DO GRAPH-COMPLETION FRONT ENDS HELP MASK-AWARE FORECASTING? A STUDY OF STRUCTURED SENSOR OUTAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Does graph completion help when a forecaster already represents missingness? We study RESOLVE, a two-step clamped completion front end, against four controls using shared structured input outages and unchanged future targets. Across Beijing Air, METR-LA, and PEMS-BAY, the primary matrix contains 45 checkpoints and 720 evaluation units. RESOLVE achieves lower mean absolute error in 17 of 24 designated sensor- and mixed-outage comparisons, but only two of six against a mask-aware diffusion-recurrent model. An equal-parameter validation ablation likewise finds small, uneven gains. Native-input estimates meet a 2% point margin against GraphGRU, whereas risk-conditioned intervals satisfy a descriptive coverage–width criterion in only 10 of 45 cells. Separate three-seed weather and feeder-energy studies yield contrasting rankings: weather favors diffusion and feature propagation, while RESOLVE has the lowest mixed-outage energy mean error. The small energy advantage over the diffusion comparator is not stable to paired-seed resampling. Prior test-aggregate exposure and unavailable primary per-origin evidence limit the primary findings to descriptive comparisons; supplementary bootstrap intervals are sensitivity checks. Overall, the evidence supports neither a universal completion advantage nor an isolated causal benefit, and highlights the need for matched architectural controls and evaluation beyond network-wide average error.

1 INTRODUCTION

Sensor histories in environmental and traffic systems are not always missing independently. Communication interruptions produce temporal blocks, sensors can disappear for extended periods, and failure modes overlap. These events remove both temporal and graph context. A method that appears robust on a whole-network average may still fail where sensing is interrupted, while a method that helps a small failed subset may worsen the larger unaffected region. Structured outages are therefore a distinct evaluation problem rather than merely a higher rate of entrywise missingness.

Graph completion is a natural response: known values are clamped while neighboring information is propagated into missing entries. Yet graph-recurrent forecasters already diffuse information internally. A front end may therefore restore useful context, duplicate learned computation, or introduce smoothing bias. We study three questions: (RQ1) Does completion improve paired forecasting error under sensor and mixed outages? (RQ2) Does it preserve clean-input performance within a prespecified 2% margin? (RQ3) Do changes localize to failed nodes, and are interval-coverage gains obtained only by widening uncertainty bands?

Our contribution is an empirical assessment, not a new completion operator. First, the primary matrix uses timeline-level failures, immutable targets, and shared masks across three datasets, five methods, and four conditions. Second, localization and a coverage–width diagnostic expose behavior hidden by network-wide averages. Third, the apparent benefit depends on the control: RESOLVE has lower mean MAE than GraphGRU in five of six designated settings, but than the diffusion-recurrent comparator in only two. A separate equal-parameter validation ablation provides a tighter architectural control. Supplementary weather and energy studies add contrasting domain evidence without enlarging the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

primary comparison family. Together, these analyses distinguish model rankings from an isolated effect of completion.

2 Related Work

Graph-based forecasters combine spatial relationships with temporal dynamics. DCRNN uses directed diffusion within a recurrent architecture (Li et al., 2018); STGCN uses spatiotemporal graph-convolutional blocks (Yu et al., 2018); and Graph WaveNet learns adaptive dependencies (Wu et al., 2019). GRU-D models missingness explicitly (Che et al., 2018), while BRITS, GRIN, and PriSTI reconstruct missing values using temporal or relational structure (Cao et al., 2018; Cini et al., 2022; Liu et al., 2023). More directly, HD-TTS forecasts with missing inputs through hierarchical spatiotemporal downsampling (Marisca et al., 2024). Thus, forecasting under missingness is an established problem; our study addresses a narrower comparison of fixed completion channels and mask-aware recurrent models.

Feature Propagation repeatedly propagates graph features while resetting observed values (Rossi et al., 2022). We adapt this operation to a forecasting input channel and compare it with a shallow damped variant. The operator itself is established; the empirical question is whether its benefit persists across matched outage settings and alternative recurrent mechanisms. We document prior aggregate exposure and incomplete inferential evidence in line with reproducibility guidance (Henderson et al., 2018; Pineau et al., 2021). The tested matrix does not include HD-TTS, GRIN, or PriSTI, and does not establish a ranking against current specialized missing-data methods.

3 Problem and Evaluation Protocol

Let $X \in \mathbb{R}^{H\times N\times F}$ be an input history and $M \in \{0,1\}^{H\times N\times F}$ its feature-level availability mask. Here H, N, F, and P denote history length, nodes, features, and forecast horizon. Each method constructs its representation from $X \odot M$ and M and predicts $\hat{Y} \in \mathbb{R}^{P\times N}$. Mask encoding is model-specific (Section 4). A separate future mask M_Y determines eligible targets. Artificial input corruption changes M, never Y or M_Y.

Failures are generated once on each split's original timeline before overlapping windows are formed. We evaluate native missingness only (`none`), contiguous temporal removals (`block`), whole-sensor removal (`sensor`), and a frozen mixture (`mixed`). Masks are shared between paired model evaluations.

A node is *input-failed* for an origin if at least one natively observed input entry was artificially hidden; all remaining nodes are *input-unaffected*. These labels localize future forecast error and are not imputation metrics. Overall MAE is reconstructed from additive error sums and target counts:

$$\mathrm{MAE}_{\mathrm{all}} = \frac{S_{\mathrm{failed}} + S_{\mathrm{unaffected}}}{n_{\mathrm{failed}} + n_{\mathrm{unaffected}}}. \tag{1}$$

Figure 1 summarizes the leakage boundary and the three representation paths used in the controlled comparisons. The same timeline-level corruption precedes every path, and evaluation always uses the unchanged future target and eligibility mask.

4 Methods

All methods use the same corrupted observations, horizon, and target mask. GraphGRU and SAFE-ST are project-defined controls; FP+GRU is a forecasting adaptation of Feature Propagation, and MASK-DCRNN is a DCRNN-style adaptation. These labels do not denote official benchmark reproductions. The primary RESOLVE model belongs to the shared GRU family below. A separate equal-parameter validation experiment adds completion to MASK-DCRNN and is reported independently.

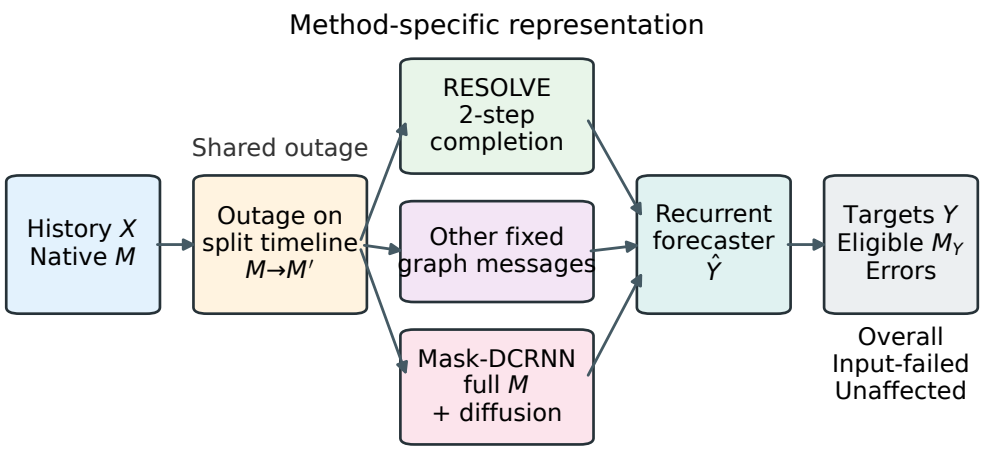

Figure 1: Evaluation pipeline. Structured failures are generated once on the original input timeline and shared across methods. RESOLVE applies two-step clamped completion; common-family baselines use alternative fixed graph messages; and MASK-DCRNN receives the full mask inside its diffusion-recurrent cell. Artificial corruption never modifies future targets or their eligibility mask.

4.1 SHARED ENCODER–GRU–HEAD FAMILY

Let $V = X \odot M$ and let $a_{btn} = F^{-1} \sum_f M_{btnf}$ be node-level availability. GraphGRU, SAFE-ST, FP+GRU, and RESOLVE share an encoder

$$\begin{aligned} q_{btn} &= [V_{btn}, G_{btn}, a_{btn}], \\ e_{btn} &= \mathrm{ReLU}(W_e q_{btn} + b_e), \end{aligned} \tag{2}$$

a node-wise PyTorch GRU of width 48, LayerNorm, and a dataset-specific direct multi-horizon head; b, t, n index batch, time, and node. Only the fixed graph-message channel G differs. The encoder receives scalar availability a, not the full feature mask; completion still uses M before encoding. The graph A is nonnegative, row-normalized, and non-trainable. Beijing Air uses nonnegative target correlations fitted on training data only; traffic datasets use frozen distance graphs.

GraphGRU takes $G = AV$. SAFE-ST renormalizes incoming mass using source availability,

$$\begin{aligned} \widetilde{A}_{btij} &= \frac{A_{ij} a_{btj}}{\max(\sum_\ell A_{i\ell} a_{bt\ell}, 10^{-6})}, \\ G_{bti:} &= \sum_j \widetilde{A}_{btij} V_{btj:}. \end{aligned} \tag{3}$$

If no available source mass remains, its message is zero. FP+GRU initializes $Z^{(0)} = V$, performs 40 parameter-free clamped steps $Z^{(k+1)} = V + (1 - M) \odot AZ^{(k)}$, and takes $G = Z^{(40)}$. "Same GRU" denotes identical architecture and parameterization, not shared fitted weights.

RESOLVE performs two damped clamped updates

$$\begin{aligned} Z^{(0)} &= V, \\ Z^{(k+1)} &= V + 0.8(1 - M) \odot AZ^{(k)}, \quad k = 0, 1, \end{aligned} \tag{4}$$

and then uses $G = AZ^{(2)}$. Depth and damping are fixed at $L = 2$ and $\beta = 0.8$ for all three primary datasets.

4.2 MASK-AWARE DIFFUSION RECURRENT COMPARATOR

MASK-DCRNN is an in-project DCRNN-style adaptation rather than an official reproduction. At step t it receives the complete feature-level concatenation $q_t = [V_t, M_t]$. With $A_f = \mathrm{row_norm}(A)$

and $A_r = \mathrm{row_norm}(A^\top)$, its order-two bidirectional diffusion is

$$\mathcal{D}_2(Q) = W_\theta[Q, A_f Q, A_f^2 Q, A_r Q, A_r^2 Q] + b_\theta. \tag{5}$$

Independent diffusion operators parameterize the gates and candidate:

$$(r_t, u_t) = \sigma\big(\mathcal{D}_2^{\mathrm{gate}}([q_t, h_{t-1}])\big), \tag{6}$$

$$\widetilde{h}_t = \tanh\big(\mathcal{D}_2^{\mathrm{cand}}([q_t, r_t \odot h_{t-1}])\big), \tag{7}$$

$$h_t = u_t \odot h_{t-1} + (1 - u_t) \odot \widetilde{h}_t. \tag{8}$$

The initial state is zero; the final state enters LayerNorm and the direct head. Hidden width is 48. This comparator has 50,934 parameters on Beijing Air and 36,828 on either traffic dataset, versus 15,654 and 14,988 in the common family. The comparison controls observations, width, optimization, and stopping, but changes recurrent operator, mask encoding, and parameter count. On single-feature traffic inputs $a = M$, so mask resolution differs only on multifeature inputs. These comparisons cannot isolate the causal effect of external completion.

The validation-only equal-parameter candidate changes only MASK-DCRNN's visible-value channel from $[V_t, M_t]$ to $[Z_t^{(2)}, M_t]$, where $Z^{(2)}$ is computed by Equation 4. Recurrent cell, head, parameter names, shapes, and counts remain identical. No learned damping, added channel, or wider state is introduced.

4.3 PROPERTIES AND SCOPE OF CLAMPED COMPLETION

For one time-feature slice, let $O = \mathrm{diag}(m)$, $U = I - O$, and $b = Ox$. The general update $z^{(k+1)} = b + \beta U A z^{(k)}$ satisfies $Oz^{(k)} = Ox$: observed entries are exactly preserved and raw hidden values cannot enter an iterate. If $q = \beta\|UA\|_\infty < 1$, it converges to

$$z^\star = (I - \beta UA)^{-1} Ox = \sum_{r=0}^{\infty} (\beta UA)^r Ox, \tag{9}$$

and initialization at Ox gives the finite-depth bound

$$\|z^\star - z^{(K)}\|_\infty \le \frac{q^{K+1}}{1-q} \|Ox\|_\infty. \tag{10}$$

For a nonnegative row-stochastic graph, $q \le \beta < 1$. The operator is equivariant to a simultaneous node permutation of A, x, and O. These properties concern its own potentially biased fixed point, not the forecasting error. For FP+GRU ($\beta = 1$), $\rho(UA) < 1$ is sufficient for convergence, but does not by itself imply the displayed sup-norm bound; that bound requires $\|UA\|_\infty < 1$. A proof and the distinction between clamped states and the final graph-message channel are given in Appendix H.

5 EXPERIMENTAL DESIGN

This section describes the primary air-quality/traffic matrix. Weather and UKPN use separate CPU protocols with different sampling intervals and seed-dependent corruption (Appendices C and D); their counts and uncertainty analyses are not pooled with the primary study.

5.1 DATASETS AND CHRONOLOGICAL PARTITIONS

We use Beijing Multi-Site Air Quality (Chen, 2017), METR-LA, and PEMS-BAY. Beijing Air uses 11 input features and 24 hourly observations to predict six hourly PM2.5 steps. Traffic inputs contain speed only, using 12 five-minute observations to predict 12 steps; zero and nonfinite speed entries are treated as native missingness. Each timeline is partitioned chronologically into 65% training, 10% validation, 10% calibration, and 15% test. Windows cannot cross boundaries. Normalization and the Beijing graph use training data only; validation selects checkpoints and the disjoint calibration split fits interval radii and risk rules. The transferred records report 5,231, 5,118, and 7,797 test origins per checkpoint for Beijing Air, METR-LA, and PEMS-BAY. The `none` condition retains native missingness; "clean" below means no additional artificial outage.

5.2 FROZEN MATRIX AND OPTIMIZATION

The matrix crosses three datasets, five methods, and training seeds 11, 42, and 73, yielding 45 checkpoints. Each checkpoint is evaluated once under `none` and under five independently seeded replicates of each structured condition, yielding

$$45 \times (1 + 3 \times 5) = 720 \tag{11}$$

evaluation units. Every method uses mixed-corruption training, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, effective batch size 128, gradient clipping at 1.0, and at most 15 epochs. Training stride is two; all later splits use stride one. The masked Smooth-L1 objective is

$$\mathcal{L} = \frac{\sum M_Y \odot \mathrm{SmoothL1}(\widehat{Y}, Y)}{\max(\sum M_Y, 1)}. \tag{12}$$

A checkpoint replaces the incumbent only after a strictly lower validation MAE; exact ties retain the earlier epoch. Patience is four. Deterministic PyTorch algorithms and cuDNN are enabled, cuDNN benchmarking and TF32 are disabled, and the backend is explicitly fixed rather than selected automatically. The restored checkpoint is packaged with its calibration artifacts and reused unchanged across all 16 evaluation units.

5.3 FAILURE CONSTRUCTION AND PAIRING

Failures are generated on the original test timeline before windows are extracted. The `block` condition samples an interval between approximately one-eighth and one-third of the test duration and hides every feature from at least one-third of nodes over that interval. The `sensor` condition removes every feature from at least one-sixth of nodes throughout the split. The `mixed` condition combines a 10% entrywise mask with block and sensor failures from one seeded stream. Component attribution follows sensor $>$ block $>$ point precedence. A scheduled removal counts only when it hides a natively observed input.

Failure seeds depend on dataset, condition, and replicate, but not on method or training seed. Consequently, every method sees the same artificial mask for a given dataset–condition–replicate. Pairing requires matching dataset, training seed, condition, failure replicate, origins, target masks and counts, artificial masks, and component membership. Neither future values nor M_Y are altered by input corruption.

5.4 ESTIMANDS AND EVIDENCE HIERARCHY

Within a cell, metrics aggregate additive errors over eligible targets:

$$\mathrm{MAE} = \frac{\sum |e|}{n}, \qquad \mathrm{RMSE} = \sqrt{\frac{\sum e^2}{n}}. \tag{13}$$

Dataset–method summaries equally weight the 15 seed–replicate metrics for each structured condition and the three seed metrics for `none`; targets are not pooled across cells. The designated primary comparisons use `sensor` and `mixed`: three datasets, two conditions, and four comparators give 24 contrasts. `Block` remains a secondary stress condition. Define $\Delta = \mathrm{MAE}_{\textsc{Resolve}} - \mathrm{MAE}_{\mathrm{comparator}}$; negative values favor RESOLVE. Relative improvement is $100(\overline{\mathrm{MAE}}_{\mathrm{comparator}} - \overline{\mathrm{MAE}}_{\textsc{Resolve}})/\overline{\mathrm{MAE}}_{\mathrm{comparator}}$, using unrounded cell means. SD describes cell dispersion, not a standard error: the 15 cells share three fitted checkpoints and five corruption masks.

The primary study has three evidence stages. Exploratory test aggregates influenced comparator selection. A subsequent equal-parameter validation-only stopping test compared the fixed completion front end with MASK-DCRNN and stopped that structural search after failing three of five designated continuation criteria. The final matrix was then completed under a frozen protocol. Freezing prevents further adaptation, but cannot restore an untouched holdout after prior test-aggregate exposure.

The frozen primary analysis specified 10,000-draw paired non-circular moving-block bootstraps from per-origin additive evidence. Robustness superiority required all 24 dataset–condition–comparator Bonferroni upper bounds to fall below zero. Clean-input non-inferiority used $\Delta_{\mathrm{NI}} =$

Table 1: Overall MAE, averaged over the final evaluation cells; lower is better. Bold indicates the lowest unrounded MAE within each dataset–condition. RMSE is reported in Appendix K.

Dataset	Model	None	Block	Sensor	Mixed
Beijing Air	GraphGRU	26.521	26.390	26.119	26.714
	FP+GRU	26.129	26.385	26.891	27.116
	MASK-DCRNN	**24.548**	**24.591**	**24.448**	**25.100**
	SAFE-ST	25.672	26.163	26.993	26.932
	RESOLVE	25.613	25.815	26.460	26.596
METR-LA	GraphGRU	3.868	4.094	4.362	4.592
	FP+GRU	3.860	4.089	4.343	4.580
	MASK-DCRNN	3.857	**4.061**	**4.254**	**4.469**
	SAFE-ST	**3.837**	4.066	4.332	4.566
	RESOLVE	3.841	4.066	4.329	4.558
PEMS-BAY	GraphGRU	2.235	2.403	2.573	2.758
	FP+GRU	2.244	2.412	2.574	2.745
	MASK-DCRNN	2.397	2.605	2.750	2.998
	SAFE-ST	**2.227**	2.396	**2.561**	**2.737**
	RESOLVE	2.228	**2.396**	2.563	2.742

$\mathrm{MAE}_{\text{RESOLVE}} - 1.02\,\mathrm{MAE}_{\text{GraphGRU}}$ and required all three Bonferroni upper bounds below zero. The lightweight transfer omitted the required per-origin NPZ shards. We therefore report complete descriptive point estimates but do not reconstruct bootstrap intervals, p-values, superiority decisions, or formal non-inferiority decisions from aggregate JSON.

6 RESULTS

6.1 OVERALL FORECASTING PERFORMANCE

Table 1 reports the descriptive means in original target units (PM2.5 concentration in μg/m^3 for Beijing Air and speed in miles per hour for traffic). RESOLVE has lower mean MAE in 17 of the 24 sensor/mixed contrasts, including 2 of 6 against MASK-DCRNN. These counts summarize setting-level means, not independent trials or statistical decisions.

The lowest-error method varies by dataset. MASK-DCRNN has the lowest MAE in every Beijing condition and all three structured METR-LA conditions, whereas SAFE-ST and RESOLVE lead the PEMS-BAY conditions. Appendix J reports changes relative to each method's own native-input result. Outage conditions are distinct mechanisms and need not induce a monotone error ordering.

6.2 PAIRED STRUCTURED-OUTAGE COMPARISONS

Figure 2 summarizes the 24 designated contrasts. Positive improvement means lower MAE for RESOLVE. The count is 5/6 against GraphGRU, 4/6 against SAFE-ST, 6/6 against FP+GRU, and 2/6 against MASK-DCRNN. Against MASK-DCRNN, RESOLVE has 8.23% and 5.96% higher MAE on Beijing sensor and mixed conditions, but 6.83% and 8.53% lower MAE on PEMS-BAY. Most improvements against other controls are below 2%. The two PEMS-BAY improvements do not satisfy the planned requirement of benefit in every component. Table 4 in the appendix provides paired differences, SDs, and within-setting cell win counts.

6.3 CLEAN INPUTS, LOCALIZATION, AND INTERVALS

Against GraphGRU under `none`, RESOLVE has MAE differences of -0.908, -0.027, and -0.006 on Beijing Air, METR-LA, and PEMS-BAY, respectively; all three meet the prespecified 2% point margin. These are point-estimate checks, not formal non-inferiority conclusions.

Table 2 shows larger forecast errors at input-failed nodes. Under sensor failure, RESOLVE MAE is 31.503 versus 25.450 on failed versus unaffected Beijing nodes, 6.969 versus 3.819 on METR-LA, and 4.249 versus 2.227 on PEMS-BAY. These are differences between node subsets, not causal outage

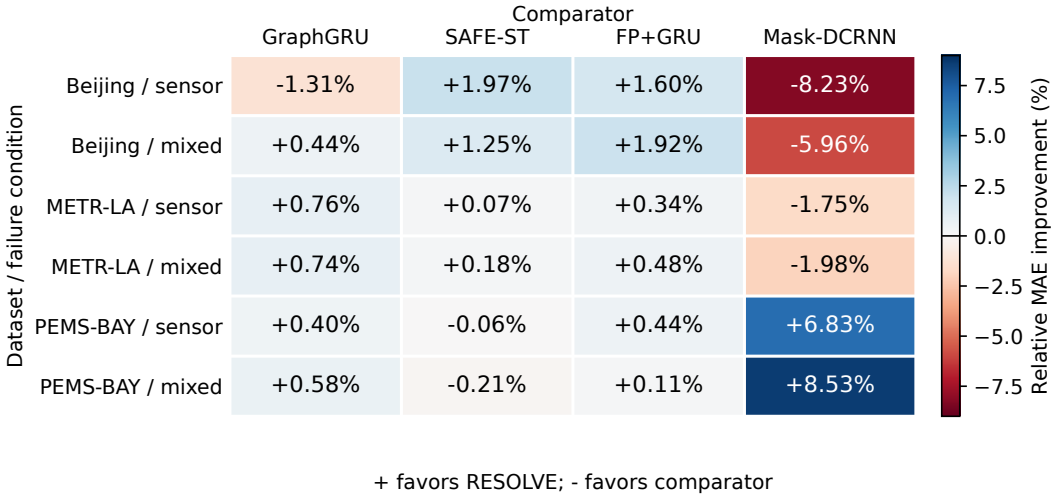

Figure 2: Relative MAE improvement of RESOLVE in the six designated dataset–condition settings. Positive blue cells indicate lower MAE for RESOLVE; negative red cells favor the comparator. Values use matched training seeds and failure replicates and are descriptive point estimates.

Table 2: Failure localization for RESOLVE. Values are overall / input-failed / input-unaffected MAE. NA denotes an empty subset.

Dataset	Condition	Overall	Failed	Unaffected
Beijing	Block	25.815	36.604	25.004
	Sensor	26.460	31.503	25.450
	Mixed	26.596	26.596	NA
METR-LA	Block	4.066	6.792	3.838
	Sensor	4.329	6.969	3.819
	Mixed	4.558	4.732	3.986
PEMS-BAY	Block	2.396	4.212	2.229
	Sensor	2.563	4.249	2.227
	Mixed	2.742	2.888	2.212

effects or evidence that completion improves either subset. The label marks any artificial removal in a node's history; it does not mean the entire history is absent. Under mixed failure every eligible Beijing node is labeled input-failed, so no unaffected estimate exists.

For interval analysis, both global-radius and risk-conditioned intervals are evaluated on the same high-risk test origins, selected using a calibration-fitted threshold (Appendix G). The descriptive rule requires coverage closer to 0.90 and a risk/global width ratio at most 1.25. Recomputing that rule yields 5/15 passes for Beijing Air, 5/15 for METR-LA, and 0/15 for PEMS-BAY. Higher coverage is accompanied by wider intervals (Figure 3); a pass is not a guarantee of nominal calibration. Stored Boolean flags disagree with the documented formula in 10 cells. Both the original flags and formula-based decisions are retained in the accompanying CSV.

6.4 EQUAL-PARAMETER VALIDATION ABLATION

The separately trained completion-augmented MASK-DCRNN improves validation MAE in both Beijing conditions and in none of the four traffic conditions. Its unweighted mean relative improvement is 0.074%. It passes the worst-setting degradation and clean-input criteria, but fails the required 5/6 breadth, 1% mean effect, and same-setting failed-node support. This triggered the recorded stopping rule (Appendix A). Because the architecture and mask channel are held fixed, this is a more direct completion ablation than the primary model comparison. Its validation-only, post-exploratory design supports stopping that tested variant, not a population-level claim of inferiority.

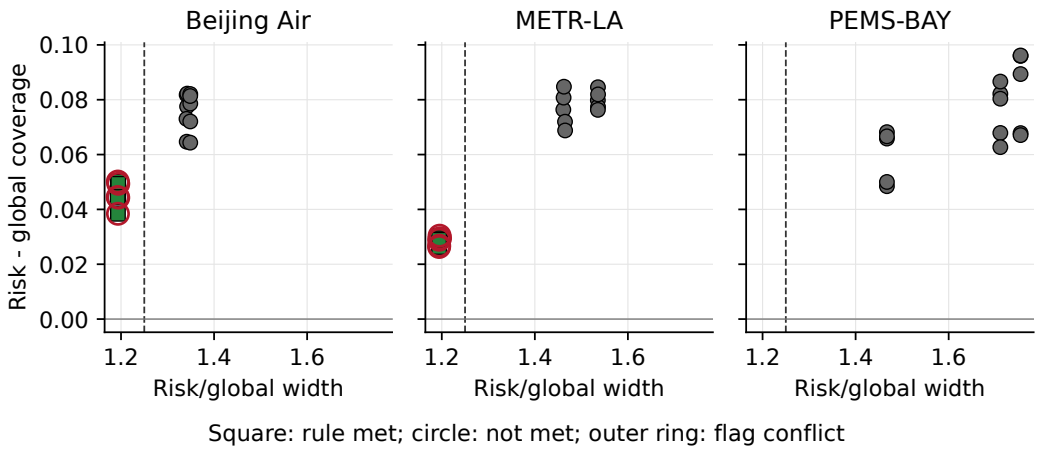

Figure 3: Risk-conditioned coverage–width trade-off across the 45 RESOLVE mixed-outage cells. Green squares satisfy the documented descriptive rule; gray circles do not. An outer ring (red in color) marks the 10 cells where the stored Boolean disagrees with formula-based recomputation. The dashed line is the maximum allowed width ratio of 1.25.

7 DISCUSSION

Comparator dependence. RESOLVE has lower mean MAE in 17 of 24 designated contrasts, but only 2/6 against MASK-DCRNN, with the direction reversing across datasets. This is a model comparison, not an isolated completion effect: diffusion, capacity, and mask encoding also differ. The equal-parameter validation ablation narrows that uncertainty, but its small and uneven gains do not support a broad completion advantage.

Contrasting supplementary domains. On the 20-station NOAA GHCNh weather benchmark (Appendix C), the depth-aligned model ($L = 2$, $\beta = 0.8$) has 7.71%/7.62% lower sensor/mixed MAE than GraphGRU, but 8.38%/15.44% higher MAE than MASK-DCRNN and 9.54%/14.82% higher MAE than FP+GRU. All ten contrasts against these latter controls remain positive under the reported block and paired-seed sensitivities. This post-hoc study has 15 current runs and 75 units; superseded depth-four runs remain separately labeled.

The 44-feeder UKPN study (Appendix D) gives a different pattern. RESOLVE has the lowest mixed-outage mean MAE among the five models: 8.087 kW versus 8.131 for MASK-DCRNN and 8.441 for FP+GRU. The paired-seed-plus-time interval remains below zero against FP+GRU, but crosses zero against MASK-DCRNN. Sensor-outage mean MAE is higher than both comparators. These 15 runs and 75 units illustrate domain- and failure-regime-dependent rankings; they do not identify why the ranking changes. Neither extension alters the primary win counts, and neither sensitivity analysis establishes confirmatory or simultaneous inference.

The study also distinguishes three endpoints often conflated in robustness reporting. Network-wide error depends on the relative size of failed and unaffected subsets; subset errors describe where forecasts are inaccurate; interval coverage must be interpreted together with width on the same eligible targets. Operator invariance alone answers none of these empirical questions. Future comparisons can retain this endpoint separation while using broader comparator coverage and independent test data.

Limitations. Prior test-aggregate exposure precludes an untouched confirmatory interpretation. Three training seeds and shared corruption limit uncertainty assessment; the synthetic mechanisms are stress scenarios, not operational prevalence estimates. Specialized missing-data forecasters such as HD-TTS are not evaluated. Primary per-origin shards remain unavailable. Each extension supplies 15 hash-valid evidence files, but not complete raw-data preprocessing artifacts; successful reconstruction does not verify historical training or establish end-to-end reproducibility. CPU execution and different graph, target, and sampling choices further preclude attributing cross-domain differences to one factor.

Seed resampling jointly varies training and corruption realizations, not training alone. Long artificial outage blocks may make the error process nonstationary; the intervals are consequently resampling sensitivities, not guaranteed coverage statements. UKPN's roughly 24-day test span particularly limits the weekly-block analysis.

8 Conclusion

Within the primary descriptive matrix, RESOLVE has lower mean MAE in 17/24 designated contrasts but only 2/6 against the diffusion-recurrent comparator. Supplementary weather disadvantages and a conditional mixed-outage energy benefit reinforce that this is not a universal ranking. Native-input point margins, localized error, and interval width answer different questions and should be reported separately. The evidence motivates matched architectural controls and retained per-origin artifacts; it does not establish general statistical superiority or an isolated causal benefit of completion.

Ethics Statement

This work studies forecasting during sensor outages in environmental and traffic systems. Such forecasts could inform decisions affecting exposure to pollution or transport reliability. The synthetic outage results and descriptive interval diagnostics do not certify safe deployment: errors are larger at input-failed nodes, and higher coverage can be obtained by widening intervals. Operational use would require monitoring on local data, explicit reporting of unavailable sensors, and validation against the consequences of forecast errors. The study uses public environmental and traffic measurements rather than individual-level records.

The supplementary energy analysis uses feeder-level active-power records from a public UKPN release. Its forecasts could affect grid operations; neither aggregate accuracy nor bootstrap sensitivity establishes operational safety or privacy guarantees.

AI Use Statement

Generative AI (OpenAI Codex) assisted with research-methodology feedback, formulation and checking of mathematical explanations, implementation of evidence-audit and bootstrap analysis code, interpretation of recorded results, literature retrieval, manuscript drafting and editing, and figure/table preparation. The disclosed analyses operate on recorded experimental artifacts rather than invented observations. Computational checks included reconstruction from additive evidence, paired-count checks, and tests of block-sum calculations against explicit resampling. These checks do not replace author verification of historical experiments or preprocessing, whose limits are reported in the paper. The human authors are responsible for reviewing all AI-assisted content and for the final claims, references, and artifacts.

Reproducibility Statement

The model equations and primary protocol are described in the main text; Appendix H gives the completion proof. Appendices C and D document supplementary configurations, evidence checks, and resampling. The accompanying analysis source regenerates tables and figures from supplied summaries and accepts external per-origin files for supplementary audits. It is not a full training release: primary per-origin shards and complete preprocessing artifacts are unavailable. These gaps constrain reproducibility and inference as explicitly stated.

References

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. BRITS: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, 2018.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

Song Chen. Beijing multi-site air quality. UCI Machine Learning Repository, 2017.

Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the gaps: Multivariate time series imputation by graph neural networks. In *International Conference on Learning Representations*, 2022.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11694.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJiHXGWAZ.

Mingzhe Liu, Han Huang, Hao Feng, Leilei Sun, Bowen Du, and Yanjie Fu. PriSTI: A conditional diffusion framework for spatiotemporal imputation. In *IEEE International Conference on Data Engineering*, 2023.

Ivan Marisca, Cesare Alippi, and Filippo Maria Bianchi. Graph-based forecasting with missing data through spatiotemporal downsampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34846–34865. PMLR, 2024. URL https://proceedings.mlr.press/v235/marisca24a.html.

NOAA National Centers for Environmental Information. Global historical climatology network hourly (GHCNh). https://www.ncei.noaa.gov/products/global-historical-climatology-network-hourly, n.d. Accessed September 12, 2026.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL https://jmlr.org/papers/v22/20-303.html.

Emanuele Rossi, Henry Kenlay, Maria I. Gorinova, Benjamin P. Chamberlain, Xiaowen Dong, and Michael M. Bronstein. On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. In *Proceedings of the First Learning on Graphs Conference*, volume 198 of *Proceedings of Machine Learning Research*, pp. 11:1–11:16, 2022. URL https://proceedings.mlr.press/v198/rossi22a.html.

UK Power Networks. Photovoltaic (PV) solar panel energy generation data. https://data.london.gov.uk/dataset/photovoltaic-pv-solar-panel-energy-generation-data-2nlqm, n.d. London Datastore. Accessed September 15, 2026.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *International Joint Conference on Artificial Intelligence*, pp. 1907–1913, 2019. doi: 10.24963/ijcai.2019/264.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *International Joint Conference on Artificial Intelligence*, pp. 3634–3640, 2018. doi: 10.24963/ijcai.2018/505.

A EVIDENCE TIERS AND STOPPING RULE

Table 3 separates the reported evaluation matrices. An evaluation unit is a checkpoint–condition–replicate combination, not an independent training run or test subject. The exploratory and validation-only stages described below are not added to these counts.

Table 3: Scope of the reported evidence. Primary structured conditions use five failure replicates; supplementary conditions use one realization per training seed. Counts exclude superseded weather runs.

Matrix	Runs	Units	Conditions per run	Uncertainty reported
Primary (three datasets)	45	720	Native + 3 structured $\times$ 5	Descriptive only
Weather (20 stations)	15	75	Native + 4 corrupted	Post-hoc sensitivity
UKPN (44 feeders)	15	75	Native + 4 corrupted	Supplementary sensitivity

The exploratory matrix contained 63 training runs and 315 dataset–condition records. Its aggregates were inspected and used to select the diffusion-recurrent comparator and formulate the subsequent question. The post-exploratory stopping test was restricted to validation data and compared a two-step damped completion front end against MASK-DCRNN with equal trainable parameter counts. It passed the worst-cell and clean-input criteria but failed breadth (2/6 wins), mean effect (0.074% versus a required 1%), and same-cell failed-node support. Because the five criteria were conjunctive, the decision was `NO_GO`; the tested primary branch was stopped. The later weather variant is separately disclosed below.

Relative validation improvements (positive favors the completion-augmented diffusion model) are $+0.404\%$ and $+1.130\%$ for Beijing sensor and mixed outages, -0.541% and -0.043% for METR-LA, and -0.222% and -0.284% for PEMS-BAY. Their unweighted mean is $+0.074\%$. These are validation results from the separately trained equal-parameter candidate, not additional primary test comparisons.

B PAIRED POINT ESTIMATES

Table 4: Paired structured-outage point estimates. Delta is RESOLVE MAE minus comparator MAE; negative values favor RESOLVE. SD is cell-level dispersion, not a standard error.

Dataset	Condition	Comparator	RESOLVE/comparator MAE	Δ mean $\pm$ SD	Wins
Beijing	Sensor	GraphGRU	26.460/26.119	0.342 ± 0.300	0/15
		SAFE-ST	26.460/26.993	-0.533 ± 0.468	10/15
		FP+GRU	26.460/26.891	-0.431 ± 1.135	6/15
		MASK-DCRNN	26.460/24.448	2.012 ± 0.675	0/15
Beijing	Mixed	GraphGRU	26.596/26.714	-0.118 ± 0.337	11/15
		SAFE-ST	26.596/26.932	-0.337 ± 0.206	14/15
		FP+GRU	26.596/27.116	-0.520 ± 1.102	7/15
		MASK-DCRNN	26.596/25.100	1.496 ± 0.599	0/15
METR-LA	Sensor	GraphGRU	4.329/4.362	-0.033 ± 0.020	15/15
		SAFE-ST	4.329/4.332	-0.003 ± 0.016	9/15
		FP+GRU	4.329/4.343	-0.015 ± 0.032	8/15
		MASK-DCRNN	4.329/4.254	0.075 ± 0.069	3/15
METR-LA	Mixed	GraphGRU	4.558/4.592	-0.034 ± 0.032	14/15
		SAFE-ST	4.558/4.566	-0.008 ± 0.013	11/15
		FP+GRU	4.558/4.580	-0.022 ± 0.031	10/15
		MASK-DCRNN	4.558/4.469	0.089 ± 0.074	1/15
PEMS-BAY	Sensor	GraphGRU	2.563/2.573	-0.010 ± 0.004	15/15
		SAFE-ST	2.563/2.561	0.002 ± 0.007	7/15
		FP+GRU	2.563/2.574	-0.011 ± 0.010	13/15
		MASK-DCRNN	2.563/2.750	-0.188 ± 0.128	15/15
PEMS-BAY	Mixed	GraphGRU	2.742/2.758	-0.016 ± 0.003	15/15
		SAFE-ST	2.742/2.737	0.006 ± 0.004	1/15
		FP+GRU	2.742/2.745	-0.003 ± 0.006	10/15
		MASK-DCRNN	2.742/2.998	-0.256 ± 0.120	15/15

C POST-HOC GHCNH GREAT LAKES EXTENSION

We performed an additional cross-domain evaluation using NOAA Global Historical Climatology Network hourly observations (NOAA National Centers for Environmental Information, n.d.). The processed benchmark contains 20 Great Lakes stations, 26,304 hourly time steps from 2022–2024, and five variables: temperature, dew-point temperature, sea-level pressure, wind speed, and relative humidity. Temperature is the target. The model receives 24 hours of history and predicts the following six hours. Chronological train/validation/calibration/test proportions are 65/10/10/15; feature-wise normalization uses training-split statistics, native missingness is retained, and missing standardized inputs are filled with zero. A row-normalized Gaussian adjacency is constructed from Haversine station distance.

This extension uses the same five model families, mixed-corruption training, and training seeds 11, 42, and 73. The initial weather release used RESOLVE at $L = 4$. Following identification of this mismatch, three new runs use $L = 2$ and damping 0.8, aligning these two settings with the primary model; their JSON configurations verify the change. The twelve comparator JSON records are unchanged. FP+GRU retains 40 steps; hidden width is 48, batch size is 128, and all runs used a CPU backend. Each current run has one evaluation under `none`, `point`, `block`, `sensor`, and `mixed`, giving 15 current runs and 75 units. The three superseded $L = 4$ runs are retained as a labeled reference, not pooled with $L = 2$ or counted as additional seeds. Configuration alignment does not convert this post-hoc extension into an untouched replication of the primary study.

Failure seeds follow $s_{\text{fail}} = s_{\text{train}} + 1000 + j$, where $j = 0, \ldots, 4$ indexes those five conditions in order. Thus each training seed has its own corruption realization; models remain paired within seed and condition. The reported mean $\pm$ sample SD over three runs jointly reflects training-seed and corruption-realization variation, not an isolated estimate of either source or a confidence interval. All five conditions appear in Table 5; origin-level bootstrap sensitivity is reported separately below.

Table 5: Post-hoc NOAA GHCNh Great Lakes results in degrees Celsius. Mean ± sample SD uses three paired runs (seeds 11, 42, 73); corruption also varies across runs. The current RESOLVE uses $L = 2$. Earlier $L = 4$ results are separately labeled and excluded from current counts. Bold marks the lowest unrounded mean among the five current models.

Model	Native	Point	Block	Sensor	Mixed
			MAE (mean ± sample SD)		
GraphGRU	1.630±0.056	1.667±0.044	1.680±0.071	1.788±0.083	1.917±0.065
MASK-DCRNN	**1.418±0.056**	**1.438±0.037**	**1.431±0.043**	1.523±0.059	**1.534±0.075**
SAFE-ST	1.556±0.021	1.627±0.046	1.584±0.020	1.658±0.034	1.735±0.013
RESOLVE ($L = 2$)	1.526±0.015	1.571±0.008	1.573±0.018	1.650±0.062	1.771±0.057
FP+GRU	1.425±0.019	1.464±0.019	1.452±0.013	**1.507±0.019**	1.542±0.013
RESOLVE ($L = 4$)	1.515±0.017	1.559±0.015	1.559±0.023	1.624±0.041	1.737±0.051

Model	Native	Point	Block	Sensor	Mixed
			RMSE (mean ± sample SD)		
GraphGRU	2.171±0.055	2.212±0.044	2.237±0.080	2.481±0.257	2.616±0.052
MASK-DCRNN	**1.914±0.083**	**1.931±0.052**	**1.928±0.065**	2.097±0.167	**2.071±0.087**
SAFE-ST	2.081±0.026	2.156±0.049	2.118±0.021	2.221±0.058	2.337±0.035
RESOLVE ($L = 2$)	2.054±0.015	2.102±0.007	2.115±0.027	2.252±0.133	2.422±0.064
FP+GRU	1.970±0.018	2.008±0.021	1.996±0.010	**2.066±0.032**	2.105±0.024
RESOLVE ($L = 4$)	2.041±0.018	2.087±0.017	2.097±0.031	2.203±0.087	2.375±0.053

For the current depth-two weather model, sensor/mixed MAE is 1.650 ± 0.062/1.771 ± 0.057 degrees Celsius. It is 7.71%/7.62% lower than GraphGRU; relative to SAFE-ST, sensor MAE is 0.46% lower but mixed MAE is 2.05% higher. Relative to MASK-DCRNN, MAE is 8.38%/15.44% higher, and relative to FP+GRU it is 9.54%/14.82% higher. The earlier depth-four runs have lower means than depth two in both conditions, but are not selected for the aligned comparison. Neither version is added to the primary 17/24 and 2/6 summaries.

The updated archive passes all 40 listed file hashes. All 15 run JSONs and 15 NPZ files are present; the latter pass ZIP integrity, declared SHA-256, finite-value, and additive subgroup checks. In particular, the previously damaged SAFE-ST seed-11 evidence is now valid and agrees with its unchanged JSON. Each run contains 3,917 consecutive hourly test origins, aligned across models and seeds. Reconstructed MAE/RMSE agree with all 75 stored evaluation units to within 4.9×10^{-7} degrees Celsius; all 25 release means and sample SDs agree within rounding. The processed numeric dataset and complete preprocessing program are still absent, so the lightweight transfer cannot independently establish station selection, native-missingness handling, or graph construction.

C.1 POST-HOC PAIRED BLOCK-BOOTSTRAP SENSITIVITY

The supplied bootstrap CSV contains ten contrasts (five conditions against MASK-DCRNN and FP+GRU). Its 30 seed-level and ten mean differences reconstruct from the additive evidence to within 10^{-12}. However, the transfer omits its bootstrap generator, random seed, draw count, and tail-probability definition. We retain that CSV as source material but do not adopt its p-values or Holm decisions. Instead, Table 6 reports an independently specified, reproducible reanalysis, not a bitwise reproduction of those supplied intervals.

Let $A^R_{s,o}$ and $A^C_{s,o}$ be sums of absolute error across valid nodes and horizons, with common target count $N_{s,o}$. The estimand for each condition and comparator is

$$\widehat{\Delta} = \frac{1}{3} \sum_{s \in \{11,42,73\}} \frac{\sum_o (A^R_{s,o} - A^C_{s,o})}{\sum_o N_{s,o}}. \tag{14}$$

Positive values mean higher RESOLVE MAE. Ratios of summed errors and counts preserve target weighting; they are not unweighted averages of origin MAEs.

We use 10,000 non-circular moving-block draws with lengths $L_b = 24$ and 168 hourly origins. Starts are uniform over $0, \ldots, n - L_b$; $\lceil n/L_b \rceil$ blocks are concatenated and truncated to $n = 3,917$. Within a draw the same indices are used for both models, all seeds, and all conditions, retaining their common weather-time variation. Ratios are recomputed within each seed and averaged equally over the three seeds. A third sensitivity additionally samples three seed pairs with replacement from those same 24-hour-block ratios. Thus it varies the empirical seed mixture without treating the models or their shared weather as independent.

The generator is NumPy PCG64 seeded with 20260913; linear-interpolated 2.5th/97.5th percentiles give pointwise 95% intervals. The packaged script fixes draw order and tests its block-sum calculation against explicit concatenation. All ten intervals remain positive in each scheme, including the four sensor/mixed contrasts. This supports the direction of the observed weather disadvantage under these resampling choices, not a claim of universal inferiority. These are post-hoc, unadjusted sensitivity intervals: three seed pairs provide limited information about training variation, block choice does not prove stationarity, and no familywise or confirmatory significance decision is asserted. The weather analysis neither supplies the missing primary shards nor changes the primary 24-comparison protocol.

Table 6: Independent weather bootstrap sensitivity for $\Delta = \mathrm{MAE}_{\textsc{Resolve}(L=2)} - \mathrm{MAE}_{\mathrm{comparator}}$ in degrees Celsius. Entries are pointwise 95% percentile intervals from 10,000 draws. MBB keeps the three seed pairs fixed; Seed + MBB also resamples seed pairs. Positive values favor the comparator.

Condition	Comparator	Δ	MBB-24	MBB-168	Seed + MBB-24
Native	MASK-DCRNN	0.108	[0.083, 0.134]	[0.080, 0.137]	[0.034, 0.167]
Native	FP+GRU	0.100	[0.083, 0.117]	[0.083, 0.117]	[0.065, 0.134]
Point	MASK-DCRNN	0.133	[0.113, 0.155]	[0.111, 0.159]	[0.094, 0.169]
Point	FP+GRU	0.107	[0.092, 0.122]	[0.090, 0.124]	[0.087, 0.128]
Block	MASK-DCRNN	0.142	[0.118, 0.168]	[0.116, 0.168]	[0.104, 0.177]
Block	FP+GRU	0.121	[0.103, 0.138]	[0.097, 0.145]	[0.100, 0.142]
Sensor	MASK-DCRNN	0.128	[0.108, 0.149]	[0.107, 0.154]	[0.095, 0.161]
Sensor	FP+GRU	0.144	[0.129, 0.159]	[0.123, 0.167]	[0.087, 0.207]
Mixed	MASK-DCRNN	0.237	[0.214, 0.262]	[0.209, 0.278]	[0.197, 0.278]
Mixed	FP+GRU	0.229	[0.205, 0.254]	[0.191, 0.278]	[0.174, 0.285]

D SUPPLEMENTARY UKPN FEEDER-ENERGY EXTENSION

This separate CPU experiment evaluates active-power forecasting using a UK Power Networks public PV-monitoring release (UK Power Networks, n.d.). It is not pooled with the frozen primary matrix or the weather extension. The handoff describes specification before outcome inspection, but does not provide an independently timestamped protocol establishing that ordering. We therefore report it as supplementary evidence without confirmatory significance claims.

Data and configuration. The handoff describes 1,079,006 source records from 48 feeder monitors at 20 substations. Training-only target coverage of at least 95% retains 44 feeders, identified by substation and feeder. The processed tensor has 23,434 ten-minute steps, 44 nodes, and five features: total active power, total reactive power, mean voltage, mean current, and frequency. The target is total three-phase active power in kW, eligible only when all three phase measurements are available. Chronological train/validation/calibration/test boundaries are 15,232/17,575/19,918 steps (65/10/10/15%). Normalization is described as training-only; native missingness is retained and missing standardized inputs are zero-filled. The graph links feeders sharing a substation, includes self-loops, and is row-normalized; it is not complete electrical connectivity. Raw measurements, the processed tensor, and the full preprocessing implementation are absent from this transfer, so these preprocessing details are attributed to the handoff rather than independently reproduced.

The verified run configurations use seeds 11, 42, and 73; hidden width 48; batch size 128; AdamW with learning rate 0.001 and weight decay 0.0001; at most 15 epochs with patience four; and a CPU backend. History 24 and horizon six denote ten-minute steps (four hours and one hour), not hourly samples. RESOLVE uses two completion steps and damping 0.8; FP+GRU uses 40 steps and MASK-DCRNN diffusion order two. Each of five models has three records, each covering native, point, block, sensor, and mixed conditions: 15 runs and 75 units. Failure seeds are training seed plus

$1000 + j$ for condition index $j = 0, \ldots, 4$, so models are paired within each seed–condition but seed SD mixes training and corruption variation.

Table 7: UKPN supplementary results across three seeds. Bold denotes lowest unrounded mean per condition, not statistical significance. Native retains native missingness.

Model	Native	Point	Block	Sensor	Mixed
			MAE (kW; mean±sample SD)		
GraphGRU	5.776±0.816	5.915±0.824	6.630±0.699	7.420±1.809	8.169±1.117
MASK-DCRNN	**5.602±0.772**	**5.754±0.811**	**6.585±0.651**	**7.001±1.686**	8.131±1.108
SAFE-ST	5.718±0.866	5.835±0.914	6.650±0.715	7.229±1.510	8.360±1.360
RESOLVE	5.758±0.874	5.915±0.887	6.638±0.759	7.290±1.490	**8.087±0.951**
FP+GRU	5.710±0.756	5.843±0.813	6.701±0.649	7.217±1.282	8.441±1.171

Model	Native	Point	Block	Sensor	Mixed
			RMSE (kW; mean±sample SD)		
GraphGRU	12.119±3.376	12.317±3.379	14.573±2.213	14.780±5.660	16.304±2.434
MASK-DCRNN	11.673±2.799	11.866±2.819	14.465±1.724	**14.049±5.120**	16.020±1.801
SAFE-ST	**11.537±2.480**	**11.677±2.448**	**14.208±1.575**	14.260±4.786	16.173±2.115
RESOLVE	11.727±2.764	11.927±2.735	14.318±1.798	14.404±4.831	**15.819±1.631**
FP+GRU	11.596±2.503	11.754±2.464	14.386±1.598	14.328±4.536	16.362±1.952

Integrity and reconstruction. All 15 JSON and 15 NPZ files are byte-identical between the two supplied archives. All 41 outer and three nested checksum entries pass. Every NPZ passes its declared hash and ZIP CRC check; additive arrays are finite, target counts align across paired models, and all 75 overall MAE/RMSE values reconstruct within floating-point rounding (maximum absolute discrepancy 3.44×10^{-6} kW). Each run contains 3,487 aligned test origins. These checks establish transfer and aggregate consistency, not historical training correctness or full end-to-end reproducibility.

Independent uncertainty sensitivity. We reconstruct all ten mean contrasts and 30 seed-level differences in the supplied bootstrap summary to 10^{-12}. Its accompanying utility implements a fixed-seed moving-block bootstrap, but not the complete hierarchical and multiplicity workflow. Its default mixed-versus-MASK-DCRNN interval is $[-0.08985, -0.00593]$ kW, slightly different from the supplied summary's $[-0.09007, -0.00629]$. We preserve that summary but do not claim exact reproduction of its resampling stream or adopt its reported p-values and Holm decisions.

Table 8 instead uses the packaged independent audit: 9,999 non-circular moving-block draws, NumPy PCG64 seed 20260914, and block lengths 144 and 1,008 ten-minute origins (24 and 168 hours). Starts are uniform over $0, \ldots, n - L_b$; concatenated blocks are truncated to $n = 3,487$. As in the weather analysis, each draw shares time indices across models, seeds, and conditions, recomputes within-seed ratios of additive error sums to counts, and averages the three paired differences equally. The seed-plus-24-hour sensitivity additionally samples three seed pairs with replacement using the same time-block draw. Linear-interpolated percentile endpoints define pointwise 95% intervals. This shared-time seed sensitivity is independently specified, not an exact rerun of the supplied hierarchical procedure. The roughly 24-day span contains only about 3.5 weekly blocks; neither block choice proves stationarity, and three seed pairs poorly characterize training variability. No simultaneous or confirmatory inference is asserted.

Mixed-outage differences are -0.044 kW against MASK-DCRNN and -0.355 kW against FP+GRU. The former has seed differences $+0.302, -0.318, -0.116$ kW and a seed-plus-time interval crossing zero; the latter favors RESOLVE in all three seeds and all reported sensitivity schemes. The block-outage advantage over FP+GRU is not stable to weekly blocks or seed resampling. Sensor means favor both comparators, and the disadvantage against MASK-DCRNN persists in each scheme. Thus the energy extension supplies a conditional mixed-outage benefit, not uniform superiority or a repair of missing primary inferential evidence.

The source package includes `audit_ukpn.py`, the generated audit JSON and tables, and the separately retained supplied bootstrap JSON. The audit accepts the external directory of 15 JSON/NPZ

Table 8: Independent UKPN sensitivity for $\Delta = \mathrm{MAE}_{\textsc{Resolve}} - \mathrm{MAE}_{\mathrm{comparator}}$ (kW). Negative values favor RESOLVE. MBB keeps seeds fixed; the last column also resamples seed pairs. All intervals are pointwise and unadjusted.

Condition	Comparator	Δ	MBB-24 h	MBB-168 h	Seed + MBB-24 h
Native	MASK-DCRNN	0.156	$[0.130, 0.186]$	$[0.136, 0.167]$	$[0.057, 0.266]$
Native	FP+GRU	0.049	$[0.027, 0.069]$	$[0.028, 0.060]$	$[-0.078, 0.169]$
Point	MASK-DCRNN	0.161	$[0.137, 0.189]$	$[0.144, 0.173]$	$[0.070, 0.250]$
Point	FP+GRU	0.072	$[0.050, 0.094]$	$[0.051, 0.086]$	$[-0.024, 0.152]$
Block	MASK-DCRNN	0.052	$[-0.030, 0.123]$	$[-0.108, 0.156]$	$[-0.195, 0.286]$
Block	FP+GRU	-0.063	$[-0.123, -0.009]$	$[-0.180, 0.019]$	$[-0.250, 0.117]$
Sensor	MASK-DCRNN	0.289	$[0.263, 0.317]$	$[0.258, 0.328]$	$[0.062, 0.490]$
Sensor	FP+GRU	0.073	$[0.047, 0.098]$	$[0.031, 0.098]$	$[-0.096, 0.315]$
Mixed	MASK-DCRNN	-0.044	$[-0.090, -0.006]$	$[-0.127, -0.006]$	$[-0.311, 0.288]$
Mixed	FP+GRU	-0.355	$[-0.397, -0.318]$	$[-0.416, -0.327]$	$[-0.645, -0.123]$

pairs; large evidence files are not duplicated in the manuscript source archive. Resampling seed pairs also changes their associated corruption realizations. Accordingly, the last interval column is not an isolated estimate of initialization uncertainty. Long block failures can induce regime changes within the test sequence; no stationarity assumption is validated by the checksum or aggregate audits.

E EFFICIENCY

The recorded latencies are descriptive measurements, not a controlled hardware benchmark. The transferred summaries do not establish identical runtime conditions across all evaluations. Small latency differences should therefore not be interpreted as intrinsic speed advantages, and these measurements are not compared with the separate CPU extensions.

Table 9: Parameters and inference milliseconds per origin (mean $\pm$ SD). Each entry summarizes 48 evaluation units.

Dataset	Model	Parameters	ms/origin
Beijing	GraphGRU	15,654	0.712 ± 0.180
	FP+GRU	15,654	0.679 ± 0.125
	MASK-DCRNN	50,934	0.791 ± 0.204
	SAFE-ST	15,654	0.682 ± 0.168
	RESOLVE	15,654	0.684 ± 0.148
METR-LA	GraphGRU	14,988	1.096 ± 0.283
	FP+GRU	14,988	1.140 ± 0.337
	MASK-DCRNN	36,828	1.224 ± 0.429
	SAFE-ST	14,988	1.197 ± 0.420
	RESOLVE	14,988	1.045 ± 0.221
PEMS-BAY	GraphGRU	14,988	1.096 ± 0.131
	FP+GRU	14,988	1.190 ± 0.311
	MASK-DCRNN	36,828	1.167 ± 0.121
	SAFE-ST	14,988	1.130 ± 0.176
	RESOLVE	14,988	1.060 ± 0.179

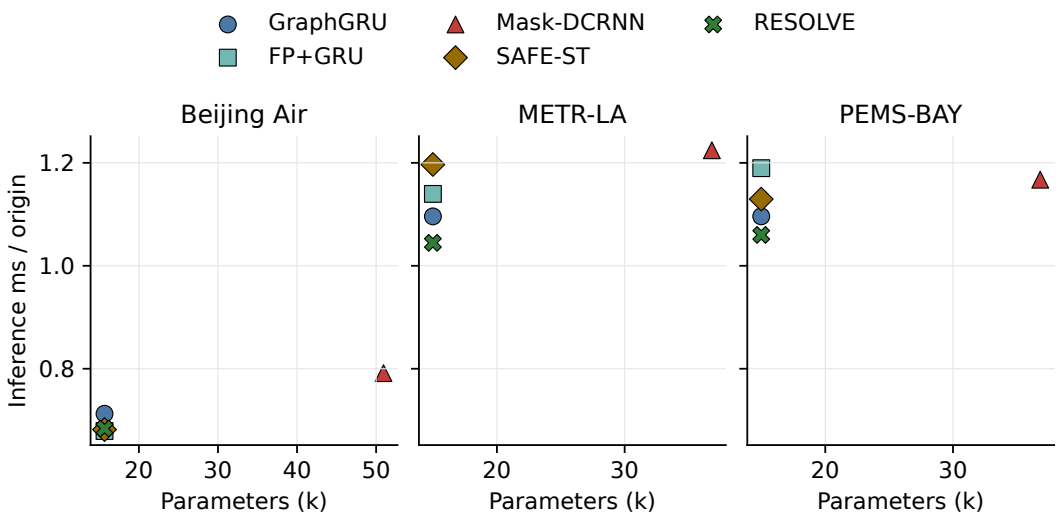

Figure 4: Descriptive efficiency profile. Parameter count is plotted against mean inference milliseconds per origin over 48 evaluation units per dataset–method combination. Runtime SDs remain available in Table 9; the plot emphasizes the parameter–latency relationship.

F Artifact completeness

All 45 expected checkpoint JSON records were present; each contains 16 evaluation units and declares final-inference eligibility. Dataset, training seed, condition, and failure-replicate identifiers were used for pairing. Each dataset had a consistent fingerprint and test-origin count across the transferred records. The summaries were regenerated only from the final-evaluation JSON files. The transfer did not include the per-origin NPZ shards referenced by the frozen bootstrap protocol; no interval was imputed from cell-level SDs.

G Coverage–width audit

For standardized input history X_o and availability mask M_o at origin o, the observable risk score is

$$r_o = 1 - \mathrm{mean}(M_o) + 0.25\,\frac{\sum_{t,n} |X_{o,tn,0}| M_{o,tn,0}}{\max(\sum_{t,n} M_{o,tn,0}, 1)}. \tag{15}$$

Feature zero is the target-history channel. For each checkpoint, absolute residuals on the disjoint mixed-corrupted calibration split yield horizon–node radii. With n eligible calibration residuals, the implementation uses NumPy's `higher` quantile at $\min\{1, \lceil 0.9(n+1)\rceil/n\}$. Four strata are defined by calibration risk quartiles; a stratum with fewer than 50 calibration origins falls back to the global radius. The radii, risk-bin edges, and the calibration 80th-percentile risk threshold are frozen before test evaluation.

On test origins satisfying $r_o \geq Q^{\mathrm{cal}}_{0.8}$, the global construction uses the global radius and the risk-conditioned construction uses the radius associated with the origin's frozen risk bin. Let C_g, C_r be their coverage and W_g, W_r their mean widths, computed on the *same* eligible targets. The documented descriptive rule (H4) is

$$|C_r - 0.90| < |C_g - 0.90| \quad \text{and} \quad W_r/W_g \leq 1.25. \tag{16}$$

This is a comparative diagnostic, not a finite-sample coverage guarantee under temporal dependence. It was designated descriptive independently of the unavailable primary bootstrap.

Table 10 summarizes Figure 3. Risk-conditioned coverage is higher in every dataset, but the mean width ratio exceeds 1.25 in all three. All 45 stored source flags are false; recomputing the documented formula gives ten positives. The currently available evaluator code instead tests a nonpositive width change, which is consistent with this disagreement, but the historical code identity has not been independently established. We preserve both flag columns and use the explicit formula for this report.

Table 10: Risk-conditioned coverage and interval width for the 15 RESOLVE mixed-outage cells per dataset. Passes use the documented formula-based rule.

Dataset	Passes	Risk origins	Global cov.	Risk cov.	Cov. change	Width ratio
Beijing Air	5/15	1768.3	0.716	0.782	0.066	1.295
METR-LA	5/15	1011.6	0.800	0.862	0.062	1.398
PEMS-BAY	0/15	1357.5	0.775	0.848	0.073	1.644

H COMPLETION PROPERTIES

Set $B = \beta UA$ and $b = Ox$. From $z^{(0)} = b$ and $z^{(k+1)} = b + Bz^{(k)}$, induction gives $z^{(K)} = \sum_{r=0}^{K} B^r b$. Since $OU = 0$, multiplying the update by O yields $Oz^{(k+1)} = Ox$ at every step. If $q = \|B\|_\infty < 1$, the Neumann series converges, and submultiplicativity bounds the omitted tail by

$$\left\| \sum_{r=K+1}^{\infty} B^r b \right\|_\infty \leq \sum_{r=K+1}^{\infty} q^r \|b\|_\infty = \frac{q^{K+1}}{1-q} \|b\|_\infty. \tag{17}$$

For a permutation matrix P, transforming (A, O, x) to $(PAP^\top, POP^\top, Px)$ transforms every iterate to $Pz^{(k)}$. These observations justify clamping, a truncation bound, and permutation equivariance. They do not establish recovery of the unobserved signal. In the primary RESOLVE implementation the encoder sees both V and $G = AZ^{(2)}$; the clamping identity applies to Z, not to the final graph-message channel G. The FP+GRU encoder instead uses $G = Z^{(40)}$.

I REPRODUCIBILITY AND AI ASSISTANCE

The accompanying source package contains aggregate summaries, figure/table generators, a consistency audit, and independent weather and UKPN bootstrap scripts. The primary JSON transfer has all 45 expected records, each with 16 units, but not their referenced per-origin shards. Aggregates cannot replace that missing bootstrap evidence. Each extension supplies 15 readable per-origin files; its script accepts the external evidence directory and regenerates summaries and intervals. These files contain additive errors and counts, not raw predictions or targets, so the audits verify internal aggregation rather than independently recomputing forecasting errors. The earlier weather depth-four summary and supplied bootstrap summaries remain separately labeled. This is a manuscript-analysis package, not a complete dataset-construction or model-training release.

OpenAI Codex assisted with code review, aggregate-consistency checks, literature and submission-guidance retrieval, manuscript editing, and figure generation from recorded results. AI tools were not used to fabricate experimental observations. The authors are responsible for verifying the AI-assisted text, analyses, references, and artifacts before submission.

J Additional Descriptive Plots

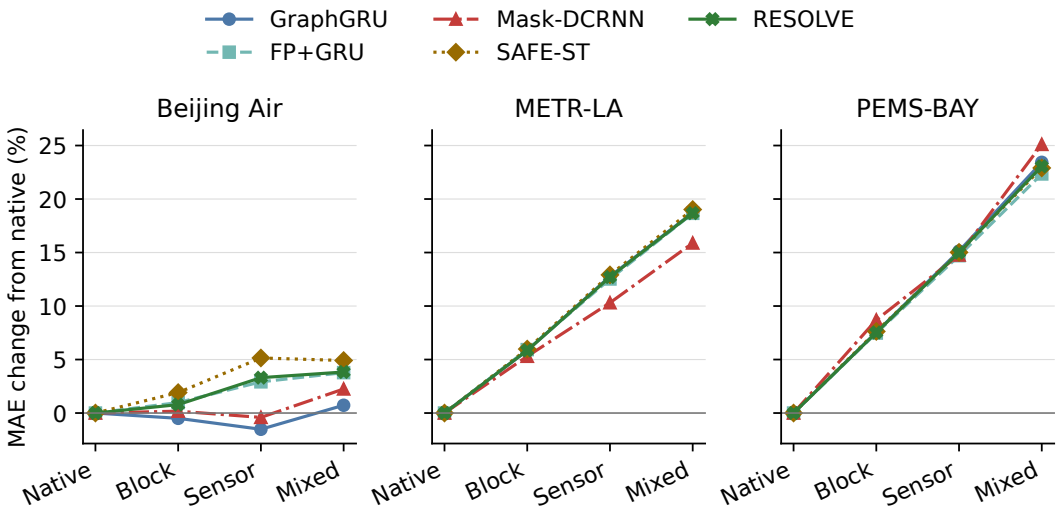

Figure 5: MAE change relative to each method's own native-input condition. Native missingness remains in that reference condition. Lines join distinct conditions for readability and do not imply a scalar ordering of outage severity.

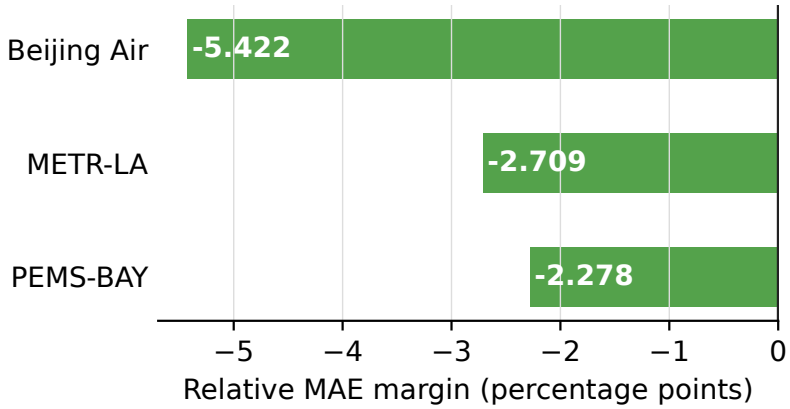

Figure 6: Dimensionless native-input margin, $100(\mathrm{MAE}_{\mathrm{RESOLVE}}/\mathrm{MAE}_{\mathrm{GraphGRU}} - 1.02)$. Negative values meet the 2% point margin. No confidence bounds or formal non-inferiority decisions are inferred.

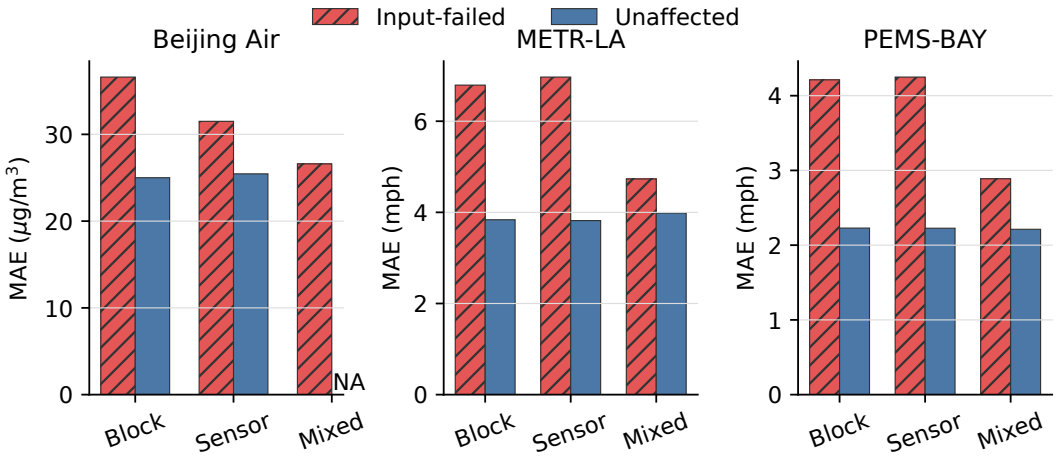

Figure 7: Descriptive mean MAE at input-failed and input-unaffected nodes. Units are μg/m^3 for Beijing Air and miles per hour for traffic. Beijing mixed has no eligible unaffected subset. These selected-subset contrasts are not causal outage effects.

K Overall RMSE

Table 11: Overall RMSE corresponding to Table 1, with identical evaluation cells and original target units. Lower is better.

Dataset	Model	None	Block	Sensor	Mixed
Beijing Air	GraphGRU	45.848	46.090	46.289	46.425
	FP+GRU	46.498	46.863	47.454	47.387
	MASK-DCRNN	43.830	44.175	44.144	44.699
	SAFE-ST	45.404	46.270	47.310	47.040
	RESOLVE	45.137	45.749	46.666	46.650
METR-LA	GraphGRU	7.456	7.766	8.231	8.494
	FP+GRU	7.567	7.851	8.239	8.456
	MASK-DCRNN	7.313	7.587	7.856	8.111
	SAFE-ST	7.449	7.753	8.198	8.424
	RESOLVE	7.457	7.753	8.209	8.431
PEMS-BAY	GraphGRU	4.738	5.011	5.270	5.638
	FP+GRU	4.850	5.103	5.361	5.667
	MASK-DCRNN	4.777	5.062	5.254	5.597
	SAFE-ST	4.728	4.992	5.258	5.571
	RESOLVE	4.720	4.984	5.247	5.588