

RiskTraf: Risk-Extrapolated Residual Learning for Multi-Variate Traffic Flow Prediction

Guangyu Wang

hulnegy@gmail.com

Dongbei University of Finance & Economics
Dalian, Liaoning, China

Zhidan Liu*

zhidanliu@hkust-gz.edu.cn

Hong Kong University of Science and Technology
(Guangzhou)
Guangzhou, Guangdong, China

Abstract

Traffic sensors commonly record flow, speed, and occupancy, but standard traffic flow forecasting benchmarks and models rarely exploit all three raw measurements reliably. Although speed and occupancy provide sensor-native traffic-state information beyond flow alone, existing releases often omit these variables, replace them with proxies, or contain logically inconsistent records. Moreover, direct empirical risk minimization over three-variable inputs may exploit regime-dependent shortcuts, as the relationships among flow, speed, and occupancy vary substantially between free-flow and congested states. We introduce **PEMSB-3V**, a public benchmark suite that preserves raw flow, speed, and occupancy measurements from PeMS detectors for flow prediction. We also propose **RiskTraf**, a model-agnostic risk-extrapolated residual plug-in. For each trained spatio-temporal backbone, RiskTraf freezes the selected checkpoint and learns a lightweight zero-start residual head from historical speed and occupancy. The residual head constructs ordered traffic-risk environments and optimizes horizon-wise flow corrections with a risk extrapolation objective, thereby mitigating regime-specific shortcut correlations without modifying the backbone. Extensive experiments demonstrate that RiskTraf consistently improves diverse forecasting backbones and outperforms debiasing and distribution-shift adaptation methods. Our code and benchmark are available at <https://github.com/Guangyu4/RiskTraf>.

CCS Concepts

• **Computing methodologies** → **Spatial and temporal reasoning**; *Neural networks*; • **Information systems** → **Data mining**.

Keywords

Traffic Flow Prediction, Multi-Variate Time Series Forecasting, Risk Extrapolation, Traffic Benchmark

ACM Reference Format:

Guangyu Wang and Zhidan Liu. 2026. RiskTraf: Risk-Extrapolated Residual Learning for Multi-Variate Traffic Flow Prediction. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*, November 7–11, 2026, Rome, Italy. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3799682.3840707>

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License. *CIKM '26, Rome, Italy*.

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2539-5/2026/11

<https://doi.org/10.1145/3799682.3840707>

1 Introduction

Traffic flow prediction aims to forecast the number of vehicles passing through road segment stations over future time intervals based on historical observations. In widely used benchmarks such as PeMS (Freeway Performance Measurement System) [8], traffic flow is typically defined as the vehicle count recorded by loop detectors within fixed time windows, e.g., every 5 minutes. As a fundamental indicator of road utilization, accurate flow prediction plays a critical role in intelligent transportation systems [43], enabling traffic authorities to anticipate congestion, optimize signal control, and allocate transportation resources more effectively [35].

Recent studies have made substantial progress in traffic forecasting by designing increasingly expressive spatio-temporal models, including graph neural networks [31, 45] and attention-based architectures [32]. However, as the field matures, performance gains from purely architectural innovations have become increasingly marginal [40], while model complexity and computational costs continue to grow. This trend has motivated researchers to exploit richer contextual information for more accurate prediction. Existing efforts have incorporated external sources such as weather conditions [15, 30] and textual information [49]. Nevertheless, incorporating such auxiliary sources often requires additional collection, storage, and maintenance, thereby increasing the data-side overhead of prediction systems [10]. More importantly, they are difficult to align with traffic measurements: weather typically affects broad regions while sensors observe localized road segments [50], and textual signals such as news are updated at much coarser temporal granularity than traffic sensors [22]. Such spatial and temporal mismatches may introduce additional noise during data fusion.

Compared with external auxiliary sources, a more natural way to enrich traffic representations is to leverage multiple variables reported by the same traffic sensor. In addition to flow, loop detectors commonly record *speed* and *occupancy*, which describe complementary aspects of traffic states. Speed captures vehicle movement dynamics, while occupancy measures the fraction of time during which detectors are occupied by vehicles. Since these variables are collected from the same sensors and at the same temporal resolution as flow, they provide sensor-native auxiliary information without the spatial and temporal misalignment of external data. From an information-theoretic perspective [3], jointly modeling heterogeneous sensor variables can provide additional state information beyond any single variable. A few studies have therefore begun to exploit flow, speed, and occupancy for traffic forecasting [24, 36, 39].

However, incorporating speed and occupancy into deep forecasting models is not a straightforward extension of flow-only

prediction. Their effective use requires satisfying two coupled requirements: obtaining reliable three-variable measurements from the same set of sensors, and learning from these auxiliary variables without relying on regime-specific shortcuts.

Challenge 1: Lack of reliable multi-variate benchmarks. Existing traffic sensor releases do not guarantee that flow, speed, and occupancy are simultaneously valid and usable. Raw loop-detector records often require strict screening based on validity thresholds and traffic-flow consistency [42], and field detectors may suffer from faults such as stuck-off, stuck-on, or hanging-on behaviors [7]. Moreover, speed and occupancy are sensitive to detector configurations, lane definitions, loop lengths, and road types. As a result, a sensor that provides usable flow measurements may still be unreliable for raw three-variable modeling. This makes it difficult to systematically evaluate methods that exploit flow, speed, and occupancy under a standardized benchmark setting.

Challenge 2: Regime-dependent auxiliary correlations. Although speed and occupancy contain useful traffic-state information, their relationships with flow are not invariant across traffic regimes. For example, under congested conditions, slower vehicles may occupy loop detectors for longer durations, increasing occupancy even when the observed flow is similar. Consequently, naively treating speed and occupancy as ordinary covariates may cause models to exploit regime-specific shortcuts rather than robust predictive patterns. Models trained with empirical risk minimization (ERM) can overfit such spurious correlations, leading to degraded generalization across traffic states. Simply discarding speed and occupancy avoids this risk but also wastes valuable sensor information. Existing distribution-shift adaptation [11, 16] and debiasing methods [53] partially address related issues, but they often rely on specialized shift or bias detectors, whose effectiveness is limited by the accuracy and availability of such detectors.

To address these challenges, we introduce both a reliable benchmark suite and a risk-aware learning framework. First, we construct the **PEMSB-3V Benchmark** consisting of four datasets: PEMS03-B, PEMS04-B, PEMS07-B, and PEMS08-B. The benchmark is district-aligned with widely used PeMS datasets while preserving raw flow, speed, and occupancy measurements reported by PeMS detectors. By conducting sensor-level validation, PEMSb-3V provides a standardized testbed for studying historical three-variable inputs with flow-only forecasting targets.

Second, we propose **RiskTraf**, a **risk** extrapolation plug-in for **traffic** prediction. Rather than using speed and occupancy as unrestricted covariates or replacing existing forecasting architectures, RiskTraf treats them as historical auxiliary signals for residual correction. Specifically, given a trained three-variable backbone, we freeze its parameters and attach a lightweight zero-initialized residual head driven by historical speed and occupancy. The residual head learns horizon-wise flow corrections across ordered low-to-high traffic-risk environments, encouraging robust improvements while reducing reliance on regime-specific auxiliary correlations. RiskTraf performs rollback when validation MAE does not improve.

The main contributions of this work are summarized as follows:

- We release **PEMSB-3V**, a district-aligned benchmark suite that enables standardized evaluation of flow prediction with historical flow, speed, and occupancy measurements.

- We formulate regime-dependent auxiliary correlations as a key obstacle in multi-variate traffic forecasting and propose **RiskTraf**, a REX-based residual plug-in that improves trained backbones without replacing their architectures.
- We conduct extensive experiments on multiple PEMSb-3V datasets and diverse spatio-temporal backbones, showing that RiskTraf achieves consistent improvements over strong baselines, debiasing methods, and distribution-shift adaptation approaches.

The rest of this paper is organized as follows. Section 2 introduces the PEMSb-3V benchmark. Section 3 presents the RiskTraf framework. Section 4 reports experimental results and analyses. Section 5 discusses related work, and Section 6 concludes the paper.

2 PEMSb-3V Benchmark

We construct **PEMSB-3V**, a benchmark suite consisting of four district-level datasets: *PEMS03-B*, *PEMS04-B*, *PEMS07-B*, and *PEMS08-B*. The suite is built from the California Department of Transportation Performance Measurement System (PeMS)¹ and its official source page². PeMS is described by Caltrans as a statewide freeway monitoring system with nearly 40,000 detectors. Unlike existing releases that replace missing variables with proxies or temporal codes, PEMSb-3V retains only detectors whose native PeMS records contain all three raw measurements: *flow*, *speed*, and *occupancy*. The overall construction pipeline is summarized in Algorithm 1.

Algorithm 1 PEMSb-3V Benchmark Construction Pipeline

Require: District ID, time range $[t_{\text{start}}, t_{\text{end}}]$, completeness threshold ρ
Ensure: Multi-variate traffic benchmark $\mathcal{D}$ and adjacency matrix $\mathbf{A}$

- 1: Query detector metadata from PeMS
 - 2: Keep only detectors with raw flow, speed, and occupancy channels
 - 3: Filter detectors whose temporal completeness is below ρ
 - 4: **for** each day in $[t_{\text{start}}, t_{\text{end}}]$ **do**
 - 5: Download 5-minute flow, speed, and occupancy records
 - 6: Remove malformed timestamps and interpolate only short gaps
 - 7: **end for**
 - 8: Aggregate the cleaned records into $\mathbf{X} \in \mathbb{R}^{T \times N \times 3}$
 - 9: Construct the road topology and compute pairwise distances
 - 10: Build $\mathbf{A}$ with a distance kernel and sparsification threshold
 - 11: Split the benchmark into train/val/test by ratio 6:2:2
 - 12: **return** $\mathcal{D}, \mathbf{A}$
-

To ensure that the benchmark is built from sensors with physically interpretable three-variable measurements, we apply a metadata-based progressive screening procedure before constructing the final sensor set. The screening follows the earliest-failed-check rule. We first perform an ID coverage check. A target sensor without a corresponding PeMS metadata record is marked as *Missing Metadata*, since its static information cannot be verified. We then perform a spatial localization check. If latitude or longitude is missing, invalid, or outside the legal geographic range, the detector is marked as *Missing Geolocation* and excluded from map-based topology construction. Next, we check static structural completeness using detector length and lane count. Records with empty, non-numeric,

¹<https://pems.dot.ca.gov/>

²<https://dot.ca.gov/programs/traffic-operations/mpr/pems-source>

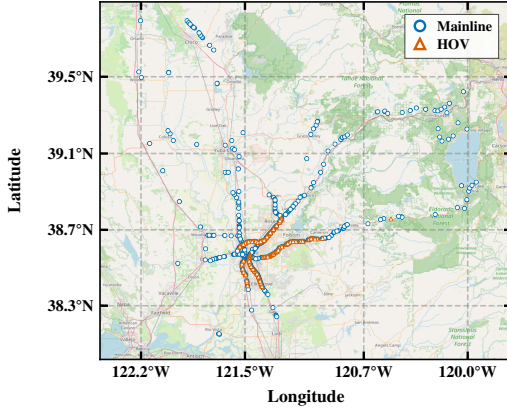


Figure 1: High-quality PEMS03-B sensors after metadata-based screening. The map shows 1,013 merged mainline/HOV detectors with complete metadata.

or non-positive values are marked as *Invalid Static Attributes*, because flow, speed, and occupancy depend on the detector segment and lane definition. Finally, we check road-type consistency. Detectors with $Type \in \{ML, HV\}$ are retained as freeway mainline or HOV sensors, whereas $Type \in \{OR, FR, FF\}$ are treated as ramp or connector sensors and excluded from the mainline subset.

Figure 1 illustrates the resulting PEMS03-B sensor footprint after metadata-based screening. Among 1,708 target sensor IDs, 41 are removed due to missing metadata and 2 due to invalid or missing geolocation. Among the remaining 1,665 geolocated records, 1,039 pass all metadata checks as complete mainline/HOV sensors, while 626 are identified as ramp or connector sensors. After merging duplicate records with the same location, road direction, and detector type, the final map contains 1,013 high-quality points, including 762 mainline and 251 HOV points. This procedure makes each discarded sensor traceable to an interpretable failure category, rather than removing sensors through opaque filtering.

The subset design follows the administrative organization of PeMS rather than an arbitrary partition. Specifically, PEMS03-B, PEMS04-B, PEMS07-B, and PEMS08-B correspond to Caltrans districts 03, 04, 07, and 08, respectively³. This preserves geographic provenance while maintaining comparability with widely used PeMS benchmarks.

We further audit representative public traffic benchmarks to examine whether they support the raw flow-speed-occupancy forecasting setting considered in this paper. As shown in Table 1, most existing benchmarks either omit speed and occupancy, replace them with proxy variables, or contain logically inconsistent records. For example, PEMS03 and PEMS07 use daily and weekly temporal encodings in place of raw speed and occupancy, while PEMS04 and PEMS08 include physically implausible speed records, such as positive speed when flow is zero and truncated low-speed ranges. These issues prevent existing benchmarks from serving as reliable testbeds for studying how raw speed and occupancy contribute to flow prediction.

³<https://cwwp2.dot.ca.gov/documentation/district-map-county-chart.htm>

Table 1: Availability and usability of raw flow, speed, and occupancy measurements in representative traffic benchmarks. ✓: usable raw measurement; ✗: not provided; ○: provided but logically inconsistent.

Source	Benchmark	Nodes	Interval	Flow	Speed	Occupancy
DCRNN [31]	METR-LA	207	5 mins	✗	✓	✗
	PEMS-BAY	325	5 mins	✗	✓	✗
LSTNet[29]	Traffic	862	1 hour	✓	○	○
STGCN[48]	PEMSD7(M)	228	5 mins	✓	○	○
	PEMSD7(L)	1,026	5 mins	✓	○	○
ASTGCN[21]	PEMSD4-I	228	5 mins	✓	○	○
	PEMSD8-I	1,979	5 mins	✓	○	○
	PEMS03	358	5 mins	✓	○	○
STSGCN[41]	PEMS04	307	5 mins	✓	○	○
	PEMS07	883	5 mins	✓	○	○
	PEMS08	170	5 mins	✓	○	○
LargeST[33]	CA	8,600	5 mins	✓	✗	✗
	GLA	3,834	5 mins	✓	✗	✗
	GBA	2,352	5 mins	✓	✗	✗
	SD	716	5 mins	✓	✗	✗
	Full_PEMS03	1,809	5 mins	✓	○	○
	Full_PEMS04	4,089	5 mins	✓	✗	✗
XXLTraffic[47]	Full_PEMS05	573	5 mins	✓	✗	✗
	Full_PEMS06	705	5 mins	✓	✗	✗
	Full_PEMS07	4,888	5 mins	✓	✗	✗
	Full_PEMS08	2,059	5 mins	✓	✗	✗
	Full_PEMS10	1,378	5 mins	✓	✗	✗
	Full_PEMS11	1,440	5 mins	✓	✗	✗
	Full_PEMS12	2,587	5 mins	✓	✗	✗
	tfNSW	27	60 mins	✓	✗	✗
TrafficDent [19]	TrafficDent*	16,972	5 mins	✓	✓	✓
PEMSB-3V	PEMS03-B	1,013	5 mins	✓	✓	✓
	PEMS04-B	2,474	5 mins	✓	✓	✓
	PEMS07-B	2,788	5 mins	✓	✓	✓
	PEMS08-B	1,515	5 mins	✓	✓	✓

*TrafficDent provides all three variables but does not apply sensor-type filtering or provide sufficient metadata to distinguish detector types.

3 Methodology

In this section, we present **RiskTraf**, a model-agnostic risk-extrapolation plug-in for multi-variate traffic forecasting. As illustrated in Figure 2, RiskTraf adopts a paired two-stage design without modifying the forecasting backbone:

- **Stage I** trains a given standard spatio-temporal backbone with 12 historical steps of traffic flow, speed, and occupancy to predict future flow.
- **Stage II** freezes the trained backbone and attaches a lightweight residual head driven by historical speed and occupancy. The residual head is optimized with a risk extrapolation objective to learn robust flow corrections across traffic-risk environments.

3.1 Problem Formulation and Paired Protocol

Let F , S , and O denote traffic flow, speed, and occupancy, respectively. For a sample ending at time $t - 1$, the historical input is

$$X_t = \{F_{t-T:t-1}, S_{t-T:t-1}, O_{t-T:t-1}\}, \quad (1)$$

where T is the input length and each variable is observed over all sensor nodes. The prediction target is future flow only:

$$Y_t = F_{t:t+H-1}, \quad (2)$$

where H is the prediction horizon. Thus, speed and occupancy are available only as historical auxiliary measurements.

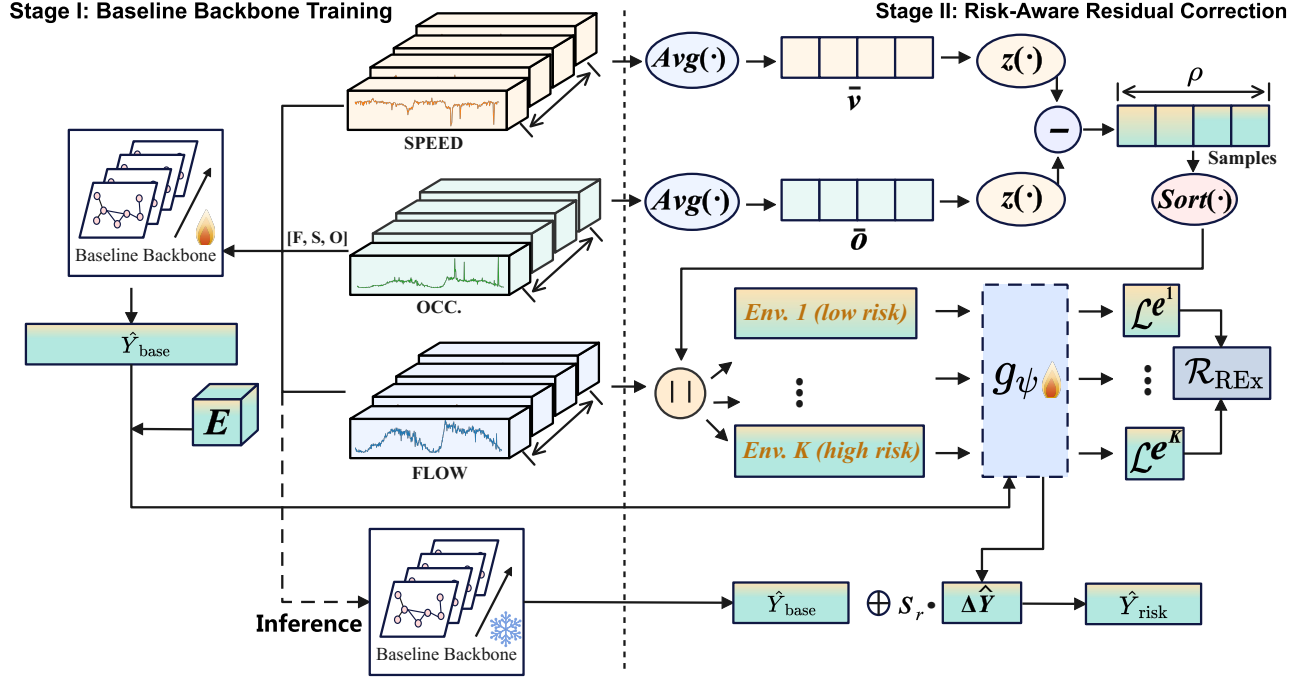


Figure 2: Overview of RiskTraf. Stage I trains a backbone to predict future flow from historical flow, speed, and occupancy. Stage II freezes the backbone, constructs traffic-risk environments from historical speed and occupancy, and learns a lightweight residual head with risk extrapolation to correct the baseline prediction.

RiskTraf is applied under a paired protocol for each dataset-backbone pair. Let B_ϕ be an arbitrary spatio-temporal backbone. The Stage-I baseline prediction is given by

$$\hat{Y}_{\text{base}} = B_\phi(X_t). \quad (3)$$

After selecting the baseline checkpoint by validation MAE, RiskTraf starts from that checkpoint, freezes the backbone parameters ϕ , and learns a residual correction head g_ψ from historical speed and occupancy:

$$\hat{Y}_{\text{risk}} = \hat{Y}_{\text{base}} + s_r \cdot g_\psi(S_{t-T:t-1}, O_{t-T:t-1}), \quad (4)$$

where s_r is a small residual scaling factor; during Stage II only the residual parameters ψ are updated.

3.2 Stage I: Baseline Backbone Training

The first stage follows the standard empirical risk minimization (ERM) setting adopted by existing traffic forecasting models. Given a historical three-variable sequence X , the backbone B_ϕ predicts the future flow sequence Y_t . The optimization objective is

$$\min_{\phi} \mathbb{E}_{(X,Y)} [\ell(B_\phi(X), Y)], \quad (5)$$

where ℓ is the masked mean absolute error (MAE) computed on inverse-scaled flow values. The checkpoint with the best validation MAE is selected as the reference checkpoint for both the vanilla backbone and the subsequent RiskTraf stage.

3.3 Stage II: Risk-Aware Residual Plug-in

The second stage attaches RiskTraf to the trained backbone obtained in Stage I. Although the Stage-I backbone has access to all three variables, its ERM objective may exploit regime-specific correlations between speed, occupancy, and flow. RiskTraf therefore freezes the backbone prediction $\hat{Y}_{\text{base}}$ and learns only a constrained horizon-wise residual correction, which keeps the plug-in model-agnostic and preserves the spatio-temporal representations learned by the backbone. RiskTraf consists of three components: a zero-start residual head, risk environment construction from historical covariates, and a REX-based objective for robust residual learning.

3.3.1 Zero-Start Residual Head. RiskTraf uses a lightweight residual head to correct the fixed backbone prediction. Since traffic sensors are not exchangeable, the same speed-occupancy pattern may imply different flow corrections at different road segments due to location, lane configuration, detector type, and upstream-downstream context. Motivated by recent studies on spatio-temporal and heterogeneity-aware embeddings [14, 32], RiskTraf introduces a minimal node-identity embedding to provide node-specific context without modifying the frozen backbone.

Specifically, RiskTraf maintains a trainable node-identity embedding table $E = [e_1, \dots, e_N]^T \in \mathbb{R}^{N \times d_e}$, where e_n denotes the embedding of sensor node n . The table is initialized at the beginning of Stage II and optimized together with ψ ; it is not copied from the backbone and requires no external sensor metadata. In addition, the head uses node-level historical summaries Q , including mean speed, mean occupancy, and occupancy-minus-speed over the input

window. The residual correction is computed as

$$\Delta \hat{Y} = g_\psi(S_{t-T:t-1}, O_{t-T:t-1}, E, Q), \quad (6)$$

where E and Q are broadcast to the batch and horizon dimensions as needed.

The last linear layer of g_ψ is initialized to zero, so RiskTraf initially reproduces the baseline prediction and any nonzero correction must be learned during the debiasing stage.

3.3.2 Risk Environment Construction. Risk environments are constructed only from historical speed and occupancy. For each sample in a training batch, we summarize the covariates over the historical window and all nodes:

$$\bar{v} = \frac{1}{TN} \sum_{t,n} v_{t,n}, \quad \bar{o} = \frac{1}{TN} \sum_{t,n} o_{t,n}, \quad (7)$$

where T is the input length and N is the number of sensor nodes. Inspired by mean alignment [18], we define a normalized traffic-risk score as

$$z(x) = \frac{x - \mu_x}{\sigma_x}, \quad \rho = z(\bar{o}) - z(\bar{v}), \quad (8)$$

where μ_x and σ_x are estimated from the training data. A larger ρ indicates higher occupancy and lower speed, corresponding to more congested or high-risk traffic states. This ordering is grounded in the fundamental diagram of traffic flow, where congestion is characterized jointly by low speed and high occupancy, so ρ ranks samples along the physical free-flow–congestion axis.

Samples are sorted by ρ and partitioned into K ordered environments as follows:

$$e^i = \{(X_j, Y_j) \mid \rho_j \in \text{quantile}_i\}, \quad i = 1, \dots, K. \quad (9)$$

This produces low-to-high traffic-risk environments without requiring incident annotations or manually defined congestion labels.

3.3.3 REx-Based Residual Learning. Given the ordered risk environments, RiskTraf trains the residual head to improve flow prediction while keeping residual errors stable across traffic regimes. For each environment e^i , we compute

$$\mathcal{L}^{e^i} = \frac{1}{|e^i|} \sum_{(X,Y) \in e^i} \ell(\hat{Y}_{\text{base}} + s_r \cdot g_\psi(S, O, E, Q), Y). \quad (10)$$

Let $\bar{\mathcal{L}} = \frac{1}{K} \sum_i \mathcal{L}^{e^i}$. The REx penalty measures the variance of environment losses:

$$\mathcal{R}_{\text{REx}} = \frac{1}{K} \sum_{i=1}^K (\mathcal{L}^{e^i} - \bar{\mathcal{L}})^2. \quad (11)$$

The full Stage-II objective is

$$\mathcal{L}_{\text{total}} = \bar{\mathcal{L}} + \lambda_r (\mathcal{R}_{\text{REx}} + \beta \cdot \mathcal{P}_{\text{pair}} + \eta \cdot \mathcal{P}_{\text{extra}}), \quad (12)$$

where β and η control the relative weights of the two order-aware penalties, defined on the ordered environment losses as

$$\mathcal{P}_{\text{pair}} = \frac{1}{K-1} \sum_{i=1}^{K-1} \left[\max(0, \mathcal{L}^{e^{i+1}} - \mathcal{L}^{e^i}) \right]^2, \quad (13)$$

$$\mathcal{P}_{\text{extra}} = \left[\max(0, \mathcal{L}^{e^K} - \text{sg}(\mathcal{L}^{e^1})) \right]^2,$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. $\mathcal{P}_{\text{pair}}$ penalizes any adjacent pair whose higher-risk environment incurs the larger residual loss, so that errors do not grow with the risk level, and

$\mathcal{P}_{\text{extra}}$ penalizes the highest-risk loss when it exceeds the detached lowest-risk reference, pushing down only the former. The penalty weight is warmed up as

$$\lambda_r = \lambda \cdot \min \left(1, \max \left(0, \frac{\tau - \tau_{\text{warmup}}}{\tau_{\text{total}} - \tau_{\text{warmup}}} \right) \right), \quad (14)$$

where τ denotes the current Stage-II epoch. This warmup allows the residual head to first learn useful corrections before enforcing stronger cross-environment consistency.

3.3.4 Validation Safeguard. Since RiskTraf is designed as a plug-in correction, it should not degrade a strong backbone. We therefore evaluate the Stage-II checkpoint on the validation set and keep the corrected model only when it improves validation MAE over the frozen baseline. Otherwise, RiskTraf rolls back to the Stage-I prediction. This keeps the plug-in conservative: it improves the backbone when useful residual patterns exist and avoids harmful corrections otherwise.

3.4 Inference

At inference, the selected Stage-I backbone first produces the baseline forecast $\hat{Y}_{\text{base}}$. If the Stage-II residual head passes the validation safeguard, RiskTraf computes $\Delta \hat{Y}$ from historical speed and occupancy, node embeddings, and node-level summaries. The final prediction is

$$\hat{Y}_{\text{risk}} = \hat{Y}_{\text{base}} + s_r \cdot \Delta \hat{Y}. \quad (15)$$

If the residual head is rolled back, the output remains the Stage-I baseline prediction.

4 Experiments

We conduct experiments on the PEMS3V benchmark suite introduced in Section 2. The experiments are designed to evaluate whether the proposed benchmark provides physically meaningful three-variable traffic data, and whether RiskTraf can improve traffic flow prediction as a model-agnostic residual plug-in. We organize the evaluation around the following research questions:

- **RQ1: Data validity.** Does PEMS3V exhibit physically meaningful flow–speed–occupancy relationships?
- **RQ2: Overall effectiveness.** Can RiskTraf consistently improve diverse spatio-temporal forecasting backbones?
- **RQ3: Method comparison.** How does RiskTraf compare with existing debiasing and distribution-shift adaptation methods?
- **RQ4: Component analysis.** How do auxiliary variables, risk penalties, and environment granularity affect RiskTraf?
- **RQ5: Practicality and interpretability.** What overhead does RiskTraf introduce, and what corrections does it learn under ambiguous traffic states?

4.1 Experimental Setup

Datasets and forecasting setting. We evaluate RiskTraf on four PEMS3V subsets: PEMS03-B, PEMS04-B, PEMS07-B, and PEMS08-B. All experiments follow the same three-variable-to-flow forecasting setting. The input contains 12 historical time steps of flow, speed, and occupancy, corresponding to one hour of observations, and the target contains the next 12 time steps of flow. All metrics

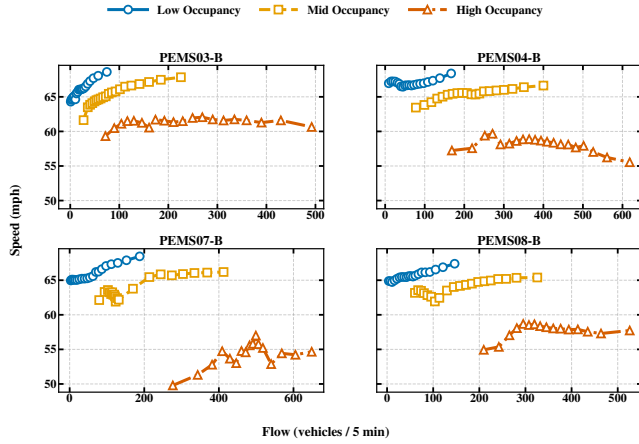


Figure 3: Flow-speed relationships under different occupancy states on the four PEMS3V subsets. Samples are grouped into low-, mid-, and high-occupancy states; each curve shows the average speed at different flow levels.

are computed on inverse-scaled flow values. We report MAE, RMSE, and MAPE, where lower values indicate better performance.

Backbones. To evaluate model-agnostic applicability, we apply RiskTraf to representative spatio-temporal forecasting backbones from different architectural generations, including STGCN [48], DCRNN [31], AGCRN [4], Graph WaveNet [46], GMAN [51], GTS [38], STEMGNN [5], STNorm [13], STWA [17], MegaCRN [26], HimNet [14], and STDN [6]. These backbones cover recurrent, graph-based, attention-based, normalization-based, and recent heterogeneity-aware designs.

Paired plug-in protocol. For each dataset-backbone pair, we first train a vanilla three-variable backbone using standard empirical risk minimization. RiskTraf then starts from the same validation-selected checkpoint, freezes the backbone, and trains only the lightweight residual head driven by historical speed and occupancy. This paired protocol ensures that any improvement comes from the RiskTraf plug-in rather than from a different backbone initialization or training recipe.

Implementation details. For RiskTraf, each training batch is sorted by the speed-occupancy risk score and partitioned into four ordered traffic-risk environments unless otherwise specified. The residual head is trained with masked MAE and the warmed-up risk extrapolation penalty described in Section 3.3. Baseline backbones and compared methods are implemented using official code when available or reproduced following the original papers, with hyperparameters selected by validation MAE under the same protocol. We use Adam for optimization, validation MAE for model selection, and rollback to the Stage-I checkpoint when the plug-in does not improve validation MAE. The batch size is 32 unless otherwise specified. All experiments are conducted on NVIDIA L20 GPUs.

4.2 Dataset Characteristics (RQ1)

We first examine whether PEMS3V preserves meaningful relationships among flow, speed, and occupancy. Figure 3 plots flow-speed curves under low-, mid-, and high-occupancy states. Across the

four displayed subsets, higher occupancy generally corresponds to lower speed, which is consistent with the loop-detector mechanism that vehicles occupy detectors for longer durations under congested conditions. For the same flow level, speed can vary substantially across occupancy states, indicating that flow alone is insufficient to identify the underlying traffic regime.

The figure also shows that occupancy is not a trivial proxy for flow. High-occupancy states span a wide range of flow levels and exhibit distinct speed-flow patterns, rather than simply corresponding to the highest flow values. Moreover, the slope and shape of the flow-speed curves differ across occupancy groups and benchmark subsets, suggesting regime-dependent auxiliary correlations. These observations support the use of speed and occupancy as informative traffic-state indicators, while motivating RiskTraf to exploit them through risk-aware residual learning rather than direct unconstrained fusion.

4.3 Overall Performance (RQ2)

Table 2 reports MAE, RMSE, and MAPE averaged over horizons 3, 6, and 12. For each backbone, the vanilla row denotes the Stage-I model trained with historical flow, speed, and occupancy, while the +RiskTraf row denotes the paired residual plug-in result.

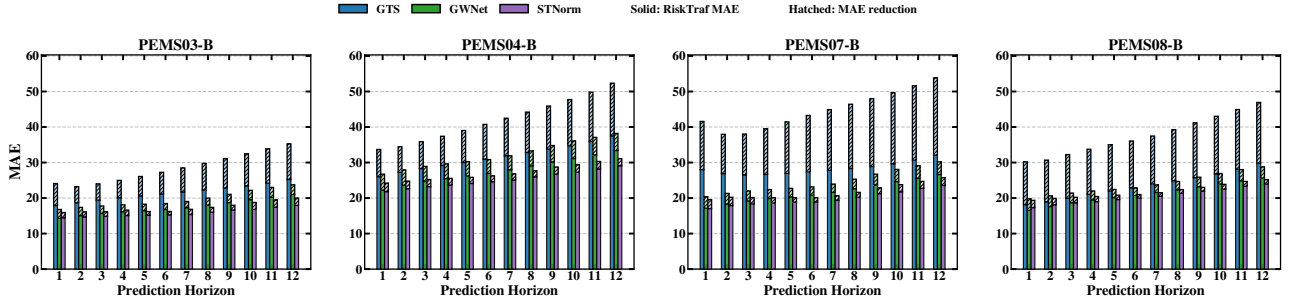
The results lead to the following observations.

- **RiskTraf consistently improves heterogeneous backbones.** Across the displayed backbones and PEMS3V subsets, adding RiskTraf reduces MAE and RMSE in nearly all cases, supporting its model-agnostic plug-in design.
- **Larger gains appear on backbones more sensitive to shifted auxiliary correlations.** DCRNN [31] and GTS [38] obtain particularly large improvements, suggesting that risk-aware residual learning is especially useful when direct three-variable ERM is unstable.
- **Strong recent backbones still benefit from the plug-in.** Competitive models such as STWA [17], MegaCRN [26], HimNet [14], and STDN [6] also improve with RiskTraf, showing that the plug-in complements rather than replaces advanced backbone architectures.
- **MAPE degradation reflects denominator sensitivity.** A few worse MAPE results do not contradict the MAE/RMSE gains, since MAPE is unstable for zero or near-zero actual values [23] and behaves like MAE weighted by inverse target magnitude [12]. Thus, better abnormal-regime corrections can still be penalized by small-flow denominators.

Figure 4 further examines whether the improvements persist at individual prediction horizons. Each bar decomposes the baseline MAE into the remaining error after applying RiskTraf and the absolute error reduced by the plug-in, where the hatched cap directly represents the removed MAE. Across all four subsets, RiskTraf reduces errors on most horizons, showing that the gains are not caused by averaging a few favorable steps. The improvement is especially visible for GTS [38], while GWNet [46] and STNorm [13] obtain smaller but still consistent reductions. This suggests that RiskTraf provides larger corrections for backbones with larger horizon-wise errors, while still refining more stable backbones.

Table 2: Main plug-in results on PEMS03-B, averaged over horizons 3, 6, and 12. Lower is better; MAPE is in percentage form. Green/red arrows mark improvement/degradation over the backbone row.

Method	PEMS03-B			PEMS04-B			PEMS07-B			PEMS08-B		
	MAE↓	RMSE↓	MAPE(%)↓	MAE↓	RMSE↓	MAPE(%)↓	MAE↓	RMSE↓	MAPE(%)↓	MAE↓	RMSE↓	MAPE(%)↓
STGCN	16.30	27.04	31.20	24.73	37.94	14.70	19.66	32.95	10.81	19.43	32.87	19.06
+RiskTraf	15.86↑	26.27↑	28.27↑	24.23↑	37.32↑	13.77↑	18.66↑	32.29↑	9.57↑	18.75↑	32.13↑	17.86↑
DCRNN	47.38	71.99	203.44	56.39	80.63	41.98	110.81	141.90	141.00	53.82	77.69	78.09
+RiskTraf	25.72↑	39.23↑	140.30↑	30.93↑	45.61↑	19.03↑	25.59↑	41.89↑	13.37↑	25.80↑	39.82↑	29.89↑
AGCRN	17.24	29.39	25.17	25.71	39.59	14.46	20.03	33.78	9.10	19.96	33.76	17.81
+RiskTraf	16.61↑	27.65↑	24.45↑	25.27↑	38.84↑	14.17↑	19.05↑	32.77↑	8.62↑	19.31↑	32.92↑	17.14↑
GWNet	19.65	33.05	25.05	32.13	48.89	17.34	24.61	41.25	10.97	23.92	39.89	19.47
+RiskTraf	17.43↑	29.07↑	23.92↑	27.75↑	42.20↑	15.35↑	21.69↑	36.84↑	9.91↑	21.22↑	35.56↑	18.22↑
GMAN	14.07	23.68	23.38	23.26	36.00	13.06	16.98	30.65	9.28	17.06	30.55	15.95
+RiskTraf	13.74↑	23.46↑	23.66↓	22.12↑	34.91↑	12.43↑	15.94↑	29.79↑	7.55↑	16.43↑	29.81↑	15.43↑
STEMGNN	16.09	26.58	26.97	26.18	40.31	14.67	45.99	79.86	74.35	20.85	34.59	21.96
+RiskTraf	15.68↑	25.80↑	26.54↑	25.29↑	38.99↑	14.18↑	40.64↑	64.59↑	59.04↑	19.73↑	33.02↑	19.40↑
STNorm	17.31	28.78	25.07	27.16	41.73	14.89	21.63	37.30	9.78	21.86	36.89	19.93
+RiskTraf	15.78↑	26.10↑	23.23↑	25.20↑	38.71↑	14.18↑	19.87↑	34.68↑	8.98↑	20.51↑	34.37↑	19.20↑
GTS	28.45	50.08	49.64	41.99	62.02	24.72	44.76	70.24	24.12	37.53	57.01	63.22
+RiskTraf	21.45↑	34.87↑	46.49↑	31.53↑	47.18↑	17.66↑	28.32↑	48.27↑	16.00↑	23.51↑	39.50↑	36.94↑
STWA	15.38	25.63	21.69	24.74	38.12	14.29	19.29	32.63	10.26	19.85	33.59	17.99
+RiskTraf	14.86↑	24.79↑	21.06↑	24.15↑	37.34↑	13.66↑	18.42↑	31.96↑	8.36↑	18.94↑	32.54↑	17.47↑
MegaCRN	15.07	25.14	20.33	24.50	37.70	13.16	18.97	33.36	7.97	18.67	32.06	15.24
+RiskTraf	14.89↑	24.82↑	20.82↓	23.98↑	37.14↑	12.99↑	18.12↑	32.44↑	7.61↑	18.22↑	31.45↑	15.43↓
HimNet	15.09	25.42	20.92	23.92	37.28	13.28	19.70	33.82	8.78	18.41	31.82	15.63
+RiskTraf	14.72↑	24.79↑	21.26↓	23.46↑	36.74↑	13.10↑	18.55↑	32.36↑	8.44↑	17.73↑	30.96↑	15.57↑
STDN	13.78	22.67	33.80	22.56	35.17	12.54	15.86	29.58	7.36	16.59	30.34	15.61
+RiskTraf	13.24↑	22.35↑	20.36↑	21.85↑	34.64↑	12.44↑	14.95↑	29.04↑	6.71↑	16.04↑	29.59↑	14.95↑

**Figure 4: Per-horizon MAE decomposition for GTS, GWNet, and STNorm on the four PEMS03-B subsets. For each horizon, the full bar denotes the baseline MAE, the solid segment denotes the MAE after adding RiskTraf, and the hatched cap denotes the absolute error reduced by RiskTraf.**

The reduction pattern is also horizon-dependent. On PEMS04-B, PEMS07-B, and PEMS08-B, the hatched caps often become more pronounced from short to medium and long horizons, although some subsets also show clear short-horizon corrections when the baseline is unstable. This pattern is consistent with the increasing uncertainty of long-range forecasting: near-term flow is constrained by recent observations, whereas later horizons require recognizing the underlying traffic regime from speed and occupancy. Therefore, the per-horizon decomposition supports that RiskTraf does not simply apply a uniform offset, but learns residual corrections that are more useful when regime-dependent uncertainty accumulates.

4.4 Comparison with Robust Forecasting Methods (RQ3)

To answer RQ3, we compare RiskTraf with representative robust traffic forecasting methods, including debiasing-based methods CauSTG [54] and ST-SSDL [18], and distribution-shift adaptation methods Dish-TS [16] and STEVE [25]. Since these methods are standalone forecasting frameworks rather than backbone models, Table 3 reports their all-horizon MAE/RMSE separately from the paired plug-in comparison in Table 2. We include STDN [6]+RiskTraf

Table 3: Comparison with robust traffic forecasting methods on PEMS3-B. Results are averaged over horizons 3, 6, and 12. The first four methods are standalone robustness methods, while the last column reports RiskTraf plugged into the strongest backbone from Table 2. Best results are in bold.

Dataset	Metric	CauSTG	ST-SSDL	Dish-TS	STEVE	RiskTraf (STDN)
PEMS03-B	MAE↓	32.82	14.94	14.51	16.95	13.24
	RMSE↓	53.95	24.91	24.29	27.24	22.35
PEMS04-B	MAE↓	53.43	23.55	23.03	24.86	21.85
	RMSE↓	75.93	37.06	35.90	36.10	34.64
PEMS07-B	MAE↓	62.79	17.32	17.72	20.62	14.95
	RMSE↓	86.79	31.52	33.37	31.91	29.04
PEMS08-B	MAE↓	54.41	17.08	18.47	18.77	16.04
	RMSE↓	82.58	30.71	32.25	29.90	29.59

as the representative RiskTraf configuration, as STDN is the strongest backbone in the main comparison.

RiskTraf obtains the best MAE and RMSE on all four PEMS3-B subsets. Compared with the strongest standalone competitor on each subset, it reduces MAE by 8.75%, 5.12%, 13.68%, and 6.09%, and reduces RMSE by 7.99%, 3.51%, 7.87%, and 1.04% on PEMS03-B, PEMS04-B, PEMS07-B, and PEMS08-B, respectively. The consistent improvements on both MAE and RMSE indicate that RiskTraf reduces not only average prediction bias but also larger forecasting deviations. Moreover, existing robust methods exhibit dataset-dependent rankings, suggesting that their robustness mechanisms may be sensitive to district-level traffic dynamics. In contrast, RiskTraf remains consistently strong across subsets while preserving the original backbone architecture. This supports the effectiveness of risk-aware residual correction for handling regime-dependent auxiliary correlations without replacing the forecasting model.

4.5 Component and Sensitivity Analysis (RQ4)

Auxiliary-variable ablation. Table 4 ablates the auxiliary measurements used for traffic-risk environment construction while keeping the backbone, residual head, and training protocol unchanged. The *w/o speed* variant constructs environments using occupancy only, the *w/o speed&occ* variant removes both speed and occupancy from environment partitioning, and the full variant uses both measurements.

The results show that occupancy alone already provides useful traffic-state information: *w/o speed* consistently outperforms *w/o speed&occ* on most backbone–dataset pairs. This is expected because occupancy directly reflects detector occupation intensity and is closely related to congestion states. However, the full *speed+occ* variant achieves the best overall performance across all backbone blocks, showing that speed provides complementary information about vehicle movement dynamics. The improvement is particularly clear for DCRNN, while stronger backbones such as GWNet and STWA obtain smaller but still consistent gains. These results confirm that RiskTraf benefits from constructing traffic-risk environments with both auxiliary measurements.

Sensitivity to λ and K . Figure 5 studies the sensitivity of RiskTraf on PEMS03-B by varying the risk penalty weight λ and the number of environments K , while keeping β and η fixed. The weight

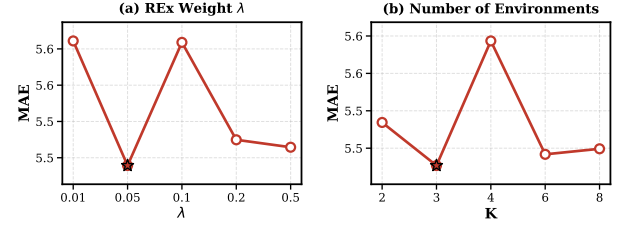


Figure 5: Sensitivity analysis of RiskTraf on PEMS03-B with respect to the risk penalty weight λ and the number of risk environments K . Stars mark the lowest MAE in each sweep.

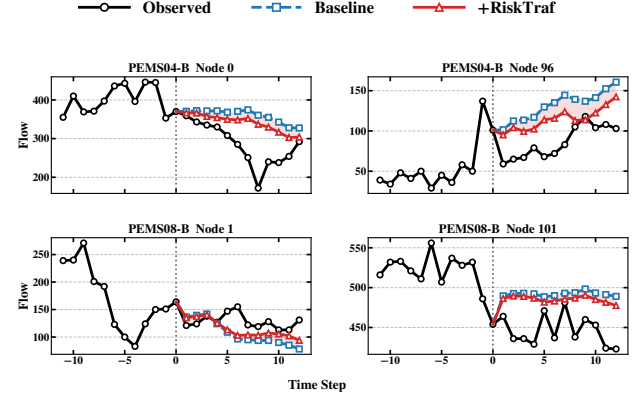


Figure 6: Targeted STAEformer forecasting cases on PEMS04-B and PEMS08-B. The black curve denotes observed flow, the red dashed curve denotes the baseline prediction, and the blue curve denotes the prediction after adding RiskTraf. The vertical dashed line marks the prediction start, and the shaded region highlights the residual correction.

λ controls the strength of the REx, pairwise, and extrapolation regularizers. A small λ weakens cross-environment constraints, whereas an overly large λ may sacrifice pointwise accuracy for environment consistency. The best performance is achieved around $\lambda = 0.05$.

The environment number K controls the granularity of traffic-risk partitioning. The non-monotonic curve shows that environment construction is not simply improved by using more partitions. Although $K = 3$ obtains the lowest MAE in this PEMS03-B sweep, we use $K = 4$ in the main experiments as a fixed quartile split for consistent evaluation across all dataset–backbone pairs. This avoids tuning K separately for each setting.

4.6 Efficiency Study (RQ5)

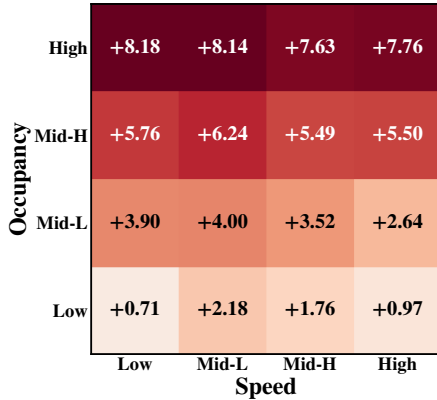
Table 5 evaluates RiskTraf from a plug-in efficiency perspective. Since RiskTraf is not designed as a standalone forecasting architecture, the relevant comparison is between each backbone and the same backbone equipped with RiskTraf. Across GTS [38], GWNet [46], and STNorm [13] on PEMS03-B, RiskTraf introduces only about 0.004M additional trainable parameters and marginal GFLOP increases. The inference latency and inference memory remain in the same practical range, indicating that the residual head adds limited deployment overhead.

Table 4: Ablation study on auxiliary measurements used for traffic-risk environment construction. Results are averaged over horizons 3, 6, and 12; MAPE is reported in percentage form. The best result within each backbone–dataset block is in bold.

Backbone	Variant	PEMS03-B			PEMS04-B			PEMS07-B			PEMS08-B		
		MAE↓	RMSE↓	MAPE(%)↓	MAE↓	RMSE↓	MAPE(%)↓	MAE↓	RMSE↓	MAPE(%)↓	MAE↓	RMSE↓	MAPE(%)↓
DCRNN	w/o speed	35.70	53.36	149.01	43.50	62.29	34.10	45.85	70.52	33.63	36.86	54.11	56.48
	w/o speed&occ	39.45	59.49	154.12	49.01	71.24	34.47	61.43	83.21	66.09	46.54	68.41	61.30
	Full (speed+occ)	25.72	39.23	140.30	30.93	45.61	19.03	25.59	41.89	13.37	25.80	39.82	29.89
GWNNet	w/o speed	19.50	32.85	24.90	31.30	47.75	17.11	23.85	40.21	10.78	23.39	39.07	19.58
	w/o speed&occ	19.52	32.88	25.16	31.80	48.46	17.14	24.48	41.09	10.96	23.71	39.60	19.63
	Full (speed+occ)	17.43	29.07	23.92	27.75	42.20	15.35	21.69	36.84	9.91	21.22	35.56	18.22
STWA	w/o speed	15.24	25.49	21.53	24.56	37.92	13.69	18.95	32.44	8.67	19.58	33.27	18.27
	w/o speed&occ	15.28	25.56	21.68	24.68	38.07	13.75	19.11	32.58	8.68	19.73	33.52	18.14
	Full (speed+occ)	14.86	24.79	21.06	24.15	37.34	13.66	18.42	31.96	8.36	18.94	32.54	17.47

Table 5: Plug-in efficiency comparison on PEMS03-B with batch size 32. Each pair compares a backbone with the same backbone equipped with RiskTraf. Training time denotes the observed wall-clock time of the recorded run, where RiskTraf trains only the residual head with the backbone frozen.

Backbone	Variant	Params (M)	GFLOPs/Batch	Infer ms/Sample	Infer Mem (MB)	Train Time (s)	Train Mem (MB)	Test MAE
GTS	Base	8.307	21.229	0.941	231.7	18.08	962.2	28.45
	+RiskTraf	8.312	21.262	1.036	239.9	6.55	274.2	21.45
GWNNet	Base	0.031	11.484	0.162	193.9	4.95	330.7	19.65
	+RiskTraf	0.035	11.518	0.159	193.9	1.72	196.8	17.43
STNorm	Base	0.042	1.685	0.069	143.1	5.27	287.8	17.31
	+RiskTraf	0.046	1.719	0.078	143.2	2.42	145.9	15.78

**Figure 7: RiskTraf correction over STAEformer on PEMS03-B. Samples are grouped by speed and occupancy bins; each cell reports $\text{mean}(\hat{Y}_{\text{RiskTraf}} - \hat{Y}_{\text{Base}})$, where larger positive values indicate stronger upward residual corrections.**

The small overhead is accompanied by clear accuracy gains. Test MAE decreases from 28.45 to 21.45 for GTS, from 19.65 to 17.43 for GWNNet, and from 17.31 to 15.78 for STNorm. These improvements show that the auxiliary risk environments and residual correction provide a favorable accuracy–efficiency trade-off. Training wall-clock time is reported as an auxiliary measurement, since it is affected by Stage-II early stopping and by the fact that only the residual head is optimized while the backbone is frozen.

4.7 Case Study (RQ5)

Trajectory-level correction. Figure 6 visualizes representative STAEformer forecasting cases on PEMS04-B and PEMS08-B. The baseline sometimes over-reacts or under-reacts after the prediction start, leading to trajectories that deviate from the observed flow. RiskTraf adjusts the baseline prediction toward the observed trajectory in these cases, especially around turning points or short-term trend changes after the prediction boundary. This shows that the residual head does not simply apply a constant shift, but can produce horizon-wise corrections that reshape the forecast trajectory.

State-dependent correction. Figure 7 further groups samples by speed and occupancy states and reports the average correction $\text{mean}(\hat{Y}_{\text{RiskTraf}} - \hat{Y}_{\text{Base}})$. The correction is positive in all bins, indicating that RiskTraf generally raises the STAEformer baseline in the selected PEMS03-B setting. More importantly, the correction magnitude is strongly associated with occupancy: low-occupancy states receive relatively small corrections, whereas high-occupancy states receive much larger corrections. In contrast, the variation along the speed axis is less monotonic. This pattern suggests that occupancy provides a stronger congestion-state signal for the residual head, while speed helps refine the correction within each occupancy regime.

Risk-environment structure. Figure 8 visualizes the risk-environment partition on PEMS03-B using t-SNE. The three environments form a coarse low-to-high risk organization with visible separation, while still retaining overlap in transition regions. The separated regions suggest that the speed–occupancy risk score captures distinct traffic regimes, whereas the overlapping regions correspond to samples whose raw traffic patterns are less clearly separable. This structure

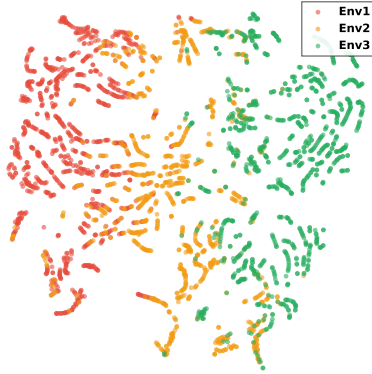


Figure 8: t-SNE visualization of PEMS03-B samples colored by RiskTraf risk environments. The three colors denote ordered environments constructed from historical speed and occupancy, showing separated regimes with overlapping transition regions.

is consistent with the motivation of REx-based residual learning: RiskTraf does not require perfectly separated environments, but encourages the residual head to remain stable across related risk regimes while still adapting to high-risk states.

Overall, these case studies provide qualitative evidence for the mechanism behind RiskTraf. The plug-in changes individual trajectories, applies larger corrections under high-occupancy traffic states, and relies on risk environments that reflect meaningful structure in the raw traffic data.

5 Related Work

5.1 Traffic Flow Prediction

Early deep approaches relied on recurrent architectures such as LSTM, which are limited in capturing complex spatial dependencies of road networks; later work turned to graph-based designs: DCRNN [31] combines diffusion convolutions with sequence-to-sequence learning, and STGCN [48] adopts a fully convolutional graph architecture, both relying on predefined adjacency matrices from road topology. Graph WaveNet [46] instead learns adaptive graphs for latent spatial dependencies, and MegaCRN [26] adds memory and meta-learning for heterogeneous patterns. STAEformer [32] showed that Transformers with spatio-temporal adaptive embeddings perform strongly, STDN [6] and HimNet [14] explored seasonality and heterogeneity, and ST-SSDL [18] aligns inputs with historical means to reduce stochastic deviations. As architectural gains become marginal [40], some works turn to richer sources such as event information (EastNet [44]) or global-scale multimodal data (Terra [9]), which require additional collection and alignment. In contrast, speed and occupancy are sensor-native measurements collected with flow, yet no existing benchmark systematically studies these raw auxiliary measurements for robust flow prediction; PEMS3-V fills this gap.

5.2 Invariant Learning and Causal Inference

Invariant learning assumes that stable mechanisms remain predictive across environments while spurious correlations vary. IRM [2]

seeks representations that admit a shared optimal classifier across environments, IRM Games [1] recasts this as Nash equilibrium finding, REx [28] minimizes the variance of risks across environments to improve extrapolation, and Group DRO [37] minimizes worst-group risk for subgroup robustness. Their effectiveness, however, hinges on how environments are defined: DomainBed [20] shows limited gains over empirical risk minimization when partitions do not align with the relevant spurious correlations. This is central in traffic prediction, where environments are not directly given. RiskTraf therefore builds ordered risk environments from historical speed and occupancy—physically interpretable indicators of traffic regimes—and applies REx-style regularization to a residual correction head.

5.3 Distribution Shift in Time Series

Non-stationarity between training and test periods is commonly handled by statistical normalization: RevIN [27] removes instance-level statistics before prediction and restores them afterward, DishTS [16] models shifts through learnable transformations, and Non-stationary Transformers [34] introduce de-stationary attention against over-stationarization, while the multi-order wavelet derivative transform [52] captures evolving non-stationary dynamics in the wavelet domain. These methods target shifts in the statistical or spectral properties of the series itself. The challenge studied here is different: the relationships among flow, speed, and occupancy themselves vary across traffic regimes, so treating speed and occupancy as ordinary covariates may introduce shortcut learning. RiskTraf instead uses them to define traffic-risk environments and learns a constrained residual correction that remains stable across these environments.

6 Conclusion and Future Works

In this paper, we study how to effectively use speed and occupancy for traffic flow prediction. While these sensor-native variables provide useful traffic-state information, direct three-variable training can exploit regime-dependent correlations and hurt generalization. To address this, we propose RiskTraf, a paired plug-in that freezes a trained three-variable backbone and learns a lightweight residual correction from historical speed and occupancy under a REx objective. By constructing traffic-risk environments, RiskTraf performs state-dependent flow correction on the frozen backbone. We also introduce PEMS3-V, a public multi-variate traffic benchmark with raw flow, speed, and occupancy measurements. Experiments show that RiskTraf consistently improves diverse spatio-temporal backbones and outperforms robust forecasting methods. Future work may explore batch-independent REx objectives and extend RiskTraf to other spatio-temporal prediction tasks with auxiliary measurements.

Acknowledgments

This work was supported in part by National Natural Science Foundations of China under Grant No. 62572416 and the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things under Grant No. 2023B1212010007.

GenAI Usage Disclosure

Generative AI tools were used only to assist with language polishing, wording refinement, and formatting during manuscript preparation. The research ideas, method design, experiments, analyses, figures, tables, and conclusions were produced and verified by the authors, who take full responsibility for the content of this paper.

References

- [1] Kartik Ahuja, Karthikeyan Shanmugam, Kush R. Varshney, and Amit Dhurandhar. 2020. Invariant risk minimization games. In *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*. JMLR.org.
- [2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization. arXiv:1907.02893 [stat.ML] <https://arxiv.org/abs/1907.02893>
- [3] Robert B Ash. 2012. *Information theory*. Courier Corporation.
- [4] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems* 33 (2020), 17804–17815.
- [5] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in Neural Information Processing Systems* 33 (2020), 17766–17778.
- [6] Lingxiao Cao, Bin Wang, Guiyuan Jiang, Yanwei Yu, and Junyu Dong. 2025. Spatiotemporal-aware Trend-Seasonality Decomposition Network for Traffic Flow Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 11 (Apr. 2025), 11463–11471. doi:10.1609/aaai.v39i11.33247
- [7] Chao Chen, Jaimyoung Kwon, John Rice, Alexander Skabardonis, and Pravin Varaiya. 2003. Detecting Errors and Imputing Missing Data for Single-Loop Surveillance Systems. *Transportation Research Record* 1855, 1 (2003), 160–167. doi:10.3141/1855-20
- [8] Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. 2001. Freeway Performance Measurement System: Mining Loop Detector Data. *Transportation Research Record* 1748, 1 (2001), 96–102. doi:10.3141/1748-12
- [9] Wei Chen, Xixuan Hao, Yuanhai Wu, and Yuxuan Liang. 2024. Terra: A Multi-modal Spatio-Temporal Dataset Spanning the Earth. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=I0zpivK0A0>
- [10] Wei Chen and Yuxuan Liang. 2025. Expand and Compress: Exploring Tuning Principles for Continual Spatio-Temporal Graph Forecasting. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=FRzCllkM71>
- [11] Xu Chen, Junshan Wang, and Kunqing Xie. 2021. TrafficStream: A Streaming Traffic Flow Forecasting Framework Based on Graph Neural Networks and Continual Learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. 3620–3626.
- [12] Arnaud de Myttenaere, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. 2016. Mean Absolute Percentage Error for regression models. *Neurocomputing* 192 (2016), 38–48. doi:10.1016/j.neucom.2015.12.114
- [13] Jinliang Deng, Xiusi Chen, Renhe Jiang, Xuan Song, and Ivor W Tsang. 2021. ST-Norm: Spatial and Temporal Normalization for Multi-variate Time Series Forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 269–278.
- [14] Zheng Dong, Renhe Jiang, Haotian Gao, Hangchen Liu, Jinliang Deng, Qingsong Wen, and Xuan Song. 2024. Heterogeneity-Informed Meta-Parameter Learning for Spatiotemporal Time Series Forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 631–641.
- [15] Stephen Dunne and Bidisha Ghosh. 2013. Weather adaptive traffic prediction using neurowavelet models. *IEEE Transactions on Intelligent Transportation Systems* 14, 1 (2013), 370–379.
- [16] Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. 2023. Dish-TS: a general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI'23)*. AAAI Press. doi:10.1609/aaai.v37i6.25914
- [17] Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. 2024. Efficient Large-Scale Traffic Forecasting with Transformers: A Spatial Data Management Perspective. *arXiv preprint arXiv:2412.09972* (2024).
- [18] Haotian Gao, Zheng Dong, Jiawei Yong, Shintaro Fukushima, Kenjiro Taura, and Renhe Jiang. 2025. How Different from the Past? Spatio-Temporal Time Series Forecasting with Self-Supervised Deviation Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=TgGH1bY6kl>
- [19] Xiaochuan Gou, Ziyue Li, Tian Lan, Junpeng Lin, Zhishuai Li, Bingyu Zhao, Chen Zhang, Di Wang, and Xiangliang Zhang. 2026. TrafficDent: A Dataset for Understanding the Interplay Between Traffic Dynamics and Incidents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://openreview.net/forum?id=sZvXXPqONQ>
- [20] Ishaan Gulrajani and David Lopez-Paz. 2021. In Search of Lost Domain Generalization. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=IQdXcDoWtI>
- [21] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 922–929.
- [22] Jingrui He, Wei Shen, Phani Divakaruni, Laura Wynter, and Rick Lawrence. 2013. Improving traffic prediction with tweet semantics. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13)*. 1387–1393.
- [23] Rob J. Hyndman and Anne B. Koehler. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting* 22, 4 (2006), 679–688. doi:10.1016/j.ijforecast.2006.03.001
- [24] Jiahao Ji, Jingyuan Wang, Zhe Jiang, Jiawei Jiang, and Hu Zhang. 2022. STDEN: Towards physics-guided neural networks for traffic flow prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 4048–4056.
- [25] Jiahao Ji, Wentao Zhang, Jingyuan Wang, and Chao Huang. 2025. Seeing the Unseen: Learning Basis Confounder Representations for Robust Traffic Prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25)*. Association for Computing Machinery, 577–588. doi:10.1145/3690624.3709201
- [26] Renhe Jiang, Zhaonan Wang, Jiawei Yong, Puneet Jeph, Quanjun Chen, Yasumasa Kobayashi, Xuan Song, Shintaro Fukushima, and Toyotaro Suzumura. 2023. Spatio-Temporal Meta-Graph Learning for Traffic Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (2023), 8078–8086. doi:10.1609/aaai.v37i7.25976
- [27] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2021. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=cGDakQo1C0p>
- [28] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghui Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REX). In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 5815–5826.
- [29] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, 105–114. doi:10.1145/3209978.3210009
- [30] Jiwan Lee, Bonghee Hong, Kyungmin Lee, and Yang-Ja Jang. 2015. A prediction model of traffic congestion using weather data. In *2015 IEEE International Conference on Data Science and Data Intensive Systems*. IEEE, 81–88.
- [31] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations (ICLR '18)*.
- [32] Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quan-jun Chen, and Xuan Song. 2023. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4125–4129.
- [33] Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhengguang Liu, Bryan Hooi, and Roger Zimmermann. 2023. LargeST: A Benchmark Dataset for Large-Scale Traffic Forecasting. In *Advances in Neural Information Processing Systems*.
- [34] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. 2022. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In *Advances in Neural Information Processing Systems*.
- [35] Zhidan Liu, Jiangzhou Li, and Kaishun Wu. 2022. Context-Aware Taxi Dispatching at City-Scale Using Deep Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems* 23, 3 (2022), 1996–2009. doi:10.1109/TITS.2020.3030252
- [36] Jiawei Lu, Chongnan Li, Xin Bruce Wu, and Xuesong Simon Zhou. 2023. Physics-informed neural networks for integrated traffic state and queue profile estimation: A differentiable programming approach on layered computational graphs. *Transportation Research Part C: Emerging Technologies* 153 (2023), 104224.
- [37] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=ryxGuJrFvS>
- [38] Chao Shang, Jie Chen, and Jinbo Bi. 2021. Discrete graph structure learning for forecasting multiple time series. *arXiv preprint arXiv:2101.06861* (2021).
- [39] Feng Shao, Hu Shao, Xin Wu, Qixiu Cheng, and William HK Lam. 2025. A physics-informed machine learning framework for speed-flow prediction: Integrating an S-shaped traffic stream model with deep learning models. *Transportation Research Part C: Emerging Technologies* 180 (2025), 105362.

- [40] Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, et al. 2024. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Transactions on Knowledge and Data Engineering* 37, 1 (2024), 291–305.
- [41] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. 2020. Spatial-Temporal Synchronous Graph Convolutional Networks: A New Framework for Spatial-Temporal Network Data Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (Apr. 2020), 914–921. doi:10.1609/aaai.v34i01.5438
- [42] Rod E. Turochy and Brian L. Smith. 2000. New Procedure for Detector Data Screening in Traffic Management Systems. *Transportation Research Record* 1727, 1 (2000), 127–131. doi:10.3141/1727-16
- [43] Guangyu Wang and Jiawei Tong. 2026. Frequency as identity: A Fourier hypernetwork for spatiotemporal forecasting. *Applied Soft Computing* 187 (2026), 114297. doi:10.1016/j.asoc.2025.114297
- [44] Zhaonan Wang, Renhe Jiang, Hao Xue, Flora D. Salim, Xuan Song, and Ryosuke Shibasaki. 2022. Event-Aware Multimodal Mobility Nowcasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 4 (2022), 4228–4236. doi:10.1609/aaai.v36i4.20342
- [45] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. 2020. Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- [46] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. AAAI Press, 1907–1913.
- [47] Du Yin, Hao Xue, Arian Prabowo, Shuang Ao, and Flora Salim. 2025. XXLTraffic: Expanding and Extremely Long Traffic Forecasting beyond Test Adaptation. arXiv:2406.12693 [cs] doi:10.48550/arXiv.2406.12693
- [48] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018)*. International Joint Conferences on Artificial Intelligence Organization, 3634–3640. <https://www.ijcai.org/proceedings/2018/0505.pdf>
- [49] Chengyang Zhang, Yong Zhang, Qitan Shao, Bo Li, Yisheng Lv, Xinglin Piao, and Baocai Yin. 2024. ChatTraffic: Text-to-Traffic Generation via Diffusion Model. *IEEE Transactions on Intelligent Transportation Systems* (2024), 1–13. doi:10.1109/TITS.2024.3510402
- [50] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* 31 (2017). doi:10.1609/aaai.v31i1.10735
- [51] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. 2020. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1234–1241.
- [52] Ziyu Zhou, Jiaxi Hu, Qingsong Wen, James T. Kwok, and Yuxuan Liang. 2026. Multi-Order Wavelet Derivative Transform for Deep Time Series Forecasting. arXiv:2505.11781
- [53] Zhengyang Zhou, Qihe Huang, Kuo Yang, Kun Wang, Xu Wang, Yudong Zhang, Yuxuan Liang, and Yang Wang. 2023. Maintaining the Status Quo: Capturing Invariant Relations for OOD Spatiotemporal Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. 3603–3614.
- [54] Zhengyang Zhou, Qihe Huang, Kuo Yang, Kun Wang, Xu Wang, Yudong Zhang, Yuxuan Liang, and Yang Wang. 2023. Maintaining the Status Quo: Capturing Invariant Relations for OOD Spatiotemporal Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) (KDD '23). Association for Computing Machinery, New York, NY, USA, 3603–3614. doi:10.1145/3580305.3599421