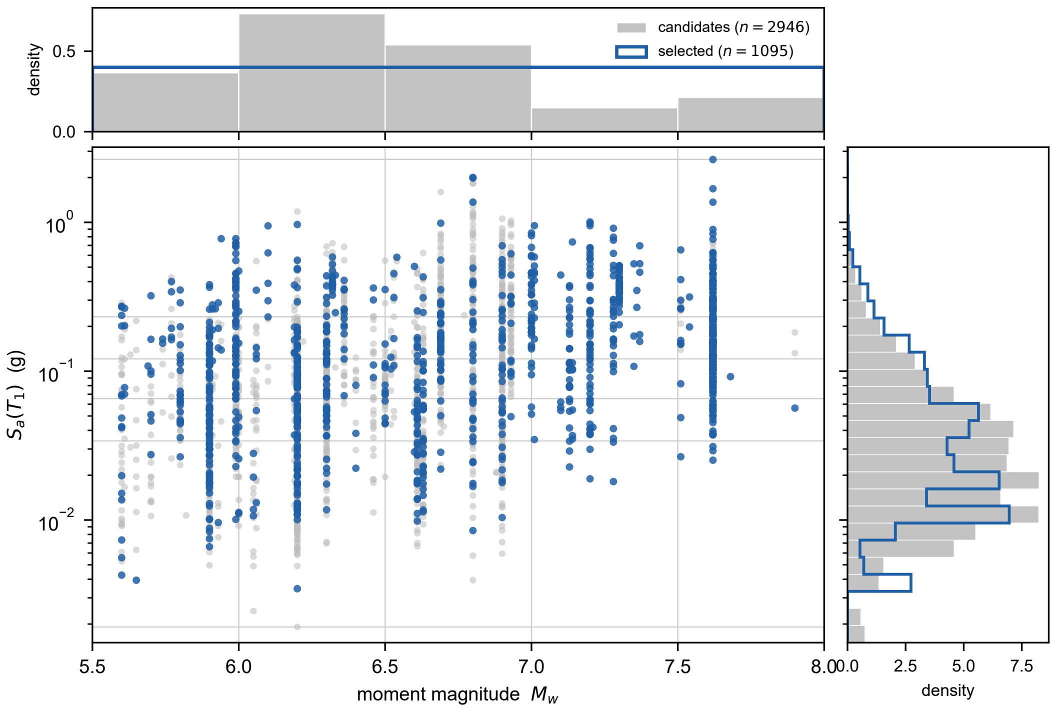
Limitations of Fourier Neural Operators for Nonlinear Structural Response Prediction under Earthquake Excitation

Supplementary Information

Kexun Li and Luis Ceferino

SUPPLEMENTARY NOTE 1: GROUND-MOTION RECORD SELECTION

We formed the 1095-record ground-motion library in two stages. In Stage 1, we applied eligibility filters to the PEER NGA-West2 records, retaining records with moment magnitude $M_w > 5.5$ and rupture distance $10 \leq R_{\mathrm{rup}} \leq 100\,\mathrm{km}$; this produced 2946 candidate records. In Stage 2, we selected a stratified subset of these 2946 candidates using five magnitude bins and five logarithmic $S_a(T_1)$ bins; we stratified on $S_a(T_1)$ because spectral acceleration at the fundamental period is a conventional intensity measure for structural response[1,2]. The smallest magnitude bin after Stage 1 contained 219 records, so we used 219 as the quota for each magnitude bin, giving $5 \times 219 = 1095$ selected records. Within each magnitude bin, the records were distributed as evenly as possible across the five $S_a(T_1)$ bins. Supplementary Figure 1 shows the Stage-1 candidate pool and the Stage-2 retained records on the $(M_w, S_a(T_1))$ plane.



SUPPLEMENTARY FIGURE 1 Two-stage selection of the ground-motion records. Stage 1 applies the magnitude and distance eligibility filters ($M_w > 5.5$ and $10 \leq R_{\mathrm{rup}} \leq 100\,\mathrm{km}$), giving 2946 candidate records (gray). Stage 2 selects the 1095 records used in the study (blue) by stratifying the candidate pool over five M_w bins and five logarithmic $S_a(T_1)$ bins. The quota is 219 records per magnitude bin, set by the smallest post-filter magnitude bin, so the final library contains $5 \times 219 = 1095$ records. The marginal histograms compare the Stage-1 candidate pool with the Stage-2 selected subset.

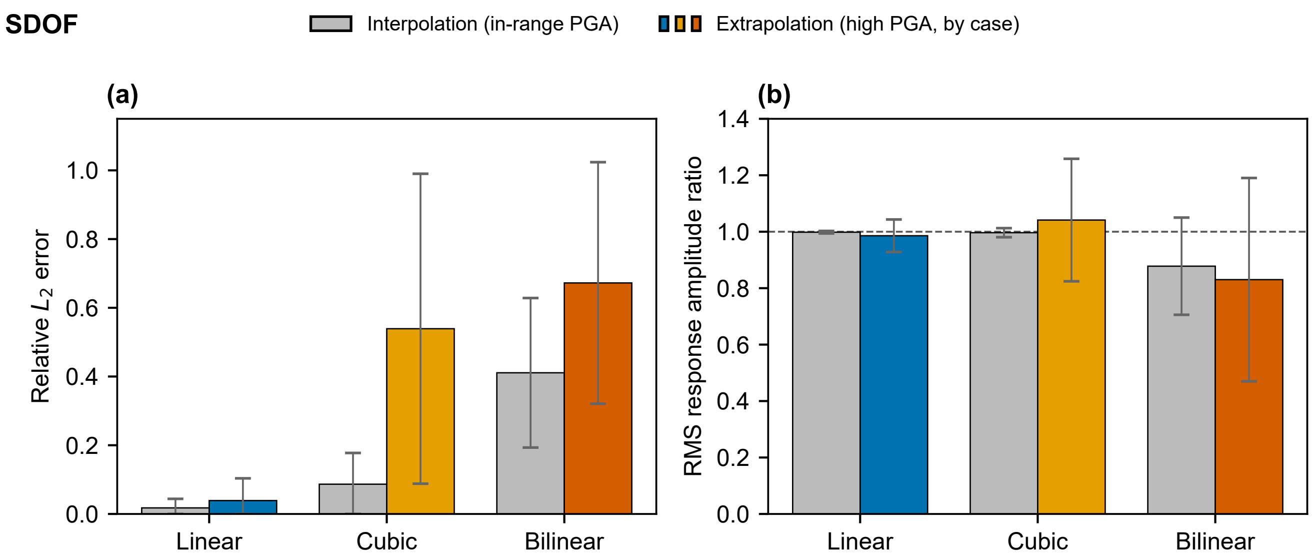
The quota of 219 records per magnitude bin was set by the limiting bin after Stage 1: the $[7.0, 7.5)$ magnitude bin contained exactly 219 candidates, while the highest bin, $[7.5, 8.0]$, contained 313, so all five magnitude bins had at least 219 candidates available after the eligibility filter. The $S_a(T_1)$ bins are equal-frequency quintiles of the candidate pool computed on a logarithmic scale, with edges at 1.91×10^{-3}, 3.40×10^{-2}, 6.54×10^{-2}, 1.21×10^{-1}, 2.32×10^{-1}, and 2.65 g. Balancing magnitude and intensity coverage in this way is consistent with ground-motion selection practice for response-history analysis[3,4]. Within each magnitude bin, the five $S_a(T_1)$ cells were filled toward equal quotas of 44 or 43 records. When a cell held fewer candidates than its quota, all of its records were taken and the shortfall was drawn from the remaining records of the same magnitude bin; no records were borrowed across magnitude bins. The three lower magnitude bins reached the balanced quotas, the bin

[7.0, 7.5) kept all 219 of its records, and the highest bin kept its sparse low-S_a cells in full and filled the remainder with higher-S_a records. Sampling was without replacement, so each of the 1095 records appears exactly once; the records come from 92 distinct earthquakes, and many records therefore share an event through different recording stations.

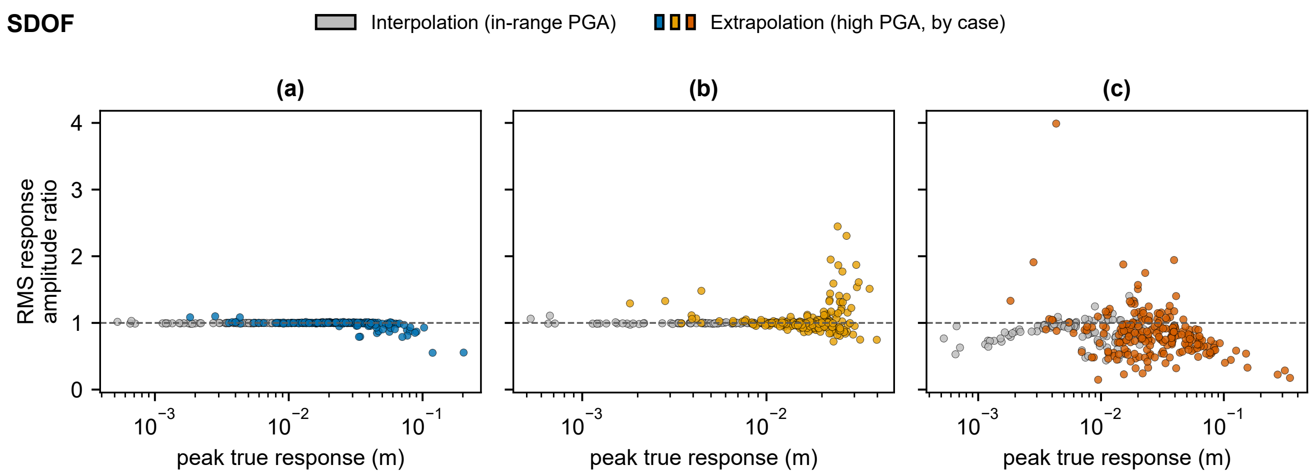
Each record enters the surrogate as a single acceleration waveform. We first rotate the two recorded horizontal components onto the horizontal principal axes, the orthogonal directions in which the component motions are uncorrelated over the record duration, and retain the major principal component, the one with the larger Arias intensity [5,6,7]. The raw histories, resampled to a common internal time step of 0.001 s during preprocessing, are low-pass filtered and decimated to a common working step of 0.02 s (50 Hz); the filter is applied before decimation to prevent aliasing, with the cutoff at the new Nyquist frequency of 25 Hz, well above all modal frequencies of the test systems. We trim each record's nearly motionless beginning and end, keeping the portion between the points where the record accumulates 0.01% and 99.99% of its total shaking energy, measured by Arias intensity, following the data-preparation procedure in [8]. Variable-length records are then zero-padded for batched training, and a binary mask identifies the valid portion of each record during normalization, training, and evaluation. Across the library, the median 5%-damped response spectrum is concentrated at short periods.

SUPPLEMENTARY NOTE 2: SDOF EXCITATION-AMPLITUDE TEST, INTERPOLATION VERSUS EXTRAPOLATION

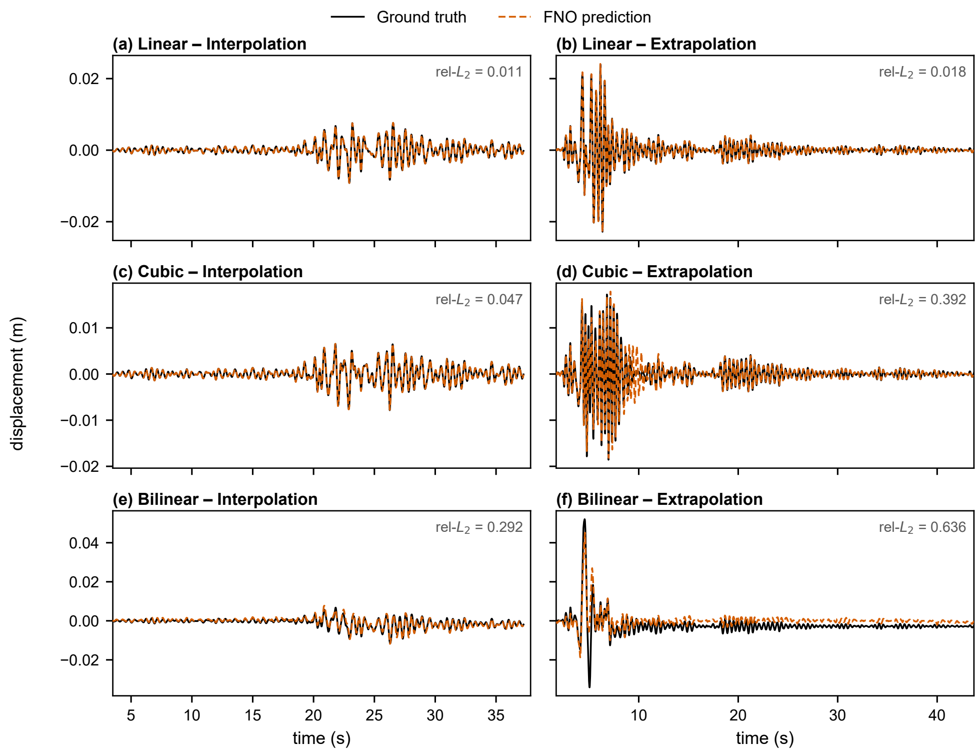
The PGA-based division of the library tests shaking stronger than the training range. One model trains on the lower-PGA records and then predicts two held-out subsets: an in-range interpolation set (n = 110) and a higher-PGA extrapolation set (n = 219). Under extrapolation, the relative L_2 error grows for every case. The RMS response amplitude ratio is the predicted RMS response amplitude divided by the true RMS response amplitude; values below one indicate underprediction of the response amplitude. This ratio stays near one for the linear and cubic cases. For the bilinear-hysteretic case, it falls from 0.88 at interpolation to 0.83 at extrapolation. Supplementary Figures 2 and 3 report the aggregate interpolation-versus-extrapolation bars and the per-record RMS-ratio scatter. Supplementary Figures 4 and 5 show the displacement histories for two further record pairs; the first pair is shown in the main text.



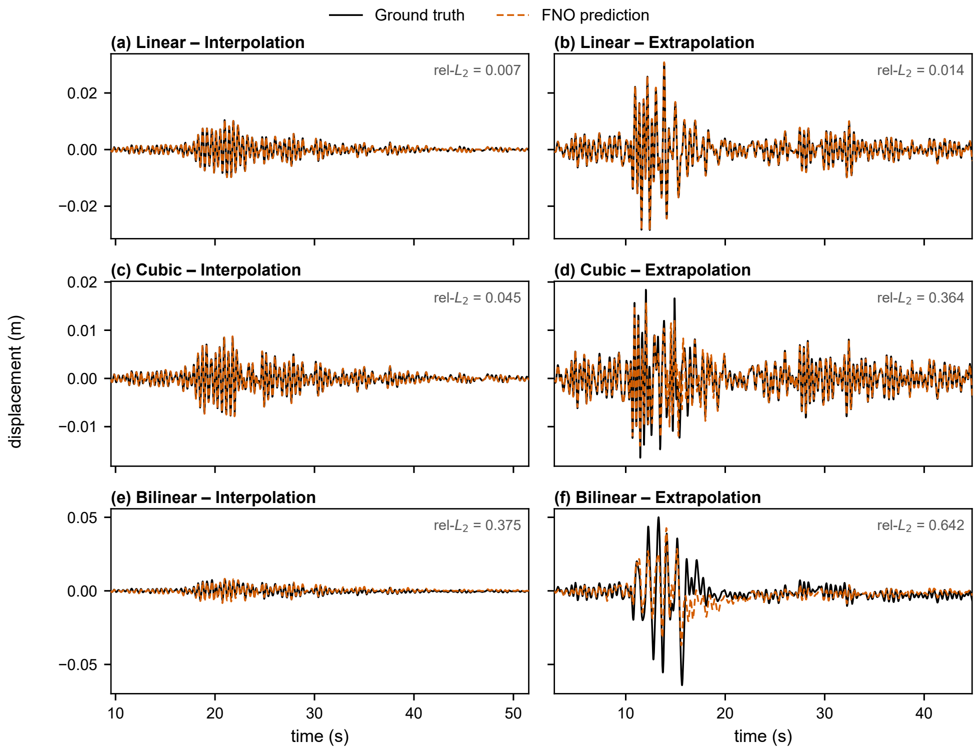
SUPPLEMENTARY FIGURE 2 SDOF excitation-amplitude test: (a) per-case relative L_2 error and (b) RMS response amplitude ratio, interpolation versus PGA extrapolation. Gray bars: interpolation (in-range PGA); colored bars: extrapolation (high PGA). Error bars show mean ± one standard deviation over the test records.



SUPPLEMENTARY FIGURE 3 SDOF per-record RMS response amplitude ratio versus true peak response, interpolation versus PGA extrapolation, for the (a) linear, (b) cubic, and (c) bilinear-hysteretic cases. For the bilinear-hysteretic case, the predicted-to-true RMS response amplitude ratio decreases as the true peak response grows.



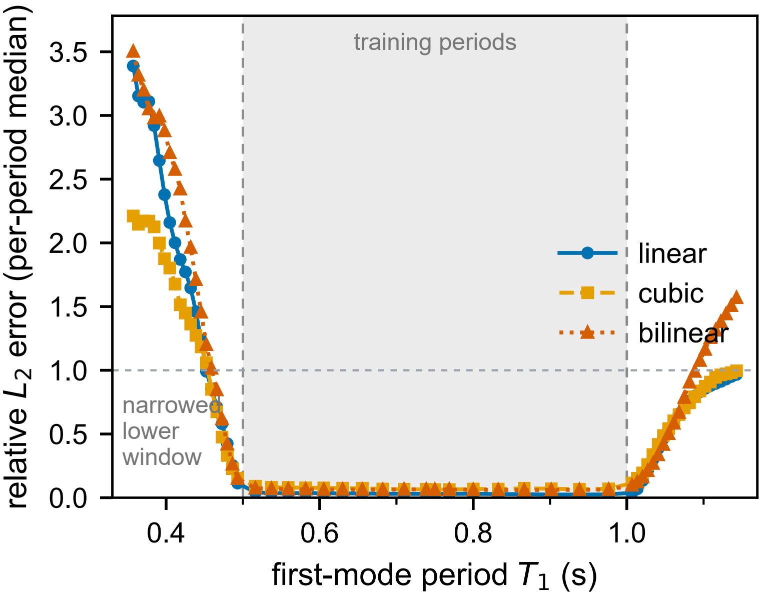
SUPPLEMENTARY FIGURE 4 SDOF displacement histories for a second record pair, comparing one interpolation record and one PGA-extrapolation record across the three restoring-force cases. The linear prediction stays close to the ground truth in both records. The cubic case tracks the interpolation record and departs further from the true response under extrapolation, and the bilinear-hysteretic case shows the strongest degradation under extrapolation, with cycle-level mismatch and a residual offset visible after the strong-motion phase.



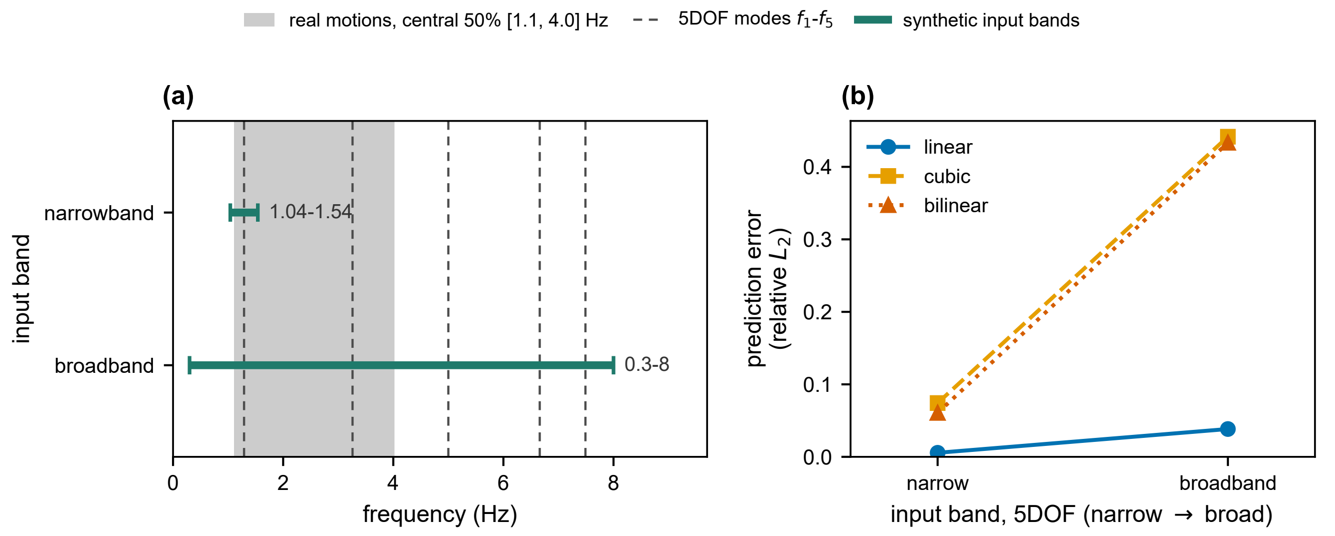
SUPPLEMENTARY FIGURE 5 SDOF displacement histories for a third record pair, comparing one interpolation record and one PGA-extrapolation record across the three restoring-force cases. The linear prediction remains close to the ground truth in both records, whereas the cubic case shows a larger cycle-level mismatch under extrapolation. The bilinear-hysteretic case has the largest visible degradation in the extrapolation record, with stronger cycle-level mismatch and a higher relative L_2 error than its interpolation counterpart.

SUPPLEMENTARY NOTE 3: 5DOF PERIOD AND INPUT-FREQUENCY TESTS

The period and input-frequency difficulties appear at 5DOF as well. The 5DOF period test applies the SDOF protocol to the first-mode period T_1, with the model trained over $[0.50, 1.00]$ s. A lower extrapolation window matching the SDOF $[0.10, 0.50)$ s window is not feasible at this scale: the fifth-mode frequency crosses the 25-Hz Nyquist limit of the 50-Hz output time grid near $T_1 = 0.23$ s. The extrapolation windows are therefore $[0.357, 0.50)$ s and $(1.00, 1.143]$ s, both about 0.14 s wide. Inside the training range, the mean relative L_2 error at held-out periods is 0.04 for the linear case, 0.09 for the cubic case, and 0.08 for the bilinear-hysteretic case. In the 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. 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 Figure 6 reports the period test, and Supplementary Figure 7 reports the input-frequency test.



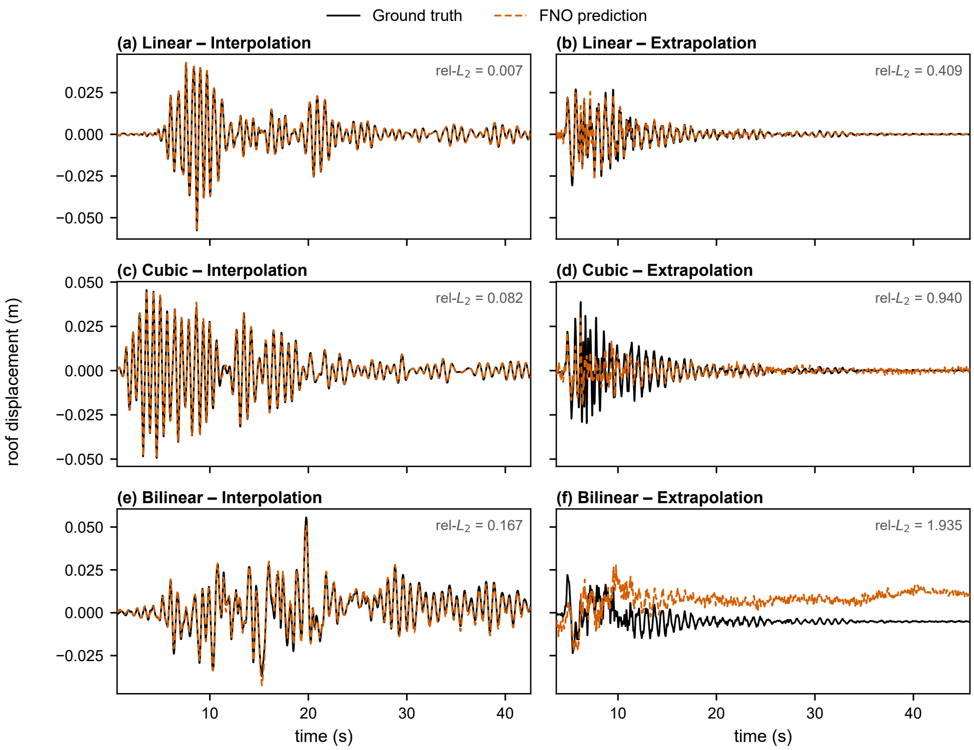
SUPPLEMENTARY FIGURE 6 5DOF structural-period test with equal-width extrapolation windows: per-period median relative L_2 error versus first-mode period T_1. The shaded band marks the training periods ($[0.50, 1.00]$ s). The lower window starts at $T_1 = 0.357$ s rather than 0.10 s because shorter T_1 pushes the higher modes beyond the Nyquist limit of the output time grid; both windows are about 0.14 s wide.



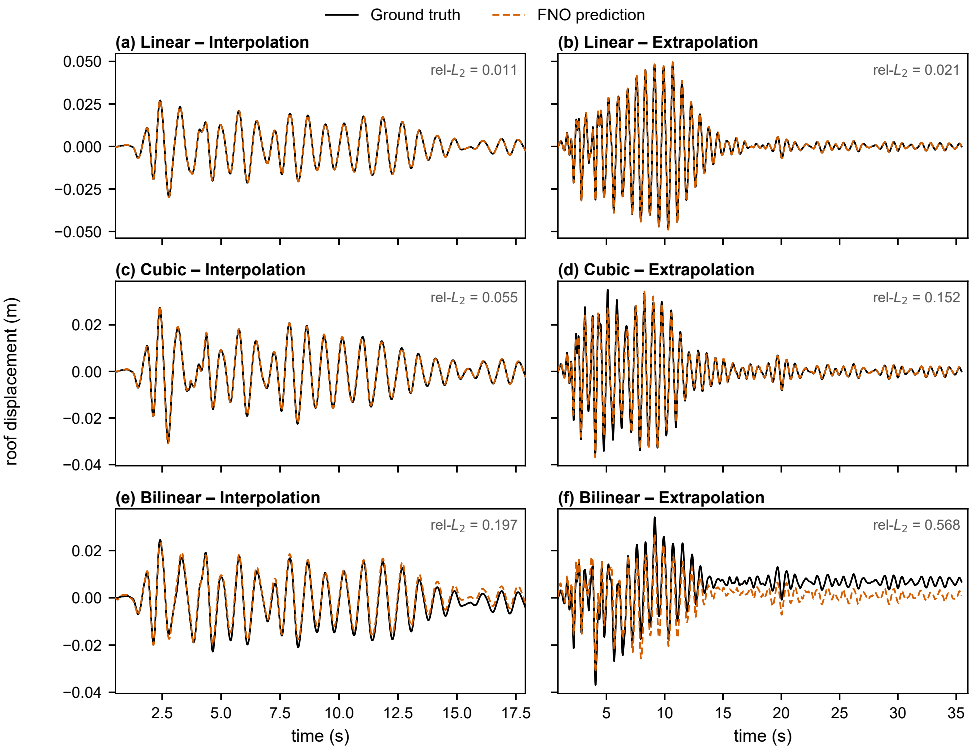
SUPPLEMENTARY FIGURE 7 5DOF input-frequency test. (a) Synthetic narrowband and broadband input bands against the central 50% band of the real records and the 5DOF modal frequencies f_1–f_5. (b) Mean relative L_2 error for the three restoring-force cases under narrowband and broadband inputs. The cubic and bilinear-hysteretic errors increase from narrowband to broadband, while the linear error remains lower.

SUPPLEMENTARY NOTE 4: 5DOF EXCITATION-AMPLITUDE TEST, INTERPOLATION VERSUS EXTRAPOLATION

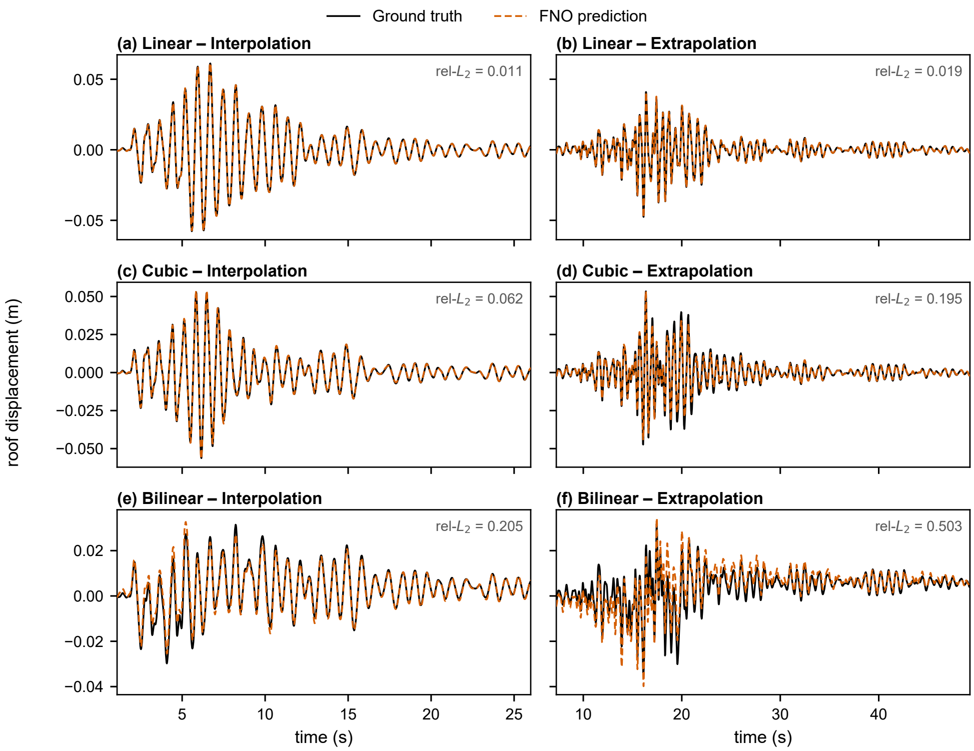
At 5DOF, amplitude extrapolation produces the most systematic underprediction observed in these tests, and it is specific to the bilinear-hysteretic case: under extrapolation the RMS response amplitude ratio is 1.00, 0.96, and 0.83 for the linear, cubic, and bilinear-hysteretic cases, respectively. The bilinear-hysteretic RMS response amplitude ratio falls from 0.98 at interpolation to 0.83 at extrapolation. This drop is the largest of the three cases, and the spread across records is the widest. Supplementary Figures 8–10 show representative roof-displacement histories; Supplementary Figures 11 and 12 report the aggregate bars and the per-record scatter.



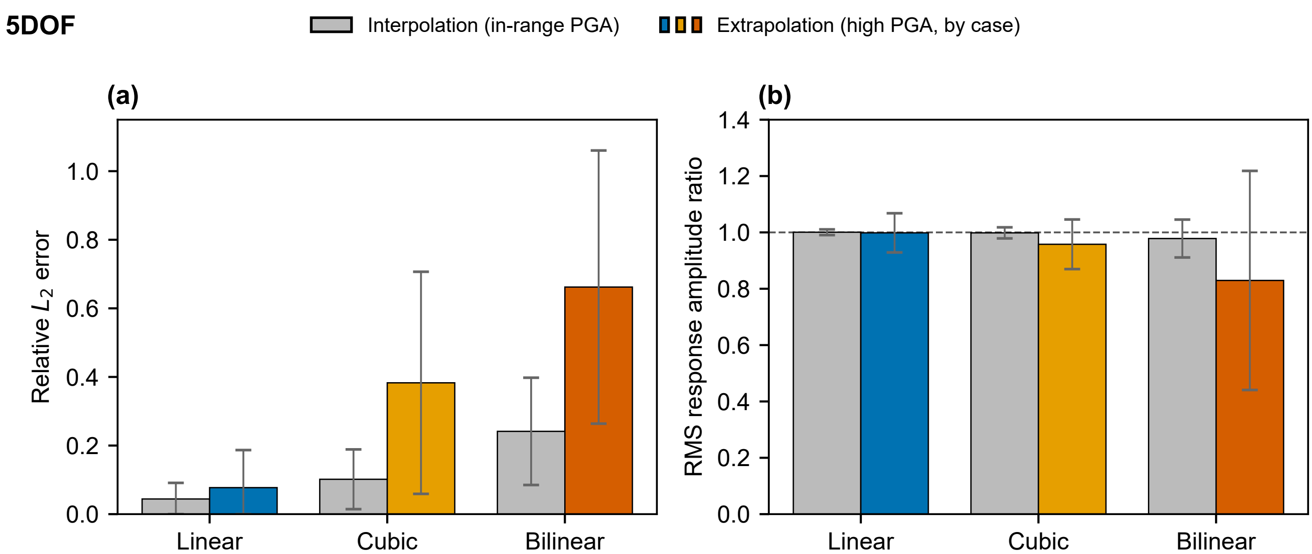
SUPPLEMENTARY FIGURE 8 5DOF roof-displacement histories for the linear, cubic, and bilinear-hysteretic cases under interpolation and PGA extrapolation, for per-case representative records. All three cases follow the ground truth on the interpolation records and degrade under extrapolation. The linear prediction shows a visible amplitude mismatch early in its extrapolation record; this record is a comparatively hard one, with a relative L_2 error well above the linear case mean. The cubic prediction departs from the true response through the strong-motion phase. The bilinear-hysteretic prediction shows the largest degradation, drifting away from the true mean level after the strong-motion phase.



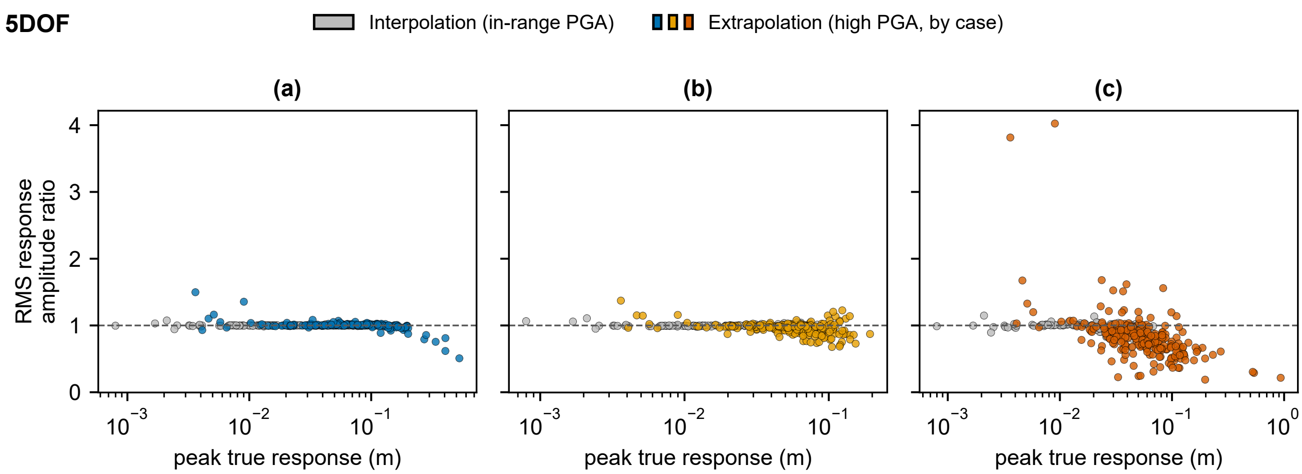
SUPPLEMENTARY FIGURE 9 5DOF roof-displacement histories for a second record pair, comparing one interpolation record and one PGA-extrapolation record across the three restoring-force cases. The linear and cubic predictions remain close to the ground truth in both records, with a moderate cycle-level mismatch for the cubic case under extrapolation, while the bilinear-hysteretic case degrades most under extrapolation, with cycle-level mismatch and an offset from the true response after the strong shaking.



SUPPLEMENTARY FIGURE 10 5DOF roof-displacement histories for a third record pair, comparing one interpolation record and one PGA-extrapolation record across the three restoring-force cases. The linear prediction stays close to the ground truth in both records, the cubic prediction shows a moderate cycle-level mismatch under extrapolation, and the bilinear-hysteretic prediction shows the largest cycle-level mismatch on the extrapolation record, with the highest annotated relative L_2 error of the three cases.



SUPPLEMENTARY FIGURE 11 5DOF excitation-amplitude test: (a) per-case relative L_2 error and (b) RMS response amplitude ratio, interpolation versus PGA extrapolation. Gray bars: interpolation (in-range PGA); colored bars: extrapolation (high PGA). Error bars show mean ± one standard deviation over the test records.



SUPPLEMENTARY FIGURE 12 5DOF per-record RMS response amplitude ratio versus true peak response, interpolation versus PGA extrapolation, for the (a) linear, (b) cubic, and (c) bilinear-hysteretic cases. For the bilinear-hysteretic case, the predicted-to-true RMS response amplitude ratio decreases as the true peak response grows.

SUPPLEMENTARY NOTE 5: ERROR DECOMPOSITION

Section 5.6 of the main text measures the residual-offset and amplitude-scale parts of each record's prediction error by recomputing the relative L_2 error ε after removing each part from both the true response x and the prediction $\hat{x}$.

Mean removal subtracts each history's time average over the unpadded portion of the record, $x_{\mathrm{dem}} = x - \bar{x}$ and $\hat{x}_{\mathrm{dem}} = \hat{x} - \bar{\hat{x}}$. Amplitude removal divides each history by its own peak absolute value, $x_{\mathrm{amp}} = x/M(x)$ and $\hat{x}_{\mathrm{amp}} = \hat{x}/M(\hat{x})$, with $M(x) = \max_t |x(t)| + \delta$, where δ is a small positive constant that avoids division by zero. When both operations are applied, amplitude removal follows mean removal.

Writing ε_{m}, ε_{p}, and $\varepsilon_{\mathrm{mp}}$ for the errors after mean removal, amplitude removal, and both operations, the percentage reductions $100\,(\varepsilon - \varepsilon_{\mathrm{m}})/\varepsilon$ and $100\,(\varepsilon - \varepsilon_{\mathrm{p}})/\varepsilon$ measure the residual-offset and amplitude-scale parts, and $100\,\varepsilon_{\mathrm{mp}}/\varepsilon$ is the remaining share. The parts need not sum to 100% because the operations overlap. The SDOF values quoted in the main text are computed from mean relative L_2 aggregates; the 5DOF residual-offset value is a median over the record-channel set.

REFERENCES

1. Tothong P, Luco N. Probabilistic seismic demand analysis using advanced ground motion intensity measures. *Earthquake Engineering & Structural Dynamics.* 2007;36(13):1837–1860.
2. Eads L, Miranda E, Lignos DG. Average spectral acceleration as an intensity measure for collapse risk assessment. *Earthquake Engineering & Structural Dynamics.* 2015;44(12):2057–2073.
3. Haselton C, Whittaker A, Hortacsu A, Baker J, Bray J, Grant D. Selecting and scaling earthquake ground motions for performing response-history analyses. In: Earthquake Engineering Research Institute Oakland, CA, USA. 2012:4207–4217.
4. Kalkan E, Chopra AK. Practical guidelines to select and scale earthquake records for nonlinear response history analysis of structures. 2010.
5. Penzien J, Watabe M. Characteristics of 3-dimensional earthquake ground motions. *Earthquake Engineering & Structural Dynamics.* 1974;3(4):365-373. doi: https://doi.org/10.1002/eqe.4290030407
6. Rezaeian S, Der Kiureghian A. Simulation of orthogonal horizontal ground motion components for specified earthquake and site characteristics. *Earthquake Engineering & Structural Dynamics.* 2012;41(2):335-353. doi: https://doi.org/10.1002/eqe.1132
7. Arias A. A measure of earthquake intensity. *Seismic design for nuclear plants.* 1970:438–483.
8. Su M, Dabaghi M, Broccardo M. Review and Validation of Stochastic Ground Motion Models: which one does it better?. arXiv preprint; 2024.