TRAFFIC24: A 24-YEAR TRAFFIC DATASET AND BENCHMARK FOR LONG-TERM FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how traffic forecasting models benefit from extended data histories and generalize across changing conditions requires reusable longitudinal datasets. We introduce Traffic24, a curated traffic dataset covering 24 years of observations in San Diego, from September 2001 to September 2025. Traffic24 contains 505.6 million station–time records at five-minute resolution, including traffic flow, speed, occupancy, truck-related measurements, and lane-level observations. Six nested historical subsets, spanning one to 24 years, support the construction of forecasting tasks with different amounts of historical data. A weather-augmented subset adds regional meteorological context aligned with traffic observations through a shared UTC time axis. Standardized schemas, explicit field definitions, and retained quality and processing indicators support consistent data access and interpretation. To establish an initial forecasting benchmark, we evaluate multiple models on selected subsets using the open-source TimeCopilot and report their predictive performance. These evaluations provide reference results for subsequent research using Traffic24. By bringing together long observational histories, structured temporal subsets, and regional weather context, Traffic24 supports research into how much historical data forecasting models can effectively use, how they generalize over time, and how environmental information can inform traffic prediction.

1 INTRODUCTION

Traffic forecasting is a core task in intelligent transportation systems: it uses historical observations and external context to characterize the temporal evolution and spatial dependence of road conditions, supporting routing, signal control, congestion management, and emergency dispatch (Chu et al., 2026; Zhang et al., 2026b). Advances in deep sequence models, graph neural networks, and Transformers have steadily improved the ability to represent complex traffic patterns (Ma et al., 2026; Zhang et al., 2025b; 2026a). Model performance, however, depends not only on network architecture but also on data coverage, temporal span, feature semantics, and quality control. For weather-aware traffic forecasting in particular, whether traffic and meteorological records can be used under a unified time axis and an explicit spatial relationship determines whether external variables form an interpretable, reproducible model input.

Existing data resources do not fully support this goal. METR-LA and the PeMS family are widely used for traffic forecasting, but most public subsets span short periods, and weather information typically must be collected and matched separately by researchers (Cheng et al., 2025; Zhang et al., 2025b). City-scale datasets such as TaxiBJ emphasize regional mobility or trajectory patterns, whose observation targets, spatial units, and temporal granularity differ from those of fixed road sensors (Lv et al., 2025; Li et al., 2025). More importantly, different studies adopt their own resampling, imputation, and station-matching rules, making the semantics and quality of weather variables difficult to compare across studies. Short observation windows also limit research on infrastructure change, evolving travel demand, sensor outages, and cross-year distribution drift.

The core challenge for long-term traffic–weather datasets is therefore not simply adding more records or fields, but building an auditable data pipeline: documenting provenance and feature semantics, performing within-modality quality control, preserving the distinction between raw observations, imputed values, and anomaly flags, and describing cross-modal relations through a unified

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

time key and station mapping (Ahmed et al., 2024; Zhang et al., 2025a; Nasser & Simon, 2025). Dataset papers should also provide technical validation for downstream tasks, showing that curated data supports stable training and discriminative evaluation, rather than presenting potential applications as already-realized performance gains.

We address this gap with Traffic24. Our contributions are threefold. First, we organize 2001–2025 observations from 261 traffic nodes and 25 weather stations in the San Diego region, retaining traffic operation, vehicle composition, and environmental features. Second, we build a reusable pipeline for timestamp normalization, missing/anomaly flagging, and a unified Parquet schema. Third, we establish a unified evaluation protocol for downstream forecasting: a 5-minute speed-prediction task comparing 29 models using MAE, RMSE, MAPE, and R^2.

2 Related Work

This section reviews prior work from two complementary perspectives: the traffic and multimodal data resources that support forecasting, and the methods that integrate traffic observations with weather and other contextual information. Together, these perspectives highlight the importance of data coverage, temporal alignment, and transparent evaluation protocols. We use them to position Traffic24 relative to existing resources and forecasting approaches.

2.1 Traffic and Multimodal Data Resources

Fixed-sensor datasets provide repeated measurements of traffic flow, speed, and occupancy, and support many established forecasting benchmarks. Trajectory-based resources capture individual or aggregate movement, but differ from fixed-sensor datasets in their spatial units, sampling patterns, and prediction tasks. Recent studies increasingly combine heterogeneous traffic observations to capture complementary spatiotemporal information (Wang et al., 2026; Lv et al., 2025; Li et al., 2025; Wu et al., 2025). These resources are not directly interchangeable: their coverage, temporal resolution, feature semantics, and processing procedures vary.

Weather provides important context for traffic conditions. Temperature, precipitation, visibility, wind, humidity, and pressure can affect road conditions, driving behavior, travel demand, and traffic dynamics, particularly during adverse weather (Shao et al., 2026; Revathi et al., 2026; Nasser & Simon, 2025). However, road sensors and weather stations often differ in spatial density and sampling frequency. Integrating their observations therefore requires clear documentation of data provenance, units, temporal semantics, quality indicators, and spatial representativeness. Such documentation helps distinguish direct observations from resampled or imputed values. Recent forecasting work likewise emphasizes explicit temporal alignment and structured fusion of traffic and contextual information (Zhang et al., 2025a). Yet auxiliary variables such as weather and events may be unavailable or inconsistently covered across locations and long time periods (Chan et al., 2025); existing multisource resources also vary in temporal coverage, station availability, metadata completeness, and processing documentation.

Traffic24 addresses this data gap by organizing long-term observations from fixed traffic sensors and providing a regional weather series aligned by UTC timestamp. Its nested historical spans support controlled investigation of data-history length, while its documented regional aggregation provides a consistent weather context. Because the weather series is regional rather than matched to individual traffic nodes, it should not be interpreted as a local measurement at each road segment.

2.2 Traffic–Weather Fusion Modeling

Classical statistical and machine-learning methods treat weather as an exogenous variable and estimate its association with traffic using regression, vector autoregression, feature selection, or tree-based models (Nguyen et al., 2022; Datla et al., 2013; Yang & Qian, 2019; Xiong et al., 2022). These methods are relatively interpretable, but may not capture nonlinear spatiotemporal dependencies. The predictive value of weather also depends on the task, time scale, and feature-selection strategy.

Deep sequence models incorporate traffic and external variables through feature concatenation, separate encoders, attention mechanisms, or explicit lag modeling (Braz et al., 2022; Zhang et al., 2023;

Wang et al., 2025). Their performance depends on accurate temporal alignment, appropriate treatment of missing data, and a well-defined evaluation protocol. In long time series, random splits can place temporally related observations in both training and test sets, producing optimistic estimates of performance on future periods.

Graph neural networks use road-network topology to model dependencies among sensors and can incorporate weather as node features or other contextual inputs (Ahmed et al., 2024; Qi et al., 2024). Applying these methods requires reliable node coordinates, network connectivity, and a documented mapping between traffic and weather observations. Without these elements, spatial performance gains may be difficult to attribute to the modeled relationships. Generative approaches have also been explored for traffic forecasting, including traffic–weather prediction under adverse conditions (Zhang et al., 2024; Wu et al., 2025). Assessing their value on long histories requires clearly specified temporal splits, out-of-time tests, extreme-weather subsets, and computational costs; dataset scale alone does not establish cross-region generalization.

Recent studies increasingly report input and output formats, feature sources, and evaluation protocols. MC-HSTA combines multisource traffic sequences with a pretraining and fine-tuning scheme (Zhang et al., 2026b); LLM-TFP describes spatiotemporal tokenizers and adapters and reports sensor and time-step dimensions for METR-LA and PEMS-BAY (Cheng et al., 2025); and DSSA-TCN details its data splits, normalization, training settings, and evaluation across multiple datasets (Zhang et al., 2025b). Other work incorporates road geometry (Franco-Panadés et al., 2026), models temporal and regional heterogeneity through adaptive graph construction (Zhang et al., 2026a), uses adaptive spatiotemporal masked pretraining (Ma et al., 2026), or jointly models temporal and spatial dependencies with attention (Chu et al., 2026). Together, these studies underscore the importance of preserving timestamps, node identities, spatial metadata, and reproducible splits alongside the observations themselves.

3 Dataset Construction

3.1 Data Sources

Traffic data originate from PeMS reports, which record road status per fixed detector station and timestamp, including flow, speed, occupancy, truck volume and share, and partial lane-level observations at 5-minute granularity; after field unification and time normalization, records can be read for arbitrary time ranges. Station activation time, device status, and valid observation length differ across stations, so long-term coverage should be understood as a collection of observation records rather than a complete, evenly spaced, gap-free station–time matrix.

Weather data come from MesoWest station observations archived by the project; the current release covers 25 weather stations and standardizes temperature, relative humidity, wind, solar radiation, precipitation, dew point, and pressure fields (Table 1). Observation period and variable completeness vary by station. Figure 1 shows the relative spatial coverage of traffic nodes and weather stations: traffic nodes are distributed along major corridors, while the sparser weather stations cover coastal, urban, and inland environments; spatial matching should therefore consider distance, shared temporal coverage, and station representativeness together, and the figure describes network coverage rather than certifying per-node weather-mapping accuracy.

3.2 Construction Pipeline

Construction centers on a unified representation of long-term observations and reusable time ranges. Traffic and weather data are separately field-unified and time-normalized; traffic records are cut by predefined historical spans, and weather is aggregated into a regional time series at 5-minute resolution and joined to traffic records by a shared UTC time key. This produces the base observation layer; it does not by itself fix the input length, prediction horizon, or split used downstream. The benchmark in Section 6 is a downstream task evaluation; it does not certify every record or establish that weather improves prediction.

Standardized records are indexed by station identifier and UTC timestamp and stored as partitioned Parquet for column- and time-range access. The traffic-only version retains 32 traffic observation and quality-related fields, including aggregate flow, occupancy, speed, truck flow, truck share, lane-

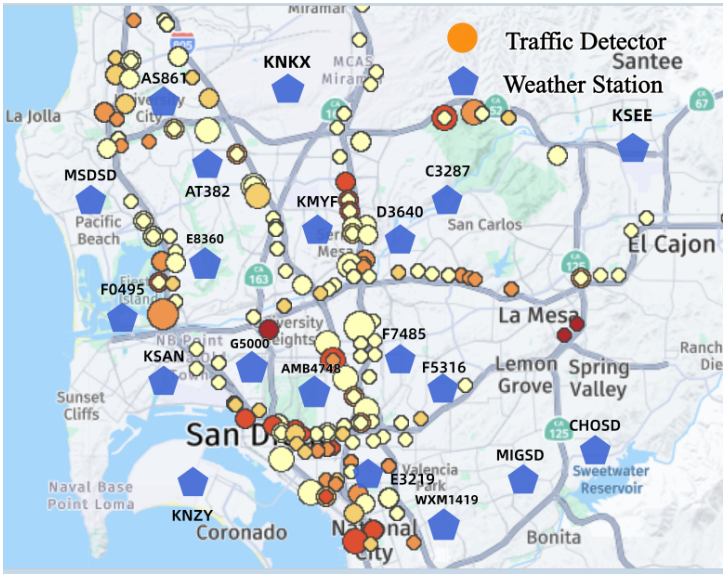


Figure 1: Spatial distribution of traffic detector nodes and weather stations in the San Diego study area.

point counts and observed ratios, and five lane-level indicators for lanes 1–5; this field count includes quality-related quantities and does not imply 32 mutually independent traffic phenomena. The traffic+weather version adds meteorological values, processing status, and station-count fields on top of the traffic schema; the base record schema and the feature set used by any particular downstream model are defined separately. Both versions are linked by UTC time key at 5-minute base granularity, using half-open intervals $[\text{start}, \text{end})$ for adjacent periods, with September 1, 2025 as the exclusive end boundary; range filtering and downstream window construction are handled separately to avoid conflating the base record span with training/validation/test splits (Section 4).

The traffic–weather version currently uses a regional weather summary for temporal linkage: at each UTC 5-minute timestamp, numeric fields are averaged and categorical fields take the mode across weather stations, then left-joined onto traffic records, preserving traffic timestamps without a weather match. All traffic stations at a given timestamp therefore share the same regional weather background; this implementation does not depend on traffic-node coordinates and is not nearest-station matching or inverse-distance-weighted interpolation. Regional aggregation reduces dependence on per-node spatial metadata but weakens local weather differences, which should be treated as a usage limitation. Weather records also retain the original point count, resampling interval, exact-observation, imputation, and out-of-range status; during aggregation, boolean flags are converted to numeric values and averaged, so their meaning becomes the share of records with that status rather than a per-observation boolean, and the weather-station count reflects how many stations contributed at that timestamp without guaranteeing equal non-missing counts per variable.

3.3 Record Schema

Each traffic record corresponds to one station at one 5-minute timestamp, with aggregate flow, occupancy, speed, truck-related quantities, and partial lane-level indicators; different fields may have different missing status, so the presence of a row does not imply all variables are valid at that timestamp. Each weather record describes the regional environmental background at the same timestamp via numeric variables, categorical status, and aggregation-related quantities; availability depends on original station configuration and period, and the weather-station count is not the valid-measurement count per variable. Weather is joined by left join, so traffic records without a weather match are retained. Cumulative precipitation, one-hour cumulative precipitation, and precipitation increment carry different temporal meanings and are not interchangeable; regional means should not be read as a single road node's local measurement. Numeric wind direction is currently arithmetic-averaged

Table 1: Core traffic and weather variables in Traffic24.

Domain	Variable	Description	Unit
Traffic	Flow	Number of vehicles passing the detector per unit time	veh / 5 min
Traffic	Speed	Average vehicle speed measured by the detector	mph
Traffic	Occupancy	Share of time a section/segment is occupied by vehicles	%
Traffic	Truck Flow	Number of trucks passing the detector per unit time	veh / 5 min
Traffic	Truck share	Fraction of observed flow attributed to trucks	%
Traffic	Lane-point count	Count of lane-level observations represented	count
Weather	Temperature	Air temperature	°C
Weather	Precipitation	Water (rain/snow/sleet/hail) falling from the sky	mm
Weather	Humidity	Water vapor content of the air	%
Weather	Wind Speed / Gust	Sustained and peak rate of air movement	m/s
Weather	Wind Direction	Compass direction the wind is blowing from	degrees
Weather	Pressure / Altimeter	Station and altimeter-corrected atmospheric pressure	Pa

Table 2: The six nested historical spans, each released as a traffic-only and a traffic+weather version (12 configurations in total). Intervals are half-open, $[\text{start}, \text{end})$.

Span	Start	End	Records	Size (traffic)	Size (traffic+weather)
1 year	2024-09-01	2025-09-01	27,430,317	0.64 GB	1.45 GB
5 years	2020-09-01	2025-09-01	134,360,205	3.08 GB	6.88 GB
10 years	2015-09-01	2025-09-01	264,243,001	6.23 GB	13.19 GB
15 years	2010-09-01	2025-09-01	379,904,539	9.07 GB	19.75 GB
20 years	2005-09-01	2025-09-01	465,709,915	11.21 GB	24.20 GB
24 years	2001-09-01	2025-09-01	505,602,607	12.14 GB	26.20 GB

rather than treated with circular statistics, which can bias values near the $0°/360°$ wrap-around — a limitation to keep in mind when selecting and interpreting weather features.

4 Dataset Composition and Versions

Traffic24 combines six historical spans (1, 5, 10, 15, 20, and 24 years) with two modality versions (traffic-only and traffic+weather), for 12 configurations (Table 2). All six windows share the end boundary September 1, 2025, with start dates September 1 of 2024, 2020, 2015, 2010, 2005, and 2001, respectively. Because the windows are nested, longer spans contain the records of shorter ones and should not be treated as statistically independent datasets. Traffic storage also includes per-year and several non-overlapping 5- and 10-year segments for staged access, separate from the 12 configurations above.

Within a span, the traffic-only and traffic+weather versions share the same start/end time and record count; the traffic+weather version is larger only because it adds weather fields, not additional rows (e.g., at 1 year, both hold 27,430,317 records, 0.64 GB vs. 1.45 GB; at 24 years, both hold 505,602,607 records, 12.14 GB vs. 26.20 GB). Record count denotes "station–5-minute-timestamp" rows, not a complete regular time axis or the number of model training samples; downstream supervised samples must be constructed separately (Section 6), and comparisons of historical span or of the weather modality's benefit should hold the evaluation period, station set, and sample-construction rule fixed, since equal row counts do not certify that weather fields are complete or that they improve accuracy.

Table 3: Traffic24 compared with commonly used traffic-forecasting datasets.

Dataset	Nodes	Time Range	Time Steps	Features
Traffic24	261	2001.09.01–2025.09.01	2,524,608	17
PeMS03	358	2018.09.01–2018.11.30	26,208	1
PeMS04	307	2018.01.01–2018.02.28	16,992	3
PeMS07	883	2017.05.01–2017.08.30	26,208	1
PeMS08	170	2017.07.01–2017.08.31	17,856	7
PeMS-Bay	325	2017.01.01–2017.05.31	43,200	3
METR-LA	207	2012.03.01–2012.06.30	34,560	3

5 Quality Control and Limitations

The construction pipeline applies schema, key, temporal-continuity, range, and cross-field checks to standardized records. These checks flag duplicate keys, time gaps, implausible values, and candidate anomalies; they do not imply that every missing value or extreme observation has been removed. The source package does not include quantitative audit results (for example, missingness by field and year, duplicate counts, or the number of flagged records), so these procedures should be read as documented quality-control rules rather than a complete empirical validation report.

Traffic quality rules include non-negativity, occupancy range, plausible speed/flow intervals, and consistency between lane-level and aggregate fields; IQR-based statistics flag candidate outliers, but whether to remove them is judged jointly with station history, neighboring windows, and sensor status to avoid discarding genuine congestion or incident responses. Weather data are checked against physical ranges, station quality flags, and temporal continuity; extreme temperature, wind, or precipitation are treated as errors only when unit, station flag, or neighboring records are inconsistent — credible extreme-weather records are retained, as they are important samples for weather-aware traffic research.

Table 3 compares Traffic24 with commonly used traffic benchmarks. Traffic24's 2,524,608 denotes the theoretical number of five-minute time steps across the 24-year window, not the 505,602,607 station–time records in Table 2; the feature count uses a unified core-input definition and does not equal the full Parquet field count. Dataset time ranges and feature counts can depend on the preprocessing version, so these comparison values should be accompanied by source citations and checked against the precise releases used.

6 Benchmark: Forecasting Task and Baselines

6.1 Problem Definition and Protocol

We define a single-step, five-minute speed-forecasting task: the model uses the preceding hour ($T = 12$ five-minute steps) to predict speed at the next step. Let $s_t \in \mathbb{R}^N$ denote the speed vector for the N traffic nodes at time t. A model f_θ maps the history window to the next speed vector:

$$f_\theta : \mathbb{R}^{T\times N} \to \mathbb{R}^N \tag{1}$$

For a prediction time t, the input comprises the preceding T speed vectors and the target is the next speed vector:

$$X_t = [s_{t-T+1}, \ldots, s_t], \quad Y_t = s_{t+1} \tag{2}$$

The reported experiment uses speed alone, $T = 12$, and 29 models at this single horizon. The split is chronological, without cross-period shuffling, and normalization parameters are estimated on training data only; exact split boundaries and the handling of missing or irregular windows are not specified in the source package. All metrics are computed after inverse normalization to mph.

Table 4: Model performance for single-step, five-minute speed forecasting (errors are in mph except MAPE; models are sorted by MAE; best values are bold).

Model	Year	MAE (mph)	RMSE (mph)	MAPE (%)	R^2
XGBoost	2016	**0.8278**	**1.5071**	1.5431	**0.9686**
LightGBM	2017	0.8342	1.5091	1.5494	0.9685
Random Forest	2001	0.8354	1.5307	1.5457	0.9676
TimesFM-3.0	2026	0.8362	1.5534	**1.5310**	0.9666
CatBoost	2018	0.8444	1.5215	1.5699	0.9680
Moirai-2.0	2025	0.8465	1.5634	1.5646	0.9662
Toto-2.0	2026	0.8686	1.5976	1.6060	0.9647
Chronos-2	2025	0.8817	1.7675	1.6831	0.9568
N-HiTS	2023	0.9067	1.6367	1.6608	0.9630
Lasso	1996	0.9131	1.6414	1.6926	0.9627
Elastic Net	2005	0.9135	1.6390	1.6974	0.9628
Linear Regression	—	0.9179	1.6264	1.6984	0.9634
Ridge	1970	0.9188	1.6265	1.6985	0.9634
PatchTST	2023	0.9189	1.6375	1.7899	0.9629
N-BEATS	2020	0.9478	1.6625	1.7258	0.9618
DOTM	2016	0.9599	1.9051	1.8537	0.9498
Median Ensemble	2025	0.9608	1.9518	1.9111	0.9473
Theta	2000	0.9622	1.9115	1.8591	0.9495
ETS	2008	0.9624	1.9609	1.9251	0.9468
CES	2022	0.9697	2.0041	1.9616	0.9445
Auto-ARIMA	2008	1.0172	2.0478	2.0553	0.9420
ADIDA	2011	1.1529	2.4539	2.2991	0.9167
IMAPA	2005	1.1529	2.4539	2.2991	0.9167
DeepAR	2017	1.4471	3.2786	3.0020	0.8513
TFT	2021	1.7134	2.2362	2.8928	0.9308
Croston	1972	1.8161	4.0239	4.0993	0.7761
Seasonal Naive	—	2.9109	6.8506	6.6194	0.3510
Prophet	2017	4.1656	6.7066	8.9102	0.3780
Historic Average	—	4.3774	7.9340	11.1391	0.1295

We evaluate point forecasts using MAE, RMSE, MAPE, and the coefficient of determination (R^2):

$$\mathrm{MAE} = \frac{1}{M}\sum_{i=1}^{M}|y_i - \hat{y}_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2}, \tag{3}$$

$$\mathrm{MAPE} = \frac{100\%}{M}\sum_{i=1}^{M}\frac{|y_i - \hat{y}_i|}{\max(|y_i|, \varepsilon)}, \qquad R^2 = 1 - \frac{\sum_{i=1}^{M}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{M}(y_i - \bar{y})^2}. \tag{4}$$

Here M is the number of evaluated node–time targets, y_i and $\hat{y}_i$ are the true and predicted speeds, $\bar{y}$ is the mean true speed, and $\varepsilon > 0$ avoids division by zero. The handling of zero-speed targets must be specified with the evaluation code because it affects MAPE.

6.2 Baselines and Results

We evaluate 29 models spanning classical statistics, linear and tree-based machine learning, deep sequence models, and pretrained time-series foundation models (Table 4), all trained and evaluated under the same 5-minute speed target and evaluation samples using the open-source TimeCopilot framework (Garza & Rosillo, 2025). All 29 models share the same input window ($T = 12$), evaluation samples, normalization, and de-normalized metric computation, so the results in Table 4 are directly comparable across model families. This baseline uses traffic speed only and is not equivalent to a paired traffic+weather comparison.

XGBoost attains the lowest MAE (0.8278 mph), the lowest RMSE (1.5071 mph), and the highest R^2 (0.9686); TimesFM-3.0 attains the lowest MAPE (1.5310%). Tree-ensemble models (XGBoost, LightGBM, Random Forest, CatBoost) and time-series foundation models (TimesFM-3.0, Moirai-

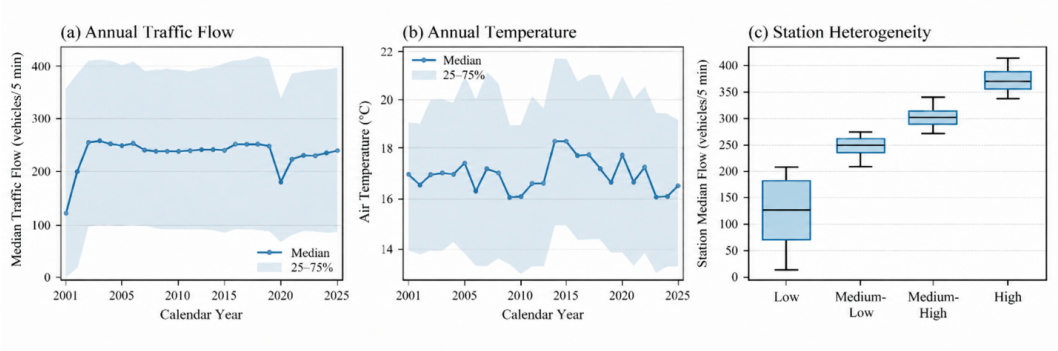


Figure 2: Exploratory summaries: (a) annual median traffic flow with the 25th–75th percentile band, (b) annual median temperature with the 25th–75th percentile band, and (c) station-level flow distributions across four groups.

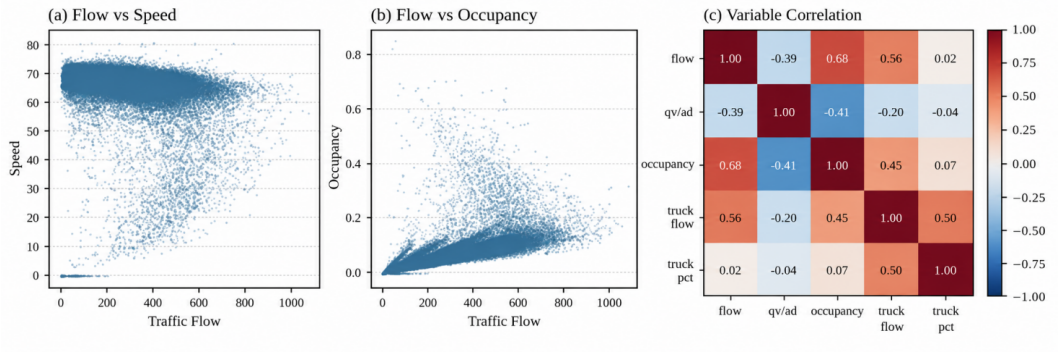


Figure 3: Exploratory traffic-variable relationships: (a) flow versus speed, (b) flow versus occupancy, and (c) pairwise correlations for five variables.

2.0) occupy the top ranks, while naive baselines (Seasonal Naive, Historic Average) and Prophet show comparatively higher error. These results show that the unified evaluation set discriminates model performance across model families but do not by themselves indicate the contribution of the weather modality.

Figures 2–3 provide exploratory summaries of temporal patterns, station-level variation, and pairwise feature relationships. These plots are descriptive: the figure does not specify the station-grouping rule or the treatment of missing and flagged values, so the apparent between-group gradient should not be interpreted as a validated population-level finding. Flow–speed and flow–occupancy relationships are also affected by congestion, detector quality, and aggregation. The summaries do not establish that a longer history improves generalization.

7 Usage Notes

We recommend reading the needed Parquet partitions by modality, year, and station, then building analysis windows from stable identifiers and the normalized time key, while also reading station metadata, native sampling granularity, units, and quality flags; reporting only numeric columns without processing status conflates genuine observations with derived records. For forecasting, we recommend a temporally ordered train/validation/test split with normalization parameters estimated only on the training period. Table 4 reports a single 5-minute speed-forecasting result; multi-step extensions should explicitly report node count N, feature count F, history window T, and each horizon H, retaining a horizon-step field. Model comparisons should hold the data partition, input window, target, normalization, and evaluation code fixed; traffic-only and traffic+weather baselines

should further share the same nodes, time window, and valid-sample set, since only under strict pairing can a performance difference be attributed to weather information.

Traffic24 supports weather-aware forecasting, anomaly analysis, out-of-time validation, and cross-year distribution-drift research. Grouped analyses should report weather thresholds, station mapping, shared temporal coverage, and group sample size; year-over-year analyses should control for the number of active stations and seasonal composition to avoid misreading missing-data patterns or sensor-network changes as traffic-system evolution. For road-safety and climate-resilience use cases, Traffic24 supplies traffic state and weather background, but without accident, construction, or control-measure variables, findings should be read as statistical association rather than causal claims about weather-induced incidents or congestion.

8 Conclusion and Future Work

Conclusion. We have presented Traffic24, a long-term traffic–weather resource for San Diego spanning September 2001 to September 2025 at five-minute base granularity, with 505,602,607 station–time records in each modality version. Exploratory summaries describe annual, diurnal, and cross-station patterns. In the traffic-only, single-step speed benchmark, XGBoost has the lowest MAE (0.8278 mph) and RMSE (1.5071 mph) and highest R^2 (0.9686), while TimesFM-3.0 has the lowest MAPE (1.5310%). The results establish reference performance under the stated evaluation setup; they do not test the effect of historical span or weather.

Future Work. Future work will extend the evaluation to multiple prediction horizons and historical spans with validated traffic–weather station mapping, and will incorporate graph-based spatial models and large-language-model prompting to further assess the contribution of the weather modality.

AI use statement

In this work, we used generative AI tools only for language translation and polishing of the manuscript text. Generative AI tools were not used for research ideation, data curation, experiments, analysis, or code. We have reviewed all AI-assisted text, and we take full responsibility for the final content of this work.

Ethics statement

Traffic24 is compiled from publicly available traffic detector reports (PeMS) and public weather station observations (MesoWest); it contains no human-subject data or personally identifiable information. The dataset is released for research use; downstream findings should be interpreted as statistical associations rather than causal claims, as discussed in the Usage Notes section. We are not aware of other ethics concerns raised by this work.

Reproducibility statement

The data construction pipeline is described in Section 3, the dataset composition and schemas in Sections 3.3 and 4, and the full evaluation protocol (split, normalization, metrics, and evaluation samples) in Section 6. All benchmark models are run with the open-source TimeCopilot framework. The dataset, anonymous download link, and exact evaluation code/configuration are not included in the source package and must be supplied or described before submission.

References

Shams Forruque Ahmed, Sweety Angela Kuldeep, Sabiha Jannat Rafa, Javeria Fazal, Mahfara Hoque, Gang Liu, and Amir H. Gandomi. Enhancement of traffic forecasting through graph neural network-based information fusion techniques. *Information Fusion*, 110:102466, 2024. doi: 10.1016/j.inffus.2024.102466.

Fernando José Braz, João C. Ferreira, Francisco Gonçalves, Kawan Weege, João Almeida, Fabiano Baldo, and Pedro Gonçalves. Road traffic forecast based on meteorological information through deep learning methods. *Sensors*, 22(12):4485, 2022. doi: 10.3390/s22124485.

Robin Kuok Cheong Chan, Joanne Mun-Yee Lim, and Rajendran Parthiban. Long-term traffic speed prediction utilizing data augmentation via segmented time frame clustering. *Knowledge-Based Systems*, 308:112785, 2025. doi: 10.1016/j.knosys.2024.112785.

Haitao Cheng, Zibin Gong, and Chang Wang. Llm-tfp: Integrating large language models with spatio-temporal features for urban traffic flow prediction. *Applied Soft Computing*, 177:113174, 2025. doi: 10.1016/j.asoc.2025.113174.

Yuquan Chu, Tingting Fu, Peng Liu, and Haksrun Lao. Sticformer: Spatio-temporal intrinsic connections transformer for traffic flow prediction. *Scientific Reports*, 16:1881, 2026. doi: 10.1038/s41598-025-31594-x.

Sandeep Datla, Prasanta Sahu, Hyuk-Jae Roh, and Satish Sharma. A comprehensive analysis of the association of highway traffic with winter weather conditions. *Procedia - Social and Behavioral Sciences*, 104:497–506, 2013. doi: 10.1016/j.sbspro.2013.11.143.

Gerard Franco-Panadés, Òscar Llatje, Laia Pagès Giralt, and Luis Jofre-Roca. Geolstm: A road geometry-aware neural network model for spatio-temporal highway traffic speed and intensity forecasting. *Expert Systems with Applications*, 305:130887, 2026. doi: 10.1016/j.eswa.2025.130887.

Azul Garza and Renée Rosillo. Timecopilot, 2025. URL https://arxiv.org/abs/2509.00616. Presented at the NeurIPS 2025 Workshop on Recent Advances in Time Series Foundation Models (BERT^{2}S), San Diego, CA, USA.

Qin Li, Zuocai Zheng, Chenyang Luo, Xuan Yang, Yong Wang, and Yuankai Wu. Joint prediction and understanding of multimodal traffic flow with a bidirectional temporal dynamic spatial hypergraph neural network model. *Transportation Research Part C: Emerging Technologies*, 180:105358, 2025. doi: 10.1016/j.trc.2025.105358.

Qinzhi Lv, Lijuan Liu, Ruotong Yang, and Yan Wang. Multimodal urban traffic flow prediction based on multi-scale time series imaging. *Pattern Recognition*, 164:111499, 2025. doi: 10.1016/j.patcog.2025.111499.

Jing Ma, Xiaobo Ma, Chen Chen, Jize Xiao, Zilong Xie, and Chunya Liu. Traffic forecasting with adaptive spatio-temporal masked pretraining. *Expert Systems with Applications*, 322:132420, 2026. doi: 10.1016/j.eswa.2026.132420.

Aram Nasser and Vilmos Simon. Traffic and weather data fusion for traffic prediction in sustainable cities. *Journal of Advanced Transportation*, 2025:1580010, 2025. doi: 10.1155/atr/1580010.

Anh Thi-Hoang Nguyen, Dung Ha Nguyen, and Trong-Hop Do. An application of vector autoregressive model for analyzing the impact of weather and nearby traffic flow on the traffic volume. In *Proceedings of the 2022 RIVF International Conference on Computing and Communication Technologies*, 2022.

Xudong Qi, Junfeng Yao, Ping Wang, Tongtong Shi, Yajie Zhang, and Xiangmo Zhao. Combining weather factors to predict traffic flow: A spatial-temporal fusion graph convolutional network-based deep learning approach. *IET Intelligent Transport Systems*, 18(3):528–539, 2024. doi: 10.1049/itr2.12401.

S. Revathi, M. P. Paulraj, and R. N. Devendra Kumar. Enhanced short-term traffic flow forecasting using a meteorological spatio-temporal transformer with multi-head attention and whale optimization. *Scientific Reports*, 16:21172, 2026. doi: 10.1038/s41598-026-51754-x.

Junqi Shao, Shen Li, Ke Zhang, Yaming Guo, and Meng Li. Leveraging llms for cross-city and cross-scenario traffic speed forecasting under abnormal weather conditions. *Transportation Research Part C: Emerging Technologies*, 188:105702, 2026. doi: 10.1016/j.trc.2026.105702.

Chaojun Wang, Shulin Huang, and Cheng Zhang. Short-term traffic flow prediction considering weather factors based on optimized deep learning neural networks: Bo-gra-cnn-bilstm. *Sustainability*, 17(6):2576, 2025. doi: 10.3390/su17062576.

Xiang Wang, Di Wu, Zirong Wang, Lingxiao Ye, Yuan Xue, and Yucong Zhang. Gadef-net: A heterogeneity-aware dual-graph framework for robust multimodal traffic forecasting. *Scientific Reports*, 16:21478, 2026. doi: 10.1038/s41598-026-51918-9.

Xiao Wu, Hua Huang, Tong Zhou, Yudan Tian, Shisen Wang, and Jingting Wang. An urban road traffic flow prediction method based on multi-information fusion. *Scientific Reports*, 15:5568, 2025. doi: 10.1038/s41598-025-88429-y.

Ming-Ming Xiong et al. Effects of rainfall on the weekday traffic flow in major cities of the beijing–tianjin–hebei region, china, in 2021. *Advances in Climate Change Research*, 13(6):858–867, 2022. doi: 10.1016/j.accre.2022.11.009.

Shuguan Yang and Sean Qian. Understanding and predicting travel time with spatio-temporal features of network traffic flow, weather and incidents. *IEEE Intelligent Transportation Systems Magazine*, 11(3):64–75, 2019. doi: 10.1109/MITS.2019.2919615.

Dingkai Zhang, Yizhe Wang, and Pengfei Wang. Adaptive graph-aware contrasting for heterogeneous traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, 188: 105685, 2026a. doi: 10.1016/j.trc.2026.105685.

Huipeng Zhang, Honghui Dong, and Zhiqiang Yang. Tsgdiff: Traffic state generative diffusion model using multi-source information fusion. *Transportation Research Part C: Emerging Technologies*, 174:105081, 2025a. doi: 10.1016/j.trc.2025.105081.

Wensong Zhang, Ronghan Yao, Xiaojing Du, Yang Liu, Rongyun Wang, and Libing Wang. Traffic flow prediction under multiple adverse weather based on self-attention mechanism and deep learning models. *Physica A: Statistical Mechanics and its Applications*, 625:128988, 2023. doi: 10.1016/j.physa.2023.128988.

Wensong Zhang, Ronghan Yao, Ying Yuan, Xiaojing Du, Libing Wang, and Feng Sun. A traffic-weather generative adversarial network for traffic flow prediction for road networks under bad weather. *Engineering Applications of Artificial Intelligence*, 137:109125, 2024. doi: 10.1016/j.engappai.2024.109125.

Zhirong Zhang, Qixiu Cheng, Zhanyi Li, Yong Fang, and Ying Li. Mc-hsta: A multi-source cross-domain hybrid spatio-temporal attention network for traffic flow prediction. *Neural Networks*, 196:108358, 2026b. doi: 10.1016/j.neunet.2025.108358.

Zhouyuan Zhang, Xin Wang, Xu Tan, and Jiatian Pi. Dssa-tcn: Exploiting adaptive sparse attention and diffusion graph convolutions in temporal convolutional networks for traffic flow forecasting. *PLOS ONE*, 20(11):e0336787, 2025b. doi: 10.1371/journal.pone.0336787.