolation, mean aggregate) and 19.8% (5DOF floor displacement, random partition; median over records and channels).

the response amplitude: its RMS response amplitude ratio, the ratio of predicted to true RMS response amplitude, is below one. Its error carries an appreciable residual-offset component, meaning the prediction oscillates about a wrong mean level. Under PGA extrapolation, the amplitude underprediction persists at both scales. Accuracy drops sharply under period and amplitude extrapolation.

The restoring-force test shows a clear accuracy ranking. At SDOF, the linear case is nearly exact, with a mean per-record relative L_2 error of 0.019; the cubic case is intermediate, at 0.277; and the bilinear-hysteretic case is the hardest, at 0.382. On the same records, the linear and cubic cases have mean RMS response amplitude ratios near one. The bilinear-hysteretic case has a mean ratio of 0.95, and this ratio varies widely across records. At 5DOF, its mean RMS response amplitude ratio is about 0.96. The bilinear-hysteretic case thus combines the largest relative L_2 error with a mean amplitude ratio below one.

The structural-period test shows that the accuracy achieved within the training range does not extend beyond the period range used for training. One model per restoring-force case is trained over periods of 0.50 to 1.00 s, with the structural period as an input, and interpolates accurately inside that range. The mean relative L_2 error increases outside the trained period range for all three restoring-force laws: the smallest increase across the cases and windows is close to an order of magnitude, and the largest exceeds two orders of magnitude. The same qualitative behavior is observed in the 5DOF test. A surrogate trained for one period range therefore cannot be assumed to transfer to a structure whose fundamental period lies outside that range, and it should be re-validated against computed responses before such use.

Broader input frequency content lowers accuracy even under interpolation. With the structure fixed and the input spectrum broadened from narrowband to broadband at matched amplitude, the linear error stays low while the nonlinear errors grow: the cubic relative L_2 error rises from about 0.02 to 0.29 at SDOF, and at 5DOF the cubic and bilinear-hysteretic errors rise from 0.07 to 0.44 and from 0.06 to 0.43. In practice, the frequency content of real ground motions is set by earthquake magnitude, distance, and site conditions. For example, motions recorded on soft soil carry frequency content different from that of motions recorded on stiff sites. These results therefore suggest that, for nonlinear structures, accuracy will also depend on the frequency content of the input motions.

The excitation-amplitude test shows two effects. First, errors grow for all three cases beyond the training PGA range. A single model per case was trained on records with PGA up to about 0.19 g and evaluated both on held-out records in that range and on stronger records with PGA up to about 1.0 g. Under extrapolation, the linear relative L_2 error more than doubles, from 0.018 to 0.039, but stays the lowest of the three; the cubic error rises from 0.087 to 0.539; and the bilinear-hysteretic error, already 0.411 within the training range, rises to 0.672. Second, the bilinear-hysteretic surrogate systematically underpredicts the response: its RMS response amplitude ratio is about 0.83 at both scales under extrapolation, an average amplitude underprediction of about 17%. The underprediction is largest for the strongest responses. The linear and cubic ratios stay near one under extrapolation. At SDOF, this bias already appears within the interpolation range. Under extrapolation, the held-out high PGA records produce larger nonlinear responses than those seen during training. Predictions beyond the training range should therefore be checked against simulation before use.

Together, the four tests locate the accuracy limits of this surrogate: in the structure, through the restoring-force law and the structural period, and in the ground motion, through the input frequency content and the excitation amplitude. The same qualitative patterns appear under the matched 5DOF tests, so these limits are not specific to the simplest structure. In practice, the training data of such a surrogate should cover the structural periods, the input frequency content, and the shaking amplitudes expected in use.

We held the TFNO configuration fixed to isolate the four factors, so error differences across tests can be attributed to the tested structural and excitation factors rather than to changes in the surrogate; tuning the network architecture was deliberately outside the scope of this study. The bilinear-hysteretic case suggests one direction for future work. Its restoring force depends on an accumulated plastic state that the operator never receives, and the error decomposition is consistent with this gap: the bilinear-hysteretic error is the only one with an appreciable residual-offset component. Hysteresis models describe exactly this quantity through internal state variables[41,42], and making such internal variables part of the learning problem, as model inputs or additional prediction targets, parallels internal-variable constitutive modeling in solid mechanics[52,53]. These ideas could be adapted to operator learning for earthquake response.

SUPPLEMENTARY INFORMATION

Supplementary Information accompanies this paper. It contains Supplementary Notes 1–5 and Supplementary Figures 1–12.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from the Department of Civil and Environmental Engineering at the University of California, Berkeley.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the DesignSafe-CI Data Depot as project PRJ-6429 at https://doi.org/10.17603/ds2-td20-nz56. The package contains the model predictions, the per-record errors, the manuscript figures, and the scripts that reproduce the reported error metrics. The ground-motion records are available from the PEER NGA-West2 database at https://ngawest2.berkeley.edu/.

REFERENCES

1. Zhang R, Chen Z, Chen S, Zheng J, Büyüköztürk O, Sun H. Deep long short-term memory networks for nonlinear structural seismic response prediction. *Computers & Structures.* 2019;220:55–68.
2. Xu Z, Chen J, Shen J, Xiang M. Recursive long short-term memory network for predicting nonlinear structural seismic response. *Engineering Structures.* 2022;250:113406.
3. Houng SE, Ceferino L. Scalable Scenario-based Earthquake Risk Modeling via Linearized Ground-Motion–Fragility Coupling and PPCA.
4. Ceferino L, Kukunoor C, Zhao J, et al. Accessing acute care hospitals in the San Francisco Bay Area after a major hayward earthquake. *Nature Communications.* 2025;16(1):9328.
5. Arora P, Ceferino L. Probabilistic and machine learning methods for uncertainty quantification in power outage prediction due to extreme events. *Natural Hazards and Earth System Sciences.* 2023;23(5):1665–1683. doi: 10.5194/nhess-23-1665-2023
6. Lallemant D, Burton H, Ceferino L, Bullock Z, Kiremidjian A. A framework and case study for earthquake vulnerability assessment of incrementally expanding buildings. *Earthquake spectra.* 2017;33(4):1369–1384.
7. Li T, Pan Y, Tong K, Ventura CE, De Silva CW. Attention-based sequence-to-sequence learning for online structural response forecasting under seismic excitation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems.* 2021;52(4):2184–2200.
8. Azimi M, Eslamlou AD, Pekcan G. Data-Driven Structural Health Monitoring and Damage Detection through Deep Learning: State-of-the-Art Review. *Sensors.* 2020;20(10). doi: 10.3390/s20102778
9. Roohi M, Hernandez EM. Performance-based post-earthquake decision making for instrumented buildings. *Journal of Civil Structural Health Monitoring.* 2020;10(5):775–792.
10. Torzoni M, Tezzele M, Mariani S, Manzoni A, Willcox KE. A digital twin framework for civil engineering structures. *Computer Methods in Applied Mechanics and Engineering.* 2024;418:116584.
11. Dang HV, Tatipamula M, Nguyen HX. Cloud-based digital twinning for structural health monitoring using deep learning. *IEEE transactions on industrial informatics.* 2021;18(6):3820–3830.
12. Fu S, Chang CC, Kaewnuratchadasorn C, et al. Nonlinear finite element model updating and fourier neural operator for digital twin of reinforced concrete containment vessel under seismic excitations. *Engineering Structures.* 2025;323:119234. doi: https://doi.org/10.1016/j.engstruct.2024.119234
13. Dokainish M, Subbaraj K. A survey of direct time-integration methods in computational structural dynamicsI. Explicit methods. *Computers & Structures.* 1989;32(6):1371-1386. doi: https://doi.org/10.1016/0045-7949(89)90314-3
14. Bathe KJ, Wilson EL. Stability and accuracy analysis of direct integration methods. *Earthquake Engineering & Structural Dynamics.* 1972;1(3):283-291. doi: https://doi.org/10.1002/eqe.4290010308
15. Zhang Q, Guo M, Zhao L, Li Y, Zhang X, Han M. Transformer-based structural seismic response prediction. *Structures.* 2024;61:105929. doi: https://doi.org/10.1016/j.istruc.2024.105929
16. Kim T, Kwon OS, Song J. Response prediction of nonlinear hysteretic systems by deep neural networks. *Neural Networks.* 2019;111:1–10.
17. Kovachki N, Li Z, Liu B, et al. Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs. *Journal of Machine Learning Research.* 2023;24(89):1–97.
18. Lu L, Meng X, Cai S, et al. A comprehensive and fair comparison of two neural operators (with practical extensions) based on FAIR data. *Computer Methods in Applied Mechanics and Engineering.* 2022;393:114778. doi: https://doi.org/10.1016/j.cma.2022.114778
19. Lu L, Jin P, Pang G, Zhang Z, Karniadakis GE. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature machine intelligence.* 2021;3(3):218–229.
20. Yang LH, Luo XL, Yang ZB, Nan CF, Chen XF, Sun Y. FE reduced-order model-informed neural operator for structural dynamic response prediction. *Neural Networks.* 2025;188:107437. doi: https://doi.org/10.1016/j.neunet.2025.107437
21. Cao Q, Goswami S, Tripura T, Chakraborty S, Karniadakis GE. Deep neural operators can predict the real-time response of floating offshore structures under irregular waves. *Computers & Structures.* 2024;291:107228. doi: https://doi.org/10.1016/j.compstruc.2023.107228
22. Li Z, Kovachki N, Azizzadenesheli K, et al. Fourier Neural Operator for Parametric Partial Differential Equations. arXiv preprint arXiv:2010.08895; 2021.
23. Kossaifi J, Kovachki N, Azizzadenesheli K, Anandkumar A. Multi-Grid Tensorized Fourier Neural Operator for High-Resolution PDEs. arXiv preprint arXiv:2310.00120; 2023.
24. Tran A, Mathews A, Xie L, Ong CS. Factorized fourier neural operators. *arXiv preprint arXiv:2111.13802.* 2021.
25. Preumont A. *Twelve Lectures on Structural Dynamics*. Springer Netherlands, 2013
26. Worden K. *Nonlinearity in structural dynamics: detection, identification and modelling*. CRC Press, 2019.
27. Xu ZQJ, Zhang Y, Luo T, Xiao Y, Ma Z. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523.* 2019.
28. Qin S, Lyu F, Peng W, et al. Toward a Better Understanding of Fourier Neural Operators from a Spectral Perspective. arXiv preprint; 2024.
29. Kalimuthu M, Holzmüller D, Niepert M. Loglo-fno: efficient learning of local and global features in fourier neural operators. *arXiv preprint arXiv:2504.04260.* 2025.
30. Oommen V, Bora A, Zhang Z, Karniadakis GE. Integrating Neural Operators with Diffusion Models Improves Spectral Representation in Turbulence Modelling. *Proceedings of the Royal Society A.* 2025;481(2309):20240819. doi: 10.1098/rspa.2024.0819
31. Kong Q, Zou C, Choi Y, et al. Reducing frequency bias of fourier neural operators in 3d seismic wavefield simulations through multistage training. *Seismological Research Letters.* 2026;97(1):272–282.

32. Lanthaler S, Stuart AM, Trautner M. Discretization Error of Fourier Neural Operators. arXiv preprint; 2025.
33. Gao W, Xu R, Deng Y, Liu Y. Discretization-invariance? On the Discretization Mismatch Errors in Neural Operators. In: International Conference on Learning Representations. 2025.
34. Sakarvadia M, Hegazy K, Totounferoush A, et al. The False Promise of Zero-Shot Super-Resolution in Machine-Learned Operators. arXiv preprint; 2026.
35. Kim T, Kang M. Bounding the Rademacher complexity of Fourier neural operators. *Machine Learning.* 2024;113(5):2467–2498.
36. Haghi R, Gaikwad B, Patra A. Fourier Neural Operators for Structural Dynamics Models: Challenges, Limitations and Advantages of Using a Spectrogram Loss. arXiv preprint; 2025.
37. Krishnapriyan AS, Gholami A, Zhe S, Kirby RM, Mahoney MW. Characterizing possible failure modes in physics-informed neural networks. arXiv preprint; 2021.
38. Goswami S, Giovanis DG, Li B, Spence SM, Shields MD. Neural operators for stochastic modeling of nonlinear structural system response to natural hazards. *Engineering Structures.* 2025;345:121284. doi: https://doi.org/10.1016/j.engstruct.2025.121284
39. Kim J, Yi rS, Wang Z. A composition of simplified physics-based model with neural operator for trajectory-level seismic response predictions of structural systems. *Structural Safety.* 2026;119:102668. doi: https://doi.org/10.1016/j.strusafe.2025.102668
40. Hu Y, Guo W, Tsang HH, Li S. Struct-WNO: Wavelet neural operator for seismic response prediction of nonlinear structural system. *Journal of Building Engineering.* 2025;111:113381. doi: https://doi.org/10.1016/j.jobe.2025.113381
41. Wen YK. Method for Random Vibration of Hysteretic Systems. *Journal of the Engineering Mechanics Division.* 1976;102(2):249-263. doi: 10.1061/JMCEA3.0002106
42. Ismail M, Ikhouane F, Rodellar J. The Hysteresis Bouc-Wen Model, a Survey. *Archives of Computational Methods in Engineering.* 2009;16(2):161–188. doi: 10.1007/s11831-009-9031-8
43. Kovachki N, Lanthaler S, Mishra S. On Universal Approximation and Error Bounds for Fourier Neural Operators. *Journal of Machine Learning Research.* 2021;22(290):1–76.
44. MacRae GA, Kawashima K. Post-earthquake residual displacements of bilinear oscillators. *Earthquake engineering & structural dynamics.* 1997;26(7):701–716.
45. Oppenheim AV, Willsky AS, Nawab SH. *Signals & systems*. Pearson Educación, 1997.
46. Liu CL, Pei SC. Closed-form output response of discrete-time linear time-invariant systems using intermediate auxiliary functions [lecture notes]. *IEEE Signal Processing Magazine.* 2020;37(5):140–145.
47. Berns K, Köpper A, Schürmann B. *Systems Theory*:85–118; Cham: Springer International Publishing . 2021
48. Chatterjee A, Chintha HP. Identification and parameter estimation of cubic nonlinear damping using harmonic probing and volterra series. *International Journal of Non-Linear Mechanics.* 2020;125:103518. doi: https://doi.org/10.1016/j.ijnonlinmec.2020.103518
49. Cichocki A, Lee N, Oseledets I, Phan AH, Zhao Q, Mandic DP. Tensor networks for dimensionality reduction and large-scale optimization part 1 low-rank tensor decompositions. *Foundations and Trends in Machine Learning.* 2016;9(4-5):249–429.
50. Liu X, Parhi KK. Tensor Decomposition for Model Reduction in Neural Networks: A Review [Feature]. *IEEE Circuits and Systems Magazine.* 2023;23(2):8-28. doi: 10.1109/MCAS.2023.3267921
51. Ancheta TD, Darragh RB, Stewart JP, et al. NGA-West2 database. *Earthquake Spectra.* 2014;30(3):989–1005.
52. Coleman BD, Gurtin ME. Thermodynamics with internal state variables. *The journal of chemical physics.* 1967;47(2):597–613.
53. Horstemeyer MF, Bammann DJ. Historical review of internal state variable theory for inelasticity. *International Journal of Plasticity.* 2010;26(9):1310-1334. Special Issue In Honor of David L. McDowelldoi: https://doi.org/10.1016/j.ijplas.2010.06.005