

Short-Term Traffic Prediction: From Statistical Models to Spatial-Temporal Deep Learning

Mohammad Sadraei

Mohammadsadraei2001@gmail.com

K. N. Toosi University of Technology

Research Article

Keywords: Short-Term Traffic Prediction, Traffic Flow Forecasting, Intelligent Transportation Systems, Spatio-Temporal Modeling, Graph Neural Networks, Transformers, Urban Traffic Networks

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11053190/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Short-term traffic prediction (STTP) is a core component of Intelligent Transportation Systems (ITS), supporting proactive traffic management, congestion mitigation, and reliable traveler information. This review synthesizes the methodological evolution of STTP from classical statistical approaches to modern spatio-temporal deep learning. Its literature is organized according to the principal modeling capabilities required by traffic forecasting: temporal dynamics, nonlinear relationships, spatial dependencies, network topology, and long-range interactions. Statistical models, classical machine learning, recurrent and convolutional neural networks, graph-based spatio-temporal models, transformer-based architectures, and hybrid frameworks are compared with respect to prediction horizon, network scale, interpretability, computational burden, and deployment feasibility. The synthesis indicates that important recent gains arise from explicitly representing spatial structure jointly with temporal dynamics, rather than from increasingly complex temporal models alone. At the same time, higher predictive capacity often introduces costs in training, inference, interpretability, and data dependence. The review therefore frames contemporary STTP as an accuracy-complexity-deployability trade-off and identifies open directions in scalable graph learning, uncertainty-aware forecasting, explainability, heterogeneous-data integration, and real-time/edge deployment. Because reported metrics are not directly comparable across studies with different datasets and experimental protocols, the review emphasizes qualitative evidence and modeling capability rather than unsupported cross-study numerical rankings.

1. Introduction

Accurate short-term traffic prediction (STTP) provides a foundation for proactive traffic management, congestion mitigation, signal-control support, and real-time traveler information. Forecast horizons commonly range from a few minutes to approximately one hour, over which traffic states can change rapidly because of recurrent demand, incidents, weather, and network interactions.

Growth in loop-detector, probe-vehicle, GPS, and connected-vehicle data has increased both the resolution and scale of traffic observations. This abundance enables data-driven prediction but also exposes limitations of models that assume linearity, stationarity, or independence among sensor locations. In large urban networks, traffic at one location is affected by upstream and downstream conditions and by non-Euclidean road topology.

Accordingly, STTP research has progressed from statistical time-series models to nonlinear machine learning, deep sequence models, graph neural networks (GNNs), and attention/transformer architectures. The resulting literature is broad, but higher model complexity does not automatically imply operational superiority: accuracy gains must be considered together with computational cost, interpretability, scalability, robustness, and deployment constraints.

Existing surveys often emphasize a particular model family or period. This review instead uses a capability-oriented taxonomy to connect methodological evolution with the practical question of when added spatial and temporal modeling complexity is justified.

1.1 Research Questions and Contributions

RQ1. How do the capabilities and limitations of major STTP model families differ in representing temporal dynamics, spatial dependencies, and nonlinear traffic behavior?

RQ2. How should STTP model families be selected according to temporal complexity, spatial/network dependence, prediction horizon, and operational constraints?

RQ3. Under what traffic-data and network conditions do explicitly spatio-temporal models provide meaningful advantages over predominantly temporal approaches?

RQ4. How does the predictive benefit of increasing model complexity compare with its computational, interpretability, scalability, and deployment costs?

2. Review Methodology

This article is structured as a critical literature review rather than a statistical meta-analysis. The scope covers the methodological development of short-term traffic prediction, with emphasis on work published from 2000 to 2025 and selected foundational studies published earlier when required to explain the origin of major model families.

Literature is organized around six methodological families: (1) statistical/state-space models, (2) classical machine learning, (3) temporal deep learning, (4) graph-based spatio-temporal models, (5) transformer/attention-based models, and (6) hybrid frameworks. Relevant studies address short-term traffic flow, speed, travel-time, or closely related network-state forecasting and provide methodological or empirical evidence useful for comparing these families.

For representative studies, the synthesis considers the modeled traffic variable, temporal mechanism, spatial representation, network setting, reported strengths and limitations, and implications for computational or operational deployment. Because datasets, train/test splits, horizons, sensor configurations, and error metrics differ across the literature, numerical results from unrelated studies are not treated as directly comparable. The review therefore does not construct synthetic MAE curves or claim a universal performance ranking.

The current manuscript does not document enough information to claim a PRISMA-style systematic search. Exact databases, search strings, search dates, duplicate removal, screening counts, and exclusion reasons should be added if the article is later presented as a systematic review. These details are deliberately not invented in this revision.

3. Capability-Oriented Taxonomy of STTP Models

A useful taxonomy distinguishes models by what they represent, rather than chronology alone. Important dimensions include temporal memory, nonlinear mapping, explicit spatial dependence, dynamic or learned topology, and long-range/global interaction. Table 1 summarizes these capabilities across the six model families.

Table 1
Capability-oriented taxonomy of short-term traffic prediction model families.

Family	Representative methods	Primary capability	Interpretability	Compute
Statistical/state-space	ARIMA/SARIMA, Kalman, VAR	Explicit temporal/state structure	High	Low
Classical ML	SVR, KNN, RF, boosting	Nonlinear mapping; engineered features	Medium	Low-Medium
Temporal deep learning	RNN, LSTM, GRU, TCN/CNN	Nonlinear temporal representation	Low-Medium	Medium
Graph spatio-temporal	STGCN, DCRNN, Graph WaveNet, STFGNN	Road-graph + temporal dynamics	Low	Medium-High
Transformer/attention	STTN, GMAN, later variants	Long-range/global interactions	Low	High
Hybrid	ARIMA-LSTM, CNN-LSTM, GCN-LSTM	Complementary mechanisms	Variable	Medium-High

3.1 Operational Mechanisms and Application Suitability

Model selection in short-term traffic prediction should not be determined solely by reported predictive accuracy. Different forecasting architectures are designed to represent different characteristics of traffic dynamics, including linearity, nonlinearity, temporal dependence, spatial interaction, network topology, and long-range spatio-temporal correlation. Consequently, the suitability of a model depends on the structure of the target transportation system, the available data, the forecast horizon, and operational constraints.

Classical statistical approaches, particularly ARIMA and SARIMA, model traffic primarily as a temporally ordered stochastic process. ARIMA represents current conditions through autoregressive relationships with previous observations, differencing to improve stationarity, and moving-average terms associated with previous errors; SARIMA additionally represents recurring seasonal structure. These properties make such models defensible for relatively stable, predominantly local traffic series with regular temporal patterns and limited spatial interaction. This can include isolated freeway or rural-highway segments, but road class alone is not a sufficient selection criterion. Their suitability decreases as traffic becomes strongly nonlinear, nonstationary, incident-driven, or dependent on conditions elsewhere in the network.

Classical machine-learning methods such as SVR, KNN, random forests, and boosting relax strict linear assumptions and can represent nonlinear mappings between engineered predictors and future traffic states. They are useful for small- to medium-scale feature-rich problems, but temporal and spatial dependence generally must be supplied through lagged variables, neighboring measurements, or other engineered features rather than learned from an explicit road-network representation.

Recurrent neural architectures, particularly LSTM and GRU, directly learn nonlinear temporal dependencies from sequential observations. Their gating mechanisms regulate the retention and updating of historical information. LSTM provides flexible memory for longer sequential dependencies, whereas GRU uses a simpler gated structure with fewer parameters and can offer a favorable accuracy-efficiency balance. These models are therefore

appropriate when temporal complexity is substantial but an explicit spatial graph is unnecessary or unavailable. Importantly, a standalone GRU or LSTM is not inherently a spatial model.

Graph-based architectures address this spatial limitation by representing sensors or road segments as nodes and physical or learned relationships as edges. STGCN combines graph convolution with temporal convolution, whereas DCRNN models directional traffic propagation through diffusion convolution embedded in a recurrent architecture. Adaptive approaches such as Graph WaveNet, AGCRN, and STFGNN relax reliance on a single fixed adjacency structure. These methods are particularly defensible for connected urban or freeway networks in which congestion propagation and upstream-downstream interaction materially influence future states.

Attention- and Transformer-based approaches extend representational capacity toward long-range and potentially global dependencies. Models such as GMAN and STTN use attention to model spatial and temporal interactions, while later approaches such as PDFormer introduce traffic-specific mechanisms for dynamic spatial dependence and propagation delay. STAEformer illustrates a complementary direction in which stronger spatio-temporal representation is combined with a comparatively standard Transformer. These models are most compelling when dependencies are distant, dynamic, or difficult to represent with local fixed-graph operations, although their computational and data requirements can be substantial.

The methodological progression from statistical forecasting to deep spatio-temporal learning should therefore be interpreted as a response to increasing representational demands rather than as a universal replacement sequence. The practical objective is to use the least complex model that adequately captures the temporal and spatial dependencies of the target system, and to justify additional complexity through gains that remain meaningful under deployment constraints. The operational mechanisms and application suitability of the major model families are summarized in Table 3. The resulting capability-oriented model-selection framework is illustrated in Fig. 1.

Table 3
Operational mechanisms and application suitability of major STTP model families.

Method/family	How it operates	Best-matched traffic/data condition	Why it is suitable	Main caution
ARIMA/SARIMA	Autoregression, differencing, moving-average errors; optional seasonal terms	Regular, predominantly local or seasonal series; isolated corridor/sensor	Transparent low-cost temporal modeling	Weak under strong nonlinearity, incidents, or network interaction
SVR/KNN/RF	Kernel regression, similarity retrieval, or tree ensembles	Small/medium feature-rich nonlinear problems	Captures nonlinear mappings without deep architectures	Temporal/spatial structure depends on engineering
LSTM	Gated recurrent memory	Strong nonlinear temporal dependence and longer sequential memory	Learns temporal dependencies directly from sequences	No inherent road-network topology; sequential compute
GRU	Simplified gated recurrent memory	Strong temporal dynamics when efficiency matters	Fewer gates/parameters than LSTM with strong temporal capacity	No inherent spatial representation
GCN/GNN	Neighborhood aggregation on a graph	Connected networks with meaningful spatial dependence	Explicitly represents road/sensor topology	Sensitive to graph definition; needs temporal mechanism for forecasting
STGCN	Graph convolution + temporal convolution	Known-topology urban/network-wide forecasting	Joint local spatial and temporal representation	Basic form relies on a fixed graph
DCRNN	Directed diffusion convolution + recurrent modeling	Propagation-sensitive directed road networks	Represents upstream-downstream diffusion over time	Higher training/inference complexity
Adaptive graph models	Learns or fuses spatial relations from data	Physical adjacency is incomplete or time-varying	Can discover functional/nonlocal dependence	Learned relations can be difficult to validate
Transformer/attention	Self-attention over temporal or spatio-temporal representations	Long-range dependencies and larger datasets	Direct modeling of distant interactions; parallelizable training	High memory/data demand; temporal attention alone is not spatial

Method/family	How it operates	Best-matched traffic/data condition	Why it is suitable	Main caution
Spatio-temporal Transformer	Spatial/graph attention + temporal attention/embedding	Large, dynamic networks with complex long-range interactions	Combines global spatial and temporal representation	High complexity; deployment and interpretability concerns

4. Evidence-Based Synthesis of Methodological Evolution

The framework is intentionally conditional rather than hierarchical: if spatial interaction is weak, a temporal model may remain preferable even in a large dataset; conversely, a connected network with strong propagation can justify explicit graph modeling even at short horizons. Model family should therefore be matched to the dominant dependency structure and deployment constraints rather than selected solely by architectural novelty.

4.1 Statistical and State-Space Foundations

ARIMA/SARIMA and Kalman-filter approaches remain important baselines because they are mathematically transparent, computationally light, and useful when traffic dynamics are sufficiently regular or explicit state estimation is required. Their principal limitation is restricted representational capacity under strong nonlinearity, nonstationarity, abrupt incidents, and large network interactions.

4.2 Classical Machine Learning as a Nonlinear Bridge

Support vector regression, KNN, random forests, and related methods relaxed strict linear assumptions and can capture nonlinear recurrent patterns. Their effectiveness depends strongly on feature design. Without explicit network representations, spatial dependence is commonly encoded indirectly through selected neighboring measurements or engineered variables.

4.3 Temporal Deep Learning

LSTM and GRU architectures improved nonlinear sequential modeling and reduced reliance on handcrafted temporal features. CNN and temporal-convolution approaches provide efficient local pattern extraction. Yet a temporal model applied independently at each detector cannot directly represent congestion propagation across a road graph, motivating joint spatio-temporal modeling.

4.4 Graph-Based Spatio-Temporal Learning

Graph neural networks treat sensors or road segments as nodes connected by physical or learned relations. STGCN (Yu et al., 2018) combines graph convolution with temporal convolution, while DCRNN (Li et al., 2018) models traffic propagation through diffusion convolution embedded in a recurrent architecture. Later approaches such as Graph WaveNet and STFGNN extend this idea through adaptive or fused graph structures. Their central benefit is explicit modeling of non-Euclidean spatial dependence; their costs include graph construction, training complexity, memory use, and reduced transparency.

4.5 Attention and Transformer-Based Models

Attention-based architectures such as STTN and GMAN model long-range and potentially global interactions without relying exclusively on recurrent state propagation. Their flexibility is attractive for large-scale forecasting, but attention mechanisms can be computationally expensive and data intensive, and superiority is dataset- and horizon-dependent rather than universal.

4.6 Hybrid and Emerging Directions

Hybrid models combine complementary inductive biases, including statistical preprocessing with neural forecasting, CNN-based feature extraction with recurrent temporal modeling, or graph modules with attention. Emerging directions include adaptive graphs, uncertainty estimation, physics-informed constraints, and efficient architectures.

5. Evidence-Based Analytical Synthesis

The original manuscript contained four conceptual error-versus-horizon plots. They have been removed because their numerical values were not tied to a common dataset, experimental protocol, or cited source. Cross-study MAE values cannot safely be combined into synthetic curves. The research questions those figures attempted to answer are retained as qualitative synthesis.

5.1 Prediction Horizon and Network Scale

For very short horizons and small or regular settings, statistical, machine-learning, and lightweight recurrent models can remain competitive because recent observations carry substantial predictive information and low latency may matter more than architectural complexity. As horizons lengthen and networks become more interconnected, models that represent spatial propagation and longer-range dependencies become increasingly relevant. This is a conditional tendency, not a universal ranking.

5.2 When Does Explicit Spatial Modeling Matter?

Spatio-temporal models are most defensible when measurements are distributed over a connected network and traffic states exhibit meaningful upstream-downstream or cross-link dependence. Their advantage is less certain for single-location forecasting, weak spatial correlations, or poorly specified graphs. Dynamic and adaptive graph learning attempts to address the last problem by learning relations from data.

5.3 Does Additional Complexity Produce Proportional Improvement?

The literature does not support a general rule that each increase in architectural complexity yields proportional accuracy improvement. Added graph layers, recurrent components, attention blocks, or learned adjacency mechanisms can increase capacity, but also optimization difficulty, memory consumption, inference cost, and hyperparameter sensitivity. The relevant comparison is marginal predictive gain per unit of operational cost.

6. Comparative Model Assessment

Table 3
Qualitative comparison of representative STTP approaches; universal 'best model' claims are intentionally avoided.

Method/family	Typical input	Principal strength	Principal limitation	Defensible use
ARIMA/SARIMA	Time series	Interpretable; low compute	Linearity/stationarity constraints	Regular series; baseline
Kalman filter	State-space observations	Online updating; noise handling	State-model assumptions	Real-time filtering; hybrids
SVR/KNN/RF	Engineered features	Nonlinear mapping	Spatial dependence engineered	Small/medium feature-rich problems
LSTM/GRU	Sequential data	Strong nonlinear temporal modeling	No inherent road graph	Temporal forecasting
CNN/TCN	Grid/sequence windows	Parallel/local feature extraction	Topology may be approximated	Regular grids/temporal convolution
STGCN	Graph + time series	Joint graph/temporal representation	Basic form uses fixed graph	Known-topology urban networks
DCRNN	Directed graph + time series	Directional diffusion modeling	Higher complexity	Propagation-sensitive networks
Graph WaveNet/STFGNN	Graph + time series	Adaptive/fused spatial relations	Tuning/compute burden	When fixed adjacency is insufficient
STTN/GMAN	Spatio-temporal attention inputs	Long-range/global interactions	High compute/data demand	Large networks/long dependencies
Hybrid models	Mixed	Complementary strengths	Complex integration	Problem-specific deployment

6.1 Application-Oriented Selection Guidance

Application-oriented model-selection guidance is summarized in Table 4.

Table 4

Application-oriented guidance based on model capabilities rather than unsupported universal rankings.

Application condition	Candidate approaches	Rationale / caution
Very short horizon / small network	ARIMA, Kalman, SVR, GRU	Low latency and simplicity may outweigh marginal gains from large architectures.
Nonlinear temporal sequence	LSTM, GRU, TCN	Strong temporal learning without requiring an explicit graph.
Connected urban sensor network	STGCN, DCRNN	Explicit spatial structure is useful when congestion propagates across linked locations.
Incomplete/flexible graph	Graph WaveNet, STFGNN, adaptive-graph variants	Learned relations can supplement fixed physical adjacency.
Long-range/global dependencies	STTN, GMAN, related attention models	Attention can represent distant interactions, subject to compute/data constraints.
Real-time/edge constrained	Kalman, compact GRU/TCN, compressed spatio-temporal models	Latency, memory, energy, and robustness must accompany accuracy.
Noisy/incomplete data	State-space filtering + robust/imputation-aware learning	Data-quality handling should be part of the forecasting pipeline.

7. Discussion: Accuracy, Complexity, and Deployability

The methodological history of STTP can be interpreted as a sequence of representational bottlenecks. Statistical models offered transparent temporal structure but limited nonlinear capacity; classical machine learning improved nonlinear mapping but often relied on engineered spatial features; recurrent deep learning strengthened temporal representation; graph models then addressed explicit network topology; and transformers expanded the range of learnable interactions.

This progression explains why graph-based and attention-based methods are prominent in contemporary research, but it does not justify treating them as universally superior. A model that reduces benchmark error but requires substantially greater training resources, has unstable inference latency, or cannot explain failures may be less useful for an operational traffic-management system.

Interpretability remains a practical concern. Traffic operators may need to understand whether a prediction is driven by recent local measurements, upstream congestion, learned long-range dependencies, or anomalous data. Explainable graph/attention mechanisms, feature attribution, calibrated uncertainty, and scenario-based validation can therefore be as important as small improvements in MAE or RMSE.

Scalability should be evaluated at both training and inference time. City-scale deployment introduces memory, communication, graph-update, and energy constraints that predictive metrics alone do not capture. Future benchmarking should report parameter counts, training and inference time, hardware, memory use, and robustness tests alongside accuracy.

Taken together, the evidence supports a conditional design principle: use the least complex model that adequately represents the temporal and spatial dependencies of the target system, and justify additional

complexity through measurable gains that remain meaningful under deployment constraints.

8. Future Research Directions

8.1 Uncertainty-Aware Forecasting

Operational systems require calibrated confidence or predictive distributions, particularly during incidents and distribution shift. Deterministic point forecasts alone do not communicate risk.

8.2 Dynamic and Transferable Spatial Representations

Road relationships change with congestion, incidents, control actions, and sensor availability. Adaptive graphs, transfer learning, and models that generalize across networks remain important.

8.3 Explainability and Failure Diagnosis

Research should connect learned spatial/temporal dependencies to traffic mechanisms and provide operator-usable explanations, not only post-hoc visualizations.

8.4 Efficient Real-Time and Edge Deployment

Compression, distillation, sparse attention, lightweight graph operations, and hardware-aware design are needed for latency- and energy-constrained deployment.

8.5 Heterogeneous and Connected-Vehicle Data

Future systems can integrate loop detectors, probe trajectories, connected vehicles, weather, incidents, events, and control states while addressing synchronization, missingness, privacy, and domain shift.

8.6 Reproducible Evaluation

Common splits, transparent preprocessing, strong baselines, ablation studies, uncertainty reporting, and computational-cost reporting are necessary for credible claims of progress.

9. Brief Conclusion

Short-term traffic prediction has evolved from interpretable statistical time-series methods toward increasingly expressive spatio-temporal learning. The central methodological change is the shift from forecasting isolated temporal sequences to representing traffic as a dynamic process over interconnected networks. LSTM/GRU models strengthened nonlinear temporal learning; graph-based models such as STGCN and DCRNN incorporated network topology; and attention-based architectures expanded long-range interaction modeling. However, predictive accuracy alone is an insufficient criterion for model selection. Computational burden, scalability, interpretability, robustness, data quality, and real-time feasibility determine whether an approach can translate from benchmark datasets to ITS operation. The next phase of STTP research should therefore emphasize uncertainty-aware, explainable, efficient, and transferable spatio-temporal models supported by reproducible evaluation.

9.1 Benchmark Datasets, Metrics, and Reproducibility

A major source of confusion in traffic-forecasting reviews is the comparison of results obtained under different datasets, sampling intervals, preprocessing pipelines, missing-data rules, graph definitions, and train/validation/test splits. A defensible review should therefore separate architectural conclusions from benchmark-specific numerical conclusions. METR-LA and PEMS-BAY are especially influential for traffic-speed forecasting, while PeMS03, PeMS04, PeMS07, and PeMS08 are frequently used for traffic-flow forecasting. Recent models such as STAEformer explicitly report results across several of these benchmark families, which makes cross-dataset robustness increasingly important.

MAE, RMSE, and MAPE remain the most common metrics. MAE is relatively easy to interpret and less dominated by large errors than RMSE; RMSE penalizes large deviations more strongly; and MAPE provides a relative error measure but can behave poorly when true traffic values approach zero. Consequently, claims of superiority should ideally be supported across more than one metric and more than one horizon.

The widely used 15-, 30-, and 60-minute horizons provide a useful lens for studying error accumulation. Published benchmark tables consistently show that forecast error rises with horizon for most methods. More importantly, the relative benefit of explicit spatio-temporal modeling becomes visible when temporal-only baselines are compared with graph-based and attention-based models under the same benchmark protocol.

Table 5
Example MAE comparison at common forecast horizons reported in a 2024 benchmark study. Values are presented only within the source's common experimental comparison and should not be generalized to unrelated protocols.

Model	METR-LA 15 min	30 min	60 min	PEMS-BAY 15 min	30 min	60 min
ARIMA	3.99	5.15	6.90	1.62	2.33	3.38
FC-LSTM	3.44	3.77	4.37	2.05	2.20	2.37
STGCN	2.88	3.47	4.59	1.46	1.86	2.21
DCRNN	2.77	3.15	3.60	1.38	1.74	2.07
STTN	2.67	3.08	3.56	1.36	1.67	1.95

Table 5 illustrates two robust observations. First, errors generally increase as the prediction horizon expands. Second, under this particular common comparison, graph-based and attention-based approaches improve substantially over ARIMA and FC-LSTM. However, the ranking among advanced models is narrower than the gap between classical and spatio-temporal families. This supports the central argument of this review: once strong spatio-temporal representation is present, additional architectural complexity can produce diminishing marginal improvements.

10. Recent Frontier: 2021–2026

The recent literature suggests that the field is moving away from the simple question of whether spatial information should be modeled and toward the harder question of how spatial relationships should be represented, updated, and constrained. Four trends are especially important: adaptive/dynamic graph learning, long-range attention, simpler but stronger representation learning, and deployment-oriented evaluation.

10.1 Adaptive and Dynamic Graph Learning

Early graph forecasting models commonly relied on a predefined adjacency matrix derived from road connectivity or sensor distance. This is physically meaningful but can be restrictive because functional traffic dependence changes with congestion, incidents, control strategies, and time of day. AGCRN demonstrated that graph relations can be inferred directly from traffic series while learning node-specific patterns, avoiding a mandatory predefined graph. MTGNN similarly introduced graph learning for multivariate time series in which dependencies are not known in advance. Graph WaveNet and later fusion/adaptive graph models further strengthened the case for combining physical structure with data-driven relational learning.

This development changes the interpretation of 'spatial modeling.' A model may use a road graph, a learned functional graph, multiple graphs, or a time-varying graph. These alternatives should not be treated as interchangeable. Physical adjacency offers interpretability and inductive bias; learned graphs can discover nonlocal dependencies; dynamic graphs can respond to changing conditions but are harder to validate and may learn unstable or spurious relations.

10.2 Long-Range Dependencies and Propagation Delay

PDFormer is an important example of the shift from generic attention toward traffic-specific attention design. It was motivated by three limitations of many earlier GNN approaches: static spatial dependencies, restricted short-range aggregation, and failure to explicitly represent propagation delay. PDFormer combines dynamic spatial self-attention, short- and long-range graph masking, and a delay-aware feature transformation. This is conceptually important because congestion effects do not propagate instantaneously across a network; the temporal lag associated with spatial influence is itself part of the forecasting problem.

GMAN had already shown that attention-based encoder-decoder architectures can model spatial and temporal interactions while reducing recursive error propagation in multi-step forecasting. Together, GMAN and PDFormer illustrate the evolution from graph convolution as the dominant spatial mechanism toward flexible attention that can represent distant and time-dependent interactions.

10.3 Diminishing Returns and Representation-Centric Design

STAEformer provides particularly relevant evidence for the research question posed in this review. Rather than introducing a highly specialized sequence of graph and attention modules, it emphasizes spatio-temporal adaptive embedding combined with a comparatively standard Transformer. Its reported performance across METR-LA, PEMS-BAY, and multiple PeMS datasets supports the argument that better representation of intrinsic spatio-temporal relations can matter as much as, or more than, adding architectural depth and bespoke modules.

This trend cautions against equating novelty with complexity. The design objective should be to encode the right dependencies with the least unnecessary machinery. For review purposes, models should therefore be compared not only by whether they use GNNs or Transformers, but by what inductive bias they introduce: fixed topology, adaptive topology, dynamic topology, local propagation, global attention, periodic embeddings, propagation delay, or external contextual information.

10.4 2025–2026 Developments and the Moving State of the Art

Recent work continues to combine graph enhancement with spatio-temporal Transformers, while 2025 survey evidence emphasizes static, adaptive, dynamic, multi-view, and hypergraph construction together with physics-

informed and hybrid architectures. A 2026 deep-learning review also highlights the expanding distinction between node-level and network-level forecasting and the emerging exploration of foundation/large-language-model ideas in transportation. These developments reinforce the need to avoid a frozen 'Transformer is the final stage' narrative. The field is better described as converging toward flexible spatio-temporal representation, with model families increasingly borrowing mechanisms from one another.

11. Accuracy–Complexity–Deployability Trade-Off

For operational ITS, the relevant optimization objective is multi-dimensional. A model may be statistically superior yet operationally inferior if it requires excessive GPU memory, cannot meet an inference deadline, fails when sensors disappear, or provides no useful indication of uncertainty. This review therefore proposes five dimensions for evaluating practical value: predictive accuracy, computational efficiency, scalability, robustness, and interpretability. These evaluation dimensions and their corresponding evidence requirements are summarized in Table 6.

Table 6
Proposed evaluation framework for moving beyond accuracy-only comparisons.

Dimension	Suggested evidence	Critical question
Predictive accuracy	MAE/RMSE/MAPE across horizons; stability across datasets	Is the gain consistent or limited to one benchmark/horizon?
Computational efficiency	Parameters, FLOPs, training time, inference latency, memory	Is the accuracy gain proportional to added resource demand?
Scalability	Behavior as nodes/sensors and sequence length increase	Can the method operate at metropolitan scale?
Robustness	Missing sensors, noise, incidents, distribution shift	Does performance degrade gracefully outside routine conditions?
Interpretability/uncertainty	Attribution, learned graph inspection, calibrated intervals	Can operators understand or appropriately trust the forecast?

A central weakness of much of the literature is that these dimensions are not reported together. Benchmark papers commonly provide predictive metrics but omit inference latency, energy use, or robustness under missing data. Conversely, operational studies may emphasize computational feasibility without comparing against the strongest forecasting baselines. Future reviews and primary studies should report both sides of the trade-off.

12. Implications for Model Selection

The evidence synthesized here does not justify a single universal best model. Instead, model choice should follow the structure of the forecasting problem. When a single sensor or a small regular corridor is forecast at a very short horizon, a compact temporal or statistical model may be sufficient. When congestion propagates across a connected urban network, graph-based models have a stronger inductive justification. When dependencies are distant, time-varying, or poorly captured by fixed adjacency, adaptive graphs or attention become more compelling. When deployment resources are constrained, compression or a simpler temporal model may be preferable even if a large Transformer achieves a modestly lower benchmark error.

This conditional perspective also clarifies why reported state-of-the-art rankings change rapidly. Differences of a few hundredths in MAE among advanced models can depend on preprocessing, random seeds, horizon aggregation, and benchmark conventions. Such small improvements should not be interpreted as evidence that an entire methodological family has become obsolete. The more durable scientific question is which representation and inductive bias are appropriate for the traffic mechanism and operational setting.

13. Research Gaps and Future Agenda

13.1 Reproducible and Fair Benchmarking

Future work should use transparent preprocessing, fixed data splits where appropriate, repeated runs, uncertainty around reported metrics, strong contemporary baselines, and explicit reporting of computational resources. Benchmark leaderboards are useful only when experimental protocols are genuinely comparable.

13.2 Robustness to Missing and Faulty Sensors

Real sensor networks contain outages, communication failures, drift, and asynchronous observations. Forecasting models should be tested under structured missingness and sensor failure, not only clean benchmark data.

13.3 Distribution Shift and Rare Events

Models trained primarily on recurrent traffic may fail during crashes, road closures, extreme weather, major events, or policy changes. Domain adaptation, continual learning, and event-aware forecasting deserve greater attention.

13.4 Calibrated Uncertainty

Point forecasts should increasingly be accompanied by prediction intervals, quantiles, or probabilistic distributions. Calibration is particularly important when forecasts feed control, routing, or safety-sensitive decisions.

13.5 Physics-Informed and Causally Plausible Learning

Traffic-flow theory can constrain data-driven models and reduce physically implausible predictions. Future graph and transformer models should explore stronger integration with conservation principles, propagation mechanisms, and causal traffic relationships.

13.6 Transferability Across Cities and Networks

Most models are trained and evaluated within one dataset. A practically valuable model should transfer across sensor layouts, cities, demand regimes, and network scales with limited retraining.

13.7 Edge and Real-Time Intelligence

Latency, memory, communication bandwidth, and energy should become first-class evaluation criteria. Distillation, pruning, sparse attention, quantization, and lightweight graph operations are promising directions.

13.8 Heterogeneous and Multimodal Data

Connected vehicles, probe trajectories, weather, incidents, events, signal states, and infrastructure sensors can provide complementary information. The challenge is not simply fusion but alignment, missingness, privacy, and avoiding leakage from unavailable future context.

14. Conclusion

This review traces short-term traffic prediction from statistical and state-space methods through classical machine learning, temporal deep learning, graph-based spatio-temporal networks, and attention/Transformer architectures. The evidence supports a representation-centered interpretation of this evolution. The major gains of modern forecasting arise from progressively richer treatment of nonlinear temporal behavior, network topology, dynamic spatial dependence, long-range interactions, and propagation effects.

At the same time, the literature does not support the assumption that increasing architectural complexity yields proportional predictive improvement. Common benchmark comparisons show a large historical gain when moving from temporal-only or classical approaches to strong spatio-temporal models, but the margins among contemporary advanced architectures can be much smaller. Recent approaches such as PDFormer and STAEformer illustrate two complementary directions: introducing traffic-specific mechanisms where they address a clear physical limitation, and simplifying architecture while strengthening spatio-temporal representation.

Accordingly, the next stage of STTP research should be judged by more than benchmark accuracy. Scalability, latency, robustness to missing data and distribution shift, uncertainty calibration, interpretability, transferability, and reproducibility are essential for bridging the gap between methodological innovation and operational ITS. The most promising future systems are likely to be those that achieve an explicit balance between representational power and deployability rather than maximizing complexity for its own sake.

Declarations

Author Contribution

Conceptualization, Methodology, Investigation, Literature Review, Data Curation, Formal Analysis, Visualization, Writing – Original Draft, Writing – Review & Editing, and Project Administration: Mohammad Sadraei.

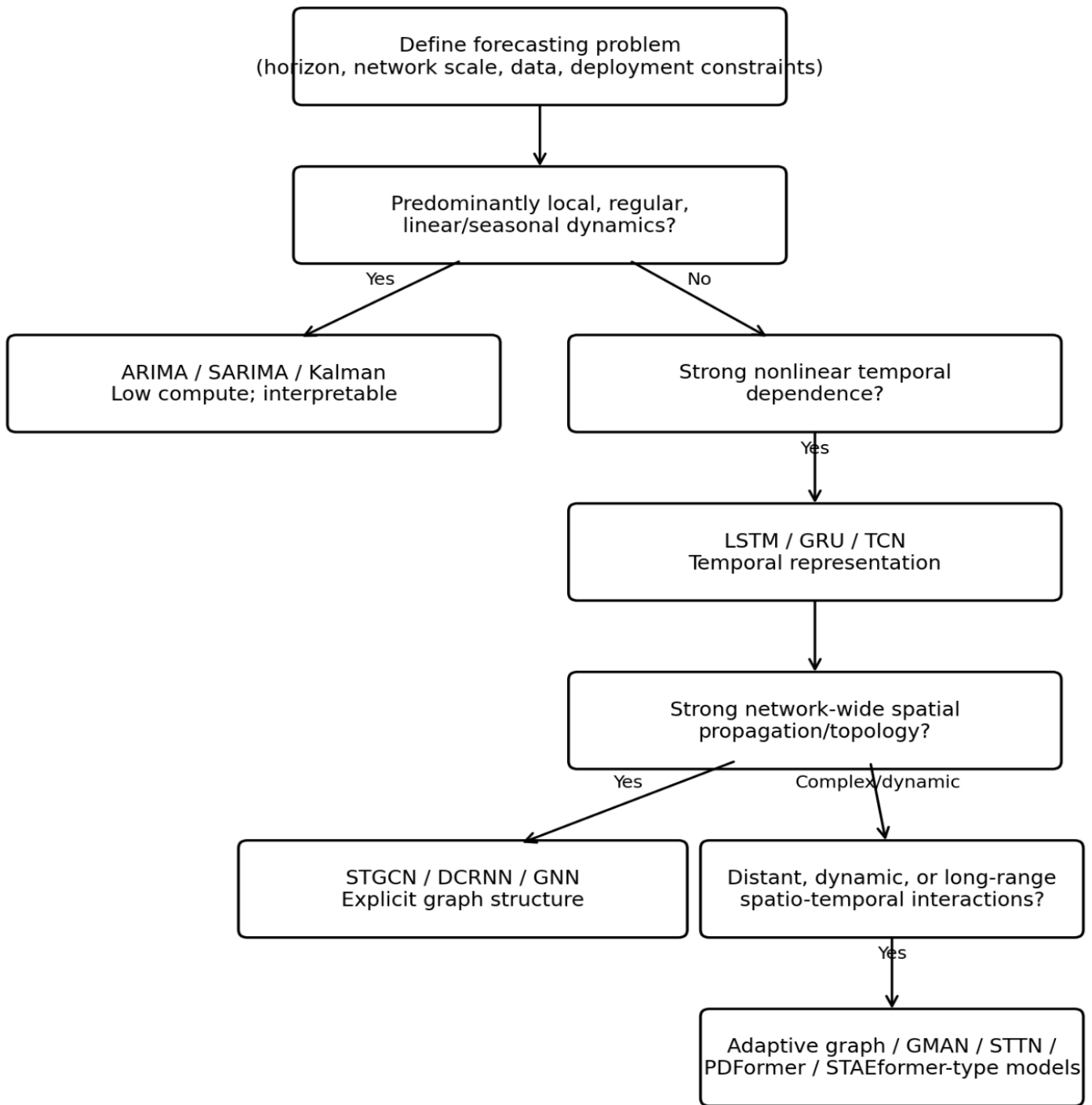
References

1. Box GEP, Jenkins GM (1970) Time Series Analysis: Forecasting and Control
2. Kalman RE (1960) A new approach to linear filtering and prediction problems. J Basic Eng 82(1):35–45
3. Sims CA (1980) Macroeconomics and reality. Econometrica 48(1):1–48
4. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297

5. Breiman L (2001) Random forests. *Mach Learn* 45:5–32
6. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9(8):1735–1780
7. Cho K et al (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation
8. Williams BM, Hoel LA (2003) Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process. *Journal of Transportation Engineering*
9. Wu C-H, Ho J-M, Lee D-T (2004) Travel-time prediction with support vector regression. *IEEE ITSC*
10. Vanajakshi L, Subramanian S, Sivanandan R (2009) Travel time prediction under heterogeneous traffic conditions using support vector regression. *Expert Systems with Applications*
11. Zheng Z, Lee D-H, Shi Q (2014) Short-term freeway traffic flow prediction: Bayesian combined neural network approach / enhanced nearest-neighbor literature as cited in the original manuscript
12. Ma X, Tao Z, Wang Y, Yu H, Wang Y (2015) Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C*
13. Lv Y, Duan Y, Kang W, Li Z, Wang F-Y (2015) Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*
14. Kumar SV, Vanajakshi L (2015) Short-term traffic flow prediction using seasonal ARIMA with limited input data. *European Transport Research Review*
15. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. *KDD*
16. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. *ICLR*
17. Vaswani A et al (2017) Attention is all you need. *NeurIPS*
18. Zhao Z, Chen W, Wu X, Chen PCY, Liu J (2017) LSTM network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*
19. Yu B, Yin H, Zhu Z (2018) Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *IJCAI*, 3634–3640
20. Li Y, Yu R, Shahabi C, Liu Y (2018) Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *ICLR*
21. Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling
22. Wu Z, Pan S, Long G, Jiang J, Zhang C (2019) Graph WaveNet for deep spatial-temporal graph modeling. *IJCAI*
23. Zheng C, Fan X, Wang C, Qi J (2020) GMAN: A graph multi-attention network for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01), 1234–1241
24. Xu M, Dai W, Liu C, Gao X, Lin W, Qi G-J, Xiong H (2020) Spatial-temporal transformer networks for traffic flow forecasting
25. Bai L, Yao L, Li C, Wang X, Wang C (2020) Adaptive graph convolutional recurrent network for traffic forecasting. *NeurIPS*
26. Wu Z, Pan S, Long G, Jiang J, Chang X, Zhang C (2020) Connecting the dots: Multivariate time series forecasting with graph neural networks. *KDD*
27. Li M, Zhu Z (2021) Spatial-temporal fusion graph neural networks for traffic flow forecasting. *AAAI*

28. Lu S, Ma X, Wang Y (2021) Combined ARIMA–LSTM model for short-term traffic flow prediction. *Alexandria Engineering Journal*
29. Jiang W, Luo J (2022) Graph neural network for traffic forecasting: A survey. *Expert Syst Appl* 207:117921
30. Jiang W et al (2023) Graph neural network for traffic forecasting: The research progress. *ISPRS Int J Geo-Information* 12(3):100
31. Jiang J, Han C, Zhao WX, Wang J (2023) PDFormer: Propagation Delay-Aware Dynamic Long-Range Transformer for Traffic Flow Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4), 4365–4373
32. Liu H, Dong Z, Jiang R, Deng J, Deng J, Chen Q, Song X (2023) Spatio-temporal adaptive embedding makes vanilla Transformer SOTA for traffic forecasting. *CIKM*, pp 4125–4129
33. Sattarzadeh S et al (2024) Extracting long-term spatiotemporal characteristics of traffic flow using attention-based convolutional transformer. *IET Intelligent Transport Systems*
34. Fan G, Sabri AQM, Rahman SSA et al (2025) Emerging trends in graph neural networks for traffic flow prediction: A survey. *Arch Comput Methods Eng* 32:4811–4855
35. Graph enhanced spatial–temporal transformer for traffic flow forecasting (2025) *Appl Soft Comput* 170:112698
36. Sun et al (2025) A self-correction Transformer network for traffic flow prediction under dynamic spatio-temporal distributions. *IET Intelligent Transport Systems*
37. A comprehensive review of traffic flow forecasting based on deep learning (2026) *Neurocomputing* 668:132269
38. Reference-quality note entries inherited from the original manuscript that were bibliographically incomplete were not silently fabricated. The newly added frontier references were selected from verifiable publisher/proceedings records. Before submission, the complete bibliography should be normalized to the exact style of the target journal and every inherited legacy citation should be checked against its source

Figures



Principle: use the least complex model that adequately represents the target dependencies.

Figure 1

Capability-oriented model-selection framework for short-term traffic prediction.