Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From sensor rankings to restoration decisions in traffic forecasting

Bailin Xiong

622250991028@mails.cqjtu.edu.cn

Chongqing Jiaotong University

Kai Du

Chongqing Jiaotong University

Jinhui Tang

Chongqing Jiaotong University

Renxiang Wang

Chongqing Jiaotong University

Article

Keywords: traffic forecasting, sensor importance, missing observations, input restoration, proxy evaluation, retrospective analysis

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11493248/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

From sensor rankings to restoration decisions in traffic forecasting

Bailin Xiong*, Kai Du†, Jinhui Tang†, Renxiang Wang†

Chongqing Jiaotong University, No. 66 Xuefu Road, Nan'an District, Chongqing, 400074, China.

*Correspondence: Bailin Xiong. E-mail: 622250991028@mails.cqjtu.edu.cn.

†Kai Du, Jinhui Tang, and Renxiang Wang contributed equally and share second authorship.

Co-author e-mails: Kai Du: 622250991054@mails.cqjtu.edu.cn; Jinhui Tang: 622250991019@mails.cqjtu.edu.cn; Renxiang Wang: 622250991037@mails.cqjtu.edu.cn.

Abstract

Sensor rankings can guide the restoration of disrupted traffic networks, but their value depends on the forecasts produced after selected readings are returned. We evaluate this decision by selecting sources on earlier observations and testing their joint restoration under shared synthetic faults. On PEMS03, the same Linear proxy selections that outperformed variance under Graph WaveNet increased mean absolute error by 0.330 vehicles per five minutes under a diffusion convolutional recurrent neural network (DCRNN). This prespecified primary contrast reversed in all three training seeds, five masks and weekly summaries. The DCRNN protocol was fixed after the Graph WaveNet results and before final output inspection. Rolling PEMS08 retained a Linear advantage under both predictors, with smaller differences against redundancy-aware alternatives. On METR-LA, the mean Linear advantage over variance was 0.000260 mph across three DCRNN seeds, with a conditional interval spanning zero; a separate long short-term memory reference also showed little separation. Exact-deletion and correlation-based policies produced further predictor-dependent outcomes. These results support evaluating source sets under the intended predictor and distinguishing the overall restoration benefit from the incremental benefit of their ranking.

Keywords: traffic forecasting; sensor importance; missing observations; input restoration; proxy evaluation; retrospective analysis

Introduction

Traffic forecasts depend on histories collected across a sensor network. Graph WaveNet (GWN) and diffusion convolutional recurrent neural networks (DCRNN) model the spatial and temporal relationships in these histories [1,2]. When sensors become unavailable and only a limited set can be recovered, source selection determines which information reaches the predictor. The relevant outcome is therefore the forecast error after those readings are restored.

Ranking sources before a fault offers a practical way to allocate a limited restoration budget. A lightweight proxy can score sources through individual deletion, and the highest-ranked sources define a fixed eligible set for later evaluation. Our Linear proxy combines recurrent encoders of values and availability with an additive graph correction; Linear names the correction, while the encoders are nonlinear. Individual deletion and joint restoration expose the predictor to different combinations of information. Agreement with a deletion ranking therefore requires a separate evaluation at the restoration outcome.

Traffic imputation methods estimate missing values using recurrent networks, attention, diffusion models and low-rank spatiotemporal structure [3,4,5,6,7,8,9,10,11]. Graph-based methods exploit relationships among incomplete sensor histories [12,13]. Source selection addresses which histories should be made available to the predictor. To isolate that decision, we return selected originally observed readings under synthetic faults and evaluate the resulting forecasts with fixed predictor weights. This controlled intervention measures the value of access to the chosen histories.

Sensor placement methods select observations for estimation or reconstruction using convex optimization, Gaussian processes and sparse data-driven sensing [14,15,16]. Feature selection considers predictive relevance and redundancy [17,18,19], while correlation-assisted attribution applies related ideas to sensor placement [20]. In restoration, both source identities and fault overlap matter. Two eligible sets of equal size can return different numbers of readings and different amounts of redundant information. A comparison must therefore report forecast error alongside the information actually restored.

Removal-based explanations and model-reliance analyses establish that importance depends on the predictor and intervention [21,22,23]. Dependence among features further affects perturbation scores [24,25]. Retraining-based explanation benchmarks [26], sequential feature acquisition [27,28,29] and proxy-based selection of training examples [30] evaluate different operations. For a fixed restoration budget, the empirical question is how earlier source scores translate into a jointly restored set that improves later forecasts. Whole-ranking agreement captures one stage of this decision.

We evaluate source selection on METR-LA, PEMS03 and PEMS08 using paired faults and shared temporal resampling [31,32]. The original GWN studies relate ranking agreement to absolute restoration outcomes. Three PEMS03 training repetitions and four rolling PEMS08 folds then test policy performance across fitted predictors and evaluation periods. Transferring the fixed GWN-selected sets to DCRNN tests whether their advantage is retained, while DCRNN-specific exact-deletion selection provides a reference-specific comparison. Three DCRNN seeds and a separate long short-term memory (LSTM) reference extend the METR-LA analysis.

The experiments address three aspects of the restoration decision: transfer of fixed source sets between predictors, comparison with redundancy-aware selection, and the size of the ranking increment relative to the total restoration benefit. PEMS03 provides a policy-ordering reversal in the prespecified DCRNN primary contrast. PEMS08 retains a Linear advantage that narrows against redundancy-aware alternatives. METR-LA shows little separation between the leading rankings despite an overall benefit from restored information. Table 1 summarizes the experiments and the role of each comparison.

Table 1. Experimental map. All policies are evaluated by returning eligible originally observed readings under shared synthetic faults. Linear denotes the additive graph-correction proxy defined above. Reference-specific exact selection uses each fitted predictor's own earlier single-source deletion scores. PEMS03 and rolling PEMS08 errors measure flow; METR-LA errors measure speed.

Cohort	Prediction references	Selection policies compared	Role in the argument
METR-LA original	GWN references 42 and 43	Linear, variance, exact; Bounded, gradients and random controls where evaluated	Connect ranking agreement, returned readings and forecast error
PEMS03	GWN and DCRNN; seeds 42–44 for each	Linear, variance, Bounded, exact, gradients, grouped and random policies	Test transfer of fixed source sets across predictor configurations
PEMS08 rolling	GWN and DCRNN; four folds for each	Linear, variance, group-aware variants, exact, Bounded, gradients and random policies	Test temporal consistency and the redundancy-aware comparator
PEMS08 original	GWN reference 42	Linear, variance, grouped variance, exact, gradients and random controls	Quantify source-group coverage and returned-entry differences
METR-LA transfer	DCRNN seeds 42–44; LSTM seed 42	27 inherited policies plus each reference's own exact set	Separate overall restoration gain from the incremental ranking benefit
Post hoc correlation-based selection	18 confirmation, rolling and METR-LA transfer references	Correlation-based mRMR-style and relevance-only; archived policy controls	Test continuous relevance–redundancy selection at the restoration outcome

Results

Earlier source rankings define the sets evaluated for restoration

Each policy is evaluated by returning its selected historical readings to a fixed predictor under the same faults (Fig. 1). Histories and prediction horizons each contain 12 five-minute steps. Source sets are selected in an earlier period and evaluated on later windows. Restoration returns originally available readings from eligible sources that intersect the additional fault. Sources outside that fault retain their existing readings, while values absent from the original availability mask remain missing.



Figure 1. Evaluation of source rankings through joint restoration in the original METR-LA and PEMS08 case studies. (a) Earlier deletion and gradient scores or fitting-period variance define a fixed source-eligibility set P, containing 21 METR-LA sources or 17 PEMS08 columns at the primary budget. Target-dependent scores use earlier labels; variance uses fitting-period observations. (b) The same later history is shown before an additional fault F, after the fault and after selective restoration. Only entries in P ∩ F ∩ M are returned, where M denotes original availability under the base mask. Blue, gray, white and orange cells denote observed, base-missing, fault-removed and returned entries. Waveforms, score bars, ranked tiles and matrix cells illustrate the design schematically. The predictor glyph represents fixed inference. (c) Policies are evaluated with unchanged Graph WaveNet weights and common target eligibility, recording prediction errors, restored entries and paired date differences. Restoration returns originally observed historical inputs under synthetic faults.

Fitting-period variance is the common model-free comparator: it uses earlier observations without querying the reference predictor or fitting a proxy. In the original full-period ranking screen, variance had the highest correlation among ten inexpensive scores for METR-LA reference 42 (0.633335). Linear value-gradient magnitude ranked highest for reference 43 and PEMS08 and is retained as an auxiliary policy (Supplementary Table S11). The Linear–variance comparison therefore provides a consistent fitting-only baseline across experiments, with the other policies showing how its relative strength varies by reference.

The primary condition combines a 10% graph-block fault with a 10% source-eligibility budget: 21 sources on METR-LA, 17 on PEMS08 and 36 on PEMS03. The fault determines how many readings an eligible set actually returns. We record forecast error and returned input occurrences, where one occurrence is one source-by-history-time value returned in a window–mask case. Overlapping windows can count the same historical reading more than once.

Original Linear proxy scores correlated with later GWN deletion losses at 0.657 and 0.656 under the two METR-LA references, and 0.748 on PEMS08. Yet the earlier Linear sets contained only nine and seven of the later METR top 21 sources, compared with 14 of the PEMS08 top 17. Later top sets serve as diagnostics and never select an evaluated policy. Linear and variance shared 16 of 21 eligible sources in METR-LA, but only two of 17 on PEMS08. These observations distinguish ranking association from the particular budgeted decision. Full ranking diagnostics and original data roles are in Supplementary Fig. S9 and Supplementary Table S33.

Linear selection improves GWN restoration on PEMS03 and rolling PEMS08

Linear selection reduced restoration error relative to variance in all three PEMS03 GWN fits (Table 2). The pooled reduction was 0.0521 vehicles per five minutes, corresponding to 0.302% of variance MAE. Exact-reference selection achieved lower error than Linear in each fit. The three training seeds were evaluated on the same 28 day blocks.

Table 2. Graph WaveNet confirmation and rolling evaluation under 10% faults and 10% source eligibility. MAEs and gains are in vehicles per five minutes; positive variance minus Linear favors Linear. PEMS03 rows share 28 evaluation day blocks and differ in training seed. PEMS08 rows use separately fitted predictors and disjoint seven-day evaluation segments with expanding training prefixes. Contrasts use unrounded errors; Supplementary section S19 gives the training and output records.

Evaluation	No restoration	Variance	Linear	Exact-reference	Variance − Linear
PEMS03 seed 42	17.3593	17.2875	17.2369	17.2105	0.0506
PEMS03 seed 43	17.2563	17.1785	17.1141	17.0906	0.0643
PEMS03 seed 44	17.3385	17.2533	17.2120	17.1738	0.0413
PEMS08 fold 1	19.0038	18.9613	18.7737	18.7737	0.1875
PEMS08 fold 2	19.5040	19.4531	19.2194	19.2180	0.2337
PEMS08 fold 3	19.0328	18.9853	18.7305	18.7408	0.2548
PEMS08 fold 4	17.9167	17.8535	17.6914	17.6684	0.1621

Rolling PEMS08 used four disjoint seven-day evaluation segments with expanding training prefixes. Pooled MAE was 18.8 for variance and 18.6 vehicles per five minutes for Linear. The unrounded difference was 0.210, a 1.11% reduction relative to variance, with all four folds favoring Linear (Fig. 2). Leaving out one fold at a time gave 0.194–0.225. The prespecified secondary comparison, variance minus group-aware variance, evaluates grouping within variance selection. The separate post hoc comparison, group-aware variance minus Linear, evaluates Linear against the grouped comparator and gave a pooled advantage of 0.136, or 0.728%. Supplementary sections S15–S19 report this direct comparison separately from the three prespecified contrasts plotted in Fig. 2.

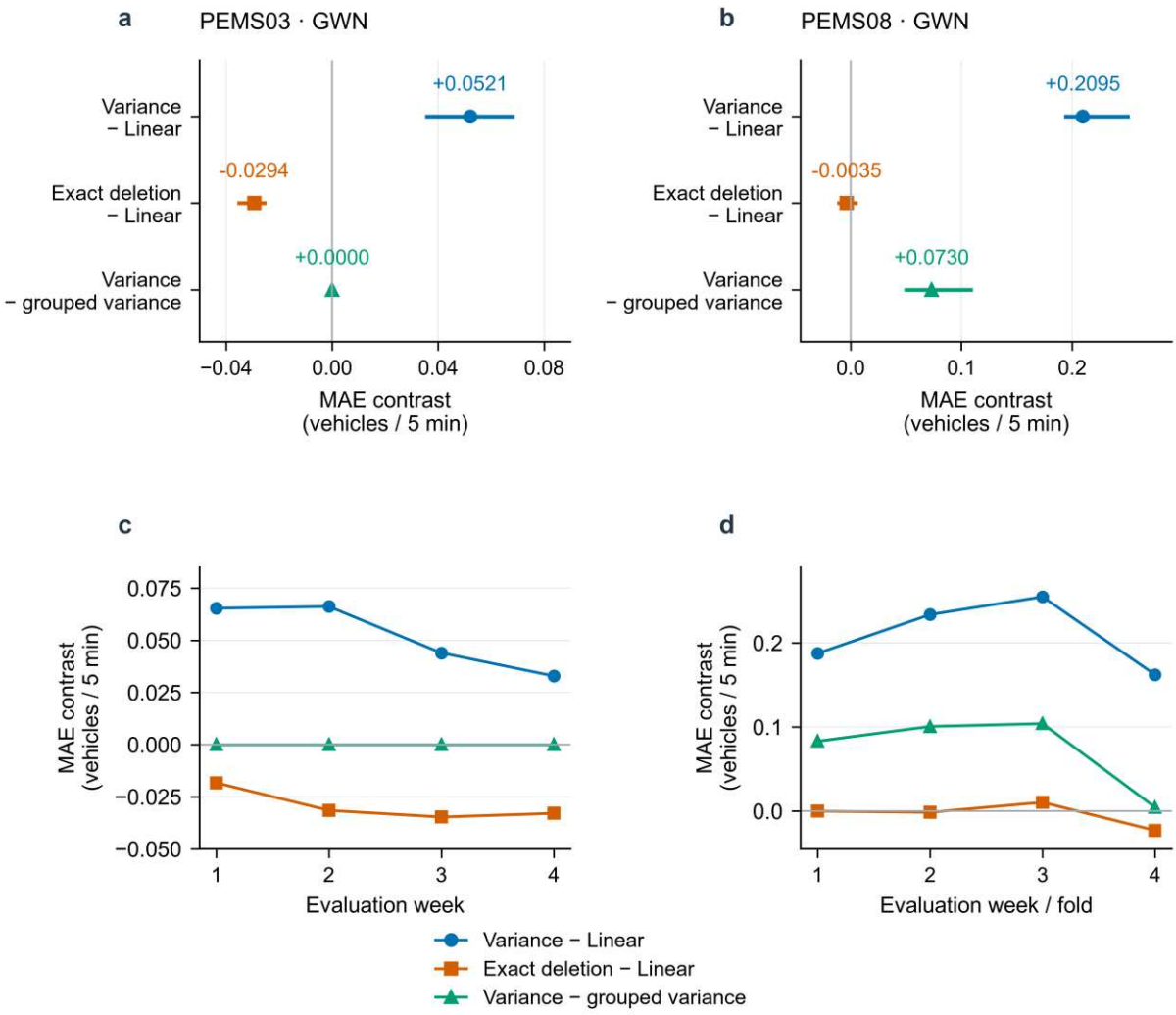


Figure 2. GWN restoration outcomes across PEMS03 fits and rolling PEMS08 folds. Panels a and b show three pooled contrasts prespecified for the added GWN confirmation and rolling protocol: variance minus Linear (primary), exact-reference minus Linear (secondary), and variance minus group-aware variance (secondary). The distinct group-aware variance minus Linear contrast was added post hoc and is not one of the plotted contrasts; its rolling result is reported separately in the text and Supplementary sections S15–S19. Positive values favor the policy named second; in particular, negative exact-reference minus Linear values favor exact-reference selection. Points are estimates and lines are conditional 95% seven-day noncircular moving-block percentile intervals from 10,000 shared draws. Exact deletion denotes exact-reference selection, and grouped variance denotes group-aware variance. Panels c and d show four weekly point contrasts for variance minus Linear, exact-reference minus Linear and variance minus group-aware variance; weeks summarize repeated temporal observations. PEMS03 means weight the three complete fits equally. PEMS08 pools four separately fitted folds over disjoint evaluation weeks and remains retrospectively exposed. Units throughout are vehicles per five minutes. Source data and three-/fourteen-day sensitivities accompany the figure.

A within-fold temporal sensitivity analysis also retained the rolling Linear–variance advantage (Supplementary sections S15–S19). Exact-reference minus Linear pooled to -0.00352, with an interval spanning zero and directions varying across folds. The identical fold-1 sets produced identical outcomes.

Thus, exact single-deletion selection provided little consistent additional benefit over Linear in the rolling evaluation.

Transferring fixed source sets to DCRNN reverses the PEMS03 advantage

The prespecified primary variance-minus-Linear comparison reversed under DCRNN on PEMS03 (Fig. 3). The DCRNN protocol was fixed after the GWN results and before final DCRNN outputs were inspected; the same frozen Linear selections that outperformed variance under GWN produced higher error under DCRNN. Under GWN, variance minus Linear MAE averaged 0.0521 vehicles per five minutes across three complete fits. Under DCRNN, it was -0.330, with a conditional 95% interval of -0.351 to -0.314. DCRNN contrasts were negative in all three fitted repetitions (-0.346, -0.185 and -0.459), every mask within each repetition and every weekly summary. Omitting one week at a time gave -0.336 to -0.318. The reversed ordering is the forecast-error comparison between the two restored source sets.

Table 3. DCRNN restoration outcomes on the original fixed source-eligibility sets and DCRNN-specific exact-deletion sets. Errors are Global MAE in vehicles per five minutes. PEMS03 equally weights three fitted repetitions; PEMS08 pools target numerators and counts over four fixed fold-specific references. Positive variance minus Linear favors Linear. Contrasts are calculated before rounding the displayed errors. Group-aware variance is an alias of variance on PEMS03.

Dataset	Variance MAE	Group variance MAE	Linear MAE	DCRNN exact MAE	Variance minus Linear
PEMS03	24.5092	24.5092	24.8391	23.8370	-0.3299
PEMS08	18.8192	18.7805	18.6643	18.6499	0.1549

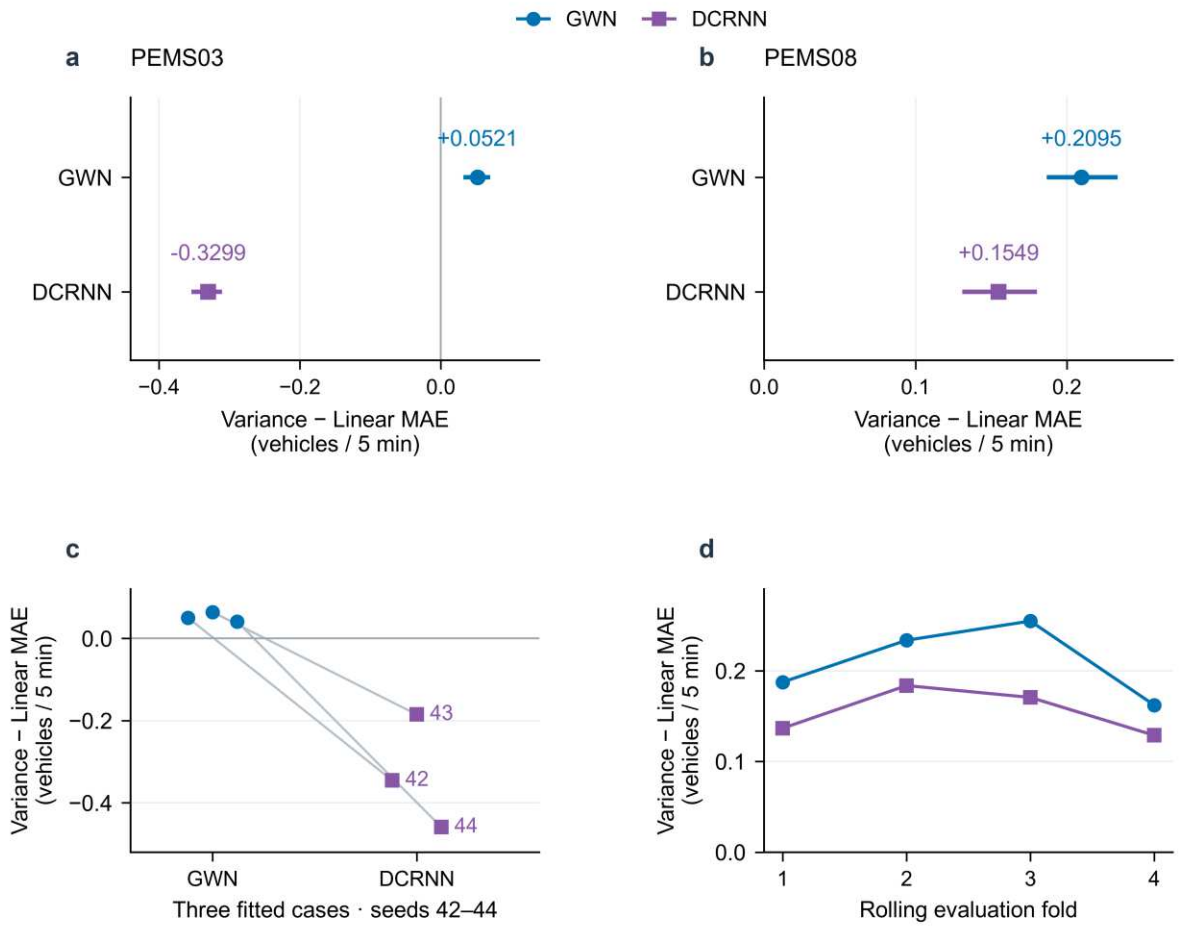


Figure 3. Predictor transfer reverses the PEMS03 ordering and retains the PEMS08 ordering. Panels a and b show variance minus Linear Global MAE on PEMS03 and rolling PEMS08. Positive values favor Linear. GWN circles and DCRNN squares use the same day-index draws within each dataset. Lines are conditional 95% percentile intervals from 10,000 shared draws: circular seven-day blocks on PEMS03, and independently sampled circular three-day blocks within each seven-day PEMS08 fold. The GWN intervals are the separately labeled common-index reaggregation, not the historical noncircular intervals retained in Fig. 2. Panel c shows the individual PEMS03 contrasts for seeds 42–44, with neutral lines connecting the cases with corresponding seed labels across references. Positions within each architecture are offset only for legibility. These fits share the same temporal observations. Panel d shows all four PEMS08 fold-level point contrasts; these are dependent temporal summaries. PEMS03 equally weights three fitted repetitions, while PEMS08 pools valid-target counts across four realized folds. Units are vehicles per five minutes. The proxy-selected source sets were transferred unchanged; reference architecture and fitted weights changed together. All estimates and intervals are taken from the saved cross-architecture analysis.

DCRNN-specific exact-deletion selection achieved lower PEMS03 MAE than both transferred policies (Table 3), improving on Linear by 1.00 vehicles per five minutes. Its sets shared only one or two of 36 sources with the inherited GWN exact sets. The exact result evaluates a newly selected, reference-specific set; the Linear–variance comparison evaluates unchanged inherited sets. Both source identities and realized returned entries differ between policies.

PEMS08 retained a Linear advantage. Under DCRNN, variance minus Linear was 0.155 (conditional interval 0.132–0.179), and the separately reported group-aware variance minus Linear contrast was 0.116

(0.106–0.127). Both directions held in all four folds and all five masks within each fold. DCRNN exact selection had lower pooled MAE than Linear by 0.0143, but fold 2 favored Linear by 0.00373. Pooling therefore conceals an exception to exact selection's direction.

With shared day-index draws, the DCRNN-minus-GWN change in the variance-minus-Linear contrast was -0.382 on PEMS03 (-0.410 to -0.357) and -0.0546 on PEMS08 (-0.0638 to -0.0455). These are paired configuration comparisons: architecture and fitted weights changed together. All seven references and their source sets were jointly frozen before final DCRNN evaluation. The unchanged policy files, absolute errors and fold/mask results are retained in Supplementary section S20.

A post hoc blockwise policy-label exchange analysis used a common sign for all fits and masks within each temporal block. The original seven-day dependence scale gave four weekly blocks and 16 exact sign patterns, yielding a two-sided reference $p = 0.125$ under blockwise sign invariance. Supplementary section S27 reports this analysis and the shorter-block sensitivities alongside the original effect estimate and temporal interval.

The PEMS03 reference configurations also differ in absolute accuracy. Mean no-restoration MAE is 17.3180 for GWN and 25.2597 for DCRNN (45.9% higher); DCRNN validation MAE was already 23.6% higher before the additional fault was introduced. A saved-output audit found that the pooled policy reversal survived omission of any single fault anchor and remained in each of four common window groups stratified by GWN no-restoration error. Realized returned entries further organize the contrast: both predictors favor Linear when it returns more entries and variance when variance returns more. In the equal-entry subset, variance minus Linear remains positive for GWN (0.0270) and negative for DCRNN (-0.108). These post hoc strata retain the original fitted configurations and their accuracy differences (Supplementary section S26).

METR-LA restoration gains exceed the incremental ranking benefit

In the original METR-LA GWN reference-42 case, returning Linear-selected histories reduced MAE from 3.23 to 3.20 mph. The unrounded gain over no restoration was 0.0305 mph, with the original paired-date interval of 0.0244–0.0380. Its smaller increment over variance was 0.000868 mph (-0.000517 to 0.00237); reference 43 gave 0.00110 mph (-0.000218 to 0.00244). Restoration helped these saved models, while the incremental choice between the two rankings remained uncertain (Supplementary Table S43).

Exact-reference selection did not consistently improve METR-LA restoration (Supplementary Fig. S10). Against Linear, its gain was -0.000823 mph for METR reference 42 (-0.00203 to 0.000342) and 0.00270 mph for reference 43 (0.000785–0.00443). The latter changed sign when the last saved checkpoint replaced the best checkpoint. Proxy-gradient selection also had higher point error than Linear under both METR references.

Additional METR-LA predictors show little separation between Linear and variance

The METR-LA extension evaluated the original 27 fixed reference-42 policies under three DCRNN training seeds and one LSTM reference. Each fitted predictor also supplied its own exact-deletion set. Evaluation retained the same 3,260 earlier selection windows, 3,042 later windows, five masks and 21-source budget. The original DCRNN seed 42 fit was followed by seeds 43 and 44; LSTM supplied a separate recurrent predictor without graph operations. The three additional predictors and their source sets were frozen before their final evaluation.

The DCRNN Linear advantage over variance was 0.000377, 0.0000724 and 0.000330 mph for seeds 42,43 and 44. Their equal-weight mean was 0.000260 mph, with a conditional 95% two-date block interval of -0.000941 to 0.00199 mph. Six of twelve date-level mean contrasts favored Linear, and leave-one-date-out estimates ranged from -0.000202 to 0.000600 mph. The additional fits therefore reproduce a small point difference while retaining uncertainty about the policy ordering on this cohort.

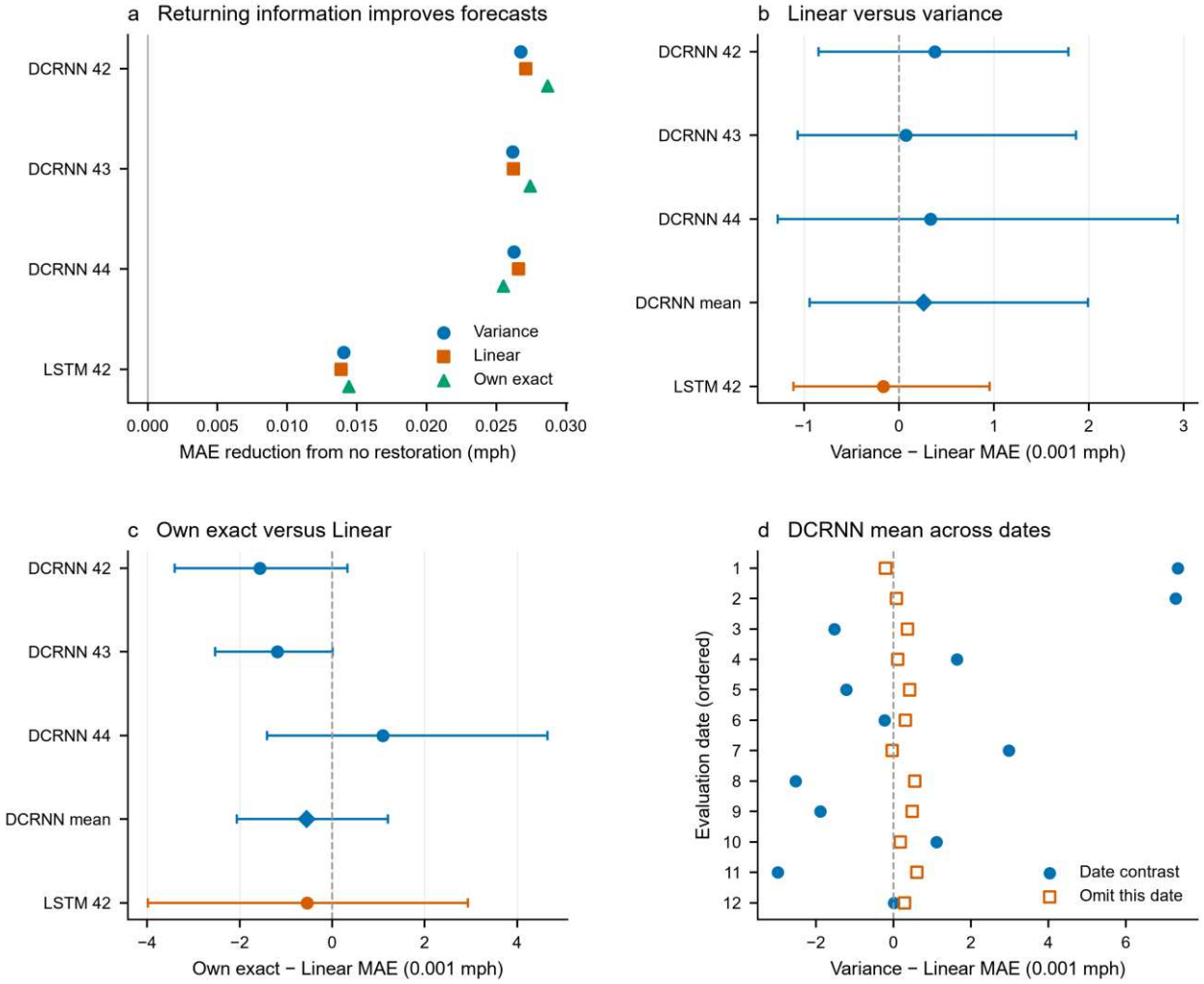


Figure 4. Restoration benefit and incremental ranking differences on METR-LA. Panel a shows saved Global MAE reductions from no restoration for variance, Linear and reference-specific exact selection; points are arithmetic differences without uncertainty bars. Panels b and c show variance minus Linear and own-exact minus Linear MAE. Lines are conditional 95% percentile intervals from 10,000 shared circular two-date draws over the twelve observed dates; positive values favor Linear in both panels. DCRNN mean equally weights seeds 42–44 within each draw, with the individual seeds retained. LSTM 42 is a separate single reference. Panel d shows each date-level DCRNN mean contrast and its leave-one-date-out estimate. Dates are ordered as in the processed archive and are repeated temporal summaries. Panel a uses mph; panels b–d use 0.001 mph. All fits use the same inherited source sets, with reference-specific exact sets selected earlier. LSTM has 943,028 parameters and DCRNN 374,273. Absolute errors, all 28 policies and mask results are in Supplementary section S23.

Absolute outcomes separate this comparison from the effect of returning information. Without restoration, DCRNN MAE was 3.20, 3.20 and 3.19 mph for seeds 42–44; with Linear restoration, it was 3.17, 3.17 and 3.16 mph. The LSTM likewise improved from 3.66 to 3.65 mph with Linear restoration. Its variance-

minus-Linear contrast was -0.000165 mph (-0.00111 to 0.000955), slightly favoring variance at the point estimate. Both policies returned the same respective entry counts across references: 264,633 for Linear and 266,610 for variance. The difference between policies is thus much smaller than the overall change from no restoration.

Reference-specific exact selection adds a second comparison. For DCRNN seeds 42 and 43, exact selection improved on Linear by 0.00156 and 0.00118 mph; seed 44 instead favored Linear by 0.00109 mph. The mean exact-minus-Linear contrast was -0.000552 mph (-0.00206 to 0.00120). Relative to the transferred GWN exact set, the mean DCRNN exact contrast was -0.000965 mph (-0.00177 to 0.000101), with seed 44 again reversing direction. Under LSTM, exact selection had slightly lower error than Linear but higher error than the transferred GWN exact set. Recomputing importance for the intended predictor therefore changes the selected decision without guaranteeing a uniform improvement in joint restoration. Supplementary section S23 reports full absolute results, per-seed and per-mask estimates, date summaries and omissions.

Broader source-group coverage narrows the PEMS08 comparison

In the original PEMS08 case, Linear selection achieved MAE of 18.2 vehicles per five minutes compared with 18.5 for variance. The unrounded advantage was 0.283 (0.257–0.310), or 1.53%. Linear returned 129,432 input occurrences and variance returned 136,032. The lower Linear error was therefore achieved with fewer returned occurrences; these counts quantify the realized information returned under each policy.

PEMS08 exact selection produced higher error than Linear by 0.0142 vehicles per five minutes (0.0111–0.0178); proxy-gradient selection also had higher point error. Thus, exact individual-deletion scoring did not identify the better joint-restoration set in this original case.

Grouping identical fitting-period sequences broadened variance selection from ten represented groups to 17. This group-aware policy reduced MAE by 0.0889 (0.0729–0.107), narrowing Linear's advantage by 31.4% to 0.194 (0.179–0.209). It also returned more entries than either original policy (Fig. 5). Grouping therefore changed both source representation and realized information access. Applying the same rule to Linear left its set and outcomes unchanged. Supplementary Table S34 defines the policy inputs, and Supplementary Table S3 retains the original R1–R7 expectations and outcomes.

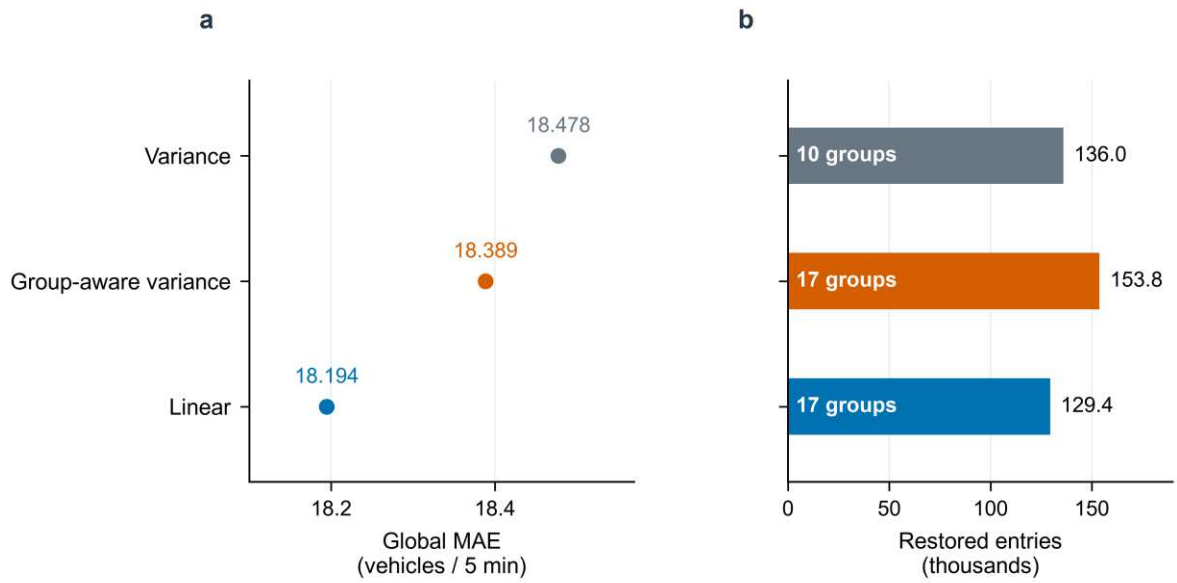


Figure 5. Group-aware variance broadens source coverage and narrows the PEMS08 comparison. (a) Absolute Global MAE for original variance, group-aware variance and Linear. (b) Returned input occurrences, shown with zero-based bars. Labels give the number of distinct fitting-period sequence groups represented in each 17-column eligible set. Group-aware variance represents 17 groups rather than ten, has lower error and returns more entries. These changes jointly compare source identities and realized information access. Applying the group-aware rule to Linear produces the same set and outcomes; this duplicate is omitted.

Correlation-based redundancy penalties have different effects across predictors and datasets

We added a post hoc correlation-based mRMR-style policy to test redundancy beyond exact duplicate groups. Fitting-period correlations define relevance to future network readings and redundancy with already selected sources; a relevance-only policy removes the redundancy penalty. Six source-set pairs were frozen and evaluated on 18 saved predictors, using the same five masks and 36/17/21-source budgets for PEMS03/PEMS08/METR-LA. This extension adds 180 policy–mask outcomes without refitting predictors or proxies (Supplementary Tables S44–S47).

On rolling PEMS08, correlation-based selection reduced MAE relative to variance by 0.0431 vehicles per five minutes under GWN and 0.0178 under DCRNN. Both retained higher error than Linear, by 0.166 and 0.137, respectively, and higher error than group-aware variance, by 0.0299 and 0.0209. Removing the redundancy penalty increased error by 0.0801 under GWN and 0.0523 under DCRNN. Thus, accounting for correlated sources improved this fitting-only baseline on PEMS08 while preserving the Linear advantage.

PEMS03 and METR-LA showed a different response. On PEMS03, the correlation-based policy had almost the same GWN error as variance (difference 0.000217), but its DCRNN error was 0.292 higher than variance. Relative to Linear, it had higher GWN error by 0.0523 and lower DCRNN error by 0.0384. The redundancy penalty also increased DCRNN error over relevance-only selection by 0.649. On METR-LA, its mean error exceeded Linear by 0.0174 mph across three DCRNN seeds and by 0.00898 mph under LSTM; it also exceeded variance and relevance-only selection. The additional comparator therefore changes the size and ordering of individual comparisons without yielding one preferred selection rule across these predictor–cohort combinations. Per-fit absolute errors, realized returned entries, set overlaps

and conditional temporal intervals are retained in Supplementary Tables S44–S47 and Supplementary Data 1.

Discussion

The restoration value of a sensor ranking depends on the predictor that receives the selected readings. The fixed Linear and variance sets reversed their ordering between GWN and DCRNN on PEMS03, retained the Linear ordering on rolling PEMS08, and produced much smaller differences on METR-LA. Evaluating the jointly restored sets connects these outcomes to the budgeted decision and reveals distinctions that whole-ranking agreement leaves unresolved.

PEMS03 provides the central finding. The same fixed source sets favored Linear under GWN and variance under DCRNN in the prespecified primary contrast, with a consistent direction across seeds, masks and weekly summaries. This extends established observations about predictor-dependent importance [21,23] to a budgeted joint-restoration decision. The saved-output analysis further locates the reversal: it persists across common error strata, after removal of any individual fault anchor, and within the subset where both policies return the same number of entries. The tested DCRNN configurations also have higher validation and final error. Architecture, fitted weights and predictive accuracy therefore define the configuration being compared; accuracy-matched transfer remains a separate question.

PEMS08 shows why the comparator matters. A fitting-only grouping rule expands variance selection from ten identical-sequence groups to 17 and narrows Linear's advantage. The correlation-based mRMR-style policy extends this analysis to graded redundancy: it improves on variance and relevance-only selection under both reference families, while Linear and group-aware variance retain lower error. Under PEMS03 DCRNN and the added METR-LA references, the same correlation penalty increases error relative to relevance-only selection. These outcomes connect relevance–redundancy criteria [18,19,20] to the downstream forecast: a more diverse source set can help, but its value depends on the information the fitted predictor uses after restoration.

METR-LA separates the overall benefit of restoring readings from the incremental benefit of their ranking. The original GWN case gains 0.0305 mph over no restoration, whereas the Linear–variance difference is below 0.001 mph. Three added DCRNN fits and a separate LSTM reference preserve this distinction, with temporal intervals spanning zero for the direct comparison. Exact single-source deletion also changes its restoration ordering across references. Thus, a useful restoration intervention can coexist with limited evidence that one ranking should determine the eligible set.

These findings suggest a practical evaluation sequence. Candidate sets are defined on an earlier fitting or selection period and compared under common faults with the intended predictor. Forecast error and returned readings are reported together, and the comparison is repeated when the predictor changes. Identical or strongly correlated fitting histories motivate redundancy-aware candidates. The source-count budget fixes eligibility; the realized fault determines how much of that eligible information becomes available.

The intervention returns originally observed histories under synthetic faults, providing a controlled measure of access to selected information. Operational extensions can combine selection with imputation error, device-repair costs and a minimum useful forecast improvement. The present temporal intervals condition on saved predictors and fixed source sets. Three PEMS03 seeds share 28 observed days; rolling PEMS08 folds share training prefixes; the later METR-LA cohort contains twelve dates and only one LSTM fit. The original METR-LA and PEMS08 branches had development exposure. These features

define the evidence available for transfer across configurations and time, while new periods and complete-pipeline refits would test broader generalization. The direct grouped-variance contrast, correlation-based policies and saved-output diagnostics are identified as post hoc analyses in their methods and supplementary records.

Source rankings become useful restoration policies when their selected sets improve the forecasts that follow restoration. The PEMS03 reversal, retained PEMS08 advantage and small METR-LA ranking increments identify three different outcomes of that decision. Direct evaluation under the intended predictor makes these differences visible before a ranking determines how a limited source budget is allocated.

Methods

Data preparation and temporal roles

The original analyses used archived METR-LA speed and PEMS08 flow releases associated with the traffic-forecasting benchmarks [1,2,33]. Their historical preprocessing and roles are described first; the additional PEMS03 and rolling PEMS08 protocol is specified separately below. Each sample uses 12 five-minute history steps to predict 12 future steps, with stride one. Roles are assigned to original time indices before windows are constructed: every future target must fall within its assigned role, although its history can precede the role's first target. Parameter fitting, training-tail calibration, development validation and historical analysis are separate roles. Calibration is not used to fit neural parameters; validation selects checkpoints. The original exclusive boundaries, counts, fitting scalers and source records are retained in Supplementary section S17.

METR-LA readings are valid when finite and nonzero. Excluding 531 historical windows with no valid targets leaves 6,313 single-source windows. PEMS08 has 17,856 time steps and 170 published columns; zero flow is valid, and 3,561 historical windows enter its single-source analysis. Published repeated columns, interpolated values and graph structures are retained. Validity conventions determine which readings enter the analysis. Speed and flow effects are reported separately in their respective units.

Joint selection and evaluation divide the historical cohort by the encoded date of the first future target. Excluding 11 windows whose targets cross the boundary in each dataset gives 3,260 earlier and 3,042 later METR-LA windows, and 1,545 earlier and 2,005 later PEMS08 windows. The later periods cover 12 and seven dates, respectively. This boundary prevents later targets from entering the particular set-selection calculation; all periods nevertheless had previous development exposure. Supplementary Tables S4, S9 specify dates and denominators.

Fixed predictors and training records

The proxy uses node-independent value/mask gated recurrent encoders with hidden dimension 32, node/calendar embeddings, a local head and a graph-output head. The local prediction incorporates the last available reading or an empty-history state. A one-hop additive graph correction uses receiver-normalized prior weights after self-loop removal, keeps at most eight sources per receiver, and is not renormalized after truncation or deletion. Each source coefficient is multiplied by its available-history fraction. Linear uses coefficient-weighted source-minus-receiver graph-head outputs; Bounded applies $g(d) = \tau \tanh(d/\tau)$, with $\tau = 2$ in standardized prediction units, before multiplying by the graph coefficient. Linear names the additive graph correction within the recurrent proxy.

Node-independent states and the fixed additive correction allow reuse of proxy computations after source deletion. Both the source prediction and its receiver contributions change; joint absolute losses are computed from the resulting predictions. The METR proxies have 6,637 parameters and the original PEMS08 Linear proxy has 6,489. Reference-model deletion uses explicit masked-input inference.

Graph WaveNet uses masked standardized values, an availability channel, and the last historical hour and weekday scaled by 23 and 6, respectively. Calendar channels remain available under faults. The adapter supplies receiver-normalized forward and reverse prior supports and adaptive adjacency [1]. Standardized values are zeroed only after availability is determined. METR references have 309,464 parameters. PEMS08 uses its own fitting scaler, 170 node embeddings and graph prior with the same adapter construction. Exact model constructors, training configurations and checkpoint hashes for the six original fitted predictors are mapped in Supplementary section S13.

The original METR reference-42 and reference-43 fits selected epochs 54 and 52 after completing 69 and 67 epochs through patience stopping. Initialization and dropout randomness differed, while training-order and base-mask seeds were shared. The original proxies and PEMS08 GWN reached the 96-epoch cap; selected epochs, stopping states and complete trajectories are recorded in Supplementary sections S11, S13 and S17. Added restoration controls use the saved weights.

Missingness, signed deletion scores and ranking

Persistent base masks remove a fixed source count throughout each history. A stateless hash of window start, source identifier, mask seed and epoch argument makes assignments independent of batch partitioning. At 30% missingness, the exact counts are 62 of 207 METR sources and 51 of 170 PEMS08 columns. Seeds 1001–1005 define five fixed conditions. Synthetic masks intersect native availability, while future-target eligibility stays unchanged.

For predictor f, E_f(j) is the valid-target absolute-error numerator after deleting source j in addition to the base mask, E_f(0) is the base-mask numerator, and C is the common target count. The Global deletion effect is

$$D_f(j) = \frac{E_f(j) - E_f(0)}{C}$$

Negative effects are retained. Own restricts targets and denominator to source j; Other uses the remaining targets and is not defined as physical downstream influence. Sources are eligible if a history is available at least once under the scoring masks. Reference deletion maps use actual masked-input inference with fixed weights in evaluation mode.

The original ten inexpensive scores include four graph-topology scores, fitting variance, missing rate and mean strength, persistence damage, and signed/magnitude Linear value gradients (Supplementary Table S1). Reference gradients, constant scores and random policies are auxiliary checks. Gradients change values while keeping availability fixed, whereas complete-history deletion changes both. Target-dependent scores are used offline on the earlier selection period. Spearman correlation uses average tied ranks over common eligible sources; constant-score correlations remain undefined. Sets use descending scores and source identifier to break ties, with size ceil(budget rate times source count). Twenty fixed random orders define both budget prefixes.

Shared faults and restoration policies

An additional fault block begins at an anchor determined by window start, mask seed and fault seed. Breadth-first traversal uses the undirected support of the maximum of the prior and its transpose, visiting neighbors by decreasing weight and then identifier. If needed, traversal resumes at the smallest unvisited source; no disconnected fallback occurred in the retained blocks. The first ceil(fault rate times source count) sources form F. These synthetic blocks follow the supplied graph support.

Let M combine native availability with the persistent 30% base mask, and let P be a globally fixed restoration-eligible set. Broadcasting source membership across history steps gives

$$M_P = M \wedge (\neg F \vee P)$$

Empty P gives no restoration. A policy reverses the additional fault within P while preserving entries already absent from M. Faults, histories, target eligibility and error denominators are shared across compared policies. The primary fault/budget rates are 10%/10%; secondary rates are 5%/5%, 5%/10% and 10%/5%. Source counts are 11/21 for METR and nine/17 for PEMS08. Restored occurrences count additional available entries relative to no restoration, including repeats across windows and masks.

Exact-reference sets use the corresponding GWN's earlier deletion scores; proxy-gradient sets use earlier Linear value-gradient magnitude. For PEMS08 group-aware selection, exact sequence equality over the first 11,249 fitting steps defines groups. Scanning the original score order skips represented groups and stops at 17 columns. The same rule applies to Linear and variance; later values do not define these sets, and no predictor column or graph edge is removed. Greedy joint-proxy selection, retained under the original policy label HJ, instead evaluates actual joint proxy predictions at each of 21 greedy steps on the earlier METR cohort. Its pre-existing continuation rule requires a gain of at least 0.003 mph over the best of Linear, variance and Bounded and a positive lower date-interval endpoint. That threshold is the continuation rule for this computational branch (Supplementary section S4).

Error aggregation and uncertainty

Global MAE divides pooled absolute-error sums by pooled valid-target counts. Primary denominators are 35,845,775 METR targets and 20,451,000 PEMS targets, shared across policies. Failed-source MAE uses the original fault set even after restoration. The sensor-tail metric averages the highest ceil(5% of valid sources) source MAEs per window and then averages window–mask occurrences; its units of aggregation differ from target-pooled MAE (Supplementary section S7). Gain is control MAE minus candidate MAE. Relative gain divides that pooled difference by control MAE. Random-policy MAEs are computed separately and then averaged equally over 20 policies.

Masks are pooled within each encoded first-target date. Paired date resampling repeats both numerators and denominators, using common draw indices within a comparison. For the original GWN case studies, we report the 2.5th and 97.5th percentiles of 2,000 draws. The subsequent GWN confirmation/rolling, DCRNN transfer and additional METR-LA analyses use 10,000 draws, as specified in their protocol subsections; the two draw counts belong to different analysis stages. The original case studies use paired date resampling, with circular two-date blocks as a sensitivity to short temporal dependence. These intervals condition on the saved predictors and sets; they are descriptive, unadjusted for multiple comparisons and do not measure independent-training uncertainty. For contrasts against the strongest inexpensive score or best HJ comparator, that comparator is reselected within each draw. Zero-denominator draws and undefined correlations are not replaced by zero.

Original METR reference-42 intervals retain their original draws. Later common-date reaggregations are labeled separately. Resampling seeds and draw files are specified in Supplementary sections S2, S14–S21 and S23–S27, allowing each reported interval to be traced to its analysis stage.

Secondary diagnostics and numerical reproducibility

Earlier/later ranking diagnostics pool saved signed deletion numerators over the exact original starts, five masks and corresponding target counts. Reconstructed Linear, variance and available Bounded sets match the original sets at both budgets. Later top sets are used only to describe correspondence. Realized-source distributions use saved active-source fields for METR; PEMS restored-read counts divided by 12 yield integer source counts, checked in all retained batches. Group-weight sensitivity separately uses full-period and fitting-prefix exact sequence groups and weights each source target by the reciprocal of group size. Inputs, graph, sets and predictions do not change. Full-period grouping is explicitly post hoc. Partial-rank diagnostics are descriptive residual-rank associations; the nominal seven-static-control PEMS design has six nonconstant directions plus an intercept (Supplementary sections S1/S9/S10).

Added restoration inference uses all original later windows and masks with fixed 32-slot batches; short batches repeat their final case, and padded rows are excluded. Dropout is disabled and BatchNorm uses saved statistics. The fixed-shape evaluation reproduced original Global MAEs within 0.000000001 dataset units and matched restoration counts exactly. Supplementary sections S14–S16 retain the numerical checks, including an earlier variable-batch attempt that exceeded the fixed 0.0001-unit sensor-MAE tolerance.

GWN confirmation and rolling evaluation

The additional protocol was fixed before final evaluation, following a local exposure audit and the authors' statement of no known relevant PEMS03 use. PEMS03 contained 26,208 five-minute rows and 358 columns. D1–42, D43–47, D48–54, D56–62 and D64–91 served as fitting, reserved calibration, validation, selection and evaluation roles, respectively; D55 and D63 were gaps. Calibration rows were not used to fit parameters or select policies. A joint hash record froze the three sets of best weights and external policies before final PEMS03 decoding. The archived freeze records identify the protocol and selections in effect before final decoding. PEMS08 retained its previously exposed status. Its four expanding folds and all inclusive day ranges are listed in Supplementary Table S13.

Both history and targets were fully contained in each new temporal role. Selection used first-target indices divisible by 12; evaluation used every valid start. The added protocol thus differs from the original target-contained windows and denser selection cohort; old and new outcomes are not pooled. Feature 0 was used, zero flow remained valid and scalers were fitted only on training values. No imputation was added. The directed Gaussian cost graph used the archived local converter and PEMS03 sensor-ID order, with construction details retained in the computational records. Calendar channels were reconstructed from nominal benchmark starts; temporal roles are reported as encoded day blocks.

Each case fitted GWN, Linear and Bounded from initialization. Adam used learning rate 0.001, weight decay 0.0001 and gradient clipping at 5; validation controlled checkpointing and a factor-0.5 plateau scheduler with patience 5 and minimum learning rate 0.00001. The raw-unit checkpoint improvement threshold was 0.00001. Training allowed at most 192 epochs, with a minimum of 30 and stopping patience of 15. Four fixed missingness views (10%, 20%, 30%, 50%) were used per training window, with 16 windows for GWN and 32 for proxies. Seed-specific initialization, masking and epoch order were repeated for PEMS03 seeds 42–44; PEMS08 used seed 42 per fold. Full states were retained across

planned stage boundaries. Table S16 records each stopping endpoint and whether it reached the epoch cap.

All added restoration comparisons used five fixed masks (1001–1005), 30% persistent missingness, 10% graph-block faults and 10% source eligibility. Policies included no restoration, Linear, Bounded, variance, Linear-gradient, exact-reference deletion, group-aware Linear/variance and 20 fixed random orders. Scores used only their assigned fitting or selection role. No later source reselection was allowed. The primary endpoint was Global MAE and the primary contrast was variance minus Linear; exact-reference minus Linear and variance minus group-aware variance were required secondary contrasts. PEMS03 pooled each seed's numerator and denominator before equally averaging its MAE; rolling PEMS08 pooled target counts across folds. Paired noncircular moving blocks of seven encoded days were sampled 10,000 times, with three- and fourteen-day sensitivities. Common draws were shared across policies and PEMS03 seeds. Conditional percentile intervals use the saved fitted models and source sets. Weekly and leave-one-week-out summaries describe concentration across the four observed weeks. The later blockwise label-exchange procedure is specified separately below.

The direct rolling PEMS08 group-aware-variance-minus-Linear contrast and temporal-boundary sensitivities were added after inspection of the completed results. They use shared policy-specific numerators and denominators with the original noncircular draws, concatenated circular blocks, and independently sampled circular blocks within each fold. Within-fold sampling preserves seven day slots per fitted configuration; length seven includes every day once. Supplementary Tables S18–S20 give all algorithms, seeds, block-length sensitivities and marginal day-inclusion weights.

All seven added GWN evaluations were repeated from raw benchmark files in a separate directory and fresh processes using the frozen inference sources, weights and policies in the same installed environment. Identity fields and counts matched exactly, and maximum per-window error differences were zero. A separate CPU reaggregation reproduced the policy summaries and intervals (Supplementary section S19).

DCRNN source-set transfer and paired cross-architecture analysis

The DCRNN protocol was frozen before its outputs were inspected and after the GWN results were known. Three PEMS03 seeds (42–44) and four PEMS08 rolling folds (seed 42) inherited the exact raw inputs, temporal roles, 12-step history/horizon, mask seeds 1001–1005, graph-block fault seed, 30% background missingness, 10% fault and 10% source-eligibility budget. Only the reference predictor was refitted; the original proxy and inexpensive policy sets were transferred unchanged. DCRNN-specific exact single-source deletion scores used only the same earlier selection windows, with the old exact-GWN set retained as an auxiliary transfer policy, yielding 29 labels including aliases.

The local DCRNN adaptation uses two 64-unit diffusion-GRU layers, diffusion order two, curriculum decay 2000 and 374,273 registered parameters. Its four channels provide masked values, availability, last-history hour and weekday, with no future labels available at inference. Pinned TensorFlow and PyTorch source commits and licenses are supplied; the local missing-input adaptation and its source versions are documented in Supplementary section S20.

Adam, weight decay, clipping, plateau scheduling, raw-unit improvement threshold, minimum 30 epochs, maximum 192 epochs and patience 15 match the frozen plan. Each optimizer update aggregates 16 windows and four missingness views using valid-target-weighted microbatches of four windows. Computation uses float32 with TF32 disabled. Full model, optimizer, scheduler and random states are retained at complete-epoch continuations. Mask 1001 alone selects the validation checkpoint; five-mask

diagnostics never repick it. Each reference must outperform same-mask last-observed persistence before final evaluation.

After all seven references and source sets were jointly frozen, all 29 policy labels were evaluated on five common masks, followed by all seven cases in a fresh directory and fresh CUDA processes in the same installed environment. Every decompressed per-window replay output matched its original exactly; the replay checks numerical consistency within that environment. Cross-architecture comparisons matched mask/start/day identities and valid/failed target denominators.

MAE pools raw error numerators over valid targets, then equally averages PEMS03 fits; rolling PEMS08 pools four realized fold segments. All policies, fitted seeds and architectures share 10,000 day-index draws.

The primary PEMS03 interval uses circular seven-day blocks over 28 days (seed 2026101910); primary PEMS08 uses circular three-day blocks independently within each seven-day fold (seed 2026101911), never crossing a fitted-model boundary. Circular 3/7/14-day concatenated sensitivities, within-fold seven-day degeneracy, and every fold/leave-one-fold-out point are retained. Fourteen-day within-fold blocks are inadmissible. No interval was selected by its favorable direction.

Historical GWN intervals are retained, and GWN reaggregation using the shared DCRNN indices is labeled separately. An independent day-multiplicity calculation reproduced the new intervals within 1e-10 units. Supplementary section S20 provides the complete resampling and replay records; S27 specifies the subsequently added label-exchange analysis.

Initial METR-LA DCRNN transfer

The later METR-LA extension fitted one additional DCRNN with seed 42 on the original reference-42 fitting, calibration and validation roles, using the original fitting-only scaler and finite-nonzero-speed validity rule. Exact boundaries, counts and scaler values are retained in Supplementary section S21. The DCRNN constructor and training settings matched the seven-case adaptation above. The selected epoch-74 checkpoint passed the mask-1001 persistence gate; five-mask validation diagnostics did not reselect it.

All 27 original METR reference-42 sets were inherited unchanged. DCRNN-specific exact deletion used all 3,260 original earlier windows and five masks, without the stride-12 subsampling used in the PEMS extension. The new reference and all 28 sets were jointly frozen before final evaluation. Five-mask inference and separate-directory CUDA inference used the same 3,042 later windows, encoded first-target dates, target validity and source budget of 21. All five decompressed per-window replay outputs matched exactly in the same installed environment. This initial extension introduced no group-aware METR policy; the subsequent reference seeds are described below.

For this extension, 10,000 paired circular two-date draws over the twelve observed dates were shared across policies and between DCRNN and the archived GWN reference-42 sufficient statistics. The primary contrast was variance minus Linear; exact-DCRNN minus Linear and exact-DCRNN minus transferred-GWN-exact were additional contrasts. Paired differences between architectures were formed within each draw. Ordinary-date reweighting and every date and leave-one-date-out point were retained. These percentile intervals condition on one fit per reference configuration, fixed source sets and the twelve dates; they do not quantify refitting uncertainty. Historical GWN intervals remain unchanged. Supplementary section S21 reports the draw seeds, all policy rows and sensitivities.

Additional METR-LA training seeds and LSTM reference

The subsequent extension added DCRNN training seeds 43 and 44 and one LSTM seed 42 on the same original METR-LA reference-42 roles. The DCRNN construction, training objective, optimizer, stopping rule, window definitions and validity mask were unchanged. A training seed changes initialization, training-window order and training masks together. The original DCRNN seed 42 and all 27 inherited policies remained unchanged.

The LSTM encodes 12 historical steps using two unidirectional layers with 128 hidden units and zero dropout. Its input at each step concatenates 207 sources and four packed channels (masked standardized value, availability, historical hour and weekday), yielding 828 input features. The final encoder output passes through a linear head to 12×207 outputs, reshaped to the common prediction format. It uses no graph in the predictor, future labels, teacher forcing or decoder curriculum; the original graph still defines the same fault. The installed PyTorch defaults initialize the model. The LSTM has 943,028 parameters versus 374,273 for DCRNN; parameter matching was not performed. Training uses the same valid-target-weighted missingness views, effective batch of 16 with microbatch 4, Adam at 0.001 with weight decay 0.0001, gradient clipping at 5, plateau decay factor 0.5 with patience 5 and minimum learning rate 0.00001, and the original minimum 30/maximum 192 epochs and early-stopping patience 15.

LSTM stopped at 47 complete epochs and selected epoch 32. DCRNN seeds 43 and 44 stopped at 99 and 93 epochs and selected epochs 84 and 78. Validation-gate and five-mask earlier selection preceded joint freezing of all three new references and their 28 policy labels. Final evaluation and separate-directory same-environment CUDA inference covered 15 case–mask combinations each; decompressed per-window outputs matched for every pair. The original common 10,000 circular two-date and ordinary-date resampling indices were retained. Three-seed DCRNN summaries average the original seed 42 and new seeds 43/44 equally within each shared draw. Every seed, mask, date and leave-one-date-out point is retained; the LSTM is summarized separately. The seed comparisons vary reference training while retaining the inherited proxy-selected policies.

Post hoc correlation-based mRMR-style comparison

For each parameter-fitting role, let $x_j(t)$ denote the observed reading of source j. Relevance is the mean absolute Pearson correlation of $x_j(t)$ with $x_k(t+h)$, averaged over every target source k and horizons $h = 1,...,12$. Both endpoints must lie within the same fitting role. Pairwise correlations use only jointly valid readings: finite nonzero speed for METR-LA and finite flow including zero for PEMS. Redundancy is the absolute same-time Pearson correlation between two source sequences over that fitting role. Each pair requires at least 24 valid observations; an undefined or constant-pair correlation contributes zero. A selectable source requires at least 24 valid fitting readings and positive fitting variance.

The first selected source maximizes relevance. Each subsequent step maximizes relevance minus the mean redundancy with the already selected sources, with penalty coefficient fixed at one. Ties resolve to the smaller source identifier. Relevance-only selection uses the same relevance values without subtracting redundancy. No coefficient, horizon aggregation or source set was tuned using restoration outcomes. These task-specific correlation definitions implement an mRMR-style difference criterion; they do not implement mutual-information MID or MIQ. Six fitting-role set pairs are shared across the corresponding reference families and seeds: one PEMS03 pair, four rolling PEMS08 pairs and one original METR-LA reference-42 pair.

After freezing both new policies, their scoring inputs, the evaluation sources and all 18 saved checkpoints, we evaluated each policy under masks 1001–1005, 30% background missingness and the unchanged 10%

graph-block fault and source budget. No predictor or proxy was retrained. Each reference first passed a 32-window replay of no restoration, Linear and variance against archived outcomes, with exact count agreement and the existing 0.0001-unit error tolerance. Evaluation uses the original fixed batches of 32, padding excluded from all outcomes. Archived policies provide the paired controls; no-restoration predictions were also rerun throughout the complete new evaluations and checked against their saved per-window counts and errors.

The additional contrasts are correlation-based selection minus Linear, variance and relevance-only selection; comparisons with group-aware variance and reference-specific exact selection are also reported. Negative differences favor the correlation-based policy. PEMS03 and METR-LA DCRNN summaries weight their three reference fits equally, while PEMS08 pools valid-target numerators and counts across four folds. LSTM remains a single separate reference. Ten thousand new common circular-block draws use seed 2026102401: seven-day blocks on PEMS03, independently sampled three-day blocks within each seven-day PEMS08 fold, and two-date blocks on METR-LA. The same cohort indices are shared across policy contrasts, seeds and reference families. These post hoc percentile intervals condition on saved fits, source sets and observed dates; historical intervals are retained separately. Complete case/mask results and reproducibility mappings accompany Supplementary section S25.

Post hoc analysis of the PEMS03 reversal

We inspected the six saved training histories, validation gates and input adapters without fitting or running prediction models. The adapters were compared on the same synthetic values, availability masks and historical calendar channels; their packed inputs matched exactly. We paired archived per-window errors for no restoration, Linear and variance by mask, start and encoded target day. Fault anchors were reconstructed from the unchanged rule: the generator with seed 20261103 + 1009 times window start + 1000003 times mask seed draws one of 358 anchors. We pooled each anchor across the three fits and five masks and repeated the contrast after omitting that anchor. Additional descriptive strata use the original four weeks, five masks and whether Linear returns more, equal or fewer entries than variance. Common error quartiles use each window–mask occurrence's mean GWN no-restoration MAE across the three seeds, then apply the same partition to DCRNN. These post hoc strata are reported as conditional point summaries.

The saved-output audit resolves policy differences by week, mask, fault anchor and realized returned-entry count. The archived errors are aggregated over prediction horizons and receiver sources, so the audit retains that resolution. Validation and final evaluation also differ in dates and fault conditions. Complete strata, source sets and training histories are in Supplementary section S26 and Supplementary Data 2.

Post hoc blockwise policy-label exchange

The primary DCRNN contrast was prespecified; the blockwise label-exchange analysis was added during revision. Saved daily numerators are pooled over masks, and the statistic is the equal mean of the three per-fit variance-minus-Linear MAE differences. Each non-overlapping temporal block receives one sign shared across all three fits and five masks. Its conditional reference distribution assumes joint invariance to these synchronized block sign changes, retaining within-block and cross-fit dependence.

The main analysis uses the original seven-day dependence scale: four consecutive weekly blocks and all 16 sign patterns. The two-sided reference p is the fraction whose absolute statistic is at least as large as observed, including the observed assignment. Fourteen-day and three-day sensitivities enumerate four and 1,024 patterns; a one-day sensitivity samples 99,999 patterns with RNG seed 2026102402 and uses the

plus-one Monte Carlo calculation. All scales are reported in Supplementary section S27. Shorter blocks impose stronger sign-invariance assumptions. The original effect estimate and temporal interval are retained.

Use of generative AI tools

Generative AI assisted with code preparation, literature organization, drafting and language revision. AI tools are not listed as authors; the authors are responsible for the content and its verification.

Data availability

The publicly available traffic datasets used in this study can be accessed at https://github.com/liyaguang/DCRNN (METR-LA), https://github.com/guoshnBJTU/ASTGNN/tree/main/data/PEMS08 (PEMS08), and https://zenodo.org/records/7816008 (PEMS03). Supporting derived data, frozen source-selection policies, experimental outputs and resampling indices are available in the unified Zenodo archive, version 1.7.1, at https://doi.org/10.5281/zenodo.23614390. Additional results and methodological details are provided in the Supplementary Information and Supplementary Data 1 and 2.

Code availability

Code, fitted checkpoints, fixed source-selection policies and derived data supporting the analyses are publicly available in the unified Zenodo archive, version 1.7.1: https://doi.org/10.5281/zenodo.23614390. The archive includes the original GWN release, correlation-based mRMR-style results, saved-output audit and blockwise label-exchange materials. Its README and manifests map the analyses to their code and outputs. Original author software is licensed under MIT; author-generated derived outputs, tables, figures and documentation are licensed under CC BY 4.0. Third-party components retain their original licenses. Raw benchmark series are obtained separately from their original sources. Computational paths and environment requirements are documented in the archive.

Author contributions

B.X. designed the experimental study, developed the experimental procedures and conducted the experiments. K.D. audited the experimental results and checked the consistency and accuracy of the reported findings. J.T. and R.W. collected and organized the research materials and prepared the manuscript. K.D., J.T. and R.W. contributed equally and share second authorship. All authors reviewed and approved the final manuscript.

Funding and acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and consent

Not applicable. This study uses previously released, aggregated traffic benchmark data and involves no human-participant or animal experiments.

References

[1] Wu, Z., Pan, S., Long, G., Jiang, J. & Zhang, C. Graph WaveNet for deep spatial-temporal graph modeling. In Proceedings of IJCAI, 1907–1913 (2019). https://doi.org/10.24963/ijcai.2019/264

[2] Li, Y., Yu, R., Shahabi, C. & Liu, Y. Diffusion convolutional recurrent neural network: data-driven traffic forecasting. In International Conference on Learning Representations (2018). https://arxiv.org/abs/1707.01926

[3] Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. Recurrent neural networks for multivariate time series with missing values. Scientific Reports 8, 6085 (2018). https://doi.org/10.1038/s41598-018-24271-9

[4] Cao, W. et al. BRITS: bidirectional recurrent imputation for time series. Advances in Neural Information Processing Systems 31 (2018). https://proceedings.neurips.cc/paper_files/paper/2018/hash/734e6bfcd358e25ac1db0a4241b95651-Abstract.html

[5] Du, W., Côté, D. & Liu, Y. SAITS: self-attention-based imputation for time series. Expert Systems with Applications 219, 119619 (2023). https://doi.org/10.1016/j.eswa.2023.119619

[6] Tashiro, Y., Song, J., Song, Y. & Ermon, S. CSDI: conditional score-based diffusion models for probabilistic time series imputation. Advances in Neural Information Processing Systems 34 (2021). https://proceedings.neurips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html

[7] Chen, X., Yang, J. & Sun, L. A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation. Transportation Research Part C 117, 102673 (2020). https://doi.org/10.1016/j.trc.2020.102673

[8] Chen, X. & Sun, L. Bayesian temporal factorization for multidimensional time series prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 4659–4673 (2022). https://doi.org/10.1109/TPAMI.2021.3066551

[9] Chen, X., Lei, M., Saunier, N. & Sun, L. Low-rank autoregressive tensor completion for spatiotemporal traffic data imputation. IEEE Transactions on Intelligent Transportation Systems 23, 12301–12310 (2022). https://doi.org/10.1109/TITS.2021.3113608

[10] Chan, R. K. C., Lim, J. M.-Y. & Parthiban, R. Missing traffic data imputation for artificial intelligence in intelligent transportation systems: review of methods, limitations, and challenges. IEEE Access 11, 34080–34093 (2023). https://doi.org/10.1109/ACCESS.2023.3264216

[11] Zhang, Y. et al. A comprehensive survey on traffic missing data imputation. IEEE Transactions on Intelligent Transportation Systems 25, 19252–19275 (2024). https://doi.org/10.1109/TITS.2024.3478816

[12] Cini, A., Marisca, I. & Alippi, C. Filling the Gaps: Multivariate Time Series Imputation by Graph Neural Networks. In International Conference on Learning Representations (2022). https://arxiv.org/abs/2108.00298

[13] Marisca, I., Alippi, C. & Bianchi, F. M. Graph-based forecasting with missing data through spatiotemporal downsampling. Proceedings of Machine Learning Research 235, 34846–34865 (2024). https://proceedings.mlr.press/v235/marisca24a.html

[14] Joshi, S. & Boyd, S. Sensor selection via convex optimization. IEEE Transactions on Signal Processing 57, 451–462 (2009). https://doi.org/10.1109/TSP.2008.2007095

[15] Krause, A., Singh, A. & Guestrin, C. Near-optimal sensor placements in Gaussian processes: theory, efficient algorithms and empirical studies. Journal of Machine Learning Research 9(8), 235–284 (2008). https://www.jmlr.org/papers/v9/krause08a.html

[16] Manohar, K., Brunton, B. W., Kutz, J. N. & Brunton, S. L. Data-driven sparse sensor placement for reconstruction: demonstrating the benefits of exploiting known patterns. IEEE Control Systems Magazine 38, 63–86 (2018). https://doi.org/10.1109/MCS.2018.2810460

[17] Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003). https://www.jmlr.org/papers/v3/guyon03a.html

[18] Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1226–1238 (2005). https://doi.org/10.1109/TPAMI.2005.159

[19] Brown, G., Pocock, A., Zhao, M.-J. & Luján, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. Journal of Machine Learning Research 13(2), 27–66 (2012). https://www.jmlr.org/papers/v13/brown12a.html

[20] Leung, S. C., Zhou, D. & Bae, H. J. Data-driven sensor placement for predictive applications: a Correlation-Assisted Attribution Framework (CAAF). Communications AI & Computing 1, 8 (2026). https://doi.org/10.1038/s44488-026-00011-1

[21] Covert, I., Lundberg, S. & Lee, S.-I. Explaining by removing: a unified framework for model explanation. Journal of Machine Learning Research 22(209), 1–90 (2021). https://www.jmlr.org/papers/v22/20-1316.html

[22] Covert, I., Lundberg, S. M. & Lee, S.-I. Understanding global feature contributions with additive importance measures. Advances in Neural Information Processing Systems 33 (2020). https://proceedings.neurips.cc/paper/2020/hash/c7bf0b7c1a86d5eb3be2c722cf2cf746-Abstract.html

[23] Fisher, A., Rudin, C. & Dominici, F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 20(177), 1–81 (2019). https://jmlr.org/papers/v20/18-760.html

[24] Aas, K., Jullum, M. & Løland, A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artificial Intelligence 298, 103502 (2021). https://doi.org/10.1016/j.artint.2021.103502

[25] Hooker, G., Mentch, L. & Zhou, S. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Statistics and Computing 31, 82 (2021). https://doi.org/10.1007/s11222-021-10057-z

[26] Hooker, S., Erhan, D., Kindermans, P.-J. & Kim, B. A benchmark for interpretability methods in deep neural networks. Advances in Neural Information Processing Systems 32 (2019). https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html

[27] Golovin, D. & Krause, A. Adaptive submodularity: theory and applications in active learning and stochastic optimization. Journal of Artificial Intelligence Research 42, 427–486 (2011). https://doi.org/10.1613/jair.3278

[28] Norcliffe, A. L. I., Lee, C., Imrie, F., Van Der Schaar, M. & Lio, P. Stochastic encodings for active feature acquisition. Proceedings of Machine Learning Research 267, 46784–46814 (2025). https://proceedings.mlr.press/v267/norcliffe25a.html

[29] von Kleist, H., Zamanian, A., Shpitser, I. & Ahmidi, N. Evaluation of active feature acquisition methods for time-varying feature settings. Journal of Machine Learning Research 26(60), 1–84 (2025). https://www.jmlr.org/papers/v26/23-1635.html

[30] Ao, X. & Chen, M. Efficient data selection for time series forecasting using a lightweight linear proxy framework. Scientific Reports 16, 16847 (2026). https://doi.org/10.1038/s41598-026-43645-y

[31] Roberts, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929 (2017). https://doi.org/10.1111/ecog.02881

[32] Künsch, H. R. The jackknife and the bootstrap for general stationary observations. Annals of Statistics 17, 1217–1241 (1989). https://doi.org/10.1214/aos/1176347265

[33] Guo, S., Lin, Y., Feng, N., Song, C. & Wan, H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. Proceedings of the AAAI Conference on Artificial Intelligence 33, 922–929 (2019). https://doi.org/10.1609/aaai.v33i01.3301922

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData1.xlsx
- SupplementaryData2.xlsx
- supplement.pdf