CWE-8
June 7-11, 2026
London, ON, Canada

Time-History Wind Response Analysis of High-Rise Buildings Using Latent Space Time-Step Operators

Chan Ho Kim [a], Thomas Kang [b]

[a] *Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea, tntndi001@snu.ac.kr*
[b] *Department of Architecture and Architectural Engineering & Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea, tkang@snu.ac.kr*

SUMMARY

One-shot neural operators are strong baselines for fixed-grid structural wind-response prediction, but they require the full input history and do not support causal step-by-step rollout. This paper studies LSR-AR, a causal latent-step autoregressive surrogate based on a learned residual state update conditioned on the current wind forcing. On the 37-story, 997-step along-wind benchmark of Goswami et al. (2025), LSR-AR attains 6.4% relative L^2 error (RelL2) with 236K parameters. As the primary causal comparison, LSR-AR reduces the test RelL2 from 16.8% (NARX-MLP, retrained) to 6.4%. In a separate continuation study using a time-embedding LSR-AR variant, stable beyond-horizon rollout is observed from sufficiently long prefixes.

Keywords: *Aerodynamic response, Autoregressive surrogate, Causal rollout, Temporal extrapolation, Latent-space dynamics, High-rise buildings*

1. MOTIVATION

Neural operators map wind-load histories to full response trajectories in a single forward pass, but they cannot be queried causally, step by step, as new loading data arrive. Causal rollout is relevant to applications such as active damper control, real-time monitoring, and temporal extrapolation beyond the training horizon. In wind-engineering practice, different design codes and wind-environment characteristics require time-series analyses of varying length, yet most neural-operator surrogates are tied to a fixed temporal grid and can only infer on sequences of the length seen during training. A new model must therefore be trained whenever the required duration changes, and an already-trained model cannot be reused even in the same wind environment to examine longer-duration behavior. Autoregressive surrogates, in principle, lift this restriction, but they face compounding prediction errors over hundreds of sequential steps. In addition, gradient vanishing and increased training cost for long time-series data make causal models more difficult to train than one-shot models.

This paper presents Latent-Step Residual Autoregression (LSR-AR), a latent-step autoregressive surrogate that achieves 6.4% RelL2 on a 37-story, 997-step along-wind benchmark with only 236K parameters. The manuscript reports a primary full-horizon causal benchmark together with a separate secondary continuation study.

2. RELATED WORK

2.1. One-shot neural operators

One-shot operator-learning surrogates map an observed loading history to the response trajectory in a single non-autoregressive pass. Within this class, Fourier Neural Operators (FNO) provide a widely used spectral baseline, while Deep Operator Networks (DeepONet) learns operator mappings through a branch-trunk decomposition. Recent wind-response work has also adapted this paradigm through Self-Adaptive FNO (SA-FNO) (Goswami et al., 2025). In the wind benchmark considered here, these full-input baselines require the available forcing history and therefore do not natively support causal rollout beyond it.

2.2. Autoregressive surrogates

Autoregressive models enable causal rollout but face error accumulation over long horizons. A common practical formulation is window-based autoregression, in which the next response is predicted from a finite history of past responses and exogenous inputs. This finite-window view underlies nonlinear autoregressive models with exogenous inputs (NARX) and provides a simple strictly causal baseline, but the flattened input dimension grows with both window length and response dimension. The NARX-MLP baseline considered here follows this formulation through a windowed multilayer perceptron (MLP) adaptation inspired by Thedy et al. (2025). By contrast, the present latent residual rollout reuses a shared latent update map at every step, maintaining a fixed latent-state dimension during long-horizon rollout.

3. METHODOLOGY

3.1. Dataset

The study uses the wind force–response benchmark of Goswami et al. (2025). The target structure is a 37-story steel frame, and the wind loading is synthesized from the Tokyo Polytechnic University (TPU) aerodynamic database (Tamura, 2012). Displacement responses are obtained by nonlinear time-history analysis in OpenSees; full structural and simulation details are given in Goswami et al. (2025). Table 1 summarizes the benchmark configuration used throughout the paper. Figure 1 shows the target structure and a representative forcing–response pair from the dataset. Figure 1 (a) depicts the 37-story steel frame, and Figure 1 (b) shows selected-floor forcing and along-wind displacement time histories from a representative test scenario.

Table 1: Benchmark summary of the 37-story along-wind response dataset

Item	Value
Structure	37-story steel frame; 150 m height; 60 m × 30 m plan
Wind condition	Along-wind loading; $\alpha = 90°$; reference wind speed $V_{\mathrm{H}} = 65$ m/s
Load synthesis	Spectral proper orthogonal decomposition (POD) from TPU aerodynamic pressure data
Time discretization	Raw: 600 s at $\Delta t_{\mathrm{raw}} = 0.02$ s ; benchmark: crop first 100 steps and stride by 30, giving effective $\Delta t = 0.6$ s and $T = 997$
Data split	720 train / 80 validation / 200 test
Input / output	37 floor-level wind forcing channels → 37 along-wind displacement channels

The task is to learn a causal input-output map from trajectory data alone:

$$\mathcal{D} = \left\{ \left(u_{0:T}^{(j)}, x_{0:T}^{(j)} \right) \right\}_{j=1}^{N} \tag{1}$$

where $u_k \in \mathrm{R}^{37}$ is the floor-level wind forcing at discrete time k and $x_k \in \mathrm{R}^{37}$ is the along-wind displacement response

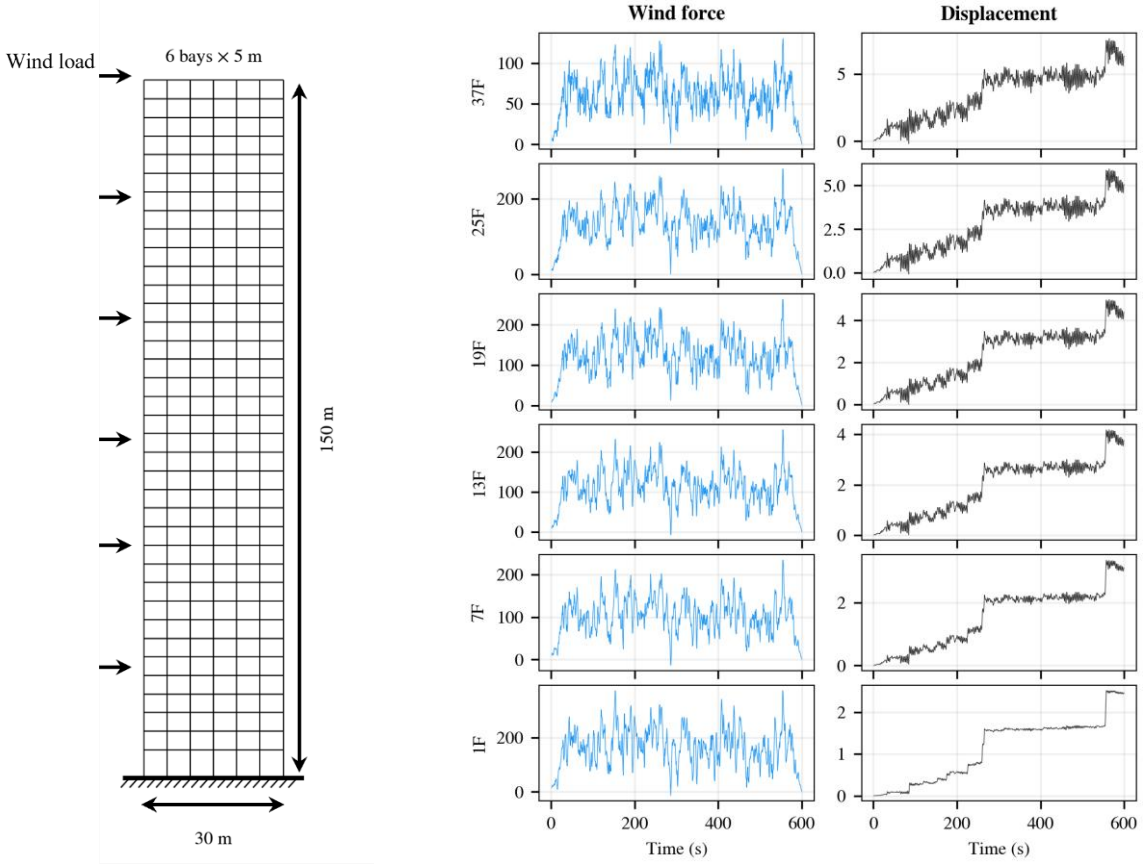

Figure 1: (a) Target structure: 37-story steel frame (150 m, 6 bays). (b) Representative test scenario: wind-force (kN) (left) and along-wind displacement (m) (right) time histories for selected floors.

The raw simulation output spans 600 s at $\Delta t_{\mathrm{raw}} = 0.02$ s (30,000 steps). For the benchmark, the first 100 steps are discarded and every 30th sample is retained, yielding an effective $\Delta t = 0.6$ s and a sequence length of $T = 997$ steps per scenario. This matches the original dataset setting of Goswami et al. (2025) and is retained for valid comparison. The benchmark split is 720 training / 80 validation / 200 test. The 200-scenario test split matches the published test cases, while the 80-scenario validation split is held out from the remaining training cases. Checkpoint selection uses validation metrics only; the test split is reserved for final reporting.

3.2. Experimental setup

3.2.1. Baseline models

Three comparison architectures are considered:

- **FNO** : one-shot Fourier Neural Operator (7.40M params).
- **SA-FNO** : Self-Adaptive FNO with adversarial frequency adjustment.
- **NARX-MLP**: adapted windowed NARX-MLP autoregressive baseline, inspired by the autoregressive formulation of Thedy et al. (2025). At each step, the next response is predicted from a sliding window of past responses ($n_{\mathrm{wr}} = 50$) and forcing ($n_{\mathrm{wg}} = 5$).

FNO and SA-FNO are one-shot operators; NARX-MLP is the only causal baseline in the present comparison. FNO and SA-FNO were evaluated using the original trained checkpoints published by Goswami et al. (2025) on the identical 200-scenario test set; no retraining was performed. Their reference metrics were recomputed from the released predictions using the definitions in this section. For roof-response diagnostics only, we additionally compare against the DeepONet and DeepFNONet outputs provided with the Goswami dataset release. NARX-MLP was retrained on the same split. At rollout start, unavailable response and forcing history in NARX-MLP are zero-padded, while models that require an observed initial state are conditioned on the first displacement sample x_0 of each scenario.

3.2.2. Evaluation metrics

All headline RelL2 values are reported as the aggregate/global relative L2 error over the full evaluated set:

$$\mathrm{RelL2} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\hat{x}^{(i)} - x^{(i)}\right\|_2}{\left\|x^{(i)}\right\|_2} \tag{2}$$

where, for each sample, the norm spans all time steps and output channels.

Pearson correlation is computed over the flattened prediction–target vector across all evaluated samples, time steps, and floors:

$$\mathrm{Corr} = \frac{\sum_{m=1}^{M}(y_m - \bar{y})(\hat{y}_m - \bar{\hat{y}})}{\sqrt{\sum_{m=1}^{M}(y_m - \bar{y})^2}\sqrt{\sum_{m=1}^{M}(\hat{y}_m - \bar{\hat{y}})^2}} \tag{3}$$

where m indexes the flattened sample–time–floor entries. Peak displacement error is the relative error of per-sample, per-floor peak absolute displacement, averaged over all samples and floors:

$$\mathrm{PeakErr} = \frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\frac{\left|p_c^{(i)} - \hat{p}_c^{(i)}\right|}{p_c^{(i)} + \varepsilon}, \qquad p_c^{(i)} = \max_t\left|x_{t,c}^{(i)}\right|, \qquad \hat{p}_c^{(i)} = \max_t\left|\hat{x}_{t,c}^{(i)}\right| \tag{4}$$

where C is the number of output channels and ε is a small constant. In addition, *Shape-Based Distance* (SBD), suggested by Paparrizos and Gravano (2015), is used here as a coarse global

diagnostic and computed over the flattened and z-normalized signal across the full evaluated set. SBD measures waveform similarity through the normalized cross-correlation after z-normalization:

$$\mathrm{SBD}(x,\hat{x}) = 1 - \max_s \mathrm{NCC}_s\left(\tilde{x},\tilde{\hat{x}}\right) \tag{5}$$

where x denotes the z-normalized signal and s is the lag index. SBD is invariant to vertical offset, amplitude scaling, and temporal shift; it ranges from 0 (identical shape) to 2 (maximally dissimilar). While RelL2 penalizes all pointwise deviations equally, SBD isolates shape fidelity from amplitude and phase errors, providing a complementary diagnostic of waveform quality.

3.2.3. Benchmark tasks

Causal models are evaluated on two tasks:

- **Full-horizon prediction.** The model is trained and evaluated on the complete 997-step sequence. This is the primary comparison task.
- **Temporal extrapolation.** The model is trained on a truncated prefix ($L < 997$steps) and rolled out over the full 997-step test sequence. The in-distribution region ($k \le L$) and extrapolation region ($k > L$) are scored separately, testing whether the learned dynamics generalize beyond the training horizon.

One-shot operators are evaluated on the full-horizon task only, as they cannot extrapolate beyond their fixed input length by construction.

3.2.4. Proposed model: LSR-AR

LSR-AR is a latent-step autoregressive surrogate that predicts the next-step structural response via autoregression in a learned latent space (Figure 2).

The dynamic response of a structure may not be fully observable from displacements alone-unobserved internal states such as velocities, internal forces, and damping forces govern the dynamics. LSR-AR addresses this by extracting a latent vector $z_0 \in \mathrm{R}^{d_z}$ from the observed displacement x_0 through an encoder. The latent vector is intended to capture not only the observable displacement but also supplementary information learned from data, serving as a surrogate for the unobserved states.

In this latent space, a latent increment field $\Delta_\theta(z_k, u_k)$ conditioned on the current forcing u_k is learned as an MLP. The increment field can be interpreted as a learned state-update field over the latent space, and the state update itself takes the form of a forward Euler step:

$$z_{k+1} = z_k + \Delta_\theta(z_k, u_k) \tag{6}$$

A decoder recovers the physical-space displacement $\hat{x}_{k+1}$ from the updated latent state z_{k+1}. No explicit time embedding is used; temporal information enters only through the forcing sequence.

The model is trained end-to-end on full-trajectory rollout loss with truncated backpropagation through time (TBPTT) (Williams and Peng, 1990). The training hyperparameters of the final selected models are listed in Table 2. Figure 2 summarizes this encoder–increment–decoder structure.

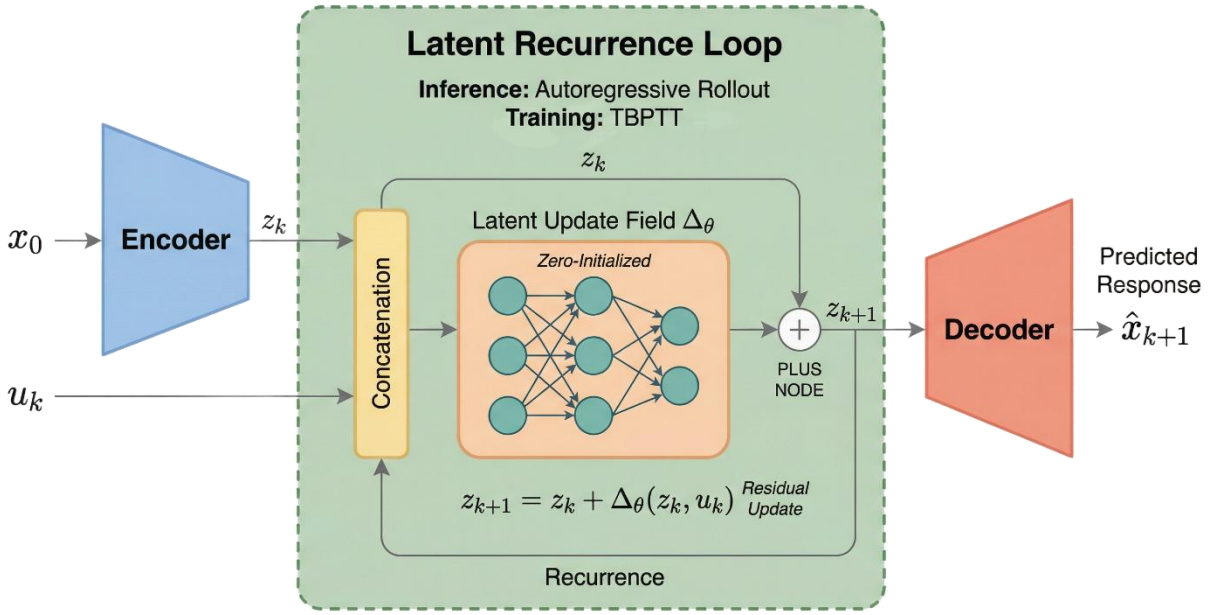

Figure 2: LSR-AR architecture: an autoregressive rollout composed of an encoder, latent increment field, and decoder.

Table 2: Hyperparameter summary for the compared evaluation setups.

	FNO	SA-FNO	NARX-MLP	LSR-AR
Parameters	7.40M	7.40M	111K	236K
Optimizer	Adam	Adam	AdamW	AdamW
Epochs	500	500	300	200
Batch size	200	200	50	32
TBPTT length	-	-	200	200

4. RESULTS

4.1. Full-horizon inference comparison

Table 3 compares the main LSR-AR configuration without explicit time embedding against the baselines. Within the retrained causal comparison, this configuration achieves 6.4% test RelL2, compared with 16.8% for NARX-MLP. Only task-aligned 37-floor comparisons are included in Table 3. Published one-shot checkpoints from Goswami et al. (2025) are shown only as contextual references; they yield 15.2% for FNO and 18.1% for SA-FNO on all 37 floors. Top-floor-only references such as DeepFNONet and DeepONet are omitted because they are not task-aligned 37-floor comparisons. The SBD column in Table 3 shows a lower value for LSR-AR (SBD = 0.006) than for the compared baselines (0.058 - 0.080). Because SBD is invariant to amplitude and temporal shift, this is consistent with better overall waveform-shape agreement under the present aggregation. Figure 3 visualizes the same all-floor comparison across RelL2, correlation, and peak displacement error.

Table 3: Main model comparison on the full 997-step along-wind benchmark using RelL2, PeakErr, Corr., and SBD

Model	Type	Params	RelL2 (%)	Peak err	Corr.	SBD
LSR-AR (no time embedding)	Causal	236K	6.4	4.2%	0.994	0.006
NARX-MLP	Causal	111K	16.8	12.0%	0.958	0.058
FNO†	One-shot	7.40M	15.2	9.6%	0.983	0.063
SA-FNO†	One-shot	7.40M	18.1	10.7%	0.977	0.080

†Published checkpoints from Goswami et al. (2025) were used.

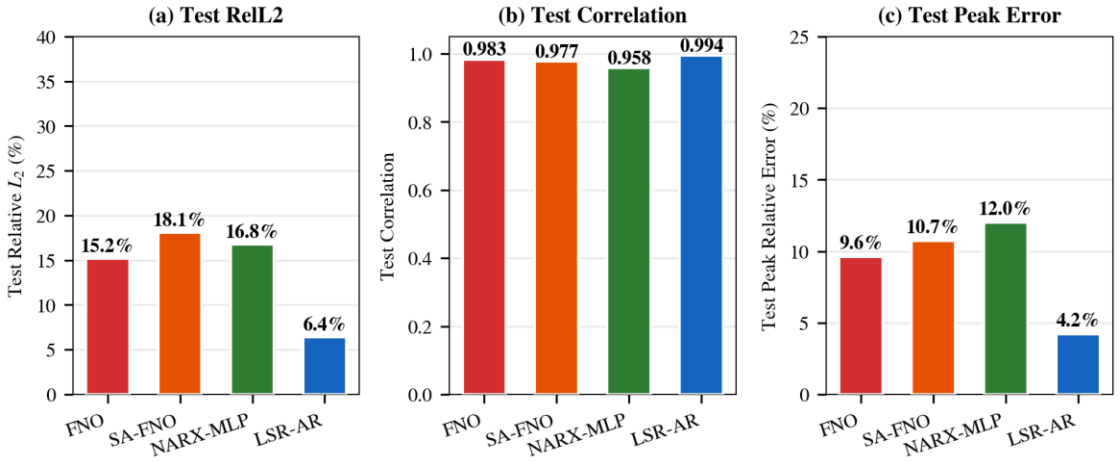

Figure 3: Model-wise comparison on 200 unseen test scenarios: (a) RelL2, (b) correlation, and (c) peak displacement error

4.2. Ablation study

Table 4 ablates the key design choices and reports test aggregate RelL2 for the selected variants. Here d_z denotes the latent dimension and d_t the time-embedding dimension.

Table 4: Selected LSR-AR ablations on latent dimension and time embedding, reported with aggregate test RelL2

Variant	Change	Params	Epochs	Best RelL2
Default ($d_t = 0$)	No time embed	236K	200	6.4%
Time embedding:				
$d_t = 32$	With time embed	244K	200	7.3%
Latent dimension (200 epochs):				
$d_z = 4$	Bottleneck	223K	200	8.3%
$d_z = 8$	Reduced	225K	200	7.6%

Table 4 summarizes the completed ablations selected by validation performance; the test metrics are shown only for final reporting. Among these variants, the default configuration without explicit time embedding yields the lowest RelL2 on this benchmark.

4.3. Peak displacement and PSD analysis

For this roof-level diagnostic, the released roof-only DeepONet and DeepFNONet references are also included. Figure 4 (a) compares empirical cumulative distribution functions (CDFs) of peak roof displacement across 200 test scenarios. LSR-AR most closely follows the ground-truth

distribution among the compared models. Figure 4 (b) shows the averaged power spectral density (PSD) at the roof. Most models, including LSR-AR, reproduce the dominant first-mode resonant peak with good fidelity. However, all models share a common artifact: a spurious secondary peak at higher frequencies that is absent or much weaker in the ground truth. One possible explanation is that the compared surrogates fit the dominant low-frequency content more readily, while high-frequency response contributes less to the overall training loss and may therefore be learned less faithfully. This behavior may be especially problematic for resonance-sensitive responses, and it should be considered explicitly in wind-response studies where high-frequency behavior is important.

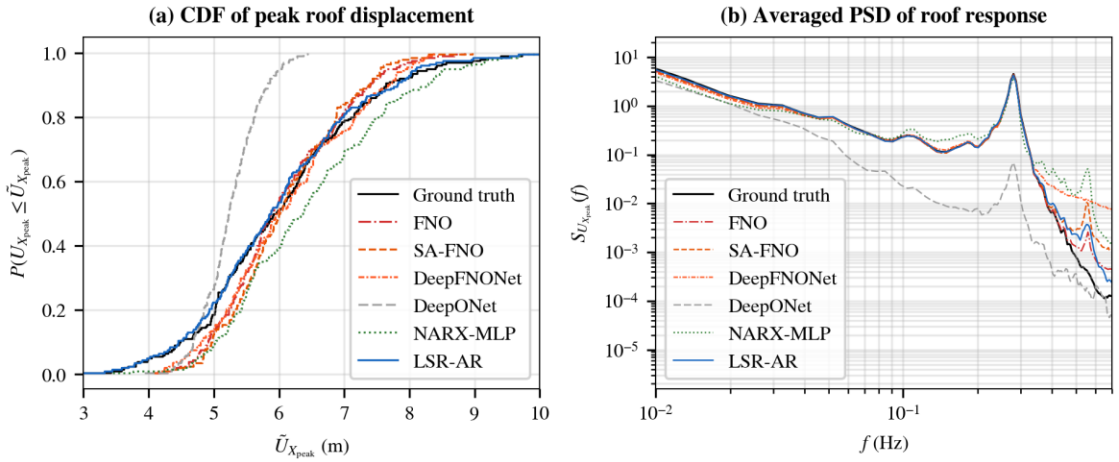

Figure 4: Roof-level diagnostics: (a) empirical CDF of peak displacement and (b) averaged PSD of the response

4.4. Separate temporal extrapolation study

To examine continuation beyond the seen training horizon, we conducted a separate rollout study using a time-embedding LSR-AR variant and NARX-MLP. Each model was trained on truncated prefixes with $L = 200$ and $L = 400$, then evaluated by rolling out the full 997-step test sequence on unseen scenarios. No additional hyperparameter tuning was performed for the $L = 200$ and $L = 400$ runs; each cropped-length experiment reused the corresponding architecture and optimization settings from the full-length setup. Unlike Table 3, Table 5 reports mean per-sample RelL2 in order to summarize scenario-wise continuation behavior, including divergent cases. For that reason, the full-train reference values in Table 5 are not directly numerically identical to the aggregate/global RelL2 values reported in Table 3.

Under this setup, both models diverged at $L = 200$, whereas both remained stable at $L = 400$. The 400-step NARX-MLP run achieved 9.9% full-rollout RelL2, lower than its own full-train result (15.2%) under the same mean per-sample summary. We do not attempt to interpret that difference further here. The time-embedding LSR-AR variant also remained stable at $L = 400$, but its full-rollout error increased to 14.7%; its in-distribution error (4.7%) nevertheless stayed below its own full-train reference (6.7%). Figure 5 shows a representative roof-floor rollout for the $L = 200$ and $L = 400$ settings.

Table 5: Temporal extrapolation results for the separate two-model continuation study, reported as mean per-sample RelL2 over the seen, extrapolation, and full regions.

Model	Train length	Seen mean per-sample RelL2	Extrapolation mean per-sample RelL2	Full mean per-sample RelL2
LSR-AR	200	3.8 %	diverge	diverge
LSR-AR	400	4.7 %	15.2 %	14.7 %
LSR-AR	997 (full)	-	-	**6.7 %**
NARX-MLP	200	7.5 %	diverge	diverge
NARX-MLP	400	8.8 %	9.9 %	**9.9 %**
NARX-MLP	997 (full)	-	-	15.2 %

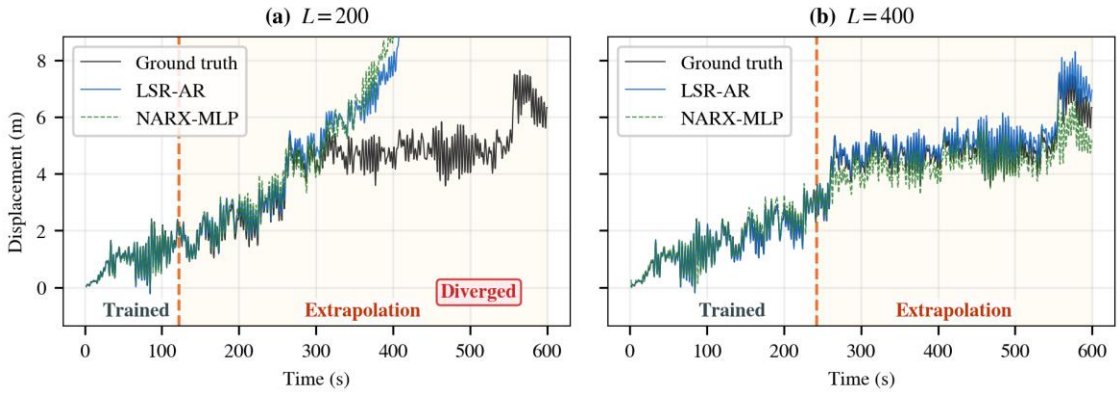

Figure 5: Representative roof-floor temporal extrapolation trajectories for (a) $L = 200$ and (b) $L = 400$

5. DISCUSSION AND CONCLUSION

This paper presented LSR-AR, a latent-step autoregressive surrogate for along-wind response prediction of a 37-story high-rise building. As summarized in Table 3 and Figure 3, the main LSR-AR configuration without explicit time embedding achieves 6.4% test RelL2 with 236K parameters. Within the full-horizon comparison of Table 3, this improves over NARX-MLP. The selected ablations in Table 4 further suggest that moderate latent compression and a simple forward-Euler residual update remain viable on this benchmark, while leaving room for higher-order integrators and larger-system studies.

A secondary continuation study of the separate time-embedding variant, together with Figure 5, showed stable beyond-horizon rollout at $L = 400$, whereas the $L = 200$ setting diverged. At the same time, the divergence observed under shorter training horizons indicates that the model cannot yet be interpreted as fully capturing the governing physics. Likewise, Figure 4 shows common high-frequency spectral artifacts across models, motivating follow-up work on cross-wind and torsional response, larger-system extrapolation, sampling sensitivity, and stronger physical constraints or regularization.

ACKNOWLEDGEMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)].

REFERENCES

Goswami, S., Giovanis, D.G., Li, B., Spence, S.M.J., Shields, M.D., 2025. Neural operators for stochastic modeling of nonlinear structural system response to natural hazards. Eng. Struct. 345, 121284. https://doi.org/10.1016/j.engstruct.2025.121284

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A., 2021. Fourier neural operator for parametric partial differential equations, in: International Conference on Learning Representations (ICLR 2021).

Lu, L., Jin, P., Pang, G., Zhang, Z., Karniadakis, G.E., 2021. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. Nat. Mach. Intell. 3, 218–229. https://doi.org/10.1038/s42256-021-00302-5

Paparrizos, J., Gravano, L., 2015. k-Shape: Efficient and accurate clustering of time series, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1855–1870. https://doi.org/10.1145/2723372.2737793

Tamura, Y., 2012. Aerodynamic database for high-rise buildings. Tokyo Polytechnic University, Global Center of Excellence Program. http://www.wind.arch.t-kougei.ac.jp/system/eng/contents/code/tpu

Thedy, J., Liao, K.W., Kim, T., 2025. Auto regressive neural network-driven reliability optimization in base-isolated building design. Results Eng. 26, 104713. https://doi.org/10.1016/j.rineng.2025.104713

Williams, R.J., Peng, J., 1990. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. Neural Comput. 2, 490–501. https://doi.org/10.1162/neco.1990.2.4.490