

Article

LSTM-TCN Forecasting for Short-Term Passenger Flow at Integrated Transport Hubs Using a Population-Mobility Proxy: Ablation Evidence from Shenzhen North Station

Xieming Zhang ^{1,*}, Zhiyun Zou ^{1,*}  and Kai Zhang ² ¹ School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China; xzm@sutpc.com² Shenzhen Urban Transport Planning Center Co., Ltd., Shenzhen 518000, China; zhangkai03@sutpc.com

* Correspondence: zyzou@hust.edu.cn

Abstract

Short-term passenger-flow forecasting at integrated transport hubs requires accurate forecasts with explicit target semantics. This study evaluates long short-term memory (LSTM) and temporal convolutional network (TCN) fusion models using two distinct 5 min targets at Shenzhen North Station: a 2023–2026 regional population-mobility crowd-density proxy and a 2025 automatic fare collection (AFC) exit-card throughput series. On the proxy task, chronological 70/15/15 splits, a 12-step purge and full-epoch neural training showed that the LSTM-only variant (M2) achieved the lowest neural mean absolute error (MAE) at 5 min, while the fused LSTM–TCN variant (M4) achieved the lowest neural MAE at 15–60 min. Ridge was the strongest conventional proxy reference at 5 min and Random Forest at 15–60 min. On AFC exits, Random Forest was strongest among the six conventional references at all four horizons, while M4 remained better than ridge at 5–60 min. The raw 5 min proxy–AFC Pearson correlation was 0.553; after weekday-by-clock de-seasonalization it was 0.505 (95% day-block confidence interval 0.457–0.538). These values support a shared-signal interpretation; calibration requires a separate mapping model. In the observed-arrival pressure test, recovery at 21:10 reduced cumulative queueing by 67.3%; delays of 10 and 15 min reduced the benefit to 63.4% and 61.7%. The workflow provides 5–60 min lead time for gate and queue preparation, staff rostering and feeder coordination.

Keywords: integrated transport hub; population-mobility proxy; long short-term memory–temporal convolutional network (LSTM–TCN); random forest benchmark; queue-timing sensitivity



Academic Editor: Andrea Prati

Received: 16 July 2026

Revised: 29 August 2026

Accepted: 4 September 2026

Published: 7 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Integrated transport hubs couple railway, urban rail, bus, taxi and pedestrian systems within a limited space. Their passenger flow evolves under the joint influence of urban activity rhythms, train timetables, holidays and facility conditions, and therefore shows periodicity, short-term inertia, discrete pulses and capacity constraints. During peak periods, concentrated train arrivals or local facility failures may shift the system from stable operation to sustained accumulation within only several observation intervals. For operations management, short-term forecasts combine point accuracy with actionable lead time for flow restriction, diversion, standby corridor activation and transport-capacity reinforcement. Forecasts can be converted into staffing and operations decisions: 5 min alerts support local gate and queue management, 15 min forecasts support staff repositioning and

temporary-channel preparation, and 30–60 min forecasts support passenger information, diversion and feeder–vehicle coordination.

Short-term passenger-flow forecasting methods can be broadly grouped into statistical models, recurrent or convolutional neural networks, attention models and spatio-temporal graph models. Statistical time-series models are compact and interpretable, and they provide useful lower-bound benchmarks before deployment. However, they are less able to represent nonlinear state transitions when train pulses, holidays and facility disturbances act together. LSTM mitigates gradient decay in long sequences through gated states [1], which is suitable for cross-peak or cross-day memory. TCN uses causal convolution, dilated convolution and residual connections for parallel local-dynamics encoding [2]. Attention mechanisms can dynamically allocate historical weights according to the forecast horizon [3]. Recent models such as Informer, TimesNet, PatchTST and STAEformer have further advanced long-sequence and traffic forecasting [4–7]. Their computational complexity, data requirements and cross-source input interpretation require scenario-specific evaluation in hub applications.

Existing rail-transit forecasting studies provide an important methodological basis, but model–data alignment remains unresolved. Recurrent networks and improved LSTM models have achieved useful performance in short-term metro passenger-flow forecasting [8–12]. Many studies rely on a single smart-card or station-flow sequence, which obscures differences among rail feeder demand, regional crowd aggregation and train-arrival pulses. Support vector regression (SVR)–LSTM and smart-card deep models for abnormal passenger flow demonstrate the value of operational data [13,14]; linking anomaly candidates to operational events still requires external logs. Sequence-to-sequence attention and multi-source fusion models improve robustness [15,16], while spatio-temporal graph convolution and graph multi-attention models strengthen network spatial dependence [17–21]. Temporal isolation, proxy-variable boundaries, prediction-interval reliability and post-capacity-exceedance consequences are rarely evaluated in one framework.

International case studies report different measurement, spatial and operating settings. London work models station interactions and special events [22] or train-level passenger load for rolling-stock and personnel planning [23]; Seoul studies preserve station heterogeneity through functional clustering [24]; Singapore work links smart-card origin–destination (OD) demand to fine-grained segment load [25]; and Toronto work explicitly compares conventional and deep models using Wi-Fi as a population proxy [26]. These cases demonstrate practical value and expose differences in transfer conditions.

This paper separates the forecasting targets, evaluation design and operational scenario. Population-mobility counts are modelled as a regional crowd-density stock proxy, whereas the 2025 AFC exit-card counts are evaluated independently as observed station throughput. Both tasks use chronological 70/15/15 splits with a 12-step purge to prevent adjacent windows from entering both training and test sets, and the final test period is reported once. Observed train-arrival pulses drive a deterministic capacity scenario in which recovery time is shifted to measure queue sensitivity. Table 1 summarizes the international comparison, and transferability is assessed through local recalibration requirements.

The study evaluates an LSTM–TCN fusion framework through a proxy ablation, an independent AFC passenger-flow validation and an observed-arrival queue-timing experiment. It examines two hypotheses and one research question:

H1 (structural-support hypothesis). *the proxy and AFC series contain strong short-term inertia and daily periodicity, so local and periodic baselines must be retained when evaluating nonlinear temporal models.*

H2 (horizon- and target-dependent model-selection hypothesis). *under strict temporal isolation, persistence, time-series ridge and same-slot-previous-day seasonal naive are competitive baselines; the relative value of LSTM-TCN depends on the target variable and forecast horizon.*

Table 1. Critical comparison of short-term passenger-flow forecasting evidence across transport systems.

Evidence/Context	Data and Approach	Practical Case	Strength	Boundary Relative to This Study
China: smart-card and station-flow studies [8–16]	LSTM, SVR, sequence-to-sequence attention and multi-source fusion	Station demand, abnormal-flow response and service planning	Rich station-level passenger measurements and nonlinear benchmarks	Usually one station/network measurement; regional crowd proxy and AFC semantics are not separated in one workflow
London Underground Central line [22]	Univariate state-space plus clustered dynamic-factor models with event effects	Real-time station-arrival prediction and special-event response	Explicit station interactions and event sensitivity	Station-arrival data and event logs differ from a hub-wide population-mobility proxy
London Victoria line [23]	Two-layer train-level passenger-flow model	Rolling-stock, personnel and facility-capacity planning	Connects forecasts to train capacity and passenger distribution	Train-load target is not a regional stock proxy or AFC exit throughput
Seoul Metro [24]	Functional clustering followed by time-series prediction using smart-card data	Service-interval adjustment, station planning and regional comparison	Preserves station heterogeneity and geographic context	Hourly station data and clustering design are not directly transferable to a single hub footprint
Singapore MRT [25]	Smart-card OD and travel-time distributions for fine-grained segment load	Timetable and segment-load management	Links OD demand to in-vehicle load	Requires network-wide OD and route-choice data unavailable in this project
Toronto subway [26]	Wi-Fi proxy compared with linear regression, exponential smoothing, ARIMA, Random Forest, N-BEATS and T-GCN	System-level demand monitoring and transfer-station calibration	Shows that conventional baselines can remain competitive and proxy quality matters	Wi-Fi penetration and station-level calibration differ from Tencent regional heat estimates
This study: Shenzhen North integrated hub	5 min population-mobility proxy + independent AFC exit cards; LSTM-TCN ablation and conventional baselines under temporal isolation	Staff rostering, queue/gate preparation, passenger information and feeder coordination	Keeps target meanings separate, adds spatial-coverage evidence, purged splits and observed-arrival queue timing	Single hub; local recalibration and multi-system validation remain necessary

International evidence shows that model value depends on the measurement target, spatial scale, event availability and operational decision. This study contributes an auditable comparison and transfer protocol for identifying when LSTM-TCN adds predictive value.

Research question (queue-timing sensitivity): under the deterministic observed-arrival scenario, how do recovery timing, peak queue, clearance time and cumulative queue person-minutes change when capacity recovery is shifted around 21:10?

The contributions of this paper are fourfold. C1 (target semantics and external AFC evidence): 105,120 timestamp-aligned AFC exit-card observations provide an external passenger-flow target and a shared-signal check against the regional crowd-density proxy, while the two measurement meanings remain separate. C2 (baseline-complete temporal-

isolation evaluation): the proxy M0–M7 ablation is expanded with seasonal naive and five conventional regressors; the AFC task includes the same conventional benchmark alongside persistence, ridge, seasonal naive and M4. All tasks use chronological splits, and neural comparisons use seed-specific loss series, Bartlett heteroskedasticity-and-autocorrelation-consistent (HAC) tests and Holm correction. C3 (target- and horizon-dependent model selection): M2 is strongest within the neural ablation at 5 min and M4 at 15–60 min for the proxy; ridge leads the proxy conventional benchmark at 5 min, Random Forest leads it at 15–60 min, and Random Forest leads the AFC conventional benchmark at all horizons. C4 (observed-arrival timing sensitivity): the queue experiment recomputes cumulative queue person-minutes from 481 observed train arrivals and reports peak time and queue-clearance time for shifted recovery clocks.

2. Materials and Methods

2.1. Study Design and Forecasting Task

Given the previous L time steps, the model predicts the population-mobility proxy count h steps ahead.

The main notation is as follows. Subscripts t or τ denote 5 min discrete times, superscript (l) denotes the network-layer index, r denotes the forecast horizon, d denotes the feature dimension, g denotes the gating vector, α denotes the temporal attention weight for horizon r , and β denotes the attention weight of feature group k . Unless otherwise stated, all standardization parameters, conditional quantiles and anomaly thresholds are estimated from the training set only. Figure 1 summarizes the implemented forecasting pipeline and its separate operational interfaces.

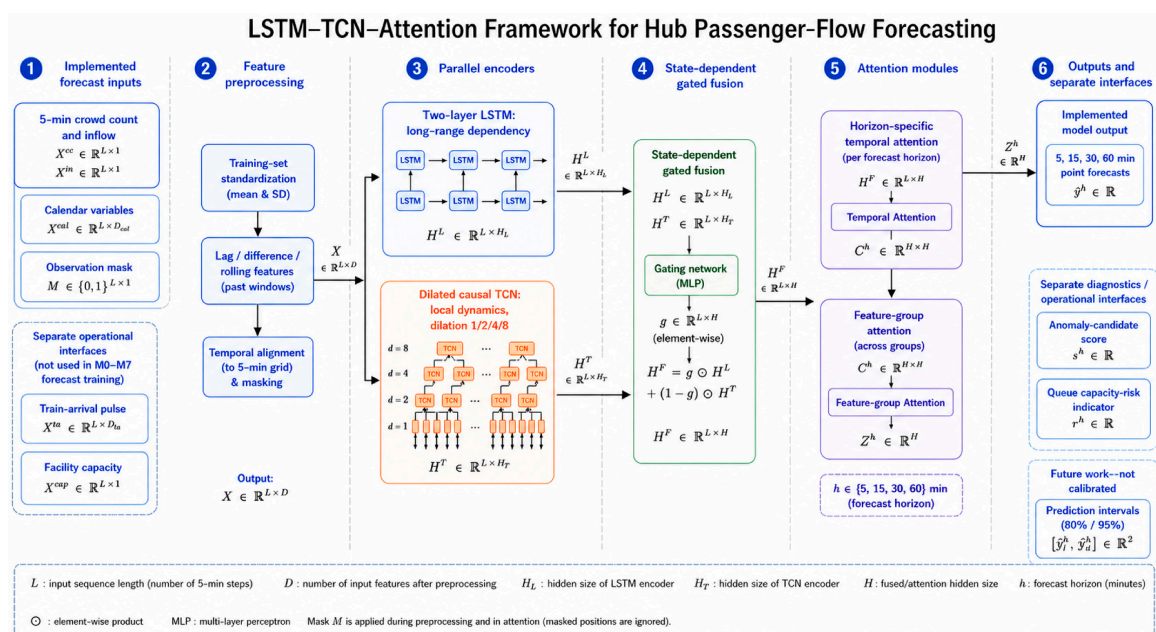


Figure 1. LSTM-TCN passenger-flow forecasting framework. Ellipses in the TCN branch indicate omitted intermediate time positions within each dilation level. Solid boxes and paths show implemented forecasting components; dashed boxes show separate diagnostic, operational, or future-work interfaces.

The solid path shows the implemented M0–M7 forecast: aligned proxy/inflow, calendar variables and the observation mask are standardized with training-set means and standard deviations before parallel LSTM and TCN encoding, learned fusion and horizon-specific prediction. The dashed boxes identify separate interfaces: train-arrival pulses and facility capacity enter the queue scenario rather than M0–M7 forecast training; anomaly

candidates are computed outside the point-forecast network; and prediction intervals remain future calibration work.

2.2. Data Sources, Study Object and Preprocessing

The data sources are organized by measurement role. Daily rail data estimate long-term magnitude, weekly structure and daily baselines. Five-minute population-mobility data—aggregated Internet-based crowd heatmap estimates—characterize short-term regional aggregation, forecasting baselines and anomaly candidates. They are treated as crowd-density proxies around the hub rather than direct gate counts. Train-service data provide railway-arrival pulses, while planning and event materials define capacity and disturbance boundaries. All analyses use cleaned and checked aggregate data; sources from different years or with different measurement definitions remain separate targets. Public information provides the urban rail operating background. Station-level rail entry/exit data, population-mobility data and train alighting flow come from project data channels, with definitions checked against public administrative sources and project documentation [27]. The population-mobility records are aggregated, desensitized regional counts with no individual identifiers or trajectories and are processed under the data-use agreement and standard privacy requirements. Table 2 summarizes the data sources, temporal resolutions, valid scales, study uses, and measurement boundaries.

Table 2. Data sources, granularity, study use and measurement boundaries.

Data Source	Time Range	Granularity	Valid Scale	Study Use
Shenzhen North Station rail entry/exit	January 2023–June 2025	Daily	894 valid days	Long-term pattern, weekly periodicity and daily baseline
Population-mobility data	January 2023–May 2026	5 min	352,224 original records; 352,167 positive observed records; 57 zero records treated as unavailable	Intra-day pattern, short-term baseline and anomaly candidate screening
Shenzhen North Station AFC exit-card counts	January–December 2025	5 min	105,120 complete records	External passenger-flow association and forecasting validation
Train-service alighting flow	May 2024	Service/5 min	481 valid services	Railway-arrival pulses and capacity scenarios
Planning and event materials	Multi-year	Event/static	Facility and event parameters	Capacity boundaries and exogenous disturbance settings

2.2.1. Population-Mobility Proxy Acquisition and Spatial Coverage

The Tencent heat-data export provides 5 min aggregated crowd-density estimates for the blue polygon shown in Figure 2. The polygon covers the Shenzhen North integrated transport hub footprint, including the station core and adjacent east/west and coach-station plazas. The provider aggregates anonymized device/network signals, while the sampling frame, penetration rate and vendor-side smoothing are proprietary.

The blue boundary defines the source-data aggregation extent and documents its spatial coverage. Reliability is evaluated as a shared-signal measure against timestamp-aligned 2025 AFC exit counts.

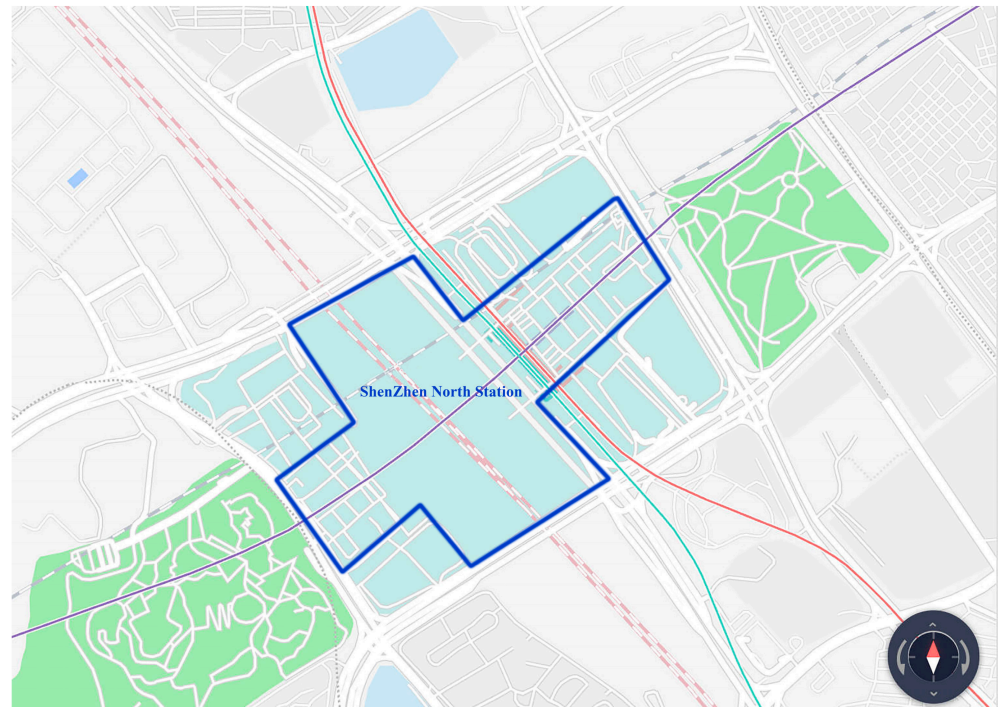


Figure 2. Tencent heat-data aggregation boundary used for the Shenzhen North integrated-hub proxy. The blue polygon denotes the aggregation extent covering the station core and adjacent plazas.

2.2.2. Software Environment and Reproducibility

The calculations were executed with Python 3.12.10, NumPy 2.4.6, pandas 3.0.3, Matplotlib 3.10.9, and PyTorch 2.13.0+cu130. Neural training used an NVIDIA RTX PRO 6000 graphics processing unit (NVIDIA Corporation, Santa Clara, CA, USA) when CUDA was available. Reproducibility manifests record source SHA-256 hashes, split counts, purge steps, seeds, model settings, GPU name and prediction outputs.

2.3. LSTM-TCN Forecasting Model with Attention Extensions

The model uses aligned proxy inputs, parallel LSTM and TCN encoders, learned fusion, horizon-specific temporal attention and state-dependent feature-group attention. M2–M7 add these modules sequentially so that long-range memory, local dynamics and reweighting can be evaluated under the same temporal split.

2.3.1. Input Variables and Tensor Representation

With a 5 min basic interval, the model receives recent crowd count, proxy inflow, lags, differences, rolling statistics, calendar encodings and a zero-record mask. Train arrivals, delays, facility states and response logs remain operational interfaces because they are not aligned with the full proxy series. The neural input uses the most recent 96 steps; the 288-step construction window is used for labels and rolling features.

$$\Delta t = 5 \text{ min}, X_t \in \mathbb{R}^{(L \times D)}$$

$$X_t = \left[X_t^{\text{flow}}, X_t^{\text{calendar}}, X_t^{\text{event}}, X_t^{\text{facility}} \right]$$

Continuous variables are standardized with training-set means and standard deviations and clipped to $[-8, 8]$. The 57 zero-valued proxy records are forward-filled, with a

leading-boundary backfill, and a zero-record mask is retained. Categorical variables use embeddings. Future information unavailable at prediction time is excluded.

$$x'_{t,d} = \frac{x_{t,d} \text{ median}_d}{1.4826 \cdot \text{MAD}_d + \varepsilon}$$

2.3.2. LSTM Long-Sequence Encoding Module

The LSTM branch receives the full historical window and controls information retention through input, forget and output gates. For layer l and time τ , the state update is defined as follows:

$$\begin{aligned} f_\tau &= \sigma(W_f x_\tau + U_f h_{\tau-1} + b_f), & i_\tau &= \sigma(W_i x_\tau + U_i h_{\tau-1} + b_i) \\ \tilde{c}_\tau &= \tanh(W_c x_\tau + U_c h_{\tau-1} + b_c), & c_\tau &= f_\tau \odot c_{\tau-1} + i_\tau \odot \tilde{c}_\tau \\ o_\tau &= \sigma(W_o x_\tau + U_o h_{\tau-1} + b_o), & h_\tau &= o_\tau \odot \tanh(c_\tau) \end{aligned}$$

The two-layer unidirectional LSTM receives the full historical window and captures cross-hour dependence using current and past observations only. Its hidden dimension is 32 and inter-layer dropout is 0.2.

2.3.3. TCN Local-Dynamics Encoding Module

The TCN runs in parallel with the LSTM and captures local slopes and peaks through four residual blocks with kernel size 3 and dilation factors 1, 2, 4 and 8. Its receptive field is 61 five-minute steps, approximately 5 h. The dilated-convolution and residual-block choices follow established sequence-modelling patterns [28–30]; weight normalization is not activated in the reported configuration [29].

$$\begin{aligned} H_{l(\tau)} &= \varphi\left(\sum_{k=0}^{K-1} W_{(l,k)} H_{(l1)(\tau-d_l k)} + b_l\right) \\ Z_l &= \text{ReLU}\left(F_{l(H_{l-1})} + P_{l(H_{l1})}\right) \end{aligned}$$

2.3.4. Dual-Branch Gated Fusion

After linear projection, the LSTM and TCN outputs have the same dimension. In M4 and M5, their sequence representations are concatenated and passed through a learned linear fusion layer. Feature-dependent gating is introduced only in M6 and M7; M4 tests dual-branch complementarity.

$$\begin{aligned} g_\tau &= \sigma\left(W_{g[H_\tau^L; H_\tau^T; E_\tau]} + b_g\right) \\ H_\tau^F &= g_\tau \odot H_\tau^L + (1 - g_\tau) \odot H_\tau^T \end{aligned}$$

For M6 and M7, the gating vector is computed along feature dimensions and performs state-dependent fusion between LSTM cross-hour memory and TCN local dynamics. In the current empirical setting, this gate and feature-group attention use only aligned proxy counts, inflow, calendar variables and zero-record mask. Missing or unaligned event and facility variables are not empirical inputs.

2.3.5. Temporal Attention and Feature Attention

The fused sequence first enters the temporal attention module. For each forecast horizon r , the model uses additive attention [31] to compute the importance of historical time step τ independently, avoiding one shared weight set across all forecast horizons.

Five-minute prediction usually focuses on the most recent steps, whereas 30–60 min prediction may assign higher weight to the same time on the previous day or to the beginning of a peak period.

$$e_{r,\tau} = v_r^T \tanh(W_r H_\tau^F + U_r q_r + b_r)$$

$$\alpha_{r,\tau} = \frac{\exp(e_{r,\tau})}{\sum_j \exp(e_{r,j})}, \quad z_r^{time} = \sum_\tau \alpha_{(r,\tau)} H_\tau^F$$

The full framework reserves a group-level attention interface for four feature groups: passenger flow, periodicity, events and facilities. Encoded vectors are scored and normalized by Softmax to form group weights and exogenous context. The reported M6 implementation evaluates aligned crowd-density proxy/inflow variables, calendar-periodic features and zero-value indicators. Event and facility groups remain empty extension interfaces in the current training.

$$\beta_r^m = \text{softmax}_{m(a^T \tanh(W_m u_r^m + b_m))}, \quad z_r^{ext} = \sum_m \beta_r^{(m)} u_r^m$$

Temporal and feature-group attention allocate weights among available historical times and aligned-feature groups. They are treated as predictive reweighting mechanisms rather than causal explanations.

2.3.6. Decoding and Loss Function

To avoid error accumulation caused by recursive prediction, the model uses direct multi-step decoding. Each forecast horizon r has an independent output head while sharing the encoder and attention context. The output head consists of two fully connected layers and uses Softplus to ensure non-negative predictions:

$$\hat{y}_{t+r} = \text{Softplus}\left(W_{2,r} \cdot \text{ReLU}\left(W_{1,r}[z_r^{time}; z_r^{ext}; h_t] + b_{1,r}\right) + b_{2,r}\right)$$

The M0–M7 objective is the uniform mean of PyTorch SmoothL1 (Huber) losses [32] over four standardized residual horizons. The final β is 0.5, selected from 0.5, 1.0 and 2.0 using validation data. M7 additionally applies the peak weight $1 + 2I(y \geq P90)$, where P90 is the 90th percentile of the training targets for each horizon.

$$L = \sum_r \omega_r [\sum_t w_t \cdot \text{Huber}_\delta(y_{t+r} - \hat{y}_{t+r})] + \lambda_c L_{cons} + \lambda_2 \|\theta\|_2^2$$

$$w_t = 1 + \lambda_p \cdot I(y_t > Q_{0.90}) + \lambda_a \cdot I\left(\left|z_t^h\right| > 3\right)$$

2.3.7. Training, Complexity and Interpretability

Model training follows a strict temporal-causal process. Data are divided chronologically into 246,335 training, 52,776 validation and 52,790 test windows, with a 12-step purge. Each neural epoch shuffles and visits all training windows once (121 batches of size 2048); maximum training length is 30 epochs and early stopping patience is 5. Validation uses all 52,776 windows. The locked run used PyTorch AdamW and cosine annealing [33,34], $\beta = 0.5$, five seeds and an NVIDIA RTX PRO 6000 when CUDA was available.

Current M0–M7 training uses only aligned crowd-density proxy, inflow, calendar, and zero-value features. Train plans and facility states enter only the separate queue scenario. Table 3 summarizes the training procedure, Table 4 lists the main hyperparameters and their rationale, and Table 5 presents the prespecified ablation scheme.

Table 3. LSTM-TCN training and rolling-prediction procedure with attention extensions.

Step	Processing Content
1 Data preparation	Clean timestamps, construct observation masks, and split training, validation and test sets chronologically
2 Feature construction	Generate lags, differences, rolling statistics, periodic encodings and zero-value indicators; describe train pulses and facility states as scenario interfaces and extensible exogenous variables
3 Window sampling	Construct samples from historical windows of length L and directly generate labels for H forecast horizons
4 Forward propagation	Compute LSTM long-sequence encoding and TCN local encoding, then fuse them through gating and dual attention
5 Loss calculation	Compute a uniform multi-horizon SmoothL1 loss; apply P90 peak weighting only in M7. No consistency regularization is active in the reported experiments.
6 Parameter update	Apply backpropagation, gradient clipping, AdamW updates and learning-rate scheduling
7 Model selection	Use validation-set composite loss and peak recall to trigger early stopping and restore the best weights
8 Rolling prediction	Update the window whenever a new 5 min observation is obtained, and output 5–60 min forecasts and risk indicators

Table 4. Main model hyperparameters and rationale.

Parameter	Value	Rationale
Input window length L	288 steps for sample construction; 96 steps for neural input	The full 5 min series constructs labels, and neural models use the most recent 96 steps as effective input
Forecast horizon H	1, 3, 6, 12 steps	Corresponding to 5, 15, 30 and 60 min.
LSTM hidden dimension	32	Two-layer unidirectional LSTM with inter-layer dropout = 0.2.
TCN channels	32	Four residual blocks with convolution kernel size 3.
Dilation factors	1, 2, 4, 8	Receptive field covers 61 time steps, about 5 h.
Attention dimension	32	Temporal attention and feature-group attention are projected to the same hidden dimension.
Batch size	2048	Each full epoch shuffles all 246,335 training windows; the final batch may be smaller.
Optimizer	AdamW (learning rate = 1×10^{-3} ; weight decay = 1×10^{-4})	AdamW (learning rate = 1×10^{-3} ; weight decay = 1×10^{-4}). Optimizer settings fixed before the final test evaluation.
Learning-rate schedule	Cosine annealing (minimum learning rate = 1×10^{-5})	Cosine annealing (minimum learning rate = 1×10^{-5}) Applied after each epoch
Early-stopping patience	5 epochs	Monitor MAE on all 52,776 validation windows.
Gradient clipping	$L_2norm \leq 5$	Suppress gradient explosion caused by abnormal peaks.
Multi-horizon weights	Uniform mean over four horizons (0.25 each)	Task-specific weights were not activated in the reported M0–M7 run.
Peak/anomaly weights	M7 only: $1 + 2I(y \geq trainingP90)$	P90 is computed separately for each horizon from the training targets; M0–M6 use unit weights.
Consistency/L2 regularization	No explicit consistency penalty; AdamW weight decay = 1×10^{-3}	Adjacent-horizon regularization was not activated in the reported run.

Table 4. *Cont.*

Parameter	Value	Rationale
Huber threshold (β)	PyTorch SmoothL1 $\beta = 0.5$ in standardized residual space	Selected from $\beta = 0.5, 1.0$ and 2.0 using validation data; the three-value M4 sensitivity check preserved the longer-horizon conclusion.
Maximum training epochs	30	Full-epoch training with early stopping; five seeds for the locked configuration.
Training batches per epoch	121 full-pass batches	246,335 training windows are visited once per epoch (batch size 2048).
Validation early-stopping cap	None; all 52,776 windows	A separate full-epoch run with a 10,000-window cap was used as a sensitivity check; the final run uses all validation windows.
GPU hardware	NVIDIA RTX PRO 6000	CUDA execution; the run manifest records the environment and source hashes.

Table 5. Prespecified model-ablation scheme.

ID	Model Configuration	Purpose
M0	Persistence model	Test the most basic short-term inertia
M1	Time-series ridge regression	Test the explanatory ability of linear lags and periodic terms
M2	LSTM residual predictor	Test long-sequence memory correction relative to the persistence baseline
M3	TCN residual predictor	Test local convolutional encoding gains
M4	LSTM + TCN with learned linear fusion	Test dual-branch complementarity
M5	M4 + horizon-specific temporal attention	Test horizon-dependent historical selection
M6	M5 + state-dependent fusion gate + feature-group attention	Test gains from proxy-count, inflow, calendar and missingness feature groups
M7	M6 + peak-weighted SmoothL1 loss	Test error changes after weighting samples above P90

The model exposes temporal-attention, feature-group and gating coefficients for diagnostic inspection. Their interpretation is restricted to predictive allocation because the current dataset lacks aligned event and facility labels; Shapley additive explanations (SHAP) are not used [35].

2.4. Anomaly-Candidate and Capacity Stress-Test Interface

Anomaly candidates are identified using local conditional thresholds estimated from the training set. The population-mobility series uses weekday-hour-minute conditional quantiles and the interquartile range (IQR) [36]. Continuous exceedance of the local upper bound defines an anomaly candidate. The yellow, orange and red labels form a proposed rule-based interface for screening candidates and linking them to capacity stress tests. If regional data are available, density-based spatial clustering of applications with noise (DBSCAN) can identify spatio-temporally continuous density clusters [37]. Complete early-warning evaluation requires verified operational event labels; the current analysis uses forecast quantiles, growth slopes and capacity load ratios for candidate screening and capacity-stress linkage.

- Yellow risk level: the forecast exceeds the conditional P90, or the 15 min growth rate is substantially higher than its historical distribution.

- Orange risk level: the forecast exceeds the conditional 95th percentile (P95) and is expected to persist for more than 10 min, or demand approaches effective service capacity.
- Red risk level: forecast demand continues to exceed effective capacity, and queue recursion indicates that accumulation cannot dissipate within the target time.

2.5. Experimental Design and Statistical Analysis

Model experiments use 5 min as the basic step. The 288-step construction window yields 351,925 supervised samples, divided into 246,335 training, 52,776 validation and 52,790 test samples after the 12-step purge. Standardization, conditional quantiles and anomaly thresholds use training data only. Hyperparameters are selected from validation data, and the held-out test period is evaluated after configuration lock.

The ablation uses the same purged split, input window, horizons and metrics for M2–M7. To test the effect of model-specific settings, M4–M7 were each rerun at one lower-rate alternative in addition to the shared base setting (learning rate 0.001, dropout 0.2): (0.0003, 0.1) for M4 and (0.0003, 0.2) for M5–M7. These bounded sensitivity runs did not overturn the M4-versus-M5–M7 ordering. The final five-seed table uses the shared base setting; this two-setting check is not an exhaustive hyperparameter search. M2 and M3 are retained as neural component baselines at the locked base setting.

M0 and M1 are deterministic baselines. Neural results are means and standard deviations over five seeds; all inferential comparisons use seed-specific loss series rather than a prediction averaged across seeds.

The primary run visits all 246,335 training windows once per epoch. In a post hoc matched M4 sensitivity control with the same 10,000-window validation cap, 12 randomly sampled batches per epoch changed test MAE by -1.3% , $+9.4\%$, $+17.2\%$ and $+32.3\%$ at 5, 15, 30 and 60 min relative to full-epoch training. The partial-epoch budget under-trained the longer-horizon forecasts.

The main metrics are mean absolute error (MAE), root mean square error (RMSE), weighted absolute percentage error (WAPE) and symmetric mean absolute percentage error (sMAPE). The classical Diebold–Mariano framework is a reference point for forecast comparison [38], but the present analysis uses paired loss differences with a Bartlett heteroskedasticity-and-autocorrelation-consistent (HAC) estimator and a 288-step bandwidth. The five seed-specific p values are summarized with a Cauchy combination [39]; Holm correction is applied separately to the M4-versus-M1 and M4-versus-M5–M7 comparison families. This procedure retains seed variability while accounting for serial correlation in rolling windows.

$$Recall_{peak} = \frac{TP_{peak}}{TP_{peak} + FN_{peak}}, \quad \Delta t_{peak} = |\operatorname{argmax}(y) - \operatorname{argmax}(\hat{y})|$$

For reproducibility, this study retains scripts for data cleaning, feature construction, proxy M0–M7 training, the seasonal-naïve baseline, AFC validation, HAC–Holm comparisons and observed-arrival queue-timing sensitivity. The reproducibility archive records source-file SHA-256 hashes, split sizes, random seeds, model settings and derived tables.

We compare the LSTM–TCN models with established machine-learning regressors using a parallel 47-feature design from the same 5 min history: current flow, proxy inflow when available, lags (1–288 steps), rolling means, recent differences, periodic encodings and the missing-record mask. The benchmark includes ridge regression, linear support vector regression (SVR), Random Forest, an extreme-gradient-boosting (XGBoost) residual booster and a squared-error Gradient Boosting model. Each model predicts the same four horizons directly, uses the same chronological split and 12-step purge, and leaves the test

set untouched. The fixed reference settings are: ridge $\lambda = 100$; linear SVR $C = 1$ and $\varepsilon = 0.1$ for 30 epochs; Random Forest 16 trees, depth 8, eight randomly selected features per split and a 20,000-row bootstrap sample; XGBoost boosting 40 rounds, depth 5 and learning rate 0.05; and Gradient Boosting 40 rounds, depth 3 and learning rate 0.05.

The benchmark is reported separately for each target because their units and meanings differ. The benchmark identifies the strongest model for each target and horizon and shows when a deep architecture adds value. No pilot experiments were conducted with Informer, PatchTST or STAEformer; a fair comparison requires model-specific tuning and an equal compute budget.

2.6. External AFC Passenger-Flow Validation Design

The automatic fare collection (AFC) experiment uses 105,120 complete 5 min exit-card observations from 1 January to 31 December 2025 as an actual passenger-throughput target. The series is divided chronologically into 73,374 training, 15,723 validation and 15,724 test samples after window construction. M0 persistence, M1 time-series ridge, M4 LSTM + TCN and a same-slot-previous-day seasonal-naïve baseline are compared. M4 uses the same 32-unit LSTM and 32-channel TCN structure, five seeds, 12 maximum epochs and 48 randomly sampled batches per epoch; no population-mobility or train-arrival values enter the AFC model. Pairwise absolute-loss tests use a Bartlett-HAC estimator with lag 288 and Holm correction across eight M4-versus-baseline horizon comparisons.

2.7. Capacity Scenario and Queueing Model

The capacity scenario uses a deterministic fluid-queue approximation and recursively updates bottleneck accumulation at a 5 min step: $Q_t = \max(0, Q_{(t-1)} + \lambda_t - \mu_t)$, with cumulative queue person-minutes calculated as 5 times the sum of Q_t . Here λ_t is the 5 min aggregation of 481 observed train arrivals from 05:50 on 5 May 2024 through 05:55 the next day (290 bins; 275,478 alighting passengers). Baseline capacity is 3384 persons per 5 min, about 90% of the observed maximum 5 min arrival volume (3760 at 20:40). Capacity reductions of 30% and 55% set μ_t to 2368.8 and 1522.8 persons per 5 min. In the recovery case, the 55% reduction remains through 21:05 and μ_t is restored to 2876.4 persons per 5 min from 21:10. Timing sensitivity shifts this recovery clock from -30 to $+30$ min in 5 min increments while keeping the observed arrival sequence and capacities fixed.

3. Results

3.1. Dataset Characteristics

The queue experiment uses 481 valid train-service records from 5 May 2024, representing 275,478 alighting passengers. Tables 6 and 7 and Figure 3 describe the separate rail and proxy data sources.

Table 6. Statistical summary of Shenzhen North Station rail feeder passenger flow.

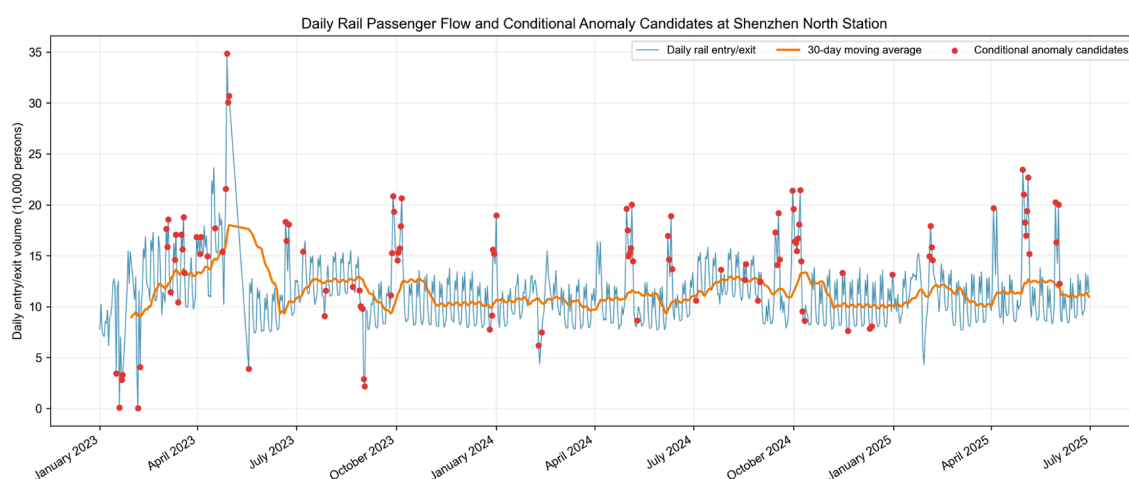
Indicator	Value
Valid daily samples	894 days
Mean daily entry/exit volume	113,856 passenger trips
Median	110,892 passenger trips
Standard deviation	33,230 passenger trips
95th percentile	170,208 passenger trips
Maximum	348,544 passenger trips
Mean entry volume	57,684 passenger trips

Table 6. *Cont.*

Indicator	Value
Mean exit volume	56,173 passenger trips
Mean directional imbalance	0.117

Table 7. Rail feeder passenger flow statistics by day of week.

Day of Week	Samples	Mean (Passenger Trips)	Median (Passenger Trips)	Standard Deviation (Passenger Trips)
Monday	128	111,230	99,945	29,268
Tuesday	127	99,171	89,632	26,458
Wednesday	127	99,294	89,845	27,061
Thursday	128	100,057	90,664	29,912
Friday	128	132,700	127,123	32,926
Saturday	128	121,318	113,443	32,305
Sunday	128	132,995	129,446	32,472

**Figure 3.** Daily rail passenger flow and conditional anomaly candidates at Shenzhen North Station.

3.2. Passenger-Flow Sensing Proxy and Descriptive Cross-Source Context

Rail feeder passenger flow shows clear day-of-week differences. Mean weekday flow is 108,500 passenger trips per day, and mean weekend flow is 127,200 passenger trips per day, giving a weekday–weekend ratio of 0.853. Mean Friday and Sunday flows are 132,700 and 133,000 passenger trips, respectively, both higher than Tuesday to Thursday.

Rail feeder statistics provide descriptive magnitude and day-of-week context, whereas the population-mobility proxy is the forecasting target in Section 3.3. The sources cover different periods and measure daily gate throughput versus 5 min regional crowd density, so their 894-day overlap supports only a feasibility check. The proxy is used as a regional-aggregation early-warning signal. Figure 4 shows the corresponding day-of-week distribution.

For valid non-zero population-mobility records, the mean real-time count is 30,600 persons, the median is 33,700 persons, the 95th percentile is 64,400 persons and the maximum is 128,200 persons. The mean real-time counts at 18:00 and 19:00 are 59,000 and 60,500 persons, respectively, and the P95 at 19:00 reaches 90,700 persons. The 19:00 interval is therefore a key window for evening-peak risk monitoring (Table 8; Figures 5 and 6).

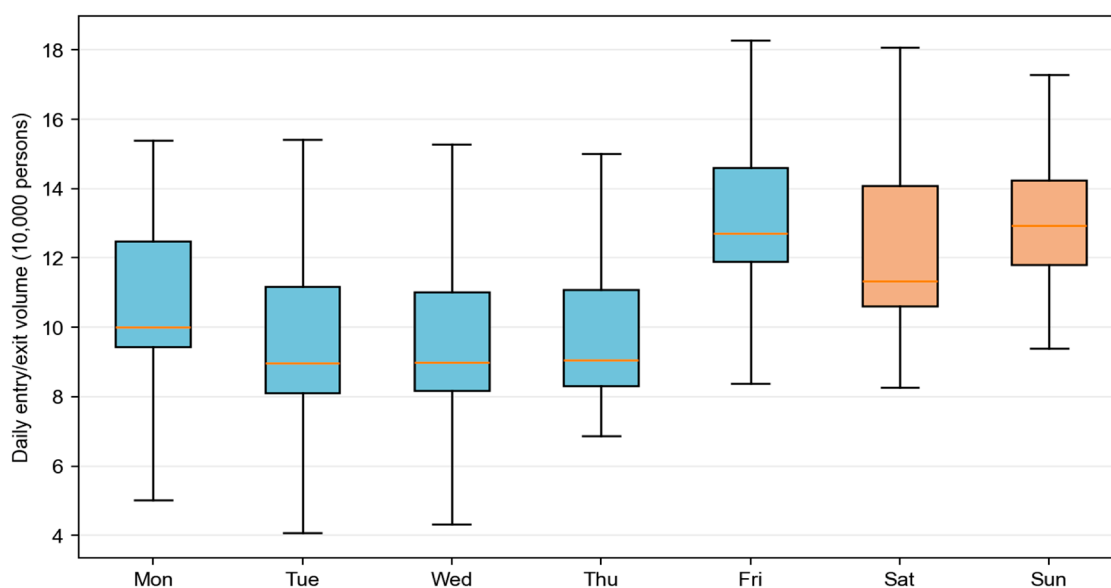


Figure 4. Distribution of rail feeder passenger flow by day of week. Blue boxes denote weekdays and orange boxes denote weekend days; orange horizontal lines mark medians, and whiskers show the non-outlier range.

Table 8. Real-time crowd count statistics for typical hours.

Hour	Mean (Proxy Count)	Median (Proxy Count)	P90 (Proxy Count)	P95 (Proxy Count)
7	23,876	22,404	33,747	38,147
8	47,539	45,392	64,336	72,464
9	46,671	44,248	62,690	70,584
12	43,897	41,698	57,738	65,518
18	58,991	55,940	78,476	88,509
19	60,470	57,320	80,202	90,661
21	35,852	33,980	47,261	53,674

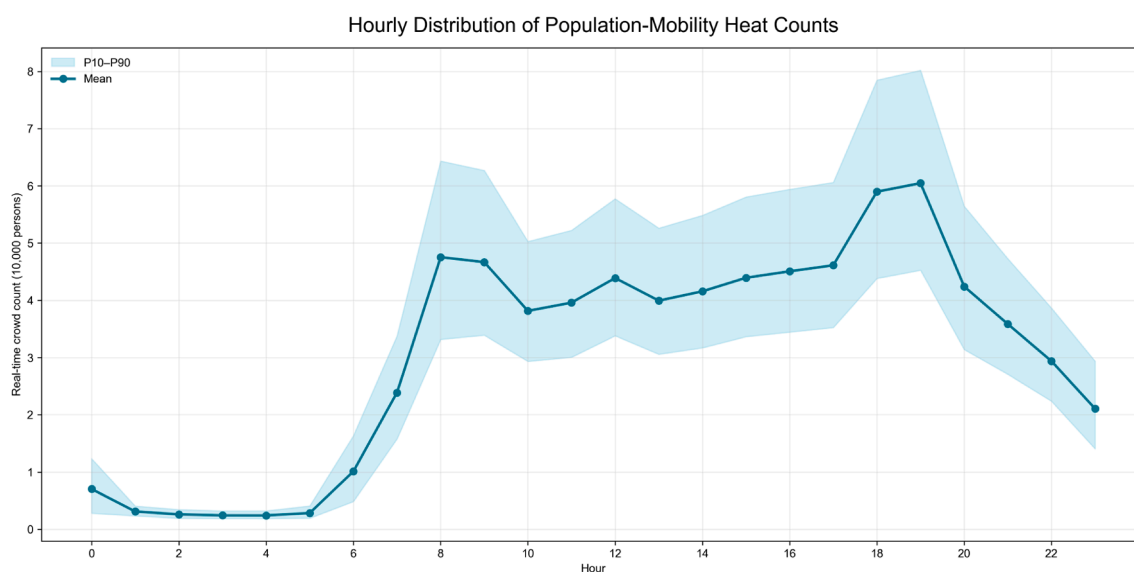


Figure 5. Hourly distribution and quantile interval of population-mobility crowd-density proxy counts.

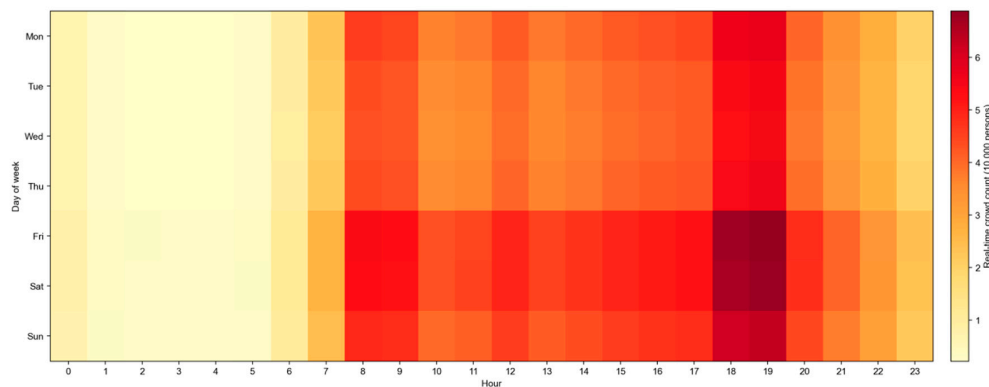


Figure 6. Day-of-week and hourly heat map of real-time crowd counts.

During the 2025 overlap, all 105,120 population-mobility observations aligned with AFC exit-card counts. The raw 5 min association was Pearson $r = 0.553$ (95% day-block bootstrap confidence interval (CI): 0.551–0.556) and Spearman $\rho = 0.877$. After removing the mean for each weekday and 5 min time slot from each series, the residual Pearson r was 0.505 (95% day-block CI: 0.457–0.538), and the residual Spearman ρ was 0.545. Raw Pearson r increased modestly to 0.568 at 60 min and to 0.872 at the daily scale (Table 9 and Figure 7). The residual association indicates shared temporal variation beyond the weekday/time-of-day pattern.

Table 9. Multi-scale external association between the population-mobility proxy and AFC exit-card passenger counts in 2025.

Aggregation	Paired Observations	Pearson r	95% Day-Block CI	Spearman ρ
5 min	105,120	0.553	[0.551, 0.556]	0.877
15 min	35,040	0.557	[0.554, 0.560]	0.880
30 min	17,520	0.560	[0.558, 0.563]	0.887
60 min	8760	0.568	[0.566, 0.571]	0.865
1 day	365	0.872	[0.832, 0.905]	0.840
5 min residual (de-seasonalized)	105,120	0.505	[0.457, 0.538]	0.545

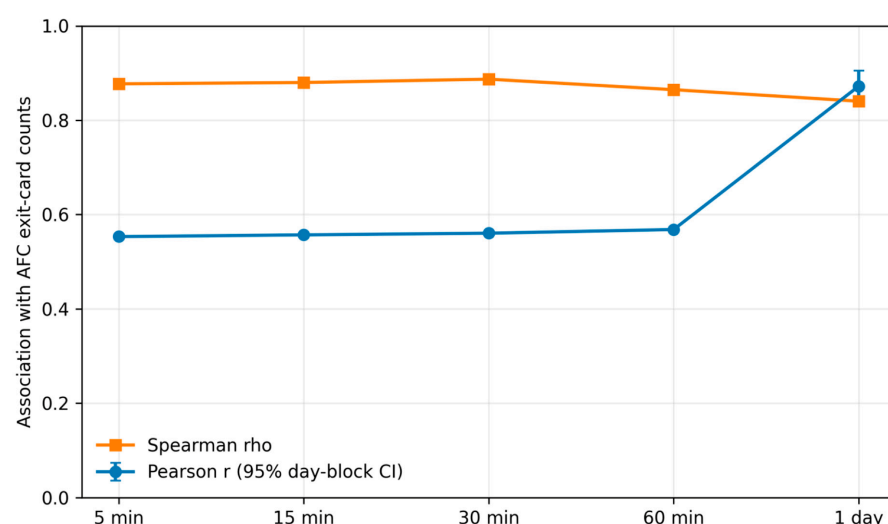


Figure 7. Multi-scale external association between the population-mobility proxy and AFC exit-card passenger counts in 2025. Error bars show 95% day-block bootstrap intervals for Pearson r .

Figure 8 compares the standardized daily trajectories over the full 2025 overlap and the raw 5 min series during the prespecified ordinary week.

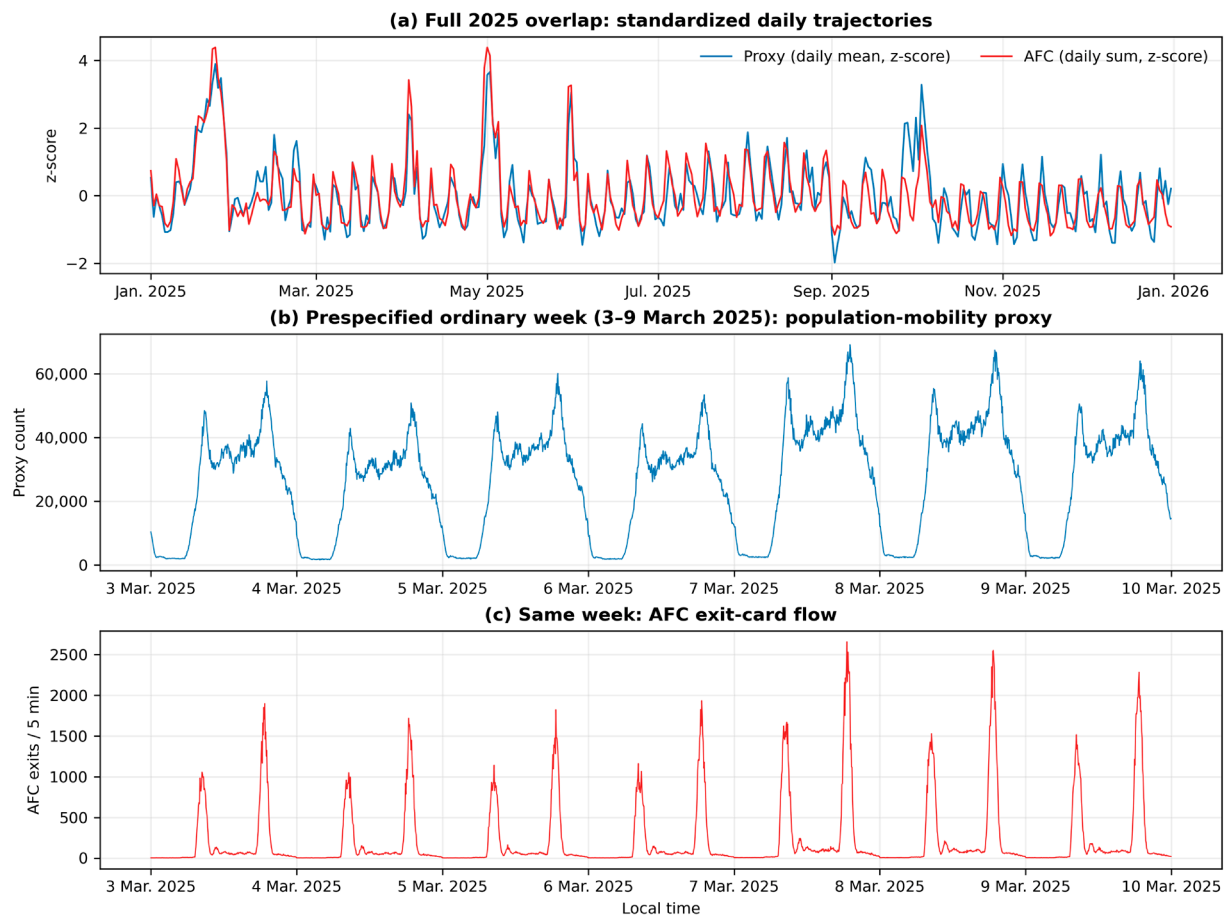


Figure 8. Side-by-side visualization of the population-mobility proxy and AFC exit-card series. Panel (a) shows standardized daily trajectories over the complete 2025 overlap; panels (b) and (c) show raw 5 min values during the same prespecified ordinary week (3–9 March 2025). The separate axes avoid implying equal units.

3.3. Forecasting Accuracy Comparison

Persistence (M0) and regularized time-series ridge regression (M1) [40] provide reproducible baselines; M2–M7 neural residual models are trained end to end on the same pipeline. Features include recent lags, short-term slopes, hour, day of week, annual cycle, month, weekend, Labour Day holiday and zero-value indicators. Training and testing are separated chronologically. Table 10 reports the full-test-set ablation results for persistence, time-series ridge regression, LSTM, TCN, LSTM + TCN, temporal attention, feature-group attention and peak-weighted loss ($n = 52,790$). All models use the same data split, input window and evaluation metrics.

At 5 min, M2 LSTM has the lowest neural MAE (1135 ± 1 proxy-count units), slightly below M5 (1150 ± 5), M4 (1155 ± 41) and ridge (1176). At 15, 30 and 60 min, M4 LSTM–TCN has the lowest neural MAE: 1321 ± 49 , 1536 ± 41 and 1634 ± 24 . These results identify a short-horizon LSTM advantage and a longer-horizon LSTM–TCN advantage within the neural ablation.

The 5 min sMAPE remains sensitive to low-count periods because small denominators amplify relative errors. MAE, RMSE and WAPE are therefore the primary accuracy criteria; sMAPE is retained as a scale-free diagnostic. The final full-epoch run reports the same metrics on the complete held-out period for every model and horizon.

Table 10. Temporally isolated proxy test results for M0–M7 and seasonal-naïve baselines across forecast horizons.

Horizon (min)	Model	Seeds	MAE (Proxy Count)	RMSE (Proxy Count)	WAPE (%)	sMAPE (%)
5	M0 Persistence	1	1326	1963	4.33	5.23
5	M1 Time-series ridge regression	1	1176	1710	3.84	7.62
5	M-SN Seasonal naïve	1	4191	6718	13.68	14.46
5	M2 LSTM	5	1135 ± 1	1752 ± 1	3.70 ± 0.00	4.36 ± 0.07
5	M3 TCN	5	1403 ± 63	3514 ± 288	4.58 ± 0.20	6.60 ± 0.06
5	M4 LSTM + TCN	5	1155 ± 41	1833 ± 187	3.77 ± 0.13	6.12 ± 0.12
5	M5 LSTM + TCN + temporal attention	5	1150 ± 5	1763 ± 4	3.75 ± 0.02	5.48 ± 0.63
5	M6 M5 + gate + feature-group attention	5	1801 ± 154	4688 ± 616	5.88 ± 0.50	7.67 ± 0.30
5	M7 M6 + peak-weighted SmoothL1	5	1717 ± 165	4362 ± 624	5.60 ± 0.54	7.90 ± 0.31
15	M0 Persistence	1	2333	3299	7.61	10.46
15	M1 Time-series ridge regression	1	1761	2476	5.75	12.10
15	M-SN Seasonal naïve	1	4190	6718	13.68	14.46
15	M2 LSTM	5	1317 ± 6	1958 ± 9	4.30 ± 0.02	5.85 ± 0.18
15	M3 TCN	5	1536 ± 67	3593 ± 275	5.01 ± 0.22	7.79 ± 0.12
15	M4 LSTM + TCN	5	1321 ± 49	2309 ± 217	4.31 ± 0.16	7.24 ± 0.11
15	M5 LSTM + TCN + temporal attention	5	1388 ± 19	2055 ± 30	4.53 ± 0.06	7.56 ± 0.68
15	M6 M5 + gate + feature-group attention	5	1912 ± 187	4656 ± 746	6.24 ± 0.61	8.98 ± 0.55
15	M7 M6 + peak-weighted SmoothL1	5	1821 ± 115	4340 ± 427	5.94 ± 0.38	9.29 ± 0.56
30	M0 Persistence	1	4005	5669	13.07	18.21
30	M1 Time-series ridge regression	1	2829	3918	9.23	19.09
30	M-SN Seasonal naïve	1	4191	6718	13.68	14.46
30	M2 LSTM	5	1501 ± 15	2316 ± 26	4.90 ± 0.05	7.36 ± 0.30
30	M3 TCN	5	1708 ± 69	3831 ± 277	5.57 ± 0.23	9.02 ± 0.17
30	M4 LSTM + TCN	5	1536 ± 41	2917 ± 171	5.01 ± 0.13	8.50 ± 0.15
30	M5 LSTM + TCN + temporal attention	5	1658 ± 53	2579 ± 80	5.41 ± 0.17	9.56 ± 0.97
30	M6 M5 + gate + feature-group attention	5	2092 ± 157	4929 ± 604	6.83 ± 0.51	10.39 ± 1.00
30	M7 M6 + peak-weighted SmoothL1	5	2056 ± 152	4800 ± 495	6.71 ± 0.50	10.52 ± 0.66
60	M0 Persistence	1	7025	9905	22.93	31.49
60	M1 Time-series ridge regression	1	4491	6293	14.66	28.29
60	M-SN Seasonal naïve	1	4190	6718	13.67	14.46
60	M2 LSTM	5	1683 ± 6	2865 ± 18	5.49 ± 0.02	9.10 ± 0.19
60	M3 TCN	5	1842 ± 47	4016 ± 169	6.01 ± 0.15	10.24 ± 0.25
60	M4 LSTM + TCN	5	1634 ± 24	3192 ± 110	5.33 ± 0.08	9.70 ± 0.28

Table 10. Cont.

Horizon (min)	Model	Seeds	MAE (Proxy Count)	RMSE (Proxy Count)	WAPE (%)	sMAPE (%)
60	M5 LSTM + TCN + temporal attention	5	1913 $\pm$ 67	3378 $\pm$ 136	6.24 $\pm$ 0.22	12.17 $\pm$ 0.77
60	M6 M5 + gate + feature-group attention	5	2279 $\pm$ 119	5254 $\pm$ 422	7.44 $\pm$ 0.39	12.27 $\pm$ 0.77
60	M7 M6 + peak-weighted SmoothL1	5	2309 $\pm$ 127	5370 $\pm$ 355	7.54 $\pm$ 0.41	12.25 $\pm$ 0.84

Note: Lower values indicate better performance. M-SN predicts each target from the same 5 min slot one day earlier. MAE and RMSE are in population-mobility proxy-count units, not AFC trips. WAPE and sMAPE are percentages; mean $\pm$ standard deviation is shown for M2–M7 over five seeds.

At the 60 min horizon, M4 LSTM–TCN has a proxy MAE of 1634 ± 24 , 63.62% lower than M1 and 76.73% lower than M0. M2 LSTM and M3 TCN have MAEs of 1683 ± 6 and 1842 ± 47 , respectively. Bartlett–HAC paired absolute-loss comparisons remain the inferential comparison for the neural ablation.

The marginal return of added modules is not monotonic. M5 is close to M4 at 5 min but has higher MAE at 15–60 min; M6 and M7 have higher MAE than M4 at every horizon. Holm-adjusted p values for M4 versus M1 at 5, 15, 30 and 60 min are 1.000, 1.35×10^{-14} , 1.10×10^{-59} and 1.08×10^{-246} ; the corresponding M4-versus-M5 values are 0.914, 0.368, 0.093 and 2.81×10^{-12} . M4 versus M6 and M7 is significant at every horizon after Holm correction ($p < 0.001$). Under the aligned proxy feature set and matched training budget, M5–M7 did not provide additional predictive value over M4.

In the held-out period, M2 is the strongest neural model at 5 min and M4 is the strongest neural model at 15–60 min. Relative to ridge, M2 reduces 5 min MAE by 3.46%, while M4 reduces MAE by 24.97%, 45.69% and 63.62% at 15, 30 and 60 min. Random Forest remains the strongest conventional reference, but full-epoch M4 is lower than the reported Random Forest MAE at 15–60 min. Model ranking is target- and horizon-dependent.

The seasonal-naïve baseline has an MAE of approximately 4190 proxy-count units at every horizon and is weaker than the final neural models at all four horizons. It retains a direct periodic reference and shows how full-epoch training changes the relative ranking of the nonlinear models. Figure 9 summarizes the horizon-specific MAE and WAPE values for all proxy-target models.

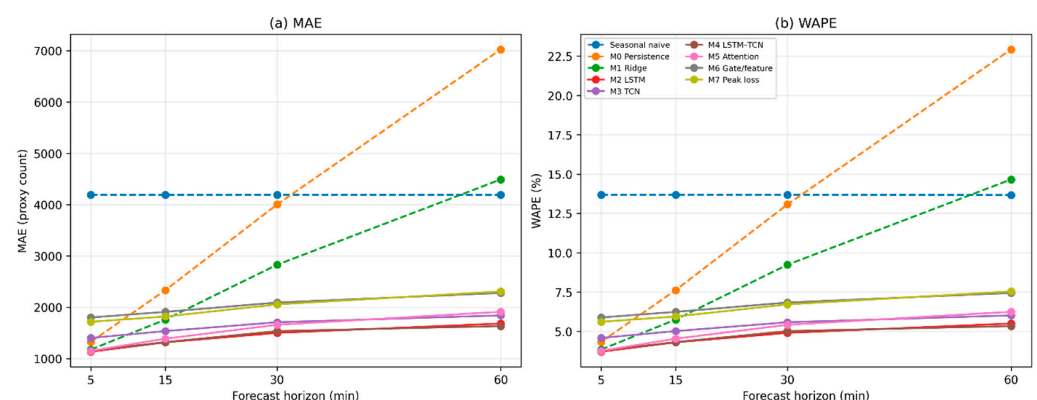


Figure 9. Model-error comparison across forecast horizons (full test set $n = 52,790$; M2–M7 are means over five random seeds). (a) MAE and (b) WAPE across forecast horizons. Dashed lines denote deterministic single-run baselines, whereas solid lines denote five-seed neural-model means. M2 is best among the neural models at 5 min; M4 is best at 15–60 min. M5–M7 do not improve on M4 at the longer horizons.

Figure 10 shows the M4 15 min point forecast, an empirical P90 residual band and signed residuals on a representative 48 h test segment. The band is a descriptive error diagnostic. A calibrated prediction interval would require a separate conformal calibration protocol [41,42]. The point-forecast analysis uses the full M0–M7 ablation and seed-aware HAC comparisons.

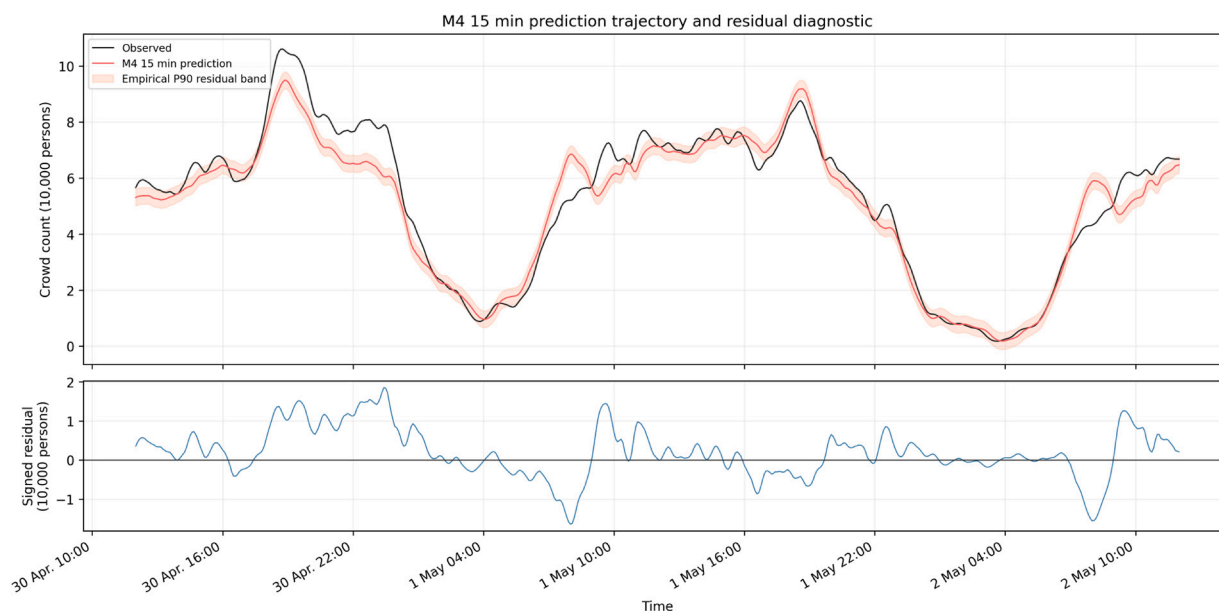


Figure 10. M4 15 min prediction trajectory and signed-residual diagnostic on a representative 48 h test segment. The black line shows observations, the red line shows the M4 point forecast, the shaded band shows the empirical P90 residual band, and the lower panel shows signed residuals.

Table 11 and Figure 11 compare the models with conventional regressors. On the proxy target, ridge remains best at 5 min, while Random Forest is the strongest conventional model at 15–60 min (MAE 1573.85, 2004.15 and 2490.56). On the independent AFC target, Random Forest is the strongest of the six conventional references at all horizons (MAE 22.27, 25.42, 30.17 and 38.26 exits per 5 min). The benchmark shows that M4 is useful at selected horizons, while Random Forest leads at the other target–horizon combinations.

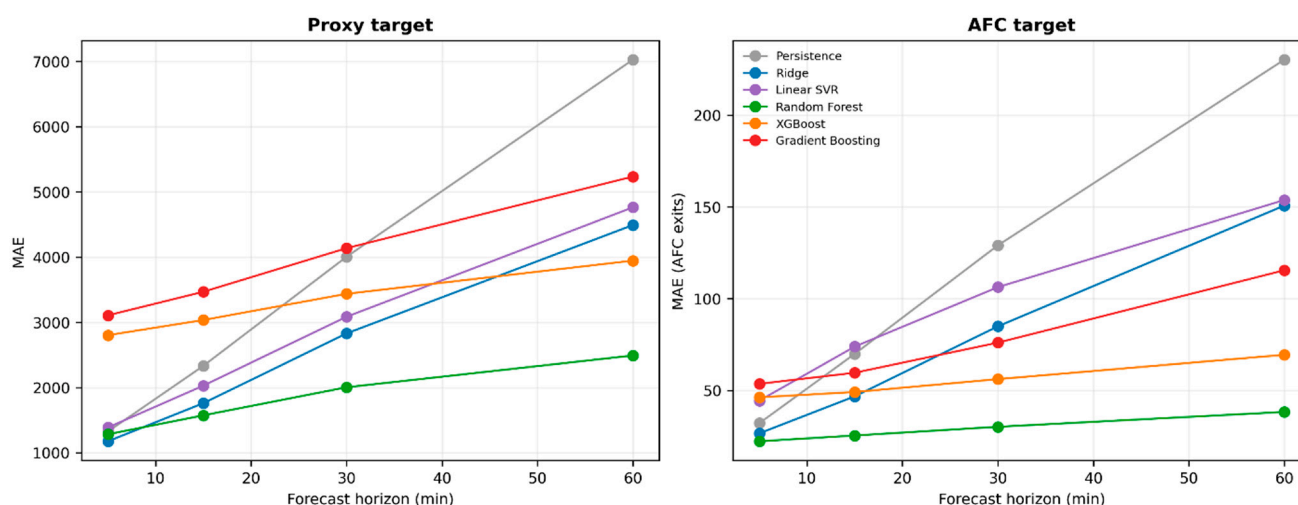


Figure 11. Conventional-regressor MAE comparison on the proxy and AFC targets.

Table 11. Conventional-regressor MAE by target and forecast horizon.

Target	Model	5 min	15 min	30 min	60 min
Proxy	Persistence	1325.57	2333.23	4004.59	7024.87
Proxy	Ridge	1175.89	1760.65	2828.93	4490.88
Proxy	Linear SVR	1387.56	2029.72	3084.37	4761.73
Proxy	Random Forest	1283.48	1573.85	2004.15	2490.56
Proxy	XGBoost	2802.24	3034.66	3436.42	3945.64
Proxy	Gradient Boosting	3106.46	3468.19	4136.39	5233.71
AFC	Persistence	32.14	69.91	129.08	230.19
AFC	Ridge	26.60	46.77	84.92	150.69
AFC	Linear SVR	44.44	73.85	106.28	153.74
AFC	Random Forest	22.27	25.42	30.17	38.26
AFC	XGBoost	46.24	49.13	56.16	69.39
AFC	Gradient Boosting	53.58	59.63	76.02	115.47

Note: MAE is in proxy-count units for the proxy target and AFC exits per 5 min for the AFC target. The XGBoost row is the self-contained XGBoost residual booster described in Section 2.5.

Full metrics and implementation settings are available from the corresponding author upon reasonable request, subject to the Data Availability Statement.

3.4. Overall Multi-Model Statistical Comparison

Across the reported candidate set, M2 has the lowest 5 min MAE and M4 has the lowest MAE at 15, 30 and 60 min. Figure 12 shows the descriptive M4-to-M1 ratio. HAC comparisons use seed-specific losses with Holm correction.

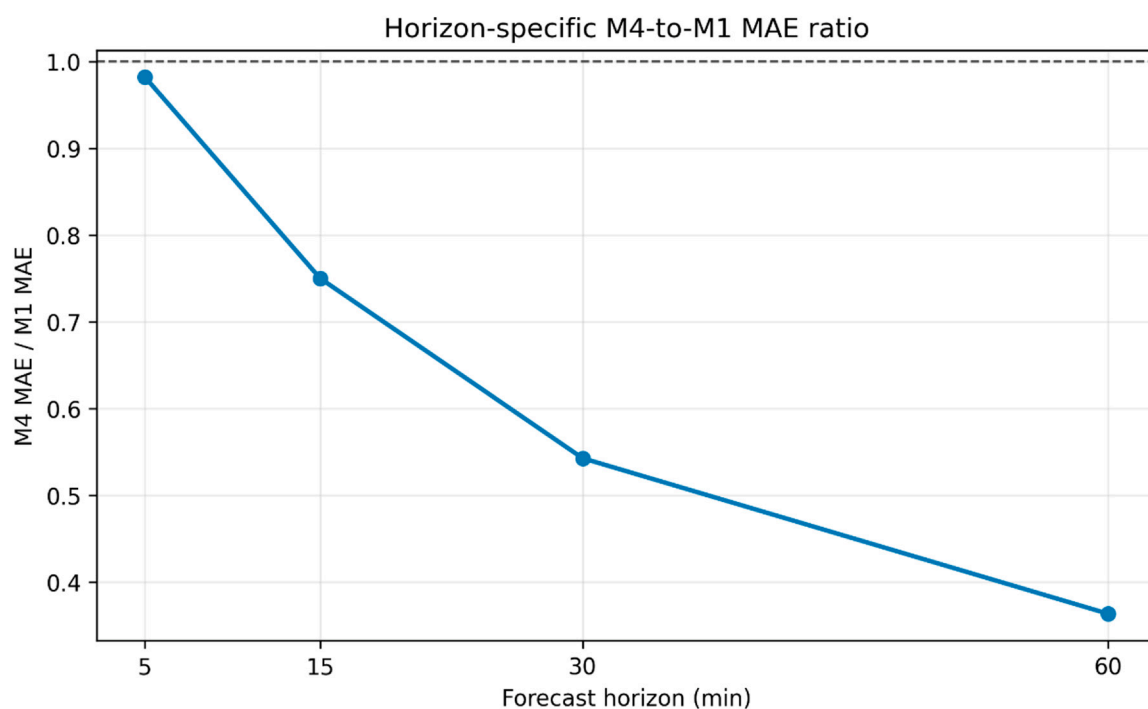


Figure 12. Horizon-specific M4-to-M1 MAE ratio in the held-out test period. The dashed horizontal line marks parity (ratio = 1); ratios above 1 favor M1 and ratios below 1 favor M4. The comparison is descriptive and does not imply rank-based inference.

M2 is the efficient neural choice at 5 min, whereas M4 is the efficient neural choice at 15–60 min. M5–M7 add parameters and training time without improving the aligned-feature forecast under the reported proxy feature set and training design.

3.5. Temporal Structure and Anomaly Candidates

The autocorrelation coefficients of the high-frequency series at 5, 15, 30 and 60 min lags are 0.995, 0.987, 0.963 and 0.888, respectively. They decrease to 0.714 after 2 h and -0.029 after 6 h. The 12 h lag is -0.440 , while the 24 h lag rises to 0.953 and the 7-day lag remains 0.919. This dual-scale pattern of strong short-term inertia and pronounced daily periodicity supports parallel modelling of long-range memory and local convolution (Figure 13).

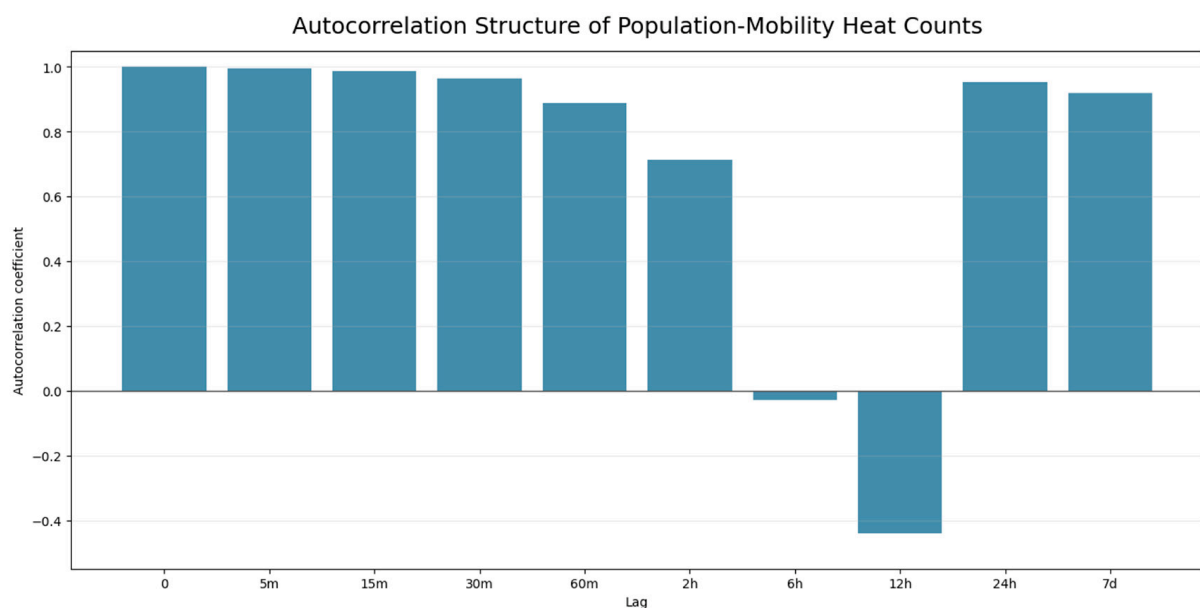


Figure 13. Autocorrelation of population-mobility real-time crowd counts at Shenzhen North Station (January 2023–May 2026, 5 min intervals).

Weekday–hour–minute conditional IQR scores identify 7173 statistical outliers. Event-level classification requires labelled operational logs, so these points are treated as anomaly candidates.

Two examples characterize different anomaly patterns in the multi-year series. The largest single-point deviation reached 128,167 persons at 19:10 on 30 April 2023, during the evening peak before the Labour Day holiday. The longest continuous exceedance ran from 23:10 on 30 April 2026 to 00:45 on 1 May 2026, with the series remaining above the local conditional baseline despite zero real-time inflow at some time points. Potential explanations include holiday retention, acquisition-definition changes and unrecorded events; operational logs are required to determine the cause. The 2026 interval lies in the held-out test period and was used only for post hoc description, not for model selection, parameter tuning or threshold estimation (Figure 14).

Anomaly scores are calculated using weekday–hour–minute conditional quantiles and IQR local thresholds. Candidate points indicate statistical deviations relative to comparable periods and require interpretation together with holidays, train operations and operational logs.

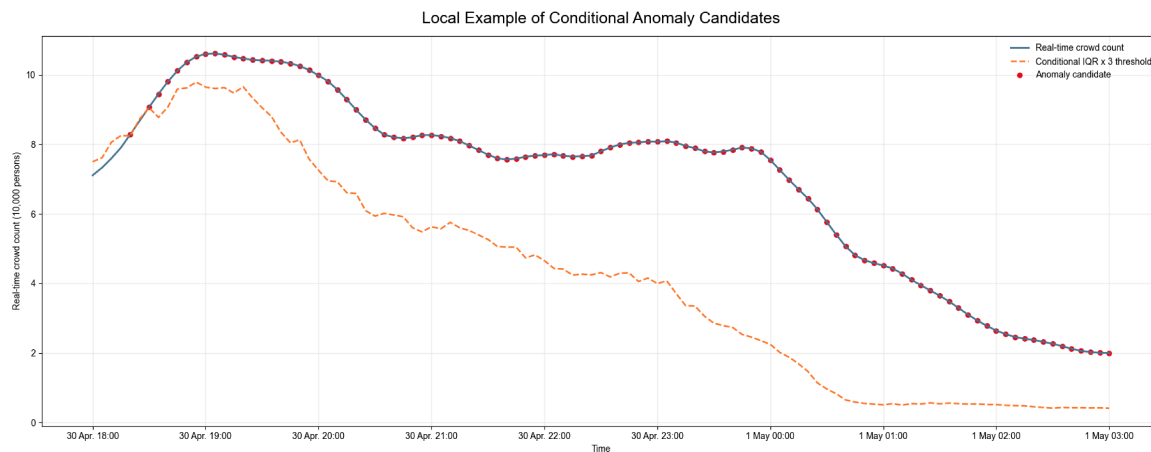


Figure 14. Conditional anomaly candidates and local baseline in population-mobility data.

3.6. Train Pulses, Capacity Scenarios, and Timing Sensitivity

3.6.1. Observed-Arrival Capacity Scenarios

Train-service data show that railway-arrival demand forms discrete pulses rather than a smooth flow. The capacity-scenario results use 5 May 2024 (481 valid services) as a high-demand pressure-test day with the peak-day arrival parameters specified in Section 2.7. Train-service records cover only May 2024, whereas the forecasting proxy spans 2023–2026, so the two sources do not overlap in the test period. This single-day experiment quantifies queue sensitivity to recovery timing under the specified deterministic capacity scenario (Table 12, Figure 15).

Table 12. Facility-capacity reduction scenario results for a peak-day pressure test based on single-day train-arrival data (5 May 2024).

Scenario	Service Capacity (Persons per 5 min)	Maximum Queue (Persons)	Cumulative Queue (Person-Minutes)
Baseline capacity	3384	376	3005
Capacity reduced by 30%	2369	1391	57,795
Capacity reduced by 55%	1523	9181	1,884,985
55% reduction; recovery at 21:10	1523 to 2876	6651	617,207

Note: The recurrence covers the complete observed arrival sequence from 05:50 on 5 May through 05:55 on 6 May (290 five-minute bins). The recovery row uses 1523 persons/5 min through 21:05 and 2876 persons/5 min from 21:10. Values are recomputed directly from the observed service records.

The train-service data contain 481 valid records; the 5 min bins are shown here to make the discrete input to the capacity recurrence explicit.

Under baseline capacity, the maximum queue is 376 persons and cumulative queueing is 3005 person-minutes. A 30% capacity reduction increases these values to 1391 persons and 57,795 person-minutes. Under a 55% reduction, maximum queueing reaches 9181 persons and cumulative queueing reaches 1,884,985 person-minutes. Restoring capacity to 85% of baseline at 21:10 reduces cumulative queueing to 617,207 person-minutes, a 67.3% reduction relative to no recovery (Figure 16).

When service capacity is reduced by 55%, the maximum queue is amplified from 376 to 9181 persons. Capacity recovery at 21:10 lowers cumulative queue person-minutes from

1,884,985 to 617,207, a 67.3% reduction. The experiment is driven by observed train-arrival pulses and a specified capacity path.

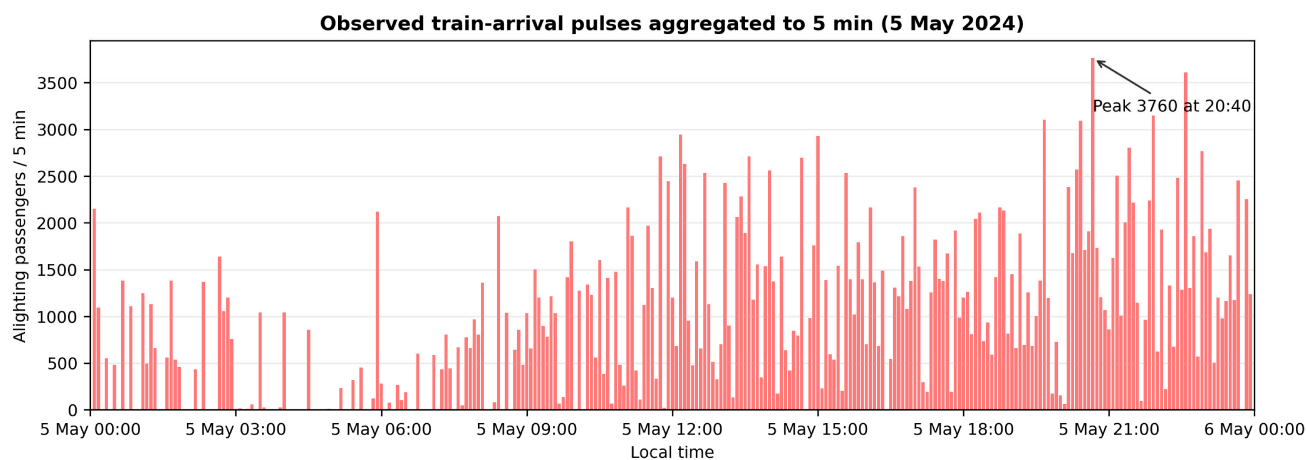


Figure 15. Observed train-arrival pulses aggregated to 5 min for the 5 May 2024 pressure-test day.

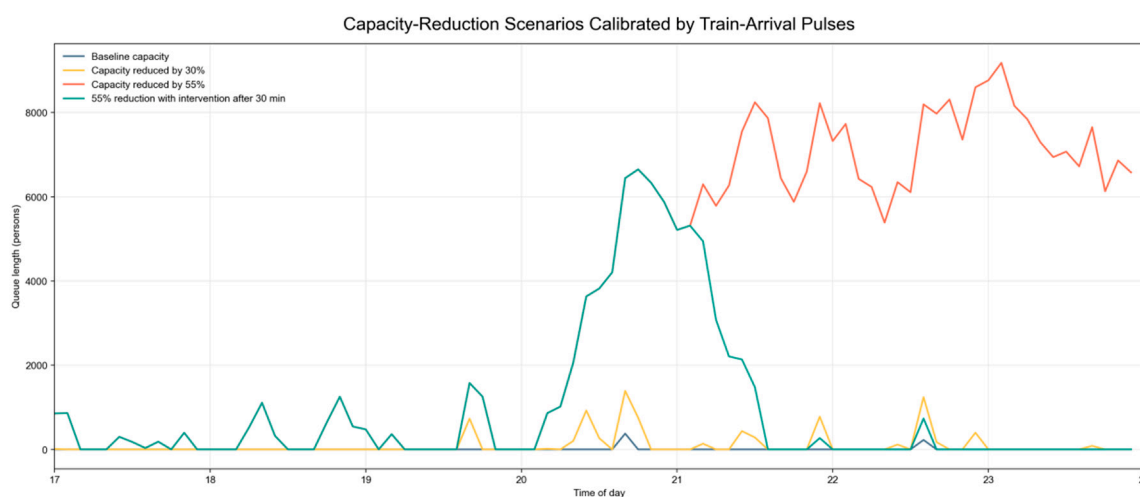


Figure 16. Capacity-reduction scenarios calibrated by observed train-arrival pulses.

3.6.2. Sensitivity of Queue Relief to Capacity-Recovery Timing

Using the same observed arrivals and capacities, moving the 21:10 recovery clock by -15 , -10 , $+10$ and $+15$ min changes the reduction in cumulative queue person-minutes from 67.3% to 72.8%, 71.3%, 63.4% and 61.7%, respectively (Table 13 and Figure 17). The maximum queue is 6651 persons at 20:45 for all five shifts. The queue clears at 21:15, 21:20, 21:35, 21:40 and 21:45, respectively; recovery timing changes queue duration rather than peak height in this scenario.

Table 13. Observed-arrival queue sensitivity to capacity-recovery timing.

Timing Shift (min)	Recovery Time	Maximum Queue	Cumulative Queue Person-Minutes	Reduction vs. No Recovery (%)	Peak Time	Queue-Clear Time
-15	20:55	6651	511,800	72.8	20:45	21:15
-10	21:00	6651	540,729	71.3	20:45	21:20
$+0$	21:10	6651	617,207	67.3	20:45	21:35

Table 13. Cont.

Timing Shift (min)	Recovery Time	Maximum Queue	Cumulative Queue Person-Minutes	Reduction vs. No Recovery (%)	Peak Time	Queue-Clear Time
+10	21:20	6651	690,390	63.4	20:45	21:40
+15	21:25	6651	722,594	61.7	20:45	21:45

Note: Negative shifts denote earlier recovery and positive shifts denote delay relative to 21:10. The observed sequence contains 481 valid services aggregated into 290 five-minute bins. The no-recovery comparator is 1,884,985 queue person-minutes; the peak occurs at 20:45 before every recovery clock in Table 13.

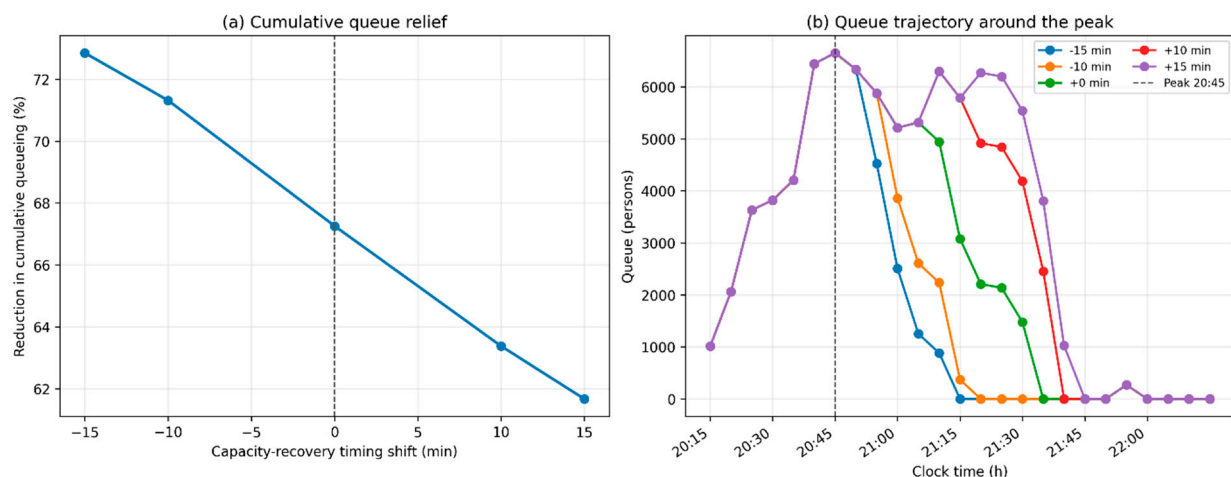


Figure 17. Queue-relief sensitivity to capacity-recovery timing and queue trajectories around the 20:45 peak. The vertical dashed line marks the 20:45 peak time.

3.7. External AFC Passenger-Flow Validation Results

On 15,724 temporally held-out AFC samples, M4 reduces MAE relative to M1 by 6.8%, 31.6%, 47.1% and 41.0% at 5, 15, 30 and 60 min (Table 14 and Figure 18). Bartlett-HAC comparisons favour M4 over M1 at every horizon after Holm correction (all adjusted $p < 0.001$). The seasonal-naïve baseline changes the interpretation: M4 is better at 5 and 15 min, but seasonal naïve is better at 30 and 60 min; all four directional differences remain significant after Holm correction, with the least decisive comparison at 15 min (adjusted $p = 0.029$). The AFC analysis compares M0, M1, M4 and seasonal naïve, the key competitive models identified in the proxy ablation (Table 10); full M0–M7 AFC replication remains future work.

Table 14. Temporally isolated AFC exit-card passenger-flow validation.

Horizon	Model	Seeds	MAE	RMSE	WAPE (%)
5	M0 Persistence	1	32.14	78.84	14.89
5	M1 Time-series ridge	1	26.44	59.11	12.24
5	M4 LSTM + TCN	5	24.63 ± 0.11	62.71 ± 0.15	11.41 ± 0.05
5	Seasonal naïve	1	35.22	97.09	16.31
15	M0 Persistence	1	69.91	159.33	32.40
15	M1 Time-series ridge	1	46.61	88.41	21.60
15	M4 LSTM + TCN	5	31.90 ± 0.74	72.18 ± 1.08	14.79 ± 0.34
15	Seasonal naïve	1	35.15	96.91	16.29

Table 14. Cont.

Horizon	Model	Seeds	MAE	RMSE	WAPE (%)
30	M0 Persistence	1	129.08	288.74	59.88
30	M1 Time-series ridge	1	84.65	147.49	39.27
30	M4 LSTM + TCN	5	44.79 $\pm$ 1.85	97.35 $\pm$ 2.76	20.78 $\pm$ 0.86
30	Seasonal naive	1	35.08	96.77	16.28
60	M0 Persistence	1	230.19	492.71	106.89
60	M1 Time-series ridge	1	149.90	250.98	69.60
60	M4 LSTM + TCN	5	88.42 $\pm$ 3.15	218.00 $\pm$ 2.23	41.06 $\pm$ 1.46
60	Seasonal naive	1	35.00	96.66	16.25

Note: AFC values are exit-card counts per 5 min. M4 reports mean $\pm$ standard deviation over five seeds. Seasonal naive uses the same 5 min slot one day earlier.

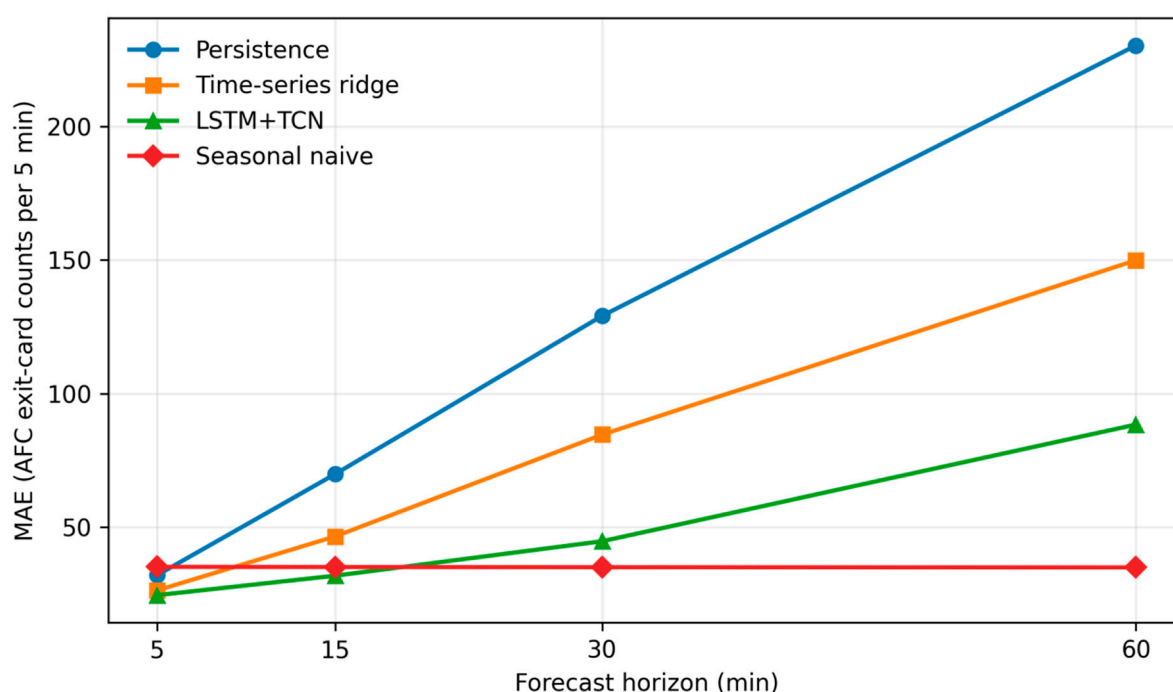


Figure 18. MAE comparison on the independent AFC exit-card passenger-flow target.

The AFC experiment confirms that model ranking depends jointly on the passenger-flow target, forecast horizon and benchmark family: M4 improves on ridge at 5–60 min, seasonal naive remains strongest at 30 and 60 min among the four models, and the added Random Forest reference is strongest among the six conventional models at all four horizons.

4. Discussion

4.1. Main Findings

The study yields three main findings. First, the population-mobility proxy and AFC exit-card passenger flow share temporal information: across 105,120 aligned 5 min observations, raw Pearson correlation was 0.553 and the weekday-by-clock de-seasonalized Pearson correlation was 0.505 (95% day-block CI 0.457–0.538). The proxy represents regional crowd stock and AFC represents station-exit throughput.

Second, model performance depends on target, forecast horizon and benchmark family. For the proxy target, ridge is best at 5 min, M4 is best among the neural ablation at

15–60 min, and Random Forest is the strongest conventional reference at 15–60 min. For the AFC target, Random Forest is strongest among the conventional references at all four horizons, while M4 remains better than ridge. These results demonstrate why learned, periodic and tree baselines should be retained together.

Third, queue relief was sensitive to capacity-recovery timing in the deterministic observed-arrival scenario. Recovery at 21:10 reduced cumulative queue person-minutes by 67.3%, whereas delays of 10 and 15 min reduced the benefit to 63.4% and 61.7%, respectively.

4.2. Relationship to Existing Methods

LSTM was originally proposed to alleviate long-term dependency learning difficulties in recurrent networks [1], and TCN studies show that causal dilated convolution can obtain long effective memory across sequence tasks [2]. The M6–M7 extensions use a gating mechanism to vary the LSTM and TCN contributions with the encoded state. Compared with global self-attention models [3] and Informer [4], the framework avoids computing a quadratic-length attention matrix over the full historical window. Compared with newer time-series models such as TimesNet, PatchTST and STAFormer [5–7], and with 2024 developments that emphasize long-short Transformer traffic representation, explicit exogenous-variable forecasting and large-language-model-based metro passenger-flow prediction under delay conditions [43–45], the study focuses on multi-source measurement constraints and a capacity interface.

Recent passenger-flow studies have largely pursued richer architectures and exogenous information. Hu et al. [46] combined a Graph Transformer with multi-gate mixture-of-experts learning and external operational variables; Wang et al. [47] integrated hybrid spatio-temporal extraction with multi-feature fusion; and Fernandez-Lobo et al. [48] used heterogeneous real-time data and LSTM-based occupancy forecasting for a multimodal transport interchange. The present study focuses on a bounded proxy feature set and asks when added temporal complexity improves accuracy and when it does not. It also links each forecast horizon to an operational action.

The M5–M7 results indicate an information-limited complexity penalty. M5 reweights the same history without adding exogenous observations; M6 applies its gate and feature-group attention only to proxy count, inflow, calendar and missingness because event and facility groups are unavailable; M7 changes the loss weighting but not the input information. Full-epoch training and the lower-rate sensitivity runs did not reverse the ordering, while M6–M7 showed larger seed variation than M4. The added modules therefore increased optimization variance without adding predictive signal in this feature set.

The population-mobility proxy and AFC exit-card counts remain separate targets and are compared only through timestamp-aligned association statistics. The external AFC experiment validates forecasting on actual passenger throughput, while the proxy ablation retains the longer multi-year regional series. Train arrivals and facility capacity enter the queue module as operational inputs and are excluded from both forecast models.

Transferability Across Countries and Transport Systems

The workflow transfers through its data roles and evaluation logic. A city needs (i) a clearly defined regional proxy or station-throughput target, (ii) a documented spatial footprint and sampling process, (iii) a 5 min or otherwise harmonized clock, (iv) a minimum persistence/seasonal/linear/tree benchmark set, and (v) operational logs for recalibrating thresholds and capacity responses. London station-arrival data, Seoul smart-card station clusters, Singapore OD/segment loads and Toronto Wi-Fi proxies can fill these roles, but their penetration, transfer-station topology, timetable regularity and fare-collection

definitions differ. Model weights, proxy-to-AFC mapping, anomaly quantiles, service-capacity parameters and horizon-to-action thresholds therefore require local estimation.

Local validation can use 5–15 min forecasts to stage staff, open or close gates, prepare temporary channels and issue targeted passenger information; 30–60 min forecasts can support train–platform coordination, diversion, bus/taxi reinforcement and cross-department rostering. Field deployment should keep a human-in-the-loop, log forecast errors and queue outcomes, and calibrate uncertainty before capacity changes.

4.3. Operational and Methodological Implications

Because the 60 min sMAPE values of M1 and M4 are 28.29% and 9.70%, respectively, the neural result is more suitable for trend-oriented proxy-risk indication than for precise gate-flow counting. For deep models to become practically useful, future work should reduce WAPE, improve peak recall, calibrate upper-risk intervals and inject proxy forecast-error distributions into the queue model.

The four forecast horizons map to progressively broader actions. A 5 min warning supports immediate on-site flow restriction and local guidance; 15 min supports staff repositioning and preparation of temporary passages; 30 min supports diversion, one-way concourse organization and feeder reinforcement; and 60 min supports cross-department staffing and bus/taxi capacity coordination. This mapping can be tested against response logs and labelled events.

Methodologically, strict temporal isolation, multiple random seeds and minimum baselines change how model complexity is judged. Proxy Table 10 shows a horizon-dependent M1-to-M4 crossover, whereas AFC Table 14 shows that seasonal naive can dominate nonlinear and linear learned models at 30–60 min. Model choice should follow target semantics, periodicity, horizon and operational cost; M4 is useful at selected proxy horizons and improves on ridge for AFC.

4.4. Limitations and Future Priorities

This study has five main limitations.

First, M5–M7 did not provide additional predictive value over M4 under the aligned proxy feature set and full-epoch training protocol. The bounded two-setting sensitivity check reduces, but does not eliminate, hyperparameter-selection uncertainty. Richer aligned exogenous inputs, a broader model-specific search and multiple temporal origins are required to test whether this ranking changes.

Second, the capacity analysis uses observed train arrivals from one high-demand day, 5 May 2024, and a deterministic fluid-queue model. Validation requires ordinary weekdays, weekends and additional holiday periods, together with facility-level simulation and operational response records.

Third, no pilot experiments with Informer, PatchTST or STAEformer were conducted on the proxy and AFC datasets. These models require a separate model-specific tuning study with equalized computational budgets and repeated temporal origins.

Fourth, the proposed anomaly-candidate and multi-level warning rules have not been evaluated against labelled operational events. Their thresholds should be calibrated and assessed using event-level precision, recall, false-alarm rates, missed-detection rates and warning lead time before operational deployment.

Fifth, Table 9 measures shared temporal signal between regional proxy counts and AFC exit throughput. A calibrated proxy-to-AFC conversion requires a separate mapping model across operating conditions and event types.

Future research should repeat both forecasting tasks across multiple rolling temporal origins, compare modern forecasting models under equal computational budgets, develop

a calibrated proxy-to-AFC mapping, extend the queue analysis to a wider range of operating days, and validate the warning and capacity-response interfaces against labelled operational records.

5. Conclusions

This study evaluates short-term hub passenger-flow forecasting through complementary proxy and AFC evidence. The multi-year population-mobility task supports a full M0–M7 ablation, the independent 2025 AFC task tests actual exit-card throughput, and the conventional benchmark compares ridge, linear SVR, Random Forest, XGBoost boosting and Gradient Boosting under the same temporal isolation.

For the proxy, M2 is best within the neural ablation at 5 min and M4 is best within the neural ablation at 15–60 min; ridge is the strongest conventional reference at 5 min and Random Forest at 15–60 min. For AFC, Random Forest is strongest among the conventional references at 5–60 min and M4 improves on ridge. Model selection should therefore be conditioned on target semantics, periodicity, horizon and the operational cost of errors.

For high-inertia proxy data, nonlinear fusion becomes useful at longer horizons. The AFC results show that a tree ensemble can outperform both linear and neural models, while daily periodic repetition remains a strong comparator at 30–60 min. Minimum baselines should follow the operational meaning and periodicity of each target.

In the observed-arrival queue scenario, recovery at 21:10 reduced cumulative queue person-minutes by 67.3%, while delays of 10 and 15 min reduced the benefit to 63.4% and 61.7%. Multi-day operational data are required to evaluate the transferability of this timing effect.

Future research should prioritize repeated-origin AFC and proxy validation, calibrated cross-source mapping, modern-model comparisons, probabilistic interval calibration, multi-day arrival scenarios and field verification of capacity-response timing.

Author Contributions: Conceptualization, X.Z. and Z.Z.; methodology, X.Z.; software, X.Z.; validation, X.Z.; formal analysis, X.Z.; investigation, X.Z. and K.Z.; resources, K.Z. and Z.Z.; data curation, X.Z. and K.Z.; writing—original draft preparation, X.Z.; writing—review and editing, X.Z., Z.Z. and K.Z.; visualization, X.Z.; supervision, Z.Z.; project administration, Z.Z. and K.Z.; X.Z. served as the principal author responsible for model design, experiments and manuscript drafting; Z.Z. served as the academic supervisor and independently checked selected results and reproducibility scripts; K.Z. contributed as a secondary author mainly through project resources, data coordination and applied interpretation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study used desensitized aggregate passenger-flow statistics and did not involve identifiable personal information.

Informed Consent Statement: Not applicable.

Data Availability Statement: The rail, population-mobility, AFC exit-card and train-service records were obtained through Shenzhen North Station project and data-cooperation channels and remain restricted by data-use agreements. Reproducible materials available upon reasonable request include data-cleaning and feature-construction scripts, M0–M7 and AFC model configurations, random seeds, seasonal-naïve outputs, proxy-AFC association tables, HAC-Holm comparisons, queue-timing scripts, derived figures and run manifests with source-file SHA-256 hashes. Requests should be directed to the corresponding author and remain subject to provider approval and the applicable data-use agreements. Raw individual operational records cannot be publicly released without provider permission.

Acknowledgments: Artificial intelligence (AI)-assisted tools were used for language polishing and manuscript formatting. All scientific content, experimental design, data analysis, and conclusions are the sole work of the authors, who take full responsibility for the integrity of the work.

Conflicts of Interest: Author Kai Zhang was employed by the Shenzhen Urban Transport Planning Center Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
- Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271. [\[CrossRef\]](#)
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
- Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [\[CrossRef\]](#)
- Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023. [\[CrossRef\]](#)
- Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023. [\[CrossRef\]](#)
- Liu, H.; Dong, Z.; Jiang, R.; Deng, J.; Deng, J.; Chen, Q.; Song, X. STAEformer: Spatio-Temporal Adaptive Embedding Makes Vanilla Transformer SOTA for Traffic Forecasting. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023), Birmingham, UK, 21–25 October 2023; ACM: New York, NY, USA, 2023; pp. 4125–4129. [\[CrossRef\]](#)
- Liu, Y.; Liu, Z.; Jia, R. DeepPF: A Deep Learning Based Architecture for Metro Passenger Flow Prediction. *Transp. Res. Part C Emerg. Technol.* **2019**, *101*, 18–34. [\[CrossRef\]](#)
- Zhang, J.; Chen, F.; Cui, Z.; Guo, Y.; Zhu, Y. Deep Learning Architecture for Short-Term Passenger Flow Forecasting in Urban Rail Transit. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 7004–7014. [\[CrossRef\]](#)
- Yang, D.; Chen, K.; Yang, M.; Zhao, X. Urban Rail Transit Passenger Flow Forecast Based on LSTM with Enhanced Long-Term Features. *IET Intell. Transp. Syst.* **2019**, *13*, 1475–1482. [\[CrossRef\]](#)
- Zhang, J.; Chen, F.; Shen, Q. Cluster-Based LSTM Network for Short-Term Passenger Flow Forecasting in Urban Rail Transit. *IEEE Access* **2019**, *7*, 147653–147671. [\[CrossRef\]](#)
- Tang, Q.; Yang, M.; Yang, Y. ST-LSTM: A Deep Learning Approach Combined Spatio-Temporal Features for Short-Term Forecast in Rail Transit. *J. Adv. Transp.* **2019**, *2019*, 8392592. [\[CrossRef\]](#)
- Guo, J.; Xie, Z.; Qin, Y.; Jia, L.; Wang, Y. Short-Term Abnormal Passenger Flow Prediction Based on the Fusion of SVR and LSTM. *IEEE Access* **2019**, *7*, 42946–42955. [\[CrossRef\]](#)
- Yang, X.; Xue, Q.; Ding, M.; Wu, J.; Gao, Z. Short-Term Prediction of Passenger Volume for Urban Rail Systems: A Deep Learning Approach Based on Smart-Card Data. *Int. J. Prod. Econ.* **2021**, *231*, 107920. [\[CrossRef\]](#)
- Hao, S.; Lee, D.-H.; Zhao, D. Sequence to Sequence Learning with Attention Mechanism for Short-Term Passenger Flow Prediction in Large-Scale Metro System. *Transp. Res. Part C Emerg. Technol.* **2019**, *107*, 287–300. [\[CrossRef\]](#)
- Li, H.; Jin, K.; Sun, S.; Jia, X.; Li, Y. Metro Passenger Flow Forecasting Through Multi-Source Time-Series Fusion: An Ensemble Deep Learning Approach. *Appl. Soft Comput.* **2022**, *120*, 108644. [\[CrossRef\]](#)
- Yu, B.; Yin, H.; Zhu, Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 3634–3640. [\[CrossRef\]](#) [\[PubMed\]](#)
- Guo, S.; Lin, Y.; Feng, N.; Song, C.; Wan, H. Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 922–929. [\[CrossRef\]](#)
- Zheng, C.; Fan, X.; Wang, C.; Qi, J. GMAN: A Graph Multi-Attention Network for Traffic Prediction. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 1234–1241. [\[CrossRef\]](#)
- Zheng, H.; Lin, F.; Feng, X.; Chen, Y. A Hybrid Deep Learning Model with Attention-Based Conv-LSTM Networks for Short-Term Traffic Flow Prediction. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 6910–6920. [\[CrossRef\]](#)
- Zhang, Y.; Sun, K.; Wen, D.; Chen, D.; Lv, H.; Zhang, Q. Deep Learning for Metro Short-Term Origin-Destination Passenger Flow Forecasting Considering Section Capacity Utilization Ratio. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 7943–7960. [\[CrossRef\]](#)
- Noursalehi, P.; Koutsopoulos, H.N.; Zhao, J. Real time transit demand prediction capturing station interactions and impact of special events. *Transp. Res. Part C Emerg. Technol.* **2018**, *97*, 277–300. [\[CrossRef\]](#)
- Zhang, Q.; Liu, X.; Spurgeon, S.; Yu, D. A two-layer modelling framework for predicting passenger flow on trains: A case study of London underground trains. *Transp. Res. Part A Policy Pract.* **2021**, *151*, 119–139. [\[CrossRef\]](#)

24. Park, Y.; Choi, Y.; Kim, K.; Yoo, J.K. Machine learning approach for study on subway passenger flow. *Sci. Rep.* **2022**, *12*, 2754. [CrossRef] [PubMed]
25. Tian, X.; Zhang, C.; Zheng, B. Fine-grained passenger load prediction inside metro network via smart card data. *Int. J. Intell. Syst.* **2024**, *2024*, 6643018. [CrossRef]
26. Da Silva, D.; Shalaby, A. Forecasting short-term subway passenger flow using Wi-Fi data: Comparative analysis of advanced time-series methods. *J. Intell. Transp. Syst.* **2026**, *30*, 108–128. [CrossRef]
27. Shenzhen Municipal Transport Bureau. Statistical Information: Monthly Reports on Transport Operation Indicators. Available online: <https://jtys.sz.gov.cn/zwgk/xgkml/tjxx/> (accessed on 25 August 2026). (In Chinese)
28. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**, arXiv:1609.03499. [CrossRef]
29. Salimans, T.; Kingma, D.P. Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. In *Advances in Neural Information Processing Systems 29*; Curran Associates, Inc.: Red Hook, NY, USA, 2016; pp. 901–909.
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
31. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the Third International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015. [CrossRef]
32. Huber, P.J. Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **1964**, *35*, 73–101. [CrossRef]
33. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019. [CrossRef]
34. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. In Proceedings of the Fifth International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2017. [CrossRef]
35. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
36. Rousseeuw, P.J.; Croux, C. Alternatives to the Median Absolute Deviation. *J. Am. Stat. Assoc.* **1993**, *88*, 1273–1283. [CrossRef]
37. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD 1996)*, Portland, OR, USA, 2–4 August 1996; AAAI Press: Menlo Park, CA, USA, 1996; pp. 226–231.
38. Diebold, F.X.; Mariano, R.S. Comparing Predictive Accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [CrossRef]
39. Liu, Y.; Xie, J. Cauchy Combination Test: A Powerful Test with Analytic p-Value Calculation under Arbitrary Dependency Structures. *J. Am. Stat. Assoc.* **2020**, *115*, 393–402. [CrossRef] [PubMed]
40. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
41. Lei, J.; G’Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-Free Predictive Inference for Regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1094–1111. [CrossRef]
42. Fontana, M.; Zeni, G.; Vantini, S. Conformal Prediction: A Unified Review of Theory and New Challenges. *Bernoulli* **2023**, *29*, 1–23. [CrossRef]
43. Luo, Q.; He, S.; Han, X.; Wang, Y.; Li, H. LSTTN: A Long-Short Term Transformer-Based Spatio-Temporal Neural Network for Traffic Flow Forecasting. *Knowl. Based Syst.* **2024**, *293*, 111637. [CrossRef]
44. Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; Long, M. TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables. In *Advances in Neural Information Processing Systems 37*; Curran Associates, Inc.: Red Hook, NY, USA, 2024.
45. Huang, P.; He, Y.; Wang, H.; Chen, J.; Luo, Q. A Prompt Refinement-Based Large Language Model for Metro Passenger Flow Forecasting under Delay Conditions. *arXiv* **2024**, arXiv:2410.15111. [CrossRef]
46. Hu, S.; Chen, J.; Zhang, W.; Liu, G.; Chang, X. Graph Transformer Embedded Deep Learning for Short-Term Passenger Flow Prediction in Urban Rail Transit Systems: A Multi-Gate Mixture-of-Experts Model. *Inf. Sci.* **2024**, *679*, 121095. [CrossRef]
47. Wang, T.; Song, J.; Zhang, J.; Tian, J.; Wu, J.; Zheng, J. Short-Term Metro Passenger Flow Prediction Based on Hybrid Spatiotemporal Extraction and Multi-Feature Fusion. *Tunn. Undergr. Space Technol.* **2025**, *159*, 106491. [CrossRef]
48. Fernandez-Lobo, A.; Benavente, J.; Monzon, A. Dynamic Management Tool for Improving Passenger Experience at Transport Interchanges. *Future Transp.* **2025**, *5*, 59. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.