Dean&Francis
ISSN 2959-6157

Highway Traffic Flow Prediction: Methods, Challenges, and Trends

Haoxuan Jing

Information School, University of Washington, Seattle, U.S.
Corresponding author: haoxuj@uw.edu

Abstract:

Traffic flow prediction is a core function of intelligent transportation systems. Accurate short-term forecasts help reduce congestion, support signal control, and improve travel reliability. This paper reviews major methods for highway traffic flow prediction in clear, simple English. We organize the literature into four families: traditional statistical models, classic machine learning (ML), modern deep learning (DL), and hybrid approaches that combine multiple ideas. Statistical models (e.g., ARIMA) are simple and fast but struggle with nonlinear and spatial effects. ML methods (e.g., SVM, random forests, shallow neural networks) capture nonlinear patterns better but often require manual feature design. DL methods (e.g., LSTM/RNN, CNN, and graph neural networks) learn complex spatiotemporal patterns directly from data and now achieve state-of-the-art accuracy on common benchmarks. Hybrids (e.g., ConvLSTM, graph-attention networks, decomposition+DL) further improve robustness and accuracy. We also summarize widely used datasets and metrics, discuss open challenges (generalization, long-horizon prediction, real-time deployment, interpretability), and outline future directions (transfer learning, adaptive graphs and attention, uncertainty estimation, and cloud–edge deployment). The discussion is supported by real, citable studies and public datasets.

Keywords: Traffic Flow Prediction; Spatiotemporal Modeling; IntelliHgent Transportation Systems; Deep Learning; Graph Neural Networks

1. Introduction

Highway traffic flow prediction is a core function of intelligent transportation systems (ITS). Accurate short-term forecasts (for example, 5–60 minutes ahead) support proactive signal control, ramp metering, dynamic routing, and incident response. Over the last decade, the research focus has shifted from classical time-series models toward data-driven learning, especially deep learning. Surveys show that deep neural networks (DNNs) often outperform traditional approaches because they can learn nonlinear patterns

and capture both temporal and spatial dependencies from large datasets [1].

This shift is also tied to the emergence of graph-based deep models for road networks. In graph formulations, sensors or road segments are nodes and physical connections (or learned relations) are edges. Pioneering works—STGCN(spatio-temporal graph convolution), DCRNN (diffusion convolution + recurrent sequence model), and their successors Graph WaveNet (adaptive graph with dilated temporal convolutions) and GMAN (graph multi-attention)—demonstrated consistent gains on widely used benchmarks such as METR-LA and PEMS-BAY [2, 3, 4, 5].

At the same time, public traffic datasets have grown in availability and scale. The California PeMS program collects real-time data from nearly 40,000 detectors across major metropolitan freeways, enabling reproducible comparisons and fueling the development of learning-based methods [6]. The METR-LA and PEMS-BAY speed datasets, released with DCRNN, remain standard graph benchmarks for method comparisons [3].

Goal and contributions. Building on these advances, this review (i) summarizes methods for highway traffic flow prediction across four families (statistics, ML, DL, and hybrid), (ii) formalizes the problem setting (graph notation, inputs, and outputs), (iii) lists common datasets and metrics, and (iv) highlights key challenges—spatiotemporal dependency modeling, external factor fusion, data quality, real-time deployment, and interpretability—to motivate future work [1, 2, 3, 4, 5].

2. Organization of the Text

This paper follows a simple structure that mirrors common Trans Tech submissions. After the Introduction (Section 1), Section 3 reviews recent literature and defines the prediction problem on road graphs. Section 4 summarizes datasets and evaluation metrics used in practice. Section 5 presents methods from statistics to deep learning and hybrids, with representative models and their pros/cons. Section 6 compares methods in a concise table/figure. Section 7 discusses practical issues and open challenges. Section 8 outlines future directions from a data–model–system perspective (including cloud–edge collaboration).

3. Literature Review and Problem Setting

3.1 Literature Screening (focus and scope)

We concentrate on peer-reviewed, English-language research from roughly 2016–2025 (with earlier classics when needed). The focus is short-term highway or urban freeway flow/speed forecasting. We include top venues such as IJCAI(STGCN), ICLR (DCRNN), IJCAI (Graph WaveNet), and AAAI (GMAN), as well as surveys that summarize methods, datasets, and open issues. This focus reflects the field's evolution from statistical and classic ML methods to graph deep learning with attention and adaptive graphs [1, 2, 3, 4, 5].

Within this scope, a common evaluation practice is to test on METR-LA and PEMS-BAY using horizons such as 15/30/60 minutes, reporting MAE/RMSE/MAPE. These datasets are derived from PeMS and distributed with the original DCRNN resources, which has helped standardize comparisons across papers [3, 6, 7].

3.2 Problem Definition (graph formulation)

When receiving the paper, we assume that the corresponding authors grant us the copyright to use We model a highway network as a directed graph $G=(V,E,A)$ where V is the set of sensors or road segments, E is the set of connections (e.g., adjacency based on road topology or learned proximity), and $A \in R^{|V|\times|V|}$ is the (possibly weighted) adjacency matrix. Let $X_t \in R^{|V|\times C}$ denote multivariate features at time t (e.g., speed, flow, occupancy; optionally exogenous variables like weather or incident flags). Given a history window $X_{t-\tau+1:t}=\{X_{t-\tau+1},\ldots,X_t\}$ and graph G, the goal is to learn a function:

$$f:?\backslash big\left(X_{t-\tau+1:t},G\backslash big\right)???\backslash big\left(\widehat{Y_{t+1:t+T}\backslash big}\right) \quad (1)$$

that predicts the next T steps for all nodes. In practice, models use spatial operators (e.g., graph convolution or diffusion on G) and temporal operators (e.g., recurrent units or temporal convolutions). DCRNN treats traffic propagation as a diffusion process over a directed graph and combines it with an encoder–decoder RNN for multi-step forecasting; STGCN uses graph convolution with gated temporal convolution blocks for efficient training; Graph WaveNet adds a learned adaptive adjacency and dilated temporal convolutions; GMAN uses multi-attention to emphasize important nodes and time steps [2, 3, 4, 5].

Targets and horizons. Most studies focus on short-term horizons (e.g., 15/30/60 minutes) where control actions are most effective. Longer horizons are possible but suffer from error accumulation and increased uncertainty. Encoder–decoder models with scheduled sampling (as used in DCRNN) and temporal dilations (as in Graph WaveNet) are common strategies to stabilize multi-step predictions [3, 4].

ISSN 2959-6157

3.3 Task Characteristics (why the problem is hard)

Heterogeneity. Traffic patterns vary across regions and time. Upstream bottlenecks, lane drops, ramp inflows, and route choices can produce different responses at different nodes. A single fixed adjacency may be too rigid; adaptive or learned graphs were introduced to cope with hidden dependencies and changing conditions [4].

Nonlinearity. Flow–speed relations change sharply near congestion. Incidents, weather, and control actions can create nonlinear dynamics. Deep models (RNN/CNN) help by learning nonlinear mappings; graph models further capture how these nonlinear effects propagate over the network [2, 3, 4].

Dynamics. Temporal patterns exhibit daily/weekly cycles but also sudden regime shifts. Methods therefore combine temporal memory (RNNs or temporal convolutions) with attention to focus on informative time steps; GMAN is a representative design for dynamic, long-range spatiotemporal attention [5].

Multi-scale structure. Local queues at ramps can interact with corridor-level waves; citywide detours can shift loads between parallel routes. Multi-scale temporal kernels (dilations) and attention across space–time are practical ways to represent such structure [4, 5].

Practical constraints. Real deployments must cope with missing/noisy sensors, tight latency budgets, and evolving networks. Public sources like PeMS enable robust data cleaning and benchmarking, but production systems often require cloud–edge designs and model compression for roadside or in-vehicle inference [6].

4. Datasets and Evaluation Metrics

4.1 Public Datasets

PeMS (California Performance Measurement System). PeMS is a statewide freeway monitoring system maintained by Caltrans. It collects real-time data from nearly 40,000 detectors across major metropolitan areas, and aggregates 30-second readings to 5-minute intervals. Many research benchmarks (including METR-LA and PEMS-BAY) are derived from PeMS. PeMS is the canonical source when describing U.S. freeway sensor networks and remains the backbone for reproducible traffic forecasting studies [6].

METR-LA. A widely used graph-based benchmark built from PeMS detectors in Los Angeles County. It contains 207 loop detectors with 5-minute speeds for March–June 2012 and is standard for evaluating spatiotemporal methods on 15/30/60-minute horizons. Many seminal models (e.g., STGCN, DCRNN, Graph WaveNet, GMAN) report results on METR-LA, enabling consistent, apples-to-apples comparisons [2, 3, 4, 5].

PEMS-BAY. A companion benchmark from the San Francisco Bay Area that provides a larger, denser sensor graph and the same 5-minute resolution. It is frequently paired with METR-LA in papers to validate generalization across two networks. Standard splits and preprocessing are distributed with reference implementations (e.g., DCRNN) [3, 7].

4.2 Evaluation Metrics and Protocols

Let y_i and y^i denote ground truth and prediction at index i in the test set of size n. The most common error metrics are:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n\backslash} big\left|y_i - \widehat{y_{i\backslash}} big\right| \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n\backslash} big\left(y_i - \widehat{y_{i\backslash}} big\right)^2} \tag{3}$$

$$\mathrm{MAPE} = \frac{100\backslash\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i + ?}\right| \tag{4}$$

Most studies evaluate multi-step forecasts (e.g., 12 steps for 60 minutes at 5-minute intervals) and report MAE/RMSE/MAPE at 15/30/60 minutes. The DCRNN release popularized unified train/validation/test splits for METR-LA and PEMS-BAY, which has helped standardize comparisons [3, 7].

5. Methods: From Statistics to Deep Learning

5.1 Statistics-Based Models

Historical Average, ARMA/ARIMA, Seasonal ARIMA, Kalman Filtering, VAR. These methods are fast, interpretable, and suitable as baselines or when data and compute are limited. They assume stationarity or weak nonlinearity and typically model only temporal dynamics at a single site, which limits robustness to incidents, regime shifts, and cross-road interactions. As the field moved to complex networks and multi-step horizons, their accuracy gap against data-driven models became evident, but they remain useful for sanity checks and deployment fallbacks. For broad reviews and context [1, 8].

5.2 Classic Machine Learning (ML)

Historical Average, ARMA/ARIMA, Seasonal ARIMA, Kalman Filtering, VAR. These methods are fast, interpretable, and suitable as baselines or when data and compute

are limited. They assume stationarity or weak nonlinearity and typically model only temporal dynamics at a single site, which limits robustness to incidents, regime shifts, and cross-road interactions. As the field moved to complex networks and multi-step horizons, their accuracy gap against data-driven models became evident, but they remain useful for sanity checks and deployment fallbacks. For broad reviews and context [1, 8].

5.3 Deep Learning (DL)

5.3.1 Sequence Models

RNN/LSTM/GRU. Recurrent networks learn temporal dependencies directly from sequences. LSTMs/GRUs reduce error over linear baselines on short-term horizons; however, they treat locations independently unless paired with spatial inputs or modules. They are often used as temporal backbones inside larger spatiotemporal architectures. Early deep studies like Lv et al. helped spark the transition to DL for traffic [9, 10].

5.3.2 Convolutional and ConvLSTM

CNNs extract spatial patterns by arranging sensors on grids or by convolving along corridors; ConvLSTM fuses CNN (space) and LSTM (time). These models perform well when a grid embedding is meaningful, but mapping irregular networks to grids can distort topology, motivating graph methods that operate directly on the road graph [(10)].

5.3.3 Graph Neural Networks (GNNs).

GNNs natively model roads as graphs, combining spatial operators(graph/diffusion convolution) with temporal operators (temporal conv or recurrent units):

STGCN uses spatiotemporal blocks (graph convolution + gated temporal convolution) for efficient end-to-end learning on traffic graphs [2].

DCRNN views traffic propagation as a directed diffusion process and pairs diffusion convolution with an encoder–decoder RNN and scheduled sampling for multi-step stability; it established strong baselines on METR-LA/PEMS-BAY [3].

Graph WaveNet adds a learned adaptive adjacency (node-embedding–based) and dilated 1-D temporal convolutions to capture long sequences and hidden spatial ties [4].

GMAN introduces multi-attention across space and time to focus on influential nodes and periods for multi-step prediction [5].

ASTGCN explores spatial–temporal attention and decouples recent/daily/weekly patterns [11, 12].

Practice note. These models report state-of-the-art results on METR-LA and PEMS-BAY under unified splits and horizons, providing robust baselines for future work [2, 3, 4, 5, 12].

5.3.4 Transformers and Attention.

Transformer-style models leverage global attention and parallelism to model very long temporal ranges; when combined with graph layers or positional encodings, they can capture network-wide patterns and long-range effects. GMAN is a representative spatiotemporal attention model tailored to traffic graphs [5].

5.4 Hybrid and Decomposition Approaches

Decomposition + Learner. Trend/seasonal/noise decomposition (or wavelet/EMD variants) feeds simpler components to DL or ML backends, improving stability and interpretability.

Model Fusion. Linear components are handled by ARIMA/VAR, while nonlinear residuals are learned by DL (e.g., ConvLSTM or GNN), reducing error and providing graceful degradation when one component fails.

Multi-branch DL. Separate spatial (CNN/GNN) and temporal (TCN/LSTM) branches with attention for relevance often yield the most robust results across unusual events and topology changes. Recent surveys document consistent improvements from such hybrids over single-backbone models [1].

6. Practical Issues and Open Challenges

Adaptive or learned graphs. Fixed distance-based adjacency can miss causal ties such as ramp interactions and parallel facilities. Graph WaveNet learns an adaptive dependency matrix from node embeddings, improving generalization and enabling long temporal receptive fields via dilated temporal convolutions. Follow-up work continues to explore learned topology and time-varying relations [4].

Multi-attention spatiotemporal encoders. Attention helps models focus on influential nodes and time steps. GMAN introduced graph multi-attention for multi-step prediction on METR-LA/PEMS-BAY; ASTGCNdecouples recent/daily/weekly patterns with spatial–temporal attention. These ideas also ease interpretability for operators [5, 12].

Stabilizing multi-step forecasts. Encoder–decoder schemes with scheduled sampling (as in DCRNN) and temporal-conv alternatives (e.g., STGCN's gated temporal convolution) help control error accumulation for 15/30/60-minute horizons—key for deployable signal control and routing [2, 3].

More broadly, recent surveys recommend combining these advances with transfer learning and pre-training to improve cross-city generalization, and with uncertainty

ISSN 2959-6157

estimation to support risk-aware decisions (e.g., when to trigger incident responses or hold a conservative green time plan) [1].

7. Future Directions: Data–Model–System Co-Design

7.1 Data Layer: Multimodality, Standards, and Privacy

Future datasets should move beyond speed/flow/occupancy to include weather, incidents, work-zones, events, and even image/text streams, because these exogenous factors often drive regime shifts in traffic. Surveys of deep learning for traffic prediction repeatedly call for standardized multimodal benchmarks and clear evaluation protocols that include these variables so results are comparable across methods [1]. This will support fair testing of models under atypical conditions (storms, lane closures, special events) and reduce overfitting to "clean" speed-only corpora.

At the same time, large freeway programs such as Caltrans PeMS show that long-running, high-coverage sensing is feasible at scale (nearly 40,000 detectors capturing 30-second data aggregated to 5-minute intervals). These programs enable reproducible research but also highlight the need for data quality controls (imputation, anomaly detection) and privacy-aware sharing when integrating third-party feeds. Building future public benchmarks on top of PeMS-like pipelines, with well-documented cleaning and split rules, will help the community converge on durable standards [6].

7.2 Model Layer: Adaptive Graphs, Attention, and Long Horizons

Methodologically, three themes stand out:

· Adaptive or learned graphs. Fixed distance-based adjacency can miss causal ties such as ramp interactions and parallel facilities. Graph WaveNet learns an adaptive dependency matrix from node embeddings, improving generalization and enabling long temporal receptive fields via dilated temporal convolutions. Follow-up work continues to explore learned topology and time-varying relations [4].

· Multi-attention spatiotemporal encoders. Attention helps models focus on influential nodes and time steps. GMAN introduced graph multi-attention for multi-step prediction on METR-LA/PEMS-BAY; ASTGCNdecouples recent/daily/weekly patterns with spatial–temporal attention. These ideas also ease interpretability for operators [5, 12].

· Stabilizing multi-step forecasts. Encoder–decoder schemes with scheduled sampling (as in DCRNN) and temporal-conv alternatives (e.g., STGCN's gated temporal convolution) help control error accumulation for 15/30/60-minute horizons—key for deployable signal control and routing [2, 3].

More broadly, recent surveys recommend combining these advances with transfer learning and pre-training to improve cross-city generalization, and with uncertainty estimation to support risk-aware decisions (e.g., when to trigger incident responses or hold a conservative green time plan) [1].

7.3 System Layer: Cloud–Edge Collaboration and MLOps

Real deployments face tight latency, compute limits at roadside units, and data drift. A practical pattern is cloud–edge collaboration: run short-horizon inference and data filtering at the edge for low delay; perform model training, long-horizon planning, and fleet-wide analytics in the cloud. Public documentation from Caltrans underscores the scale and continuity of freeway sensing (tens of thousands of stations), which argues for automated monitoring, retraining, and validation (MLOps) rather than ad-hoc updates. Lightweight variants of graph models (or distilled students of SOTA models) are natural candidates for edge inference [6].

8. Conclusion

We reviewed highway traffic flow prediction from statistics and classic ML to deep learning and hybrids, with emphasis on graph neural networks that operate directly on road topology. Representative models—STGCN, DCRNN, Graph WaveNet, GMAN/ASTGCN—have set strong baselines on METR-LA and PEMS-BAY, thanks to better handling of spatiotemporal dependencies and multi-step horizons. Looking ahead, we argue for data–model–system co-design: richer multimodal datasets and standardized protocols; adaptive-graph and attention-based architectures with transfer/uncertainty tooling; and cloud–edge pipelines that keep inference fast and retraining reliable. With these pieces in place, traffic centers can deliver more accurate, robust forecasts that translate into smoother operations and safer, greener mobility.

References

[1] Yin, X.; Wu, G.; Wei, J.; Shen, Y.; Qi, H.; Yin, B. Deep Learning on Traffic Prediction: Methods, Analysis and Future Directions. IEEE Transactions on Intelligent Transportation Systems, 23(8), 2021: 4927–4943.

[2] Yu, B.; Yin, H.; Zhu, Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting.

IJCAI 2018, pp. 3634–3640.
[3] Li, Y.; Yu, R.; Shahabi, C.; Liu, Y. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. ICLR 2018 (arXiv:1707.01926).
[4] Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Chang, X.; Zhang, C. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. IJCAI 2019 / arXiv:1906.00121.
[5] Zheng, C.; Fan, X.; Wang, C.; Guo, J. GMAN: A Graph Multi-Attention Network for Traffic Prediction. AAAI 2020, 34(01): 1234–1241.
[6] Caltrans. Performance Measurement System (PeMS) Data Source. (Official program overview; nearly 40,000 individual detectors; 30-second data aggregated to 5-minute intervals.)
[7] DCRNN GitHub repository. Reference implementation and standardized splits for METR-LA/PEMS-BAY (including preprocessing and evaluation protocols).
[8] Vlahogianni, E.I.; Karlaftis, M.G.; Golias, J.C. Short-term traffic forecasting: Where we are and where we're going. Transportation Research Part C, 43 (2014): 3–19.
[9] Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.-Y. Traffic Flow Prediction with Big Data: A Deep Learning Approach. IEEE Transactions on Intelligent Transportation Systems, 16(2), 2015: 865–873.
[10] Zhao, Z.; Chen, W.; Wu, X.; Chen, P.C.; Liu, J. LSTM Network: A Deep Learning Approach for Short-Term Traffic Forecast. IET Intelligent Transport Systems, 11(2), 2017: 68–75.
[11] Ma, X.; Dai, Z.; He, Z.; Ma, J.; Wang, Y.; Wang, Y. Learning Traffic as Images: A Deep CNN for Large-Scale Transportation Network Speed Prediction. Sensors, 17(4), 2017: 818.
[12] Guo, S.; Lin, Y.; Feng, N.; Song, C.; Wan, H. Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. AAAI 2019.