A Comparative Study of Machine Learning Methods for Block-Missing CV Speed Imputation

Xiaofeng Chen[1], Wen Zhang[2], Adel W. Sadek[3], and Chunming Qiao[4]

[1]M.S. Student, Dept. of Computer Science and Engineering, University at Buffalo, SUNY, Buffalo, NY 14260 (corresponding author). E-mail: xchen326@buffalo.edu

[2]Ph.D., Dept. of Computer Science and Engineering, University at Buffalo, SUNY, Buffalo, NY 14260.

[3]Professor, Dept. of Civil, Structural and Environmental Engineering, University at Buffalo, SUNY, Buffalo, NY 14260.

[4]Professor, Dept. of Computer Science and Engineering, University at Buffalo, SUNY, Buffalo, NY 14260.

ABSTRACT

Reliable highway speed data are fundamental to traffic management, safety monitoring, and intelligent transportation systems (ITS), yet large-scale connected-vehicle (CV) deployments often experience extended block-missing periods, spanning weeks to months, in which speed records are entirely absent while weather and road-surface covariates remain fully observed. This paper presents a realistic, weather-aware benchmark for imputing block-missing highway speed under heterogeneous seasonal and extreme winter-storm conditions. We curate 10-minute aggregated freeway speeds from Buffalo, New York, from September 2022 to March 2023, paired with 36 co-located Road Weather Information System (RWIS) variables and 5 calendar features, 41 auxiliary inputs in total. To stress-test generalization across seasonal regimes, we design a three-round cross-seasonal protocol spanning autumn, mild-winter December 12–18, and peak-winter December 19–25, which includes the December 23–25 Buffalo Blizzard. We evaluate nine deep learning imputation models alongside three statistical

temporal and five classical machine-learning baselines, with multi-seed verification. Across 14,904 evaluation timesteps, SAITS attains the lowest weighted mean absolute error (W-MAE) of 3.15 ± 0.05 km/h, narrowly leading the strongest classical baseline, HistGradient Boosting (HGBT), at 3.36 km/h; three gradient- and forest-based baselines occupy ranks 2–4 and outperform every other deep model. iTransformer is competitive on Rounds 1–2 but degrades sharply on the storm-included Round 3, finishing rank 11. Masked-attention with random sub-mask training, as in SAITS, is thus the most robust learned architecture on this dataset, while classical gradient-boosted trees are highly competitive and easier to deploy. Because storm-peak averages rest on far sparser CV sampling, we treat the storm-included benchmark as the primary stress test and report a matched storm-excluded sensitivity comparison, which lowers the best W-MAE to about 2.47 km/h and confirms that the storm period, not any single model, is a major driver of imputation difficulty.

Keywords: Traffic Speed Imputation; Block Missing Data; Machine Learning; Score-Based Diffusion; Multivariate Time Series; Weather-Traffic Interaction; Seasonal Generalization; Deep Learning Benchmark.

PRACTICAL APPLICATIONS

This study offers transportation agencies practical guidance on reconstructing block-missing highway speed records by combining continuously observed road-weather covariates with a benchmark of deep learning and classical imputation methods. Across three seasonal training–test splits, including a December round containing the December 23–25 Buffalo Blizzard, the benchmark identifies which architectures reliably reconstruct missing speeds and which produce physically implausible outputs under severe winter conditions. Self-attention imputation with random sub-mask training achieved the lowest pooled error among the learned models, while three classical gradient-boosted-tree and random-forest baselines remained close behind, indicating that simple tabular models are still competitive for weather-conditioned speed imputation. Agencies completing historical speed records for performance reporting, safety audits, or corridor planning can use the benchmark to se-

lect methods matching their seasonal data coverage and accuracy requirements, balancing accuracy against training and deployment cost. Because the released benchmark retains per-cell imputation provenance, completed series can be audited and reused for downstream metrics with known assumptions rather than treated as ground truth outside the training distribution.

INTRODUCTION

Accurate, complete highway speed data underpin traffic monitoring, incident detection, travel-time estimation, and intelligent transportation systems (Guide 2001; Wang et al. 2019), yet large-scale speed records are routinely incomplete: sensors fail, communication links drop, and connected-vehicle (CV) sources cover networks unevenly and intermittently. Unlike sparse or random point-level gaps that interpolation handles adequately, *block-missing* events leave entire periods, from days to months, with all speed measurements for a segment absent (Chan et al. 2023; Bae et al. 2018); adjacent observations cannot bridge them, so the model must rely on indirect signals, primarily weather and temporal covariates, to reconstruct the missing trajectories.

This is compounded in cold-climate cities, where ice accumulation, reduced visibility, and snow loading induce abrupt, non-stationary speed reductions: a model trained on moderate weather may extrapolate poorly to extreme winter, and vice versa. Quantifying this *seasonal generalization gap* requires a protocol that tests behavior under different seasonal train/test splits rather than a single aggregate hold-out.

We construct such a benchmark from a Buffalo freeway CV dataset paired with high-resolution RWIS weather records from September 2022 to March 2023. Buffalo's climate, which oscillates between mild autumn, moderately cold early winter, and severe peak-winter conditions, is an ideal testbed; the speed series has large block-missing intervals while weather and temporal features stay fully observed. Unlike standard forecasting and imputation benchmarks, which evaluate only ordinary regimes, we deliberately retain a historically severe storm, the December 23–25, 2022 Buffalo Blizzard, inside the evaluation window, so the

benchmark measures how models behave when the test period includes an extreme-weather event rather than only routine conditions.

We evaluate nine deep learning models drawn from the major methodological families active in time-series imputation research, as well as three statistical temporal baselines that exploit diurnal and weekly traffic periodicity. For each model, we employ a three-round evaluation protocol in which Rounds 2 and 3 probe opposite winter-generalization directions and Round 1 is a symmetric autumn/spring evaluation, yielding 14,904 total evaluation timesteps across three distinct seasonal regimes.

Our main contributions are as follows.

1. A three-round cross-seasonal protocol for block-missing speed imputation that explicitly probes asymmetric winter generalization across 14,904 evaluation timesteps and three seasonal regimes, unlike benchmarks built on a single fixed train–test split.
2. A like-for-like comparison of nine deep learning models against five classical and three temporal baselines, all conditioned on 41 auxiliary features, comprising 36 RWIS weather and road-surface variables and 5 calendar features, and reported in km/h using mean absolute error (MAE), root mean square error (RMSE), mean squared error (MSE), and mean absolute percentage error (MAPE). SAITS attains the lowest W-MAE, 3.15 ± 0.05 km/h, narrowly leading the best classical baseline, HGBT, at 3.36 km/h.
3. Evidence of a systematic winter-generalization asymmetry: Round 3, the peak-winter storm-included round, is the hardest for all deep models, with the increase over Round 1 varying widely, and iTransformer, competitive on Rounds 1–2, fragile on Round 3 because its December training segment ends Dec 18. We also assess weather–speed correlation preservation, on which CSDI, FGTI, SAITS, and TimesNet rank best and iTransformer worst.
4. A data-quality interpretation, not merely a ranking: during the blizzard each 10-minute average rests on far fewer raw CV observations than normal, so storm-period

ground truth is a high-stress, low-support regime rather than measurement error. We therefore retain the storm-included benchmark as the primary stress test and report a matched storm-excluded sensitivity analysis in the Appendix, with the detailed protocol and ranking tables reproduced in full as Tables 5 and 6. Excluding the Dec. 23–25 window lowers the best W-MAE from about 3.15 to about 2.47 km/h. Across both settings, strong classical models, notably HGBT, remain competitive with the best deep models.

The remainder of the paper surveys related work, describes the dataset and three-round protocol, presents and analyzes the benchmark results, and concludes with limitations and future work.

LITERATURE REVIEW

We organise prior work by method family, following each family from its origins to the imputation- and traffic-specific variants closest to our benchmark.

Missing-Data Patterns and Block-Missing

Rubin's framework classifies missingness by *cause* into Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) (Rubin 1976), and imputation accuracy depends jointly on the method and this pattern (Kang 2013; Haji-Maghsoudi et al. 2013; Aljuaid and Sasi 2016), yet this cause-based view does not describe how missing values are arranged in time and space. Chan et al. add a structure-based taxonomy for transportation data: random, fiber, and block missing (Chan et al. 2023). Block missing denotes large contiguous gaps in time and/or space, typically caused by sensor faults, detector removal, or communication loss, that strip away the adjacent reference points simple interpolation relies on (Liang et al. 2021; Xing et al. 2022; Bae et al. 2018; Zhang et al. 2024), and its dimensionality further affects difficulty (Duruflé et al. 2021; Umar and Gray 2023). We study a univariate block-missing regime: speed is absent over multi-day intervals while co-located weather and temporal covariates remain fully observed. This

is a conditional-imputation setting in which the missing variable is inferred from auxiliary context (Nevalainen et al. 2009; Lee and Carlin 2010; Chaudhry et al. 2019).

Statistical and Classical Machine-Learning Imputation

The earliest tools are statistical heuristics such as mean substitution, linear interpolation, spline fitting, and time-of-day or weekday-time-of-day averaging; these are robust and scalable but inaccurate once gaps exceed a few observations (Henrickson et al. 2015; Tak et al. 2016). Chained-equation predictive mean matching exploits inter-detector spatial correlation and holds up at high missingness (Henrickson et al. 2015; Li et al. 2014; Li et al. 2018), building on the multiple-imputation foundation (Rubin 1976; Lee and Carlin 2010; Chaudhry et al. 2019). Tree ensembles are the dominant tabular workhorse: Random Forest (Breiman 2001) provides a strong nonparametric same-time estimator, and histogram and standard gradient-boosted trees (HGBT, GBDT) add cheap sequential residual fitting (Tak et al. 2016; Goehry et al. 2023; Xing et al. 2022), while linear models (Ridge, LinearRegression) suit locally smooth speed–weather relations. All treat timesteps independently and propagate no temporal context, so under a fully absent block they reduce to a weather-conditional regression on co-located covariates, which in our regime actually favours them.

Recurrent Imputation Networks

The first deep imputation wave used recurrent neural networks (RNNs) to propagate observed values across time. BRITS (Cao et al. 2018) processes a series bidirectionally with a delayed-gradient mechanism that handles missingness inline, fusing both directions at each step. Such models excel on short, random medical and sensor gaps; under block missing every in-gap speed is simultaneously absent, so the recurrence has no speed signal to propagate and collapses to a weather-conditional decoder.

Graph-Based Spatiotemporal Imputation

Graph models treat each sensor or segment as a node and fill gaps by spatial message passing. GRIN (Cini et al. 2022) couples graph convolutions with bidirectional gated re-

current units (GRUs) to borrow spatially-adjacent observed signals, and later work adds attention-style layers and learned adjacency (Marisca et al. 2022). These methods need several correlated series with partly observed values in the same window; in our single-segment setting the graph degenerates to one node and loses its main advantage, consistent with GRIN's mid-pack result.

Self-Attention and Transformer-Based Imputation

The Transformer (Vaswani et al. 2017) replaced recurrence and convolution with multi-head self-attention, giving every timestep constant-path access to every other and suiting long-range dependencies. SAITS (Du et al. 2023) is the attention model in our benchmark most explicitly designed for imputation: it jointly optimises an Observed Reconstruction Task over all observed cells and a Masked Imputation Task over an artificially masked subset with grouped diagonally-masked self-attention, and the masked half exposes it to many mask patterns during training, which provides implicit regularisation against block-missing test conditions. Three forecasting-derived transformers are adapted by holding out the target: TimesNet (Wu et al. 2023) reshapes the series into 2-D tensors along DFT-identified periods for intra- and inter-period convolution; PatchTST (Nie et al. 2023) tokenises channel-independent non-overlapping patches; and iTransformer (Liu et al. 2024) attends along the feature dimension rather than time, which is powerful when inter-feature relations dominate but discards the temporal attention that helps under out-of-distribution regimes.

Diffusion-Based Imputation

Diffusion models are the most active deep lineage for time-series imputation since 2021. From the original formulation (Sohl-Dickstein et al. 2015), DDPM (Ho et al. 2020) set a weighted-variational-bound objective and DDIM (Song et al. 2021) a faster non-Markovian sampler. CSDI (Tashiro et al. 2021) introduced conditional score-based imputation, denoising a target subset while conditioning on randomised observed positions. Extensions follow: PriSTI (Liu et al. 2023) adds a differential-equation prior; SSSD (Alcaraz and Strodthoff 2023) swaps the denoiser for structured state-space sequence models (S4) (Gu et al. 2022)

for longer sequences; and FGTI (Yang et al. 2024) splits low-frequency trend, obtained via Fourier representations, from high-frequency residuals and pre-masks block-missing positions for direct optimisation. For traffic specifically, FastSTI (Cheng et al. 2024) reaches competitive accuracy with order-of-magnitude fewer denoising steps, and Pei et al. target MNAR block missing across adjacent segments with a self-supervised diffusion model (Pei et al. 2024); both, however, assume multi-segment spatial context and do not test the seasonal-shift regime our protocol targets. Across the three diffusion variants we evaluate, namely CSDI, FGTI, and SSSD, sampling stochasticity yields measurable seed-to-seed variance, which we report quantitatively.

Weather-Informed Traffic Speed Modelling

Weather strongly shapes traffic speed, and conditioning on meteorology improves accuracy in adverse conditions: models using environmental and temporal predictors outperform traffic-only ones (El Esawey 2020; Kyrtsoglou et al. 2021; Shabarek et al. 2020; Deb et al. 2019; Kar and Feng 2023), with precipitation, visibility, and road-surface state carrying the strongest signals (Lai et al. 2022; Chou et al. 2019). This matters most under block missing, where recent speed is unavailable for extrapolation. We extend this direction by conditioning every model on the full auxiliary feature set and stress-testing on a peak-winter Buffalo Blizzard window far from the training distribution.

Positioning Relative to Existing Benchmarks

Existing traffic-imputation benchmarks mostly use random- or fiber-missing loop-detector data (Li et al. 2018; Li et al. 2014), whose spatial redundancy graph models can exploit. Block missing (Chan et al. 2023) instead removes whole temporal segments of the target, eliminating temporal and spatial reference points; few studies benchmark a broad deep-learning family under this regime, and most omit weather covariates despite their strong adverse-condition signal (Shabarek et al. 2020; Lai et al. 2022; Zhang et al. 2024). As detailed in the contributions above, our benchmark addresses both gaps with a broad deep and classical roster under a strictly block-missing, weather-conditioned, three-round cross-seasonal

protocol, measuring both pointwise accuracy and seasonal out-of-distribution robustness over a storm-included Round 3 window.

BENCHMARK DESIGN

Speed Data

Speed observations are derived from a commercial Wejo connected-vehicle (CV) Global Positioning System (GPS) ping feed for the Buffalo, NY metropolitan area between 2022-09-27 and 2023-03-31 (186 calendar days). Each raw record is one GPS ping, comprising a captured timestamp, latitude, longitude, instantaneous vehicle speed, heading, vehicle-type code, and a journey identifier; the feed does not contain a persistent unique-vehicle key, so we report distinct trip identifiers rather than unique-vehicle counts and we cannot compute a CV market penetration rate (see the Limitations subsection).

In an upstream pre-processing pipeline, CV pings were spatially joined to OpenStreetMap (OSM) road segments and the dataset was restricted to segments with posted speed limit $\geq$ 89 km/h ($\geq$ 55 mph), that is, motorway and motorway-link carriageways with a small share of trunk-link. The retained network consists of 198 directional OSM segments on seven expressway corridors covering the Buffalo metro area within an approximate 26×16 km bounding box: *I-90*, *I-190*, *I-290*, *NY-33*, *NY-400*, *NY-5*, and *US-219*. Directional segment lengths range from 22.5 m (0.014 mi) to 4,697.6 m (2.92 mi) with mean 729 m (0.45 mi) and median 519 m (0.32 mi); the total directional centerline length is approximately 144 km (89.7 mi). On a representative day (2022-09-28), the feed captured approximately 53,000 distinct journey IDs on these corridors; comparable daily counts (48,000 to 54,000 trips) were observed across the study period.

For each 10-minute interval t the bin-level speed used by all models in this paper is the arithmetic mean of every instantaneous CV speed sample whose GPS ping in t fell on one of the 198 selected segments. The result is a single network-wide 10-minute average speed series of length $186 \times 144 = 26{,}784$ samples; the model imputes this single series rather than a separate series per OSM segment. Of the 26,784 bins, 58.84% are missing, which is the

block-missing pattern this paper addresses. The upstream PySpark pre-processing code is available upon reasonable request (see the Data Availability Statement); only the processed 10-minute series is distributed with the manuscript.

The Dec 23–25, 2022 Buffalo Blizzard storm peak is observed in the data and is part of the storm-included three-round protocol (Round 3 test set). Naturally missing periods outside the held-out evaluation windows are used only for qualitative full-period imputation (see below); they are not included in any quantitative metric because they lack ground-truth speed observations.

Connected-vehicle sampling density during the storm. Each 10-minute average during the Blizzard is supported by far fewer raw observations than in normal periods: our sample-coverage audit finds normal bins supported by roughly 25,000–28,000 raw GPS samples versus only on the order of one hundred samples per bin, with a pooled median of about 90, over Dec. 23–25. Storm-period averages therefore remain directionally meaningful but are lower-confidence per-bin ground truth, which is why we treat the storm-included benchmark as the primary stress test and the storm-excluded benchmark as a matched *sensitivity* comparison. The data comprise three distinct observation blocks, namely autumn in October 2022, early winter in December 12–25, 2022, and late winter–spring in March 2023, separated by extended block-missing gaps over Oct 29–Dec 10 and Dec 27–Feb 28; a scatter of every valid 10-minute observation across the full study period is provided in Fig. 1. The spatial relationship between the CV speed coverage and the co-located RWIS weather station is shown in Fig. 2.

Weather and Road-Surface Features

Road-weather data come from a single New York State Department of Transportation (NYSDOT) Road Weather Information System (RWIS) station, "Service Center Rd" in the Town of Amherst, NY (north of the I-290 / NY-33 interchange complex), reporting at 10-minute resolution synchronized with the speed series.

The raw RWIS feed comprises 24 weather columns: 20 continuous numerical variables

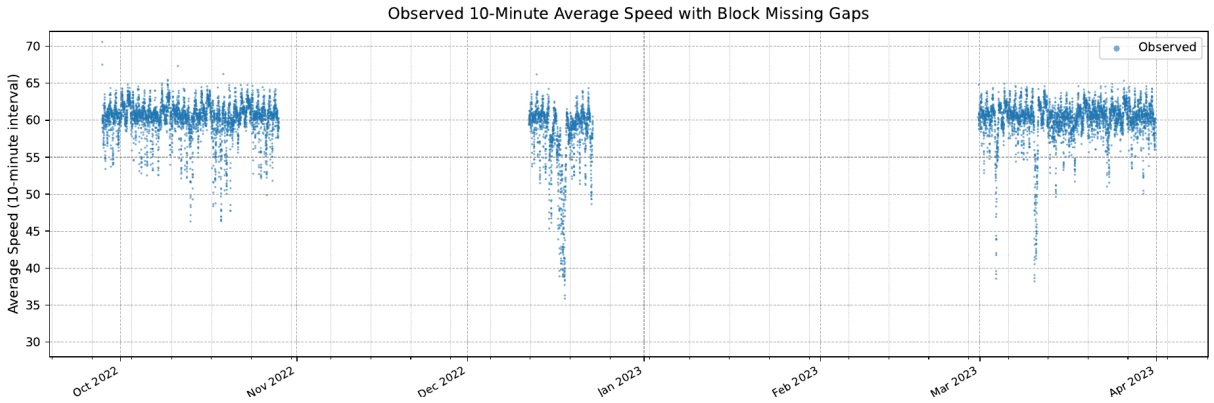

FIG. 1. Scatter plot of all valid 10-minute average speed observations from 2022-09-27 to 2023-03-31. Three observation blocks are separated by extended block-missing gaps (Oct 29–Dec 10 and Dec 27–Feb 28), motivating the weather-informed imputation task. The Dec 23–25 winter storm peak is observed and is part of the storm-included three-round protocol (Round 3 test set). Naturally missing periods outside the held-out evaluation windows are used only for qualitative full-period imputation and are not included in any quantitative metric.

(pavement and air temperatures, dew point, surface grip level, water / ice / snow layer thicknesses, relative humidity, rain intensity, 10-minute and spot wind speed and direction, maximum wind speed, visibility, and five rolling-window precipitation totals at 1 h, 3 h, 6 h, 12 h, and 24 h), three multi-class categorical state variables (*Surface State*: dry / moist / wet / slushy / snowy / icy; *Rain State*: none / light / medium / heavy / l.snow / m.snow; and *Alarm Status*: no warn / ice alarm / ice warn), and one binary indicator (Rain on/off). The continuous variables and the binary indicator are passed through unchanged; the three categorical variables are expanded by *one-hot encoding* into $6 + 6 + 3 = 15$ binary level columns, so that a categorical variable with k levels is replaced by k binary indicator columns, exactly one set to 1 per row, and models ingest the levels directly without imposing an artificial ordering on what are unordered states. Five derived calendar features are then appended: an integer time-of-day index, the sine and cosine of bin-of-day, a weekend flag, and a U.S. federal-holiday flag computed via the `holidays` Python package. The model input matrix therefore has 42 columns per 10-minute bin: 1 speed target, 21 numeric weather, 15 one-hot categorical levels, and 5 calendar features. An equivalent 31-column variant used by

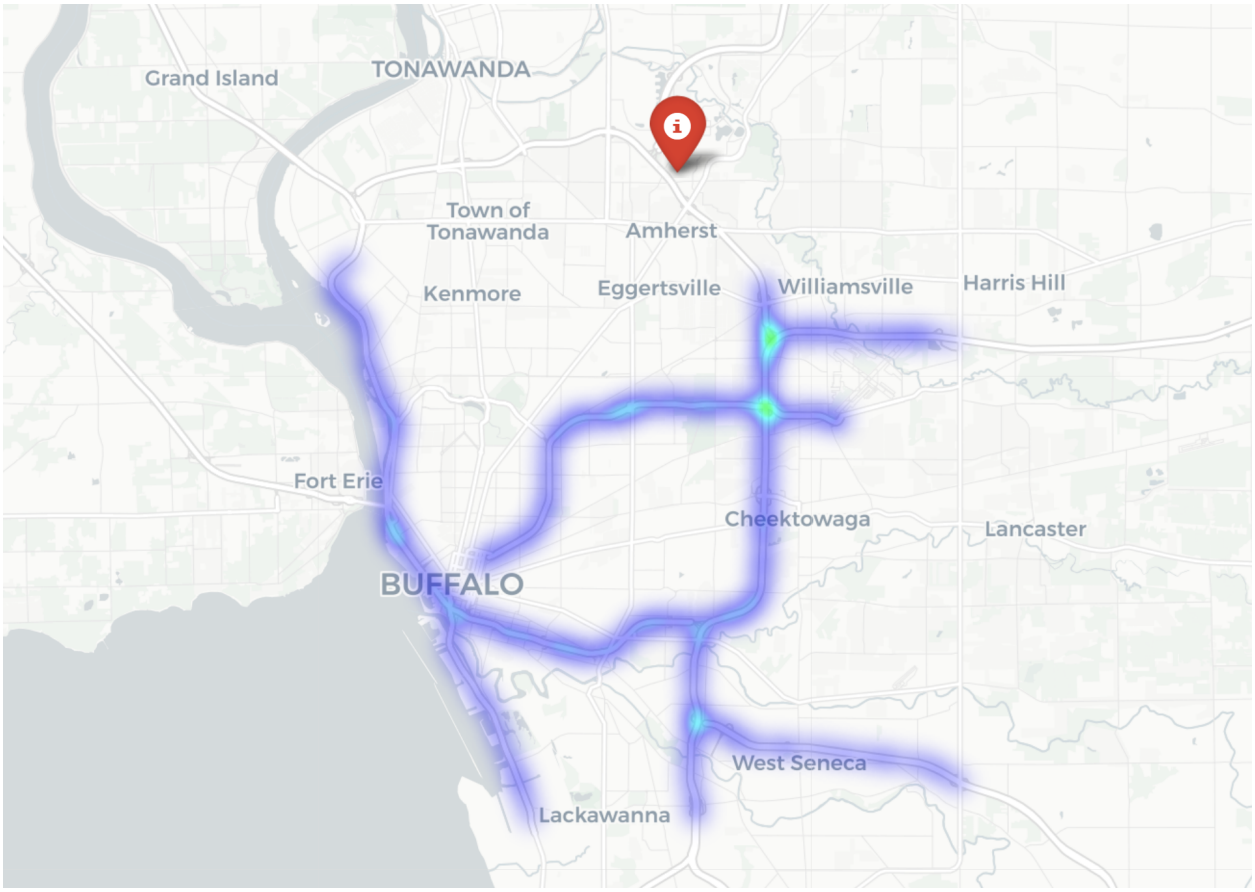

FIG. 2. Coverage of the data sources used in this study. The density heatmap shows the spatial distribution of Wejo connected-vehicle speed pings along the seven Buffalo, NY expressway corridors (I-90, I-190, I-290, NY-33, NY-400, NY-5, and US-219); the red marker is the NYSDOT RWIS "Service Center Rd" weather station (Amherst, NY), the sole source of the 10-minute weather covariates.

some legacy training pipelines preserves the same content but keeps the four categorical and binary columns as ordinal integers rather than one-hot expansions, and both variants yield identical evaluation cells.

The dataset's temporal structure motivates the temporal baselines used in this study: the diurnal speed pattern, aggregated within five non-overlapping two-week windows spanning the study period (Fig. 3a), and the full weekly cycle at 10-minute resolution (Fig. 3b) motivate the time-of-day and weekday-time-of-day baselines, respectively.

Fig. 4, the canonical 20-feature ground-truth correlation profile, lists the Pearson correlation between observed speed and 20 of the strongest weather/road-surface features computed

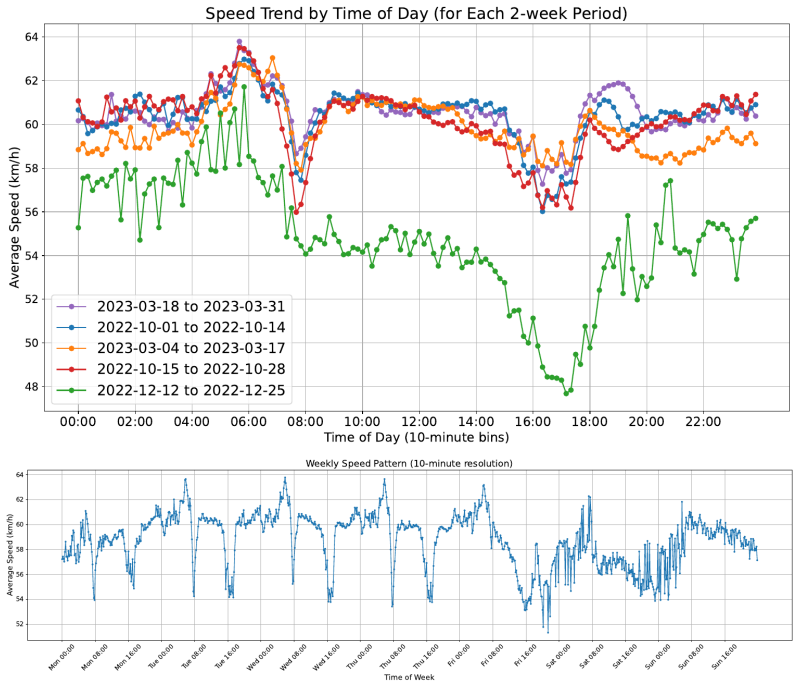

FIG. 3. Temporal speed structure. (a) Speed trend by time of day, averaged within each of five non-overlapping two-week windows; the Dec 12–25 winter window shows a distinctly lower and noisier diurnal profile, especially the late-day drop visible only in the winter trace, that motivates the cross-seasonal evaluation protocol. (b) Weekly speed pattern at 10-minute resolution; clear weekday morning- and evening-peak structures and a flatter weekend profile motivate the WdTodMean (weekday-time-of-day mean) temporal baseline.

over the full study period; Table 1 summarizes the top five.

These correlations confirm that weather carries a strong predictive signal for speed reconstruction, particularly in winter when pavement grip and precipitation accumulation dominate driver behavior.

Three-Round Cross-Seasonal Evaluation Protocol

To assess model robustness under distinct seasonal regimes, we define a three-round evaluation protocol in which training and evaluation periods are non-overlapping. Table 2 summarizes the splits.

Round 1 tests generalization over periods with strong diurnal and weekly regularity,

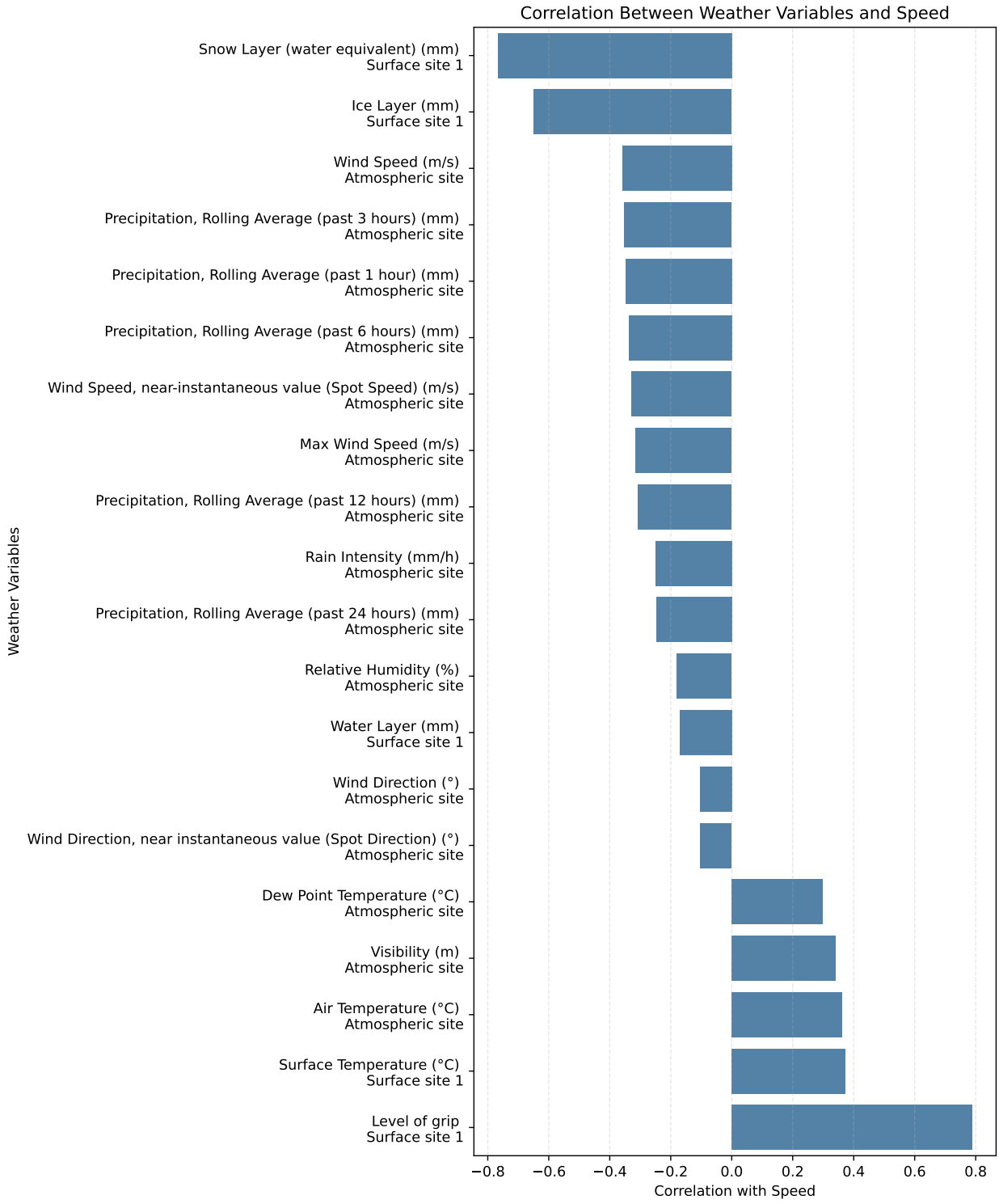

FIG. 4. Ground-truth Pearson correlation between observed speed and 20 road-weather features over the full study period. "Level of grip" ($+0.78$) and "Snow Layer" (-0.78) are the strongest signals. The 20 features visualized here define the feature subset used by the structural-consistency analysis (Fig. 7).

namely autumn October and early spring March; the December winter data of Dec 12–25 is included in training.

Round 2 withholds mild-winter December 12–18 from training and includes it in evaluation, while training on the more severe late-winter segment of Dec 19–25. This creates a severe-to-mild generalization scenario.

Round 3 reverses this logic, training on mild-winter December 12–18 and evaluating on peak-winter December 19–25, which includes the Dec 23–25 Buffalo Blizzard storm peak, and so constitutes a mild-to-severe generalization scenario.

Models and Configurations

We evaluate seventeen methods drawn from the families surveyed in the Literature Review section above. They span three groups; Table 3 reports the unified per-model configuration selected by the screening procedure, and Table 4 reports the full ranking.

Statistical temporal baselines (three). The parameter-free TrainMean (TrMean), Time-of-day mean (TodMean), and Weekday-time-of-day mean (WdTodMean) fill each missing cell with a training-set average and establish a temporal-regularity floor that learned models must exceed: TrMean uses the global training mean, TodMean the per-10-minute interval mean (0–143) pooled across days, and WdTodMean the mean at the same interval *and* day-of-week.

Classical machine-learning baselines (five). HistGradient Boosting (HGBT), Gradient-Boosted Decision Trees (GBDT), Random Forest, Ridge regression, and Linear Regression operate on the same 42-column feature matrix one timestep at a time and exploit the strong same-time weather-speed conditioning available at every cell.

Deep learning imputation models (nine). iTransformer, SAITS, CSDI, FGTI, TimesNet, BRITS, GRIN, PatchTST, and SSSD span the major methodological families active in time-series imputation, whose mechanisms are detailed in the Literature Review. All nine receive the same masked $(T_{\text{seq}}, 42)$ input tensor and use the hyperparameters listed in Table 3.

Coverage of cyclical and multi-frequency structure. Freeway speed exhibits strong diurnal, weekly, and seasonal cycles, and the roster deliberately spans the multi-frequency modelling approaches that target this structure, namely TimesNet (Wu et al. 2023), PatchTST (Nie et al. 2023), iTransformer (Liu et al. 2024), and SAITS (Du et al. 2023), whose mechanisms are detailed in the Literature Review; the WdTodMean baseline isolates the contribution of cyclical priors alone.

Deep learning model configurations. All nine deep learning models receive the same input tensor of shape $(T_{\text{seq}}, 42)$, where $T_{\text{seq}} = 144$ timesteps covers one calendar day and the 42 columns comprise the speed target and the 41 auxiliary features described above. The masking policy follows a *full-speed masking* strategy: the entire speed channel is masked during both training and inference, so the model must reconstruct speed from the auxiliary features and temporal structure alone, matching the block-missing scenario where no adjacent speed observation is available.

Table 3 summarizes the unified per-model configuration selected by the two-stage screening procedure (Stage A on Round 1+Round 3 and Stage B Round 2 verification, both decided by W-MAE minimization). One configuration per model is used throughout the paper; no per-round mixed configurations are used in headline results.

Evaluation Metrics

We evaluate all models with four standard error metrics on each round's held-out test segment listed in Table 2. Let y_t and $\hat{y}_t$ denote the true and imputed speed in km/h at evaluation cell t, where t ranges over every test-day 10-minute position in the segment whose speed is observed; cells with missing ground truth, such as the 24 sensor-outage cells on 2022-10-28, are excluded from all metrics. With T valid cells in the round of interest:

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T}\left|y_t - \hat{y}_t\right|, \quad \text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}, \tag{1}$$

$$\mathrm{MSE} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2, \quad \mathrm{MAPE}(\%) = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{y_t - \hat{y}_t}{y_t}\right|. \tag{2}$$

All metrics are computed in km/h. The original mph records, and all reported MAE and RMSE values including W-MAE and W-RMSE, are converted by the factor 1.60934; the unit-free MAPE and correlation-preservation scores are reported without conversion. The weighted MAE, W-MAE, pools the per-round MAE by evaluation-set size,

$$\text{W-MAE} = \frac{N_1 \cdot \mathrm{MAE}_1 + N_2 \cdot \mathrm{MAE}_2 + N_3 \cdot \mathrm{MAE}_3}{N_1 + N_2 + N_3}, \tag{3}$$

with $N_1 = 4{,}296$ and $N_2 = N_3 = 5{,}304$ for a total of 14,904 cells; W-RMSE is computed analogously by pooling mean squared error and taking the square root.

Beyond pointwise accuracy, we report a correlation-preservation score, Corr-Pres, for which lower is better: the mean absolute difference between the Pearson correlation profile of imputed speed against the 20 weather features of Fig. 4 and the ground-truth profile on the same test-period rows,

$$\text{Corr-Pres}(m) = \frac{1}{20}\sum_{k=1}^{20}\left|\rho_k^{(m)} - \rho_k^{(\mathrm{gt})}\right|. \tag{4}$$

Each deep model is trained and evaluated with three seeds $\{42, 1337, 2024\}$; each classical baseline is trained with five seeds $\{1, 2, 3, 4, 5\}$. We report the per-round and pooled metrics as mean $\pm$ standard deviation across seeds.

Robustness. Separately from pointwise accuracy, we report robustness in a strictly empirical sense through three observables. The first is the degradation in MAE and RMSE from Round 1 to Round 3, which measures sensitivity to seasonal and storm distribution shift. The second is the seed-to-seed standard deviation of the per-round and pooled metrics, a coarse repeatability indicator because the deep models use only three seeds. The third is the Corr-Pres distance of Eq. 4, which measures preservation of the weather–speed structure.

Because storm-period ground truth rests on far sparser connected-vehicle sampling, the Round 3 error increases reflect stress-test performance under lower-support ground truth rather than model fragility alone.

RESULTS AND DISCUSSION

Table 4 reports the headline ranking by W-MAE (and W-RMSE) for all evaluated methods on the storm-included three-round protocol, with per-round MAE columns and mean ± std across seeds (one standard deviation: deep models 3 seeds, classical baselines 5 seeds). Detailed per-round MAE and RMSE values are provided in the released artifact package.

Overall Performance Summary

Three findings stand out from Table 4.

First, SAITS is the headline winner, with W-MAE $= 3.15 \pm 0.05$ km/h and the tightest seed-to-seed variability of any deep model; its margin over the strongest classical baseline, HGBT, is roughly five times its own across-seed standard deviation, and it also leads on W-RMSE. Second, classical gradient-boosted trees and random forests rank 2–4, beating every deep model except SAITS. HGBT is deterministic (zero across-seed standard deviation), while GBDT and RandomForest retain small stochasticity; strong same-time weather conditioning and tree-ensemble efficiency on tabular data favour these baselines, and Ridge and LinearRegression still beat five of the nine deep models. Third, the field spreads sharply at the bottom: iTransformer is competitive on Rounds 1–2 but collapses on the storm-included Round 3 at rank 11, only PatchTST and SSSD clearly fail to beat the WdTodMean baseline with GRIN at the threshold, and the diffusion models CSDI and FGTI drop from single-seed Stage A ranks #3/#4 to multi-seed ranks #6/#5 once seed variability is included. These behaviors are analyzed below.

Interpretation of Temporal Baselines

The three mean baselines form a hierarchy of increasingly informative temporal priors, from TrMean to TodMean to WdTodMean (see the Models and Configurations subsection), and the results confirm that temporal structure is a strong source of predictability: weekday-time-of-day conditioning lowers Round 1 MAE from 3.25 for TrMean to 2.77 for WdTodMean, and the gap widens in Round 2 when the evaluation includes the Dec 12–18 winter segment, where weekday-specific conditioning is especially valuable. The top learned models add further gains over WdTodMean on Rounds 1–2, but this advantage vanishes on the storm-included Round 3.

Round 1: Autumn–Spring Evaluation

Round 1 evaluates on October 14–28 and March 1–15, periods of moderate temperatures, dry pavement, and strong diurnal periodicity, and is the easiest round. The best learned models, FGTI, CSDI, and SAITS, cluster tightly, and the deterministic HGBT is competitive while training in under a second per round. GRIN and PatchTST score at or below the WdTodMean baseline and SSSD only matches TrMean, showing that neither graph structure nor patch tokenization helps on this block-missing task.

Round 2: Severe-to-Mild Winter Generalization

Round 2 (severe-to-mild) shows both model spread and seed variability exceeding Round 1. SAITS leads, with HGBT and GBDT close behind, while the diffusion models fluctuate widely because a seed 1337 divergence inflates CSDI and FGTI variance, motivating the per-seed reporting.

Round 3: Storm-Included Peak-Winter Generalization (incl. Dec 23–25)

The storm-included Round 3 evaluates on October 14–28, Dec 19–25, and March 1–15. Including the Dec 23–25 Buffalo Blizzard storm-peak days makes Round 3 the hardest round for every model (Table 4, R3 column): SAITS has the lowest R3 MAE among deep models, ahead of TimesNet, BRITS, CSDI, and FGTI, while iTransformer, GRIN, PatchTST, and

SSSD trail. The classical baselines degrade roughly 2× from Round 1 yet remain comparable to or better than every deep model except SAITS. iTransformer's collapse is concentrated on the storm peak: the Dec 23–25 cells are 8% of the R3 test set but contribute roughly 57% of its R3 absolute error.

The contrast between SAITS's modest Round 1→Round 3 degradation and iTransformer's much larger rise, driven almost entirely by Dec 23–25 cells, characterizes the seasonal generalization split on this benchmark.

Per-day imputed-speed profiles for each round are provided as part of the released artifact package; in the main text we present a per-round MAE summary across all models (Fig. 5) alongside the SAITS three-panel prediction-comparison figure (Fig. 6) as a qualitative illustration of the rank-1 model's behavior across rounds, and of how even SAITS does not fully capture the Dec 23–25 blizzard speed collapse.

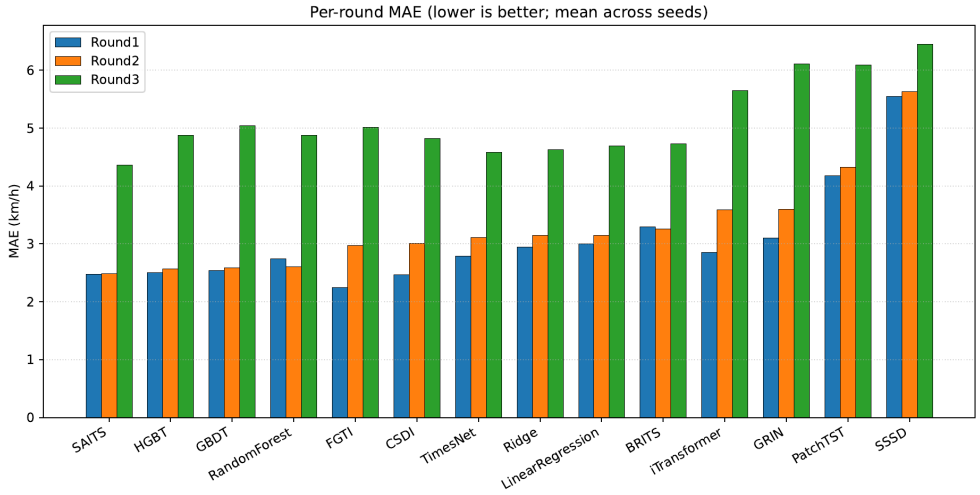

FIG. 5. Per-round MAE across the three evaluation rounds, ordered by overall W-MAE rank. Every deep model except SAITS exhibits a sharp Round 3 degradation once Dec 23–25 (Buffalo Blizzard storm peak) is included in the test set; iTransformer in particular jumps from Round 1 MAE 2.85 to Round 3 MAE 5.65 km/h. SAITS remains the most consistent deep model across rounds.

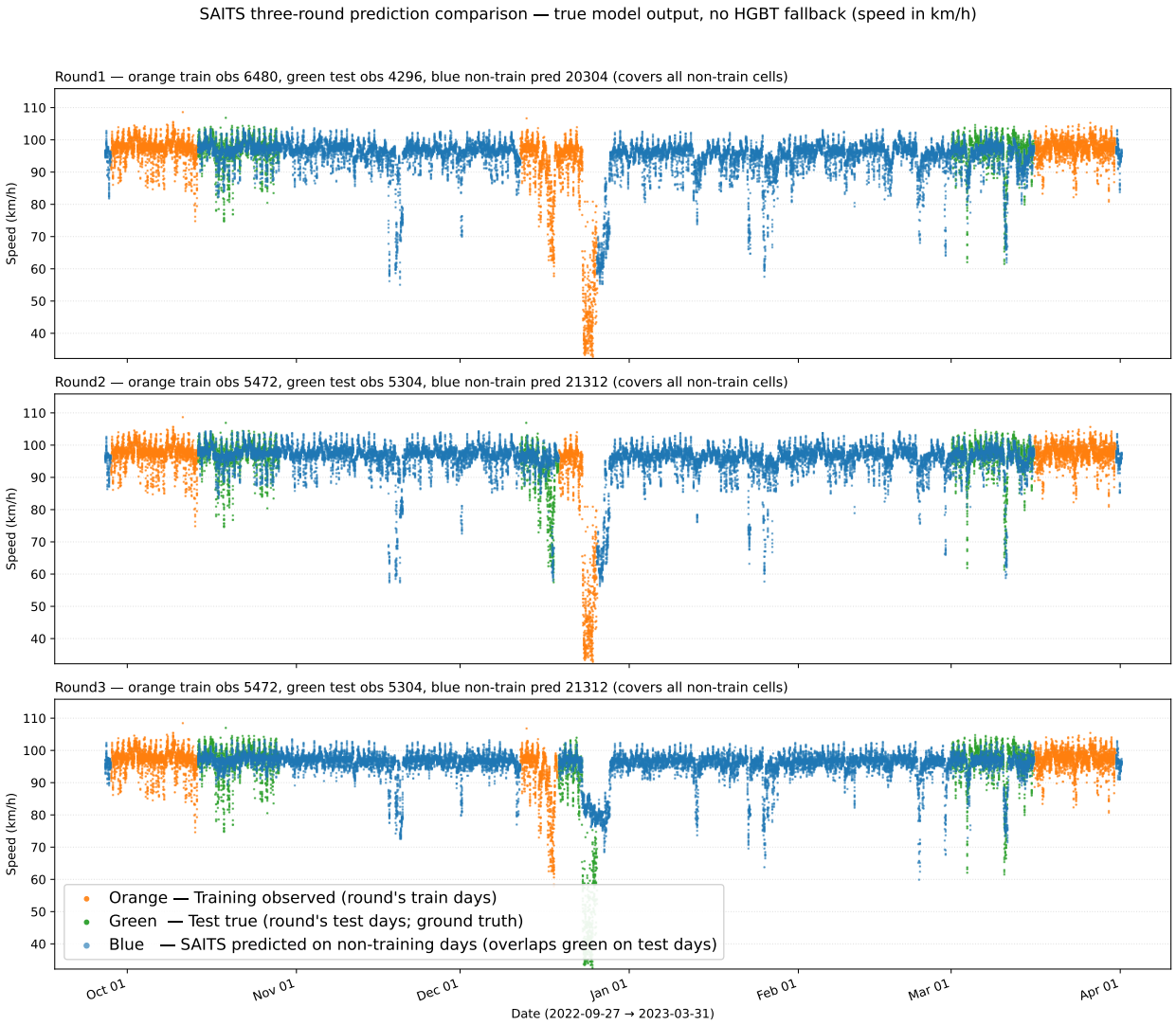

FIG. 6. SAITS three-round prediction comparison on the full study period (2022-09-27 to 2023-03-31), one round per row sharing a common time axis. Orange: observed speed on training days; green: observed speed on test days (ground truth); blue: SAITS imputed speed on every non-training cell (test days and the long originally-missing windows). The Round 3 panel shows the extreme-weather failure mode: the December training segment ends Dec 18, so on the Dec 23–25 storm peak the green ground truth drops to ∼40 km/h while the blue prediction stays in the ∼89–97 km/h band—even the rank-1 model does not capture the storm-peak collapse. Per-round MAE values are in Table 4.

Weather–Speed Structural Consistency

Beyond pointwise errors, we assess whether imputed speeds preserve the physical speed–weather relationships, scored by the Corr-Pres distance (Eq. 4) over the 20 weather features of Fig. 4; the full per-round profile plot is in Fig. 7. Two findings emerge. CSDI, TimesNet, FGTI, and SAITS preserve the ground-truth correlation profile best in Round 1, with HGBT close behind, whereas iTransformer has the *worst* Corr-Pres distance in every round and degrades from Round 1 to Round 3, indicating that its feature-token representation struggles to preserve the weather–speed dependence structure under this block-missing setting.

Cross-Round Comparison and Seasonal Robustness

Per-round MAE (Fig. 5) confirms that SAITS is the most consistent deep model across rounds. Among deep models, CSDI, FGTI, and TimesNet show the highest seed sensitivity, while the classical baselines are near-deterministic. A single-season sensitivity analysis (Table 7), which isolates the contribution of the three-season training pool, found that winter-only training produces substantially larger errors on the Dec 19–25 storm-peak window than the corresponding cross-season Round 3 models, supporting the cross-season design.

Model Family Analysis

Beyond the per-model ranking (Table 4), the families differ in *why* they succeed or fail. **Sequence imputation models (SAITS, BRITS):** SAITS's joint ORT + MIT loss with grouped attention and random sub-mask augmentation provides implicit regularization, so it degrades gracefully on the storm-included Round 3 and attains the lowest deep-model R3 MAE; BRITS's bidirectional GRU is outperformed by attention once weather conditioning is added. **Inverted-attention (iTransformer):** feature-first attention gives a competitive same-time estimate on Rounds 1–2 but, with its December training segment ending Dec 18, collapses on the storm-peak Round 3. **Diffusion (CSDI, FGTI, SSSD):** CSDI and FGTI lead the single-seed Stage A ranking but drop to ranks 5–6 under multi-seed verification, because sampling stochasticity lets a single divergent denoising trajectory shift the median MAE; SSSD is worst overall, as 20,000 iterations are insufficient for its S4 backbone to con-

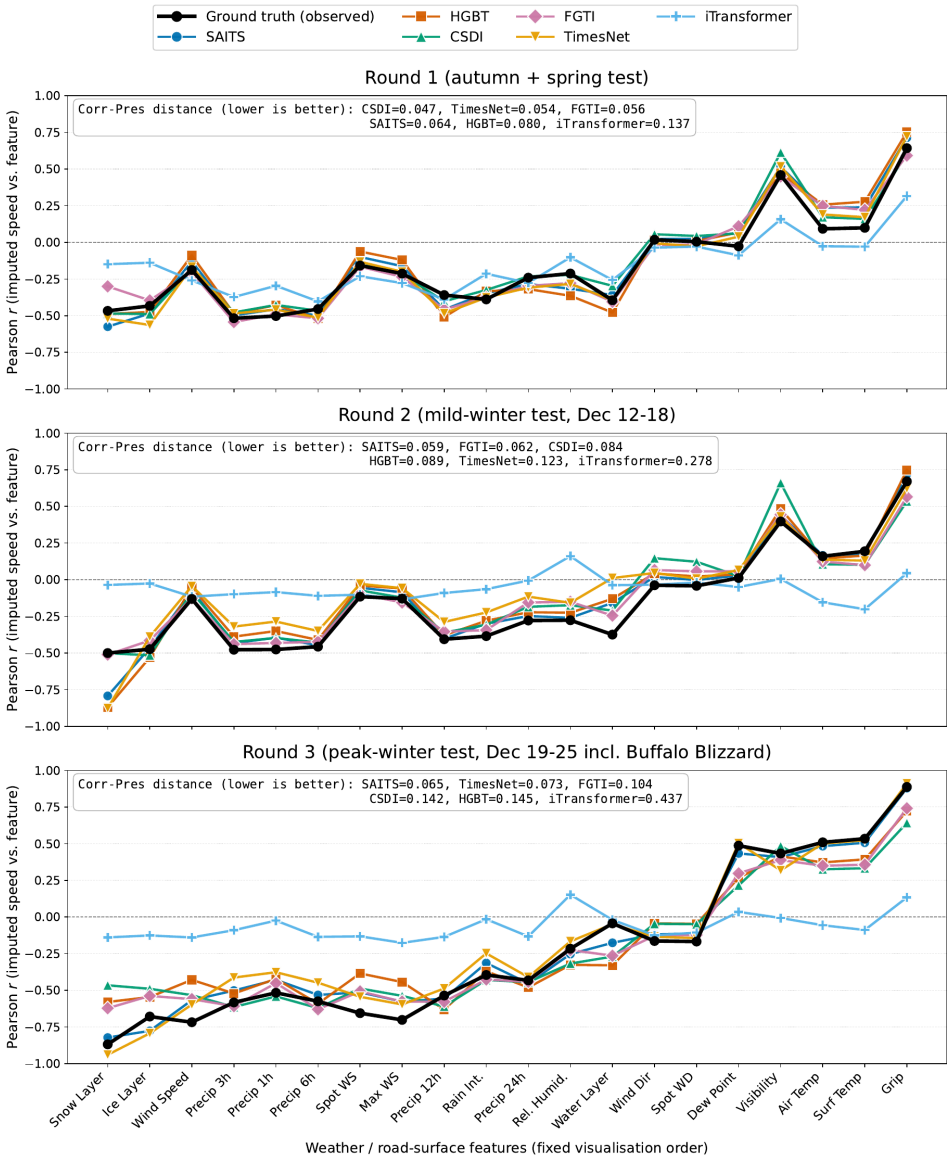

FIG. 7. Imputed-vs-ground-truth weather–speed correlation profiles for the top-six models across the three rounds. Black line: Pearson correlation between observed speed and each of the 20 weather and road-surface features (the subset visualized in Fig. 4) on the round's test-period rows. Coloured lines: each model's imputed-speed correlation on the same rows. The Corr-Pres distance (lower is better) is annotated per panel. Lines are visual guides to compare correlation-profile shapes across a fixed weather-feature order; the features are not a continuous physical axis.

verge. **Frequency/forecasting-adapted (TimesNet, PatchTST):** TimesNet is competitive in Round 1 but degrades sharply in Round 3, and PatchTST's non-overlapping patches and channel-independent design suppress the cross-feature weather–speed conditioning that other models exploit. **Graph (GRIN):** with all speed nodes simultaneously absent, message passing has no observed speed to propagate, so GRIN reduces to a non-spatial model. **Classical baselines:** tabular efficiency plus strong same-timestep weather conditioning lifts HGBT, GBDT, and RandomForest to ranks 2–4, ahead of every deep model except SAITS.

Companion Full-Period Missing-Speed Imputation

Beyond the 14,904 scored timesteps, the study period contains long naturally missing stretches over Oct 29–Dec 10 and Dec 27–Feb 28. For qualitative inspection we release model-specific full-period imputations, which are unscored because these periods lack ground truth, with per-cell provenance retained so observed, evaluated, and imputed cells stay distinguishable.

CONCLUSION

This paper presents a systematic benchmark of nine deep learning imputation models for block-missing highway speed data, using a three-round cross-seasonal protocol on a Buffalo, NY freeway dataset paired with 41 auxiliary weather and calendar features. Rather than proposing a new model, the contribution is a realistic, weather-aware benchmark that deliberately retains a historically severe winter storm inside the evaluation window, with a data-quality interpretation of that choice.

SAITS is the most robust model overall, with W-MAE $= 3.15 \pm 0.05$ km/h across 14,904 evaluation points and three deep-model seeds: it leads every deep model, narrowly beats the strongest classical baseline HGBT, and shows the smallest Round 1-to-Round 3 degradation among the top models. Classical gradient-boosted trees and random forests remain highly competitive, taking three of the top four places; HGBT is a strong deterministic baseline to try before deeper architectures. iTransformer is competitive on Rounds 1–2 but fragile on the storm-included Round 3, finishing rank 11. Pointwise accuracy and correlation preservation

prove complementary: SAITS wins on both criteria and is the recommended choice when the weather–speed relationship must be retained. Because storm-peak averages rest on far sparser connected-vehicle sampling, the storm-included benchmark is the primary stress test; a matched storm-excluded comparison reported in the Appendix lowers the best W-MAE to about 2.47 km/h and reduces error for every model, confirming the storm period as a major driver of benchmark difficulty rather than a defect of any single model.

Limitations

The benchmark covers a single urban network (Buffalo, NY) over one winter season, so generalization to other climates and year-to-year variability is not assessed, since same-month records from earlier years were unavailable. The Wejo extract carries per-trip identifiers but no persistent vehicle key and no co-located stream count, so we report distinct trip counts of roughly 5×10^4 per day rather than a CV market penetration rate. The storm-included Round 3 window contains the Dec 23–25 Buffalo Blizzard, during which CV sampling is far sparser than on normal days (see the Speed Data subsection); we therefore treat the storm-included benchmark as the primary stress test and report the storm-excluded analysis (the Appendix) as a matched sensitivity check.

Future Work

With additional archived CV and RWIS records spanning more winter seasons, corridors, and weather stations, future work could evaluate spatial weather heterogeneity and broader seasonal generalizability and, where segment-level ground truth becomes available, extend the network-wide task to per-segment imputation under conditions more favourable to graph-based models. As this study targets point-imputation accuracy, future work could also assess uncertainty calibration for probabilistic imputers such as CSDI and FGTI.

APPENDIX. STORM-EXCLUSION SENSITIVITY ANALYSIS

The primary benchmark includes the Dec 23–25 Buffalo Blizzard storm peak into the Round 3 test window. Because that window is by far the most sparsely sampled period in

the dataset (see the Speed Data subsection) and corresponds to an officially declared state of emergency with a full driving ban and roads closed for roughly five days (Guidehouse 2023), we ran a controlled *sensitivity* analysis, not an alternative headline protocol and not a replacement for the storm-included benchmark, in which the Dec 23–25 storm window is excluded from all training and evaluation. The two benchmarks share the same model roster, multi-seed protocol, evaluation scorer, and true-km/h reporting; they differ only in that the storm-excluded protocol removes Dec 23–25 from every window, which shrinks N_3 from 5,304 to 4,872 and the total from 14,904 to 14,472 cells, and re-selects two per-setting configurations, namely CSDI at 128 channels and SAITS at n_{groups}=5, lr=5×10^{-4}, that screening prefers once the storm no longer penalizes higher-capacity diffusion. The storm-excluded round windows and the full multi-seed ranking are reported below in Tables 5 and 6.

Excluding the storm improves every method and contracts the spread across models: the best storm-excluded W-MAE is $\approx$ 2.47 km/h, attained jointly by CSDI at 2.471 ± 0.029 and SAITS at 2.473 ± 0.016, versus the storm-included best of 3.15 km/h from SAITS; HGBT moves from 3.36 to 2.51 and the weakest deep model SSSD from 5.90 to 3.50 km/h. Two points follow. *(i)* Although the storm window is only a small fraction of evaluation cells, it dominates imputation difficulty: removing it lowers the best pooled error by about 0.68 km/h, or about 22%, and the Round 3 collapse of the inverted-attention transformer disappears, with iTransformer becoming a flat mid-field model rather than collapsing to rank 11. *(ii)* The best learned model is protocol-dependent: SAITS leads with the storm included, whereas CSDI, re-selected at 128 channels for the storm-free setting, ties it once the sparse, low-confidence storm bins no longer penalize its higher-capacity configuration. The two benchmarks should therefore be read as a matched comparison under a common protocol, not a single fixed-configuration ablation; together they bound the practical accuracy an agency can expect with and without an extreme, data-sparse winter-storm window.

DATA AVAILABILITY STATEMENT

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request. Processed speed data (with only timestamp and speed information), along with the RWIS data, can also be made available upon reasonable request.

ACKNOWLEDGMENTS

The research described in this paper was inspired by a project funded by the New York State Department of Transportation (NYSDOT), but the specific work reported in this paper was not funded by NYSDOT. The University at Buffalo (UB) provided partial institutional and computational support for this work.

SUPPLEMENTAL ARTIFACTS

This technical report incorporates, as Figs. 1, 3, 5, and 7 and Tables 5–7, all figures and tables that accompanied the journal submission as Supplemental Material. The full per-seed results, trained-model configuration files, and processing code underlying this benchmark are additionally archived at `https://doi.org/10.5281/zenodo.20779591`.

REFERENCES

Alcaraz, J. M. L. and Strodthoff, N. (2023). "Diffusion-based time series imputation and forecasting with structured state space models." *Transactions on Machine Learning Research*.

Aljuaid, T. and Sasi, S. (2016). "Proper imputation techniques for missing values in data sets." *2016 International Conference on Data Science and Engineering (ICDSE)*, IEEE, 1–5.

Bae, B., Kim, H., Lim, H., Liu, Y., Han, L. D., and Freeze, P. B. (2018). "Missing data imputation for traffic flow speed using spatio-temporal cokriging." *Transportation Research Part C: Emerging Technologies*, 88, 124–139.

Breiman, L. (2001). "Random forests." *Machine Learning*, 45(1), 5–32.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., and Li, Y. (2018). "Brits: Bidirectional recurrent imputation for time series." *Advances in Neural Information Processing Systems*, Vol. 31.

Chan, R. K. C., Lim, J. M.-Y., and Parthiban, R. (2023). "Missing traffic data imputation for artificial intelligence in intelligent transportation systems: review of methods, limitations, and challenges." *IEEE Access*, 11, 34080–34093.

Chaudhry, A., Li, W., Basri, A., and Patenaude, F. (2019). "A method for improving imputation and prediction accuracy of highly seasonal univariate data with large periods of missingness." *Wireless Communications and Mobile Computing*, 2019(1), 4039758.

Cheng, S., Osman, N., Qu, S., and Ballan, L. (2024). "FastSTI: A fast conditional pseudo numerical diffusion model for spatio-temporal traffic data imputation." *IEEE Transactions on Intelligent Transportation Systems*, 25(12), 20547–20559.

Chou, C.-H., Huang, Y., Huang, C.-Y., and Tseng, V. S. (2019). "Long-term traffic time prediction using deep learning with integration of weather effect." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 123–135.

Cini, A., Marisca, I., and Alippi, C. (2022). "Filling the g_ap_s: Multivariate time series imputation by graph neural networks." *International Conference on Learning Representations*.

Deb, B., Khan, S. R., Hasan, K. T., Khan, A. H., and Alam, M. A. (2019). "Travel time prediction using machine learning and weather impact on traffic conditions." *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, IEEE, 1–8.

Du, W., Côté, D., and Liu, Y. (2023). "Saits: Self-attention-based imputation for time series." *Expert Systems with Applications*, 219, 119619.

Duruflé, H., Selmani, M., Ranocha, P., Jamet, E., Dunand, C., and Déjean, S. (2021). "A powerful framework for an integrative study with heterogeneous omics data: from univariate statistics to multi-block analysis." *Briefings in Bioinformatics*, 22(3), bbaa166.

El Esawey, M. (2020). "Using spatio-temporal data for estimating missing cycling counts: a multiple imputation approach." *Transportmetrica A: transport science*, 16(1), 5–22.

Goehry, B., Yan, H., Goude, Y., Massart, P., and Poggi, J.-M. (2023). "Random forests for time series." *REVSTAT-Statistical Journal*, 21(2), 283–302.

Gu, A., Goel, K., and Ré, C. (2022). "Efficiently modeling long sequences with structured state spaces." *International Conference on Learning Representations*.

Guide, T. M. (2001). "Traffic monitoring guide." *Federal Highway Administration*.

Guidehouse (2023). "New york state preparedness and response efforts — blizzard of 2022: After-action review." *After-action review*, New York State Division of Homeland Security and Emergency Services (NYS DHSES), Albany, NY (8).

Haji-Maghsoudi, S., Haghdoost, A.-a., Rastegari, A., and Baneshi, M. R. (2013). "Influence of pattern of missing data on performance of imputation methods: an example using national data on drug injection in prisons." *International journal of health policy and management*, 1(1), 69.

Henrickson, K., Zou, Y., and Wang, Y. (2015). "Flexible and robust method for missing loop detector data imputation." *Transportation Research Record*, 2527(1), 29–36.

Ho, J., Jain, A., and Abbeel, P. (2020). "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems*, Vol. 33, 6840–6851.

Kang, H. (2013). "The prevention and handling of the missing data." *Korean journal of anesthesiology*, 64(5), 402–406.

Kar, P. and Feng, S. (2023). "Intelligent traffic prediction by combining weather and road traffic condition information: a deep learning-based approach." *International Journal of Intelligent Transportation Systems Research*, 21(3), 506–522.

Kyrtsoglou, A., Asimina, D., Triantafyllidis, D., Krinidis, S., Kitsikoudis, K., Ioannidis, D., Antypas, S., Tsoukos, G., and Tzovaras, D. (2021). "Missing data imputation and meta-analysis on correlation of spatio-temporal weather series data." *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, 0496–0502.

Lai, Y.-C., Chen, S.-Y., Wang, S.-F., and Lin, B.-S. (2022). "A weather-based traffic predic-

tion system using big data techniques." *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, IEEE, 379–383.

Lee, K. J. and Carlin, J. B. (2010). "Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation." *American journal of epidemiology*, 171(5), 624–632.

Li, L., Zhang, J., Wang, Y., and Ran, B. (2018). "Missing value imputation for traffic-related time series data based on a multi-view learning method." *IEEE Transactions on Intelligent Transportation Systems*, 20(8), 2933–2943.

Li, Y., Li, Z., and Li, L. (2014). "Missing traffic data: comparison of imputation methods." *IET Intelligent Transport Systems*, 8(1), 51–57.

Liang, Y., Zhao, Z., and Sun, L. (2021). "Dynamic spatiotemporal graph convolutional neural networks for traffic data imputation with complex missing patterns." *arXiv preprint arXiv:2109.08357*.

Liu, M., Huang, H., Feng, H., Sun, L., Du, B., and Fu, Y. (2023). "Pristi: A conditional diffusion framework for spatiotemporal imputation." *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, IEEE, 1927–1939.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Long, M., and Zhang, W. (2024). "itransformer: Inverted transformers are effective for time series forecasting." *The Twelfth International Conference on Learning Representations*.

Marisca, I., Cini, A., and Alippi, C. (2022). "Learning to reconstruct missing data from spatiotemporal graphs with sparse observations." *Advances in Neural Information Processing Systems*, 35, 32069–32082.

Nevalainen, J., Kenward, M. G., and Virtanen, S. M. (2009). "Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification." *Statistics in medicine*, 28(29), 3657–3669.

Nie, Y., Nguyen, N. H., Sinthong, P., and Kalagnanam, J. (2023). "A time series is worth 64 words: Long-term forecasting with transformers." *International Conference on Learning*

Representations.

Pei, L., Cao, Y., Kang, Y., Xu, Z., and Liu, Q. (2024). "Spatiotemporal imputation of traffic emissions with self-supervised diffusion model." *IEEE Transactions on Neural Networks and Learning Systems* Early Access.

Rubin, D. B. (1976). "Inference and missing data." *Biometrika*, 63(3), 581–592.

Shabarek, A., Chien, S., and Hadri, S. (2020). "Deep learning framework for freeway speed prediction in adverse weather." *Transportation Research Record*, 2674(10), 28–41.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. (2015). "Deep unsupervised learning using nonequilibrium thermodynamics." *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2256–2265.

Song, J., Meng, C., and Ermon, S. (2021). "Denoising diffusion implicit models." *International Conference on Learning Representations*.

Tak, S., Woo, S., and Yeo, H. (2016). "Data-driven imputation method for traffic data in sectional units of road links." *IEEE Transactions on Intelligent Transportation Systems*, 17(6), 1762–1771.

Tashiro, Y., Song, J., Song, Y., and Ermon, S. (2021). "Csdi: Conditional score-based diffusion models for probabilistic time series imputation." *Advances in Neural Information Processing Systems*, 34, 24804–24816.

Umar, N. and Gray, A. (2023). "Comparing single and multiple imputation approaches for missing values in univariate and multivariate water level data." *Water*, 15(8), 1519.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is all you need." *Advances in Neural Information Processing Systems*, Vol. 30, 5998–6008.

Wang, H., Tang, X., Kuo, Y.-H., Kifer, D., and Li, Z. (2019). "A simple baseline for travel time estimation using large-scale trip data." *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2), 1–22.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., and Long, M. (2023). "Timesnet: Temporal 2d-

variation modeling for general time series analysis." *International Conference on Learning Representations*.

Xing, J., Wu, W., Cheng, Q., and Liu, R. (2022). "Traffic state estimation of urban road networks by multi-source data fusion: Review and new insights." *Physica A: Statistical Mechanics and its Applications*, 595, 127079.

Yang, X., Sun, Y., Yuan, X., and Chen, X. (2024). "Frequency-aware generative models for multivariate time series imputation." *Advances in Neural Information Processing Systems*, 37, 52595–52623.

Zhang, Y., Kong, X., Zhou, W., Liu, J., Fu, Y., and Shen, G. (2024). "A comprehensive survey on traffic missing data imputation." *IEEE Transactions on Intelligent Transportation Systems*.

TABLES

TABLE 1. Top-5 weather–speed Pearson correlations (full study period, ground-truth speed only). Sign and magnitude are reproduced from Fig. 4.

Feature	Physical interpretation	Pearson r
Level of grip	Wet/icy pavement → lower speed	+0.78
Snow Layer (water eq.)	Accumulated snow → lower speed	−0.78
Ice Layer	Ice on pavement → lower speed	−0.65
Surface Temperature	Cold pavement → lower speed	+0.37
Visibility	Poor visibility → slower speeds	+0.34

TABLE 2. Three-round evaluation protocol (storm-included). Rounds 2 and 3 probe opposite directions of December-winter generalization. The peak-winter days Dec 23–25 (Buffalo Blizzard storm peak) are retained in the Round 3 test set and in the Round 1 and Round 2 training windows. Per-round N_{eval} are recomputed from the on-disk test files. The storm-excluded sensitivity analysis is reported separately (see the Appendix and Table 5); this table describes the primary storm-included benchmark split.

Round	Training periods	Evaluation periods	N_{eval}
1	Sep 28–Oct 13, Dec 12–25, Mar 16–30	Oct 14–28, Mar 1–15	4,296
2	Sep 28–Oct 13, Dec 19–25, Mar 16–30	Oct 14–28, Dec 12–18, Mar 1–15	5,304
3	Sep 28–Oct 13, Dec 12–18, Mar 16–30	Oct 14–28, Dec 19–25, Mar 1–15	5,304

Note: Total evaluation timesteps $4{,}296 + 5{,}304 + 5{,}304 = 14{,}904$. N_{eval} excludes 24 genuinely missing speed points on 2022-10-28 that are shared across all three rounds' test sets.

TABLE 3. Selected per-model configurations (deep learning + classical baselines). All deep models are adapted for univariate block-missing speed imputation conditioned on 41 auxiliary features (36 RWIS weather/road-surface variables and 5 calendar features).

Model	Key hyperparameters	Training
Deep learning		
iTransformer	$d = 256$, 3 layers, 4 heads, $d_{\text{ff}} = 512$	100 ep., patience 10
SAITS	$d = 256$, $n_{\text{groups}} = 2$, lr 10^{-3}	1000 ep., early-stop 50
CSDI	64 diff. channels, 50 steps	200 ep., lr 10^{-3}
FGTI	$d = 256$, 256 diff. channels	200 ep., lr 5×10^{-4}
TimesNet	$d = 64$, top-$k = 3$, 2 layers	200 ep., lr 10^{-3}
BRITS	hidden 256	200 ep., lr 10^{-3}
GRIN	hidden 64	200 ep., lr 10^{-3}
PatchTST	$d = 128$, patch 12, stride 8	3 layers, 200 ep.
SSSD	64 res. channels, 8 layers	20,000 iterations
Classical baselines		
HGBT	`HistGradientBoostingRegressor`, max_depth 8	lr 0.05, max_iter 300
GBDT	`GradientBoostingRegressor`, max_depth 4	lr 0.05, $n = 200$
RandomForest	$n_{\text{est}} = 200$, default depth	`sklearn` defaults
Ridge	$\alpha = 10$	closed-form fit
LinearRegression	—	closed-form fit

TABLE 4. Final multi-seed results on the storm-included three-round benchmark. Lower is better; all MAE and RMSE values are in km/h, converted from the stored mph speed records by a factor of 1.60934. Mean ± std across seeds (deep: [42, 1337, 2024]; classical: [1, 2, 3, 4, 5]). W-MAE = $(4296\,\text{MAE}_{R1} + 5304\,\text{MAE}_{R2} + 5304\,\text{MAE}_{R3})/14904$**; W-RMSE analogously. Per-model selected configurations are listed in Table 3.**

Rank	Model	Seeds	R1 MAE (km/h)	R2 MAE (km/h)	R3 MAE (km/h)	W-MAE (km/h)	W-RMSE (km/h)
Deep models							
1	SAITS	3	2.4732 ± 0.1427	2.4829 ± 0.0317	4.3584 ± 0.0304	3.1475 ± 0.0420	6.5410 ± 0.0660
5	FGTI	3	2.2478 ± 0.1236	2.9699 ± 0.5800	5.0145 ± 0.0800	3.4894 ± 0.1762	7.8977 ± 0.0618
6	CSDI	3	2.4623 ± 0.1584	3.0083 ± 0.8372	4.8216 ± 0.2821	3.4963 ± 0.2828	8.7400 ± 1.5726
7	TimesNet	3	2.7835 ± 0.2102	3.1105 ± 0.1455	4.5829 ± 0.4263	3.5402 ± 0.2364	6.8812 ± 0.7094
10	BRITS	3	3.2934 ± 0.2173	3.2534 ± 0.1965	4.7300 ± 0.1670	3.7905 ± 0.1101	7.0438 ± 0.0908
11	iTransformer	3	2.8542 ± 0.0678	3.5858 ± 0.1822	5.6478 ± 0.0523	4.1086 ± 0.0863	8.7196 ± 0.0443
12	GRIN	3	3.1036 ± 0.2076	3.5927 ± 0.0351	6.1023 ± 0.0209	4.3449 ± 0.0787	9.3456 ± 0.0352
13	PatchTST	3	4.1783 ± 0.6429	4.3211 ± 0.4849	6.0878 ± 0.1442	4.9086 ± 0.2419	9.4634 ± 0.0930
14	SSSD	3	5.5440 ± 0.4852	5.6272 ± 0.1870	6.4514 ± 0.0195	5.8965 ± 0.1957	9.9198 ± 0.2166
Classical baselines							
2	HGBT	5	2.5043	2.5624	4.8728	3.3679	7.5158
3	GBDT	5	2.5418 ± 0.0034	2.5857 ± 0.0010	5.0366 ± 0.0134	3.4453 ± 0.0053	7.7409 ± 0.0241
4	RandomForest	5	2.7393 ± 0.0148	2.6036 ± 0.0114	4.8739 ± 0.0676	3.4507 ± 0.0214	7.3421 ± 0.1027
8	Ridge	5	2.9449	3.1447	4.6304	3.6159	6.9087
9	LinearRegression	5	2.9979	3.1461	4.6880	3.6521	6.9400

TABLE 5. Storm-excluded three-round evaluation protocol. Training and evaluation periods are non-overlapping and the Dec 23–25 Buffalo Blizzard window is excluded from every training and evaluation window; Round 3's late-winter test window is Dec 19–22 only. Per-round N_{eval} are recomputed from the on-disk test files. This is the storm-excluded counterpart of the storm-included main-text protocol (Table 2).

Round	Training periods	Evaluation periods	N_{eval}
1	Sep 28–Oct 13, Dec 12–22, Mar 16–30	Oct 14–28, Mar 1–15	4,296
2	Sep 28–Oct 13, Dec 19–22, Mar 16–30	Oct 14–28, Dec 12–18, Mar 1–15	5,304
3	Sep 28–Oct 13, Dec 12–18, Mar 16–30	Oct 14–28, Dec 19–22, Mar 1–15	4,872

Note: Total evaluation timesteps $4{,}296 + 5{,}304 + 4{,}872 = 14{,}472$. N_{eval} excludes 24 genuinely missing speed points on 2022-10-28 shared across all three rounds' test sets.

TABLE 6. Final multi-seed ranking on the storm-excluded three-round protocol (true km/h, lower is better). W-MAE is the cell-count weighted mean absolute error with storm-excluded weights 4,296/5,304/4,872 (total 14,472). Mean ± std across seeds; deterministic models show no ± std. Compare with the storm-included headline table in the main text (Table 4): the best storm-excluded W-MAE is ≈ 2.47 km/h (CSDI ≈ SAITS) versus the storm-included best of 3.15 km/h (SAITS).

Model	Type	R1	R2	R3	W-MAE	W-RMSE
CSDI	deep	2.28 ± 0.17	2.63 ± 0.04	2.46 ± 0.17	2.47 ± 0.03	4.44 ± 0.02
SAITS	deep	2.24 ± 0.05	2.90 ± 0.08	2.21 ± 0.05	2.47 ± 0.02	4.23 ± 0.12
HGBT	classical	2.37	2.69	2.45	2.51	4.66
GBDT	classical	2.55 ± 0.01	2.63 ± 0.00	2.48 ± 0.00	2.55 ± 0.00	4.67 ± 0.01
RandomForest	classical	2.68 ± 0.01	2.74 ± 0.01	2.62 ± 0.02	2.68 ± 0.01	4.84 ± 0.02
FGTI	deep	2.41 ± 0.02	3.29 ± 0.06	2.44 ± 0.06	2.74 ± 0.01	4.75 ± 0.12
TimesNet	deep	2.53 ± 0.04	3.31 ± 0.11	2.60 ± 0.08	2.84 ± 0.06	4.69 ± 0.14
BRITS	deep	2.75 ± 0.19	3.11 ± 0.04	2.66 ± 0.22	2.85 ± 0.14	4.71 ± 0.16
iTransformer	deep	2.67 ± 0.02	3.20 ± 0.00	2.67 ± 0.05	2.86 ± 0.03	4.95 ± 0.02
Ridge	classical	2.96 ± 0.00	2.99	2.87 ± 0.00	2.94 ± 0.00	4.74
LinearRegression	classical	3.02 ± 0.00	3.03 ± 0.00	2.92	2.99 ± 0.00	4.81
GRIN	deep	2.95 ± 0.14	3.57 ± 0.00	2.95 ± 0.03	3.17 ± 0.05	5.46 ± 0.04
PatchTST	deep	3.06 ± 0.19	3.62 ± 0.17	3.10 ± 0.03	3.28 ± 0.09	5.39 ± 0.09
SSSD	deep	3.36 ± 0.02	3.73 ± 0.01	3.36 ± 0.10	3.50 ± 0.03	5.66 ± 0.06
TodMean	temporal	2.80	3.18	2.72	2.91	5.09
TrMean	temporal	3.25	3.60	3.18	3.35	5.44
WdTodMean	temporal	2.77	2.93	2.77	2.83	4.95

TABLE 7. Single-season sensitivity MAE (km/h, lower is better). Each model was trained on one season's training window only and evaluated on that season's held-out test window: Autumn (train 2022-09-28 to 10-13, test 10-14 to 10-28; $N = 2{,}136$), Winter (train 2022-12-12 to 12-18, test 12-19 to 12-25, including the Dec 23–25 storm peak; $N = 1{,}008$), and Spring (train 2023-03-16 to 03-30, test 03-01 to 03-15; $N = 2{,}160$). Hyperparameters are taken verbatim from the main benchmark (a sensitivity check, not a re-tuning); seed 42 for all models, with iTransformer also run at seeds 1337 and 2024 (negligible variation, $\sigma_{\mathrm{MAE,Winter}} = 0.13$ km/h). iTransformer is the three-seed mean; all other rows are single-seed. Winter-only training degrades by roughly 5–10× relative to Autumn- or Spring-only and is roughly 2–4× worse than the same models' storm-included Round 3 MAE in the main multi-seed benchmark (Table 4), confirming the benefit of the cross-season three-round training schema.

Model	Family	Autumn-only	Winter-only	Spring-only
SAITS	Attention (imputation)	2.00	11.27	2.27
HGBT	Classical	1.85	12.36	2.48
RandomForest	Classical	1.75	13.86	2.48
LinearRegression	Classical	2.69	11.70	2.43
TimesNet	Attention (forecasting)	2.20	12.89	2.66
iTransformer	Attention (forecasting)	3.11	19.28	3.41
BRITS	Recurrent	2.08	12.68	2.75
GRIN	Graph	2.72	18.56	3.07
CSDI	Diffusion	2.16	18.11	2.53
SSSD	Diffusion	2.77	19.07	3.20