

Article

Short-Term Metro Passenger OD Demand Forecasting Based on Low-Rank Tensor Network Extended Kalman Filter

Aijing Su ^{1,*}, Bing Wu ¹ and Xiaoxing Fang ²

¹ Key Laboratory of Road and Traffic Engineering of Ministry of Education, School of Transportation Engineering, Tongji University, Shanghai 201804, China; wubing@tongji.edu.cn

² Research Center of Applied Electromagnetics, School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; fang_xiaoxing@nuist.edu.cn

* Correspondence: 1910910@tongji.edu.cn

Abstract

Accurate short-term metro origin–destination (OD) demand forecasting is essential for intelligent passenger flow management and urban rail transit operation. However, forecasting large-scale metro OD demand remains challenging due to its high dimensionality, nonlinear spatiotemporal dependencies, and demand uncertainty. To address these challenges, this paper proposes a Tensor Network Extended Kalman Filter (TNEKF) framework for short-term metro OD demand forecasting. First, metro OD demand is formulated as a nonlinear dynamic state-space prediction problem, where a multi-input multi-output Volterra model is adopted to characterize the nonlinear relationship between historical passenger demand and future OD flows. An Extended Kalman Filter (EKF) is then developed to recursively estimate the latent model parameters and continuously refine demand prediction using newly available observations. To improve computational efficiency for high-dimensional OD systems, both the latent state vector and covariance matrix are represented using low-rank tensor network structures, and all recursive filtering operations are implemented through tensor-network contractions without explicitly constructing large-scale matrices. Experiments on real-world smart-card data from the Hangzhou metro system demonstrate that the proposed method consistently outperforms ARIMA, conventional EKF, and several state-of-the-art spatiotemporal prediction models in terms of MAE, RMSE, and MAPE. Compared with the best-performing baseline of the whole-day scenario, the proposed method reduces MAE, RMSE, and MAPE by 30.2%, 9.8%, and 6.3%, respectively. Furthermore, the proposed framework exhibits strong robustness under disruption scenarios, demonstrating its effectiveness and scalability for large-scale metro OD demand forecasting.

Keywords: metro demand prediction; Volterra system; extended Kalman filter; tensor network; low-rank

Academic Editors: Wolfgang Kainz and Maria Antonia Brovelli

Received: 29 May 2026

Revised: 15 July 2026

Accepted: 17 July 2026

Published: 17 July 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. 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

1.1. Background

Urban rail transit systems have become an indispensable component of modern metropolitan transportation owing to their large passenger capacity, operational efficiency, and environmental sustainability. As metro networks continue to expand and travel demand becomes increasingly dynamic, intelligent operational strategies, including train

scheduling, capacity allocation, passenger flow management, and service optimization, have become essential for improving operational efficiency and passenger service quality [1]. The effectiveness of these strategies largely depends on accurate short-term passenger demand forecasting, which provides timely information for real-time operational decision making.

Passenger demand in metro systems can generally be represented at two spatial resolutions, namely station-level passenger flow and origin–destination (OD) demand. Station-level prediction focuses on inflow and outflow at individual stations [2], whereas OD demand explicitly describes passenger movements between every origin–destination station pair across the entire metro network [3]. Compared with station-level passenger flow, OD demand contains substantially richer spatial interaction information and therefore supports more refined operational decisions, such as train dispatching, dynamic timetable adjustment, and passenger flow control. However, the corresponding prediction problem becomes considerably more challenging because the OD matrix simultaneously exhibits high dimensionality, nonlinear temporal evolution, strong spatial coupling, and stochastic variations caused by passenger travel behavior.

From a data representation perspective, metro OD demand naturally forms a multi-dimensional structure involving origin stations, destination stations, and time intervals. Tensor representations provide an effective way to preserve these intrinsic spatial–temporal relationships and have been successfully applied to traffic data completion and prediction problems [4–6]. Nevertheless, directly performing recursive estimation on high-dimensional tensor representations usually incurs rapidly increasing computational and memory costs, making large-scale online prediction difficult [7–9].

1.2. Problem Statement

Existing passenger demand forecasting methods can generally be divided into two categories. Statistical approaches, including autoregressive integrated moving average (ARIMA) models [10,11] and Kalman filtering (KF) techniques [12], are widely used due to their interpretability and computational efficiency. In particular, the extended Kalman filter (EKF) offers an effective Bayesian framework for nonlinear dynamic estimation and has been successfully applied to traffic state estimation and related transportation systems [13–18]. On the other hand, deep learning approaches have demonstrated remarkable predictive performance by learning complex spatiotemporal features directly from historical traffic data [19–23]. Although these methods have significantly improved forecasting accuracy, several limitations remain when they are applied to large-scale metro OD demand prediction. Nevertheless, existing recursive estimation, tensor-based, and deep learning approaches still struggle to simultaneously achieve recursive uncertainty propagation, computational efficiency, and scalability for large-scale metro OD forecasting.

1.3. Proposed Approach and Contributions

To address these challenges, this paper develops a recursive Bayesian forecasting framework for short-term metro OD demand prediction based on an EKF. Unlike existing tensor-network Kalman filtering methods that were originally developed for nonlinear system identification in Volterra systems [24,25], the proposed framework is specifically designed for recursive metro OD forecasting. A multi-input multi-output Volterra state-space model is introduced to characterize the nonlinear evolution of metro OD demand, while low-rank tensor network representations are employed solely as an efficient computational tool for scalable recursive estimation. Specifically, the latent state vector and covariance matrix are represented using Matrix Product States (MPS) and Matrix Product Operators (MPO), respectively, allowing all recursive EKF computations to be carried out

directly in low-rank tensor-network form without explicitly constructing exponentially large matrices.

The major contributions of this paper are summarized as follows.

- (1) A recursive Bayesian forecasting framework for short-term metro OD demand prediction is proposed. The proposed framework formulates metro passenger demand forecasting as a nonlinear state-space estimation problem and develops a Volterra-based observation model to characterize the nonlinear spatiotemporal relationship between historical and future OD demand, enabling recursive prediction and uncertainty propagation for network-wide passenger demand.
- (2) A low-rank tensor-network implementation is developed to enable scalable recursive estimation for high-dimensional metro OD systems. The latent state and covariance are represented using MPS and MPO, respectively, allowing all EKF operations to be performed directly in tensor-network form without explicitly constructing large-scale matrices.
- (3) Experiments on a real-world Hangzhou metro dataset demonstrate the effectiveness of the proposed framework. Compared with representative statistical and deep learning methods, the proposed TNEKF achieves superior forecasting accuracy, stronger robustness under demand fluctuations and disruption scenarios, and significantly improved computational scalability for large-scale metro OD demand prediction.

1.4. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews the related work on metro passenger demand forecasting and tensor-network-based recursive estimation. Section 3 introduces the tensor network preliminaries that provide the computational foundation for the proposed method. Section 4 presents the proposed TNEKF framework for short-term metro OD forecasting, including the dynamic state-space formulation, recursive Bayesian estimation, analytical Jacobian construction, and low-rank tensor-network implementation. Section 5 presents the experimental settings, numerical results, and discussion. Finally, Section 6 concludes the paper and discusses future research directions.

2. Literature Review

Accurate short-term metro passenger demand forecasting has attracted considerable attention over the past decade because it provides essential support for intelligent transportation management and real-time operational decision-making. Existing studies can generally be categorized into four methodological groups: conventional statistical and state-space methods, deep learning approaches, tensor-based traffic prediction methods, and tensor-network-based recursive estimation methods.

2.1. Conventional Statistical and State-Space Methods

Traditional statistical models have been widely adopted for short-term traffic prediction because of their clear mathematical formulation and relatively low computational complexity. Shekhar and Williams [10] proposed an adaptive seasonal ARIMA model for short-term traffic flow forecasting. Their method effectively captures temporal periodicity but assumes approximately linear system dynamics, making it difficult to represent highly nonlinear passenger demand. Moreira-Matias et al. [11] combined streaming data with statistical learning to predict taxi passenger demand. Although the proposed framework demonstrated good real-time performance, it primarily focused on temporal prediction and did not explicitly model network-wide spatial interactions among OD pairs.

Kalman filtering techniques have also been extensively applied to transportation systems because they naturally incorporate uncertainty into recursive estimation. Wang and Papageorgiou [13] introduced an EKF for freeway traffic state estimation, demonstrating that recursive Bayesian estimation can effectively improve robustness under noisy observations. Subsequently, Wang et al. [14] further extended EKF to large-scale freeway surveillance systems, while Tampere and Immers [15] introduced dynamic parameter estimation within the EKF framework. Van Hinsbergen et al. [16] proposed a localized EKF strategy to improve computational scalability.

Although EKF-based methods provide an elegant recursive estimation framework with uncertainty quantification, these studies mainly consider low-dimensional traffic states.

2.2. Deep Learning Methods for Passenger Demand Forecasting

With the rapid growth of large-scale traffic data, deep learning has become the dominant paradigm for passenger demand prediction. Lv et al. [19] demonstrated that deep neural networks significantly outperform conventional statistical models for traffic flow forecasting. Ma et al. [26] proposed a convolutional bidirectional LSTM framework for network-wide metro ridership prediction, successfully integrating local spatial information with temporal dependencies. To further improve temporal modeling capability, Wei et al. [27] introduced an attention-enhanced LSTM model for subway passenger flow forecasting, while Fu et al. [28] investigated GRU networks for traffic prediction with reduced computational cost.

Graph neural networks have further improved the modeling of spatial relationships. Han et al. [29] proposed a spatiotemporal graph convolutional network for station-level metro passenger prediction. Yu et al. [30] developed the classical STGCN architecture by integrating graph convolution with temporal convolution. Chen et al. [31] combined graph convolution and bidirectional LSTM for metro ridership prediction, whereas Rong et al. [32] proposed a hierarchical GCN-LSTM framework for urban OD prediction.

Although deep learning approaches have achieved impressive prediction accuracy, they generally operate as deterministic nonlinear mapping models. The latent dynamic state is not explicitly modeled, and uncertainty propagation is usually absent. Consequently, model robustness may deteriorate under rapidly changing passenger demand or unexpected disruptions.

2.3. Tensor-Based Traffic Prediction

Because metro OD demand naturally possesses multi-dimensional spatial-temporal structures, tensor representations have attracted increasing attention in intelligent transportation research. Tan et al. [7] proposed a dynamic tensor completion method for short-term traffic prediction by exploiting low-rank tensor structures in traffic data. Chang et al. [9] developed a tensor Levenberg–Marquardt algorithm for multi-relational traffic prediction, demonstrating that tensor representations effectively preserve correlations among multiple traffic variables.

These studies verify the effectiveness of tensor modeling for traffic prediction. However, the computational cost of large-scale tensor operations still increases substantially as tensor dimensions grow, limiting their applicability in real-time metro OD prediction.

2.4. Tensor-Network Recursive Estimation

To alleviate the computational burden associated with high-dimensional tensors, tensor network representations have recently been introduced into recursive filtering algorithms.

Batselier and Wong [24,25] extended the tensor network Kalman filter to matrix-output Volterra system identification by representing the state vector and covariance matrix in low-rank tensor-network form. Their work demonstrated that tensor networks can significantly reduce computational complexity while maintaining estimation accuracy for nonlinear system identification problems.

Although this work provides an important computational foundation for recursive Bayesian estimation in high-dimensional systems, its primary objective is nonlinear MIMO Volterra system identification. The state variables, observation models, and estimation objectives are fundamentally different from those encountered in metro OD demand forecasting. In particular, metro passenger demand exhibits strong spatial coupling among OD pairs and continuously evolving travel patterns, requiring a forecasting-oriented state-space formulation rather than system parameter identification. Therefore, directly applying existing tensor-network Kalman filtering techniques to metro OD prediction remains challenging.

2.5. Research Gap and Motivation

To provide a comprehensive comparison between existing approaches and the proposed framework, Table 1 summarizes representative methods from different research categories in terms of research objectives, modeling techniques, advantages, limitations, and their differences from this work. This comparison reveals several important research gaps.

- (1) Statistical recursive estimation methods provide uncertainty quantification but cannot efficiently handle the extremely high-dimensional nonlinear state encountered in metro OD forecasting.
- (2) Deep learning approaches achieve high predictive accuracy but generally lack recursive Bayesian updating and explicit uncertainty propagation.
- (3) Existing tensor-based traffic prediction methods generally suffer from high computational complexity, which limits their applicability to large-scale real-time forecasting.
- (4) Existing tensor-network Kalman filtering methods were developed for nonlinear system identification rather than metro passenger demand prediction and therefore cannot directly exploit the unique spatiotemporal characteristics of metro OD demand.

To bridge these gaps, this paper proposes a TNEKF for short-term metro OD forecasting. Different from existing tensor-network Kalman filtering methods, the proposed approach formulates metro OD prediction as a recursive Bayesian forecasting problem. A Volterra-based nonlinear state-space model is developed to characterize passenger demand evolution, while low-rank tensor-network representations are introduced to efficiently implement recursive EKF estimation for large-scale metro OD systems.

Table 1. Comparison of representative methods and the proposed approach.

Reference	Research Objective	Method	Advantages	Limitations	Difference from This Work
Shekhar & Williams [10]	Traffic flow forecasting	ARIMA	Simple and interpretable	Linear assumption	No nonlinear OD dynamics modeling
Wang & Pageorgiou [13]	Traffic state estimation	EKF	Recursive Bayesian estimation	Low-dimensional state	No tensor representation
Ma et al. [26]	Metro ridership prediction	CNN-BiLSTM	Strong spatiotemporal learning	Deterministic prediction	No uncertainty modeling
Yu et al. [30]	Station passenger prediction	STGCN	Captures graph topology	Station-level prediction	No uncertainty propagation

Chen et al. [31], Rong et al. [32]	Urban OD prediction	LSTM-GCN, DBSTNet	Good OD prediction accuracy	No recursive Bayesian updating	No uncertainty prop- agation
Tan et al. [7]	Traffic prediction	Tensor comple- tion	Exploits tensor struc- ture	High computa- tional cost	Full rank
Chang et al. [9]	Multi-relational traffic prediction	Tensor LM	Tensor regression	High computa- tional cost	Full rank
Batselier et al. [24,25]	MIMO Volterra identification	Tensor Network Kalman Filter	Low-rank recursive estimation	System identifica- tion	Our work reformu- lates the framework for metro OD fore- casting

3. Tensor Network Preliminaries

Before presenting the proposed TNEKF, this section briefly introduces the tensor network notation required throughout the paper. Only the tensor operations directly involved in the proposed forecasting algorithm are presented. More comprehensive treatments of tensor networks can be found in [24,25].

In this paper, bold calligraphic letters (e.g., $\mathcal{A}$, $\mathcal{B}$) are used to denote tensors, which are multi-dimensional arrays and can be viewed as higher-order generalizations of vectors and matrices. Matrices, vectors, and scalars are represented using standard notation.

A d -th order tensor (or equivalently, a d -dimensional tensor) is denoted as $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \dots \times n_d}$, where each element is indexed by $\mathcal{A}(i_1, i_2, \dots, i_d)$ or equivalently $a_{i_1 i_2 \dots i_d}$ with $i_k \in [1, n_k]$. The identity matrix of order n is denoted by $\mathbf{I}_n$, and the all-ones vector of length n is denoted by $\mathbf{1}_n$.

To map a multi-index $n_1, n_2, \dots, n_d$ to a single linear index $[n_1 n_2 \dots n_d]$, we adopt the following indexing rule:

$$[n_1 n_2 \dots n_d] = n_1 + \sum_{k=2}^d (n_k - 1) \prod_{l=1}^{k-1} N_l \quad (1)$$

where $1 \leq n_k \leq N_k$ ($k = 1, \dots, d$), and the resulting linear index satisfies $1 \leq [n_1 n_2 \dots n_d] \leq \prod_{k=1}^d N_k$.

Given a vector $\mathbf{x} \in \mathbb{R}^{N^d}$, it can be equivalently reshaped into a d -th order tensor $\mathcal{X} \in \mathbb{R}^{N \times N \times \dots \times N}$. The MPS representation expresses this high-dimensional tensor as a sequence of third-order tensors. Specifically, each element $\mathbf{x}([n_1 n_2 \dots n_d])$ (or equivalently $\mathcal{X}(n_1, n_2, \dots, n_d)$) can be decomposed as

$$\mathcal{X}(n_1, n_2, \dots, n_d) = \sum_{r_2, \dots, r_d=1}^{R_2, \dots, R_d} \mathbf{X}^{(1)}(1, n_1, r_2) \mathbf{X}^{(2)}(r_2, n_2, r_3) \dots \mathbf{X}^{(d)}(r_d, n_d, 1) \quad (2)$$

where $\mathbf{X}^{(k)} \in \mathbb{R}^{R_k \times N \times R_{k+1}}$ are the MPS core tensors, and $R_2, R_3, \dots, R_d$ are referred to as the MPS ranks, with boundary conditions $R_1 = R_{d+1} = 1$. The MPS representation stores the original high-dimensional tensor in terms of these low-order core tensors. If all ranks are bounded by R , the total storage complexity is $O(dNR^2)$, where d is the tensor order and N is the size of each mode. This representation effectively alleviates the curse of dimensionality associated with exponentially growing vector or tensor representations. As long as the MPS rank R remains moderate, the storage cost is significantly reduced, enabling efficient representation and computation in high-dimensional spaces.

Given a matrix $\mathbf{A} \in \mathbb{R}^{N^d \times M^d}$, it can be equivalently reshaped into a $2d$ -th order tensor $\mathcal{A} \in \mathbb{R}^{N \times \dots \times N \times M \times \dots \times M}$. The Matrix Product Operator MPO representation expresses this tensor as a sequence of d fourth-order core tensors

$$\mathbf{A}^{(k)} \in \mathbb{R}^{R_k \times N \times M \times R_{k+1}}, k = 1, \dots, d \quad (3)$$

where R_k are the MPO ranks with boundary conditions $R_1 = R_{d+1} = 1$. The $[n_1 \dots n_d], [m_1 \dots m_d]$ -th entry of the matrix can be written as

$$\mathcal{A}([n_1 \cdots n_d], [m_1 \cdots m_d]) = \sum_{r_1, \dots, r_d} \prod_{k=1}^d \mathbf{A}^{(k)}(r_k, n_k, m_k, r_{k+1}) \quad (4)$$

where $1 \leq n_k \leq N$ and $1 \leq m_k \leq M, k = 1, \dots, d$.

4. Tensor Network Extended Kalman Filter for Metro OD Forecasting

4.1. Dynamic State Representation of Metro OD Demand

Short-term metro OD demand forecasting aims to predict passenger movements between all origin–destination station pairs during the next time interval based on historical observations. The proposed framework is developed for an urban metro network, which is represented as a directed graph consisting of N_s stations connected by rail lines. During each observation interval, passenger trips recorded by the Automatic Fare Collection (AFC) system are aggregated into a network-wide OD demand matrix. It is assumed that the AFC data have been preprocessed into consecutive time-dependent OD matrices, while the physical metro topology remains unchanged throughout the forecasting horizon. Under these conditions, passenger demand evolves continuously over time owing to variations in travel behavior, resulting in strong temporal dependence as well as spatial correlations among different OD pairs.

From a modeling perspective, the metro OD forecasting problem can therefore be regarded as the evolution of a high-dimensional nonlinear dynamic system. Rather than directly learning a deterministic mapping from historical observations to future demand, the proposed framework introduces a latent state to describe the time-varying characteristics governing passenger demand evolution. This latent-state representation enables recursive prediction while explicitly accounting for stochastic demand variations and model uncertainty.

Consider a metro network consisting of N_s stations. The passenger demand during time interval t is represented by the OD matrix $\mathbf{D}_t \in \mathbb{R}^{N_s \times N_s}$, where the (i, j) -th element denotes the number of passengers travelling from origin station i to destination station j . Given the passenger demand observed during the previous p time intervals, the objective is to predict the OD demand matrix at the next interval. To preserve the temporal dependency of historical observations, the OD matrices are seen as a third-order tensor

$$\mathcal{D}_t = \{\mathbf{D}_t, \mathbf{D}_{t-1}, \dots, \mathbf{D}_{t-p+1}\} \in \mathbb{R}^{N_s \times N_s \times p} \quad (5)$$

whose three modes correspond to origin stations, destination stations, and historical time intervals, respectively.

However, for recursive forecasting, the historical tensor is rearranged into a historical demand vector

$$\mathbf{u}(t) = [1, \mathbf{d}^T(t), \mathbf{d}^T(t-1), \dots, \mathbf{d}^T(t-p+1)]^T \in \mathbb{R}^{mp+1} \quad (6)$$

where

$$\mathbf{d}(t) = \text{vec}(\mathbf{D}_t) \in \mathbb{R}^m \quad (7)$$

and $m = N_s \times N_s$ denotes the total number of OD pairs. The vectorization procedure is illustrated in Figure 1. This representation provides a unified description of historical network-wide passenger demand while maintaining the correspondence between temporal observations and OD relationships.

The conventional regression models directly establish a mapping between historical observations and future demand. Different from this, the proposed framework introduces a latent parameter state to characterize the time-varying nonlinear relationship between historical OD demand and future passenger demand. The parameter state is recursively

updated over time so that the state-space prediction model can gradually adapt to continuously changing travel patterns.

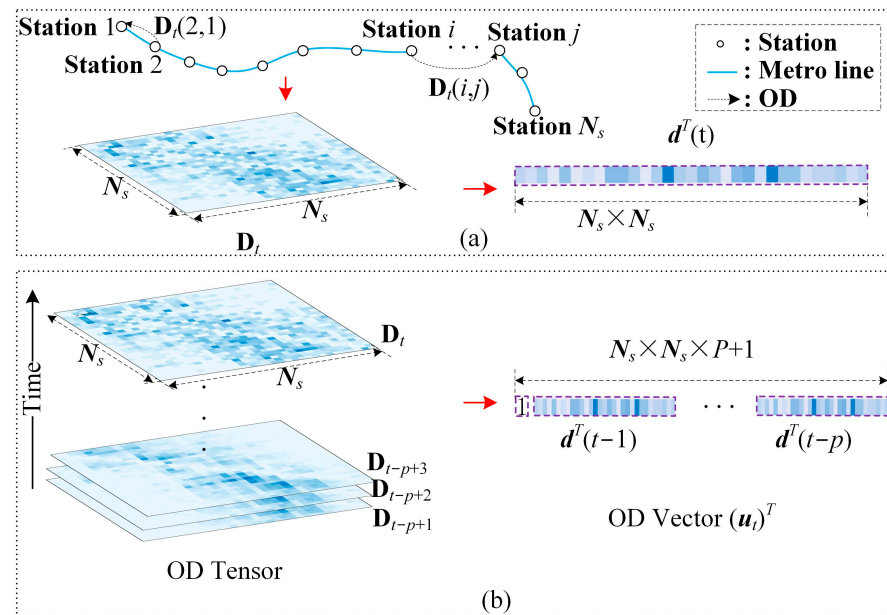


Figure 1. The vectorization procedure. (a) Convert the OD matrix $\mathbf{D}_t$ to demand vector $\mathbf{d}(t)$. (b) Convert the historical OD tensor to historical demand vector.

Let $\boldsymbol{\theta}_t \in \mathbb{R}^L$ denote the latent state vector describing the dynamic characteristics of metro OD demand, where L is the total number of unknown model parameters. For a time-varying discrete-time MIMO Volterra model with l outputs, order d , and memory length p , the state dimension is $L = lN^d$, where $N = pm + 1$. Here, p denotes the number of historical time intervals included in the input vector for capturing temporal dependencies of OD demand. The state evolution is modeled by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mathbf{w}_t \quad (8)$$

where $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ represents zero-mean Gaussian process noise describing uncertainties in the temporal evolution of latent demand characteristics and unmodeled variations in passenger travel patterns.

The future OD demand is predicted through a nonlinear mapping between the latent demand state and the historical demand vector,

$$\hat{\mathbf{d}}_{t+1} = f(\mathbf{u}_t, \boldsymbol{\theta}_t) + \mathbf{v}_t \quad (9)$$

where $\hat{\mathbf{d}}_{t+1} \in \mathbb{R}^l$ is the predicted OD demand vector ($l = m$), $f(\cdot)$ denotes a nonlinear forecasting function, and $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ represents observation uncertainty arising from passenger demand fluctuations.

The Volterra expansion [24] is employed here solely as a flexible nonlinear state-space prediction model capable of approximating complex spatiotemporal dependencies in metro passenger demand. Accordingly, the nonlinear mapping in (9) can be expressed as

$$\hat{\mathbf{d}}_{t+1} = \mathbf{H}(\mathbf{u}_t)\boldsymbol{\theta}_t + \mathbf{v}_t \quad (10)$$

where $\mathbf{H}(\mathbf{u}_t)$ is the nonlinear regression matrix constructed from the Volterra expansion of the historical demand vector. Each row of $\mathbf{H}(\mathbf{u}_t)$ corresponds to one OD pair, while different rows share the same historical demand representation, enabling the model to

capture both nonlinear temporal evolution and cross-OD interactions within a unified framework.

The process noise and observation noise are assumed to be mutually independent Gaussian white noise processes,

$$E \begin{bmatrix} \mathbf{w}_t \\ \mathbf{v}_t \end{bmatrix} \begin{bmatrix} \mathbf{w}_s^T & \mathbf{v}_s^T \end{bmatrix} = \begin{bmatrix} \mathbf{Q} & 0 \\ 0 & \mathbf{R} \end{bmatrix} \delta_{ts} \quad (11)$$

The above formulation establishes metro OD forecasting as a nonlinear dynamic state-space model, where the latent demand state captures the temporal evolution of hidden demand characteristics. This representation provides the foundation for the recursive Bayesian estimation framework introduced in the following subsection.

4.2. Recursive Demand Prediction Using Extended Kalman Filtering

The dynamic demand model established in the previous subsection provides a nonlinear representation of metro passenger demand evolution. Since both the latent demand state and passenger observations are affected by stochastic disturbances, deterministic forecasting may accumulate prediction errors over time. To improve forecasting robustness, the proposed framework formulates short-term OD demand prediction as a recursive Bayesian estimation problem.

At each prediction interval, the latent demand state is represented by a Gaussian posterior distribution

$$p(\boldsymbol{\theta}_t | \mathbf{d}_{1:t}) = \mathcal{N}(\mathbf{m}_t, \mathbf{P}_t), \quad (12)$$

where $\mathbf{m}_t \in \mathbb{R}^{LN^d}$ denotes the posterior estimate of the latent demand state after assimilating all historical observations up to time t , while $\mathbf{P}_t \in \mathbb{R}^{LN^d \times LN^d}$ quantifies the corresponding estimation uncertainty.

Given the posterior estimate at time t , the latent demand state is first propagated to the next prediction interval through the state transition model. Since the proposed demand evolution model assumes a random-walk process for the latent state, the prior state estimate is obtained as

$$\mathbf{m}_{t+1|t} = \mathbf{m}_t \quad (13)$$

and the associated covariance is predicted by

$$\mathbf{P}_{t+1|t} = \mathbf{P}_t + \mathbf{Q} \quad (14)$$

where $\mathbf{Q}$ is the covariance matrix of the process noise.

The predicted latent demand state is subsequently transformed into the forecasted OD demand using the nonlinear demand prediction model

$$\hat{\mathbf{d}}_{t+1|t} = f(\mathbf{u}_t, \mathbf{m}_{t+1|t}) \quad (15)$$

Equation (15) provides the one-step-ahead prediction of all OD pairs before incorporating newly observed passenger demand. After the passenger demand at time $t + 1$ becomes available, the prediction residual is computed as

$$\mathbf{r}_{t+1} = \mathbf{d}_{t+1} - \hat{\mathbf{d}}_{t+1|t} \quad (16)$$

where $\mathbf{d}_{t+1}$ is the recorded OD demand vector.

The prediction residual reflects the discrepancy between the predicted and observed passenger demand and is used to update the latent demand state recursively. According to Bayesian estimation theory, the posterior state estimate is obtained by

$$\mathbf{m}_{t+1} = \mathbf{m}_{t+1|t} + \mathbf{K}_{t+1} \mathbf{r}_{t+1} \quad (17)$$

where $\mathbf{K}_{t+1}$ denotes the Kalman gain.

The corresponding posterior covariance is updated as

$$\mathbf{P}_{t+1} = (\mathbf{I} - \mathbf{K}_{t+1}\mathbf{H}_{t+1})\mathbf{P}_{t+1|t} \quad (18)$$

where $\mathbf{H}_{t+1} \in \mathbb{R}^{l \times lN^d}$ is the Jacobian matrix of the nonlinear forecasting function with respect to the latent state. The prior prediction generated from historical passenger demand is continuously refined whenever newly observed OD demand becomes available, thereby preventing long-term error accumulation and improving forecasting robustness under dynamically changing travel conditions.

The Kalman gain determines the balance between model prediction and newly observed passenger demand. It is computed as

$$\mathbf{K}_{t+1} = \mathbf{P}_{t+1|t}\mathbf{H}_{t+1}^T(\mathbf{H}_{t+1}\mathbf{P}_{t+1|t}\mathbf{H}_{t+1}^T + \mathbf{R})^{-1} \in \mathbb{R}^{lN^d \times l} \quad (19)$$

where $\mathbf{R}$ denotes the covariance matrix of observation uncertainty.

Finally, the posterior state estimate $\mathbf{m}_{t+1}$ is propagated to the next prediction interval, enabling recursive short-term forecasting of metro OD demand as new passenger observations continuously arrive.

4.3. Efficient Jacobian Construction for Recursive Metro OD Forecasting

The recursive EKF framework presented in the previous subsection requires the Jacobian matrix of the nonlinear forecasting function for linearization around the predicted state. Since the proposed demand prediction model is constructed from a multi-input multi-output Volterra representation [24], the Jacobian can be derived analytically. According to the nonlinear demand prediction models (10) and (15), the corresponding one-step prediction is

$$\hat{\mathbf{d}}_{t+1|t} = \mathbf{H}(\mathbf{u}_t)\mathbf{m}_{t+1|t} \quad (20)$$

Since the observation model is linear with respect to the latent parameter vector, the Jacobian is directly obtained as

$$\mathbf{H}_{t+1} = \left. \frac{\partial f(\mathbf{u}_t, \theta)}{\partial \theta} \right|_{\theta=\mathbf{m}_{t+1|t}} = \mathbf{H}(\mathbf{u}_t) \quad (21)$$

For the metro OD forecasting problem, the historical demand vector is shared by all OD pairs, whereas each row of the Volterra regression matrix corresponds to one destination output. Therefore, the Jacobian exhibits a highly structured block form,

$$\mathbf{H}_{t+1} = \begin{bmatrix} \mathbf{h}_1(\mathbf{u}_t) \\ \mathbf{h}_2(\mathbf{u}_t) \\ \vdots \\ \mathbf{h}_l(\mathbf{u}_t) \end{bmatrix} \quad (22)$$

where $\mathbf{h}_i(\mathbf{u}_t) \in \mathbb{R}^{1 \times L}$ denotes the Volterra regression vector associated with the i -th OD pair.

Unlike a general nonlinear observation model, the Jacobian rows share the same historical demand representation and differ only in their corresponding Volterra coefficients. Therefore, the nonlinear basis needs to be evaluated only once for all OD outputs, significantly reducing the computational cost of Jacobian construction. Moreover, the resulting structured Jacobian naturally facilitates the low-rank tensor representation introduced in the following subsection.

4.4. Low-Rank Tensor Network Implementation

Although the analytical Jacobian derived in Section 4.3 avoids the expensive numerical differentiation required in conventional EKF implementations, the recursive estimation still involves high-dimensional latent states and covariance matrices. In the proposed Volterra-based OD forecasting model, the dimension of the latent parameter vector is determined by the number of OD pairs, memory length, output number, and nonlinear expansion order. Therefore, for large-scale metro networks, directly storing and updating the state vector and covariance matrix becomes the dominant computational bottleneck.

To address this issue, a low-rank tensor network representation is introduced to reformulate the EKF recursion. The key idea is to exploit the inherent low-dimensional structure of the Volterra parameter space and represent high-dimensional states and operators using tensor networks [24] rather than explicit vector and matrix forms. Specifically, the posterior state vector is represented by an MPS, while the covariance matrix and linear transformation operators are represented by MPOs. This low-rank representation preserves the recursive Bayesian estimation mechanism while avoiding the storage and computation of full-dimensional quantities.

Instead of treating $\mathbf{m}_t \in \mathbb{R}^{lN^d}$ as a high-dimensional vector, it is reshaped into a high-order tensor $\mathcal{M}_t \in \mathbb{R}^{N \times N \times \dots \times N \times l}$ and decomposed into an MPS form,

$$\mathcal{M}_t = \mathbf{M}^{(1)} \mathbf{M}^{(2)} \dots \mathbf{M}^{(d+1)} \quad (23)$$

where each tensor core is defined as

$$\mathbf{M}^{(k)} \in \mathbb{R}^{r_{k-1} \times n_k \times r_k} \quad (24)$$

where r_k denotes the tensor rank controlling the information coupling between adjacent tensor modes. If $1 \leq k \leq d$, $n_k = N$ and if $k = d + 1$, $n_k = l$. If $k = 1$, $r_{k-1} = 1$, and if $k = d + 1$, $r_k = 1$. When the latent parameter tensor exhibits strong correlations, the required tensor ranks remain small, allowing the original high-dimensional state to be represented with significantly fewer parameters.

Similarly, the posterior covariance matrix $\mathbf{P}_t \in \mathbb{R}^{lN^d \times lN^d}$ is reformulated as an MPO,

$$\mathcal{P}_t = \mathbf{P}^{(1)} \mathbf{P}^{(2)} \dots \mathbf{P}^{(d+1)} \quad (25)$$

where each MPO core satisfies

$$\mathbf{P}^{(k)} \in \mathbb{R}^{r_{k-1} \times n_k \times n_k \times r_k} \quad (26)$$

Based on these representations, the EKF prediction and update operations are performed directly in the tensor-network domain, and (13) corresponds to a direct propagation of the MPS representation,

$$\mathcal{M}_{t+1|t} = \mathcal{M}_t \quad (27)$$

while the covariance prediction implemented through MPO addition is written as

$$\mathcal{P}_{t+1|t} = \mathcal{P}_t \oplus \mathbf{Q} \quad (28)$$

During the measurement update, the analytical Jacobian $\mathbf{H}_{t+1} \in \mathbb{R}^{l \times lN^d}$ obtained in (22) is also expressed in an MPS,

$$\mathcal{H}_t = \mathbf{H}^{(1)} \mathbf{H}^{(2)} \dots \mathbf{H}^{(d+2)} \quad (29)$$

where each tensor core is defined as

$$\mathbf{H}^{(k)} \in \mathbb{R}^{r_{k-1} \times n_k \times r_k} \quad (30)$$

If $1 \leq k \leq d$, $n_k = N$, and if $k = d + 1$ or $d + 2$, $n_k = l$. If $k = 1$, $r_{k-1} = 1$, and if $k = d + 2$, $r_k = 1$.

Therefore, the Kalman gain

$$\kappa_{t+1} = \mathcal{P}_{t+1|t} \otimes \mathcal{H}_{t+1}^T \otimes (\mathcal{H}_{t+1} \otimes \mathcal{P}_{t+1|t} \otimes \mathcal{H}_{t+1}^T \oplus \mathbf{R})^{-1} \quad (31)$$

is realized through a sequence of tensor contractions among MPOs, without explicitly assembling the full covariance and Jacobian matrices.

After obtaining the Kalman gain, the posterior state update

$$\mathcal{M}_{t+1} = \mathcal{M}_{t+1|t} \oplus \kappa_{t+1} \otimes r_{t+1} \quad (32)$$

is performed as an MPS update, where the residual information is injected through tensor contractions.

Likewise, the covariance update implemented using MPO multiplication and addition operations is written as

$$\mathcal{P}_{t+1} = \mathcal{P}_{t+1|t} \oplus (-\kappa_{t+1} \otimes \mathcal{H}_{t+1} \otimes \mathcal{P}_{t+1|t}) \quad (33)$$

During the recursive process, tensor contractions may gradually increase the tensor ranks, leading to unnecessary computational growth. To maintain a bounded computational complexity, singular value decomposition (SVD)-based truncation is applied after each tensor operation. Consequently, the tensor ranks remain controlled throughout the forecasting process.

The proposed low-rank tensor network implementation provides two essential advantages for large-scale metro OD forecasting. First, it avoids explicit storage of the full Volterra parameter vector and covariance matrix, significantly reducing memory requirements for high-dimensional OD systems. Second, the EKF recursion is converted from dense matrix operations into local tensor contractions [24], where the computational complexity is governed primarily by the tensor ranks and tensor order rather than the original state dimension. Therefore, the proposed tensor-network EKF framework can efficiently handle metro networks with a large number of OD pairs and high-order nonlinear demand representations while maintaining the recursive adaptability of Bayesian forecasting. As illustrated in Figure 2, the prediction process of our algorithm is presented.

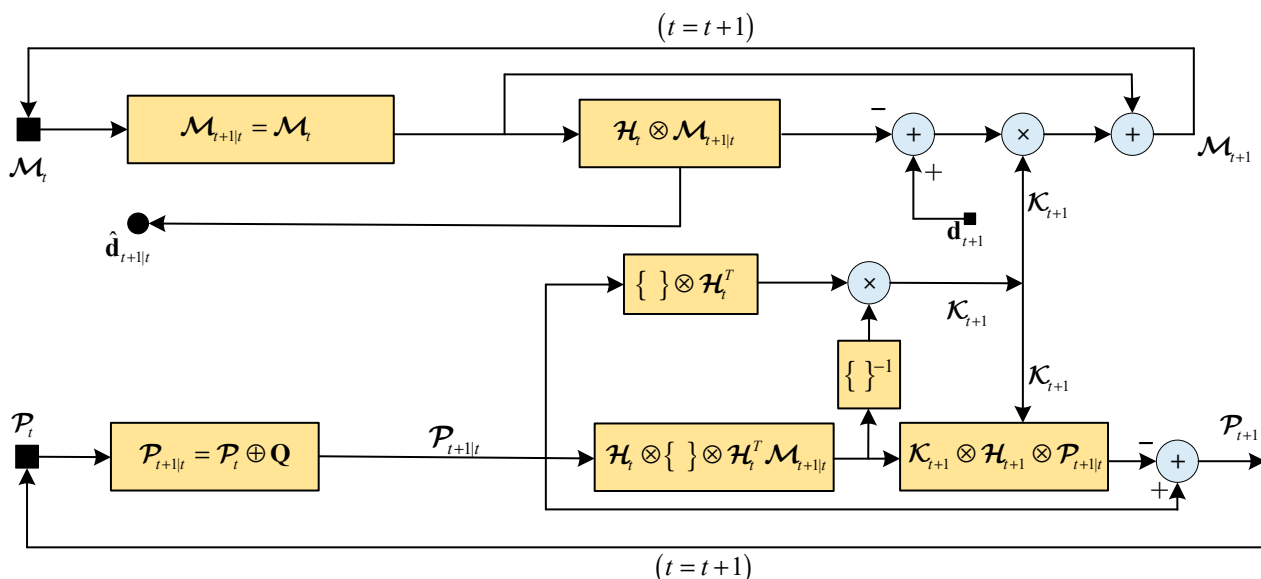


Figure 2. Diagram of the prediction process of the TNEKF algorithm. Yellow boxes represent computational modules, and blue circles denote tensor-network operation nodes.

5. Numerical Results

In this section, we present experimental results on real-world datasets to validate the effectiveness of the proposed TNEKF model. All experiments were conducted on a personal computer equipped with 64 GB RAM, an Intel® Core™ i7-13700KF CPU (3.40 GHz).

5.1. Dataset

To validate the proposed method for short-term metro demand forecasting, smart card transaction data from the Hangzhou metro system are used in this study. As illustrated in Figure 3, the experimental dataset is constructed from a subset of 32 stations selected from Metro Lines 1 and 4. Line 2 is also shown to illustrate the overall network connectivity of the study area. These stations form an interconnected subnetwork with multiple transfer relationships and diverse passenger flow characteristics, which is suitable for evaluating complex spatiotemporal demand patterns.

The dataset (<https://tianchi.aliyun.com/dataset/21904#1>, accessed on 30 May 2026) spans 25 consecutive days, from 1 January to 25 January 2019, and records metro operations between 6:00 and 22:00 each day. On average, approximately 235,000 passenger trips are recorded daily. Each trip entry contains essential information, including timestamp, line identifier, station identifier, travel status (entry or exit), and an anonymized user ID.

To prepare the data for modeling, a three-stage preprocessing procedure is implemented. First, data quality control is performed to eliminate incomplete, duplicated, and logically inconsistent records, such as trips with invalid station indices or incorrect temporal ordering. Second, full passenger trajectories are reconstructed by pairing entry and exit records based on user identifiers and status information. Finally, the cleaned trip data are aggregated into fixed 10 min intervals to produce time-indexed OD demand matrices for subsequent analysis.

After preprocessing, the dataset is organized into a three-dimensional tensor $\mathcal{X}$, where each slice $X_t \in \mathbb{R}^{32 \times 32}$ denotes the OD demand matrix at time interval t . For model evaluation and validation, the dataset is chronologically divided into three subsets. The first 16 days are used for training, the following 4 days serve as the validation set for hyperparameter tuning and model selection, and the final 5 days are reserved for out-of-sample testing to evaluate predictive performance. The forecasting task considered in this study is one-step-ahead network-wide OD demand prediction under normal metro operating conditions. Given the OD demand during the previous six 10 min intervals, the objective is to predict the OD demand matrix for the next 10 min interval.

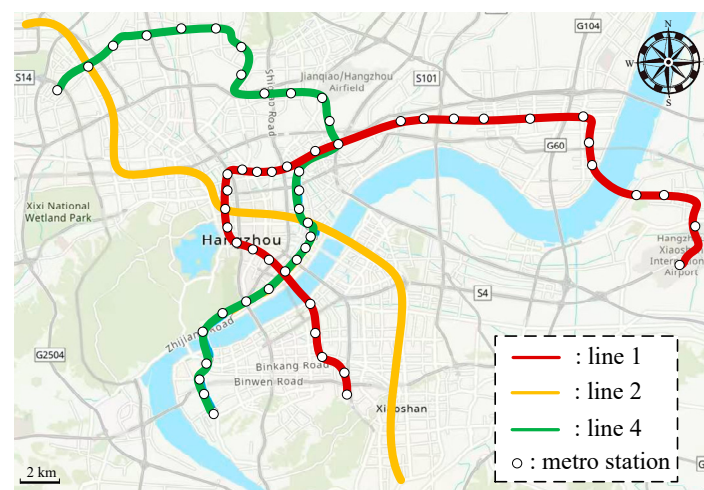


Figure 3. Geographic layout of the Hangzhou metro network considered in this study, including Metro Lines 1, 2, and 4. White circles denote metro stations.

5.2. Performance Measures

To quantitatively evaluate the predictive performance of the proposed models, three widely used metrics are adopted, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Weighted Mean Absolute Percentage Error (wMAPE). Their mathematical definitions are provided as follows:

$$MAE = \frac{1}{T} \sum_{t=1}^T \frac{1}{N_s^2} \sum_{j=1, j \neq i}^{N_s} \sum_{i=1}^{N_s} |\hat{x}_{pred}(i, j, t) - x(i, j, t)| \quad (34)$$

$$REMS = \frac{1}{T} \sum_{t=1}^T \sqrt{\frac{1}{N_s^2} \sum_{j=1, j \neq i}^{N_s} \sum_{i=1}^{N_s} (\hat{x}_{pred}(i, j, t) - x(i, j, t))^2} \quad (35)$$

$$wMAPE = \frac{1}{T} \sum_{t=1}^T \frac{\sum_{j=1, j \neq i}^{N_s} \sum_{i=1}^{N_s} |\hat{x}_{pred}(i, j, t) - x(i, j, t)|}{\sum_{j=1, j \neq i}^{N_s} \sum_{i=1}^{N_s} x(i, j, t)} \quad (36)$$

where $\hat{x}_{pred}(i, j, t)$ and $x(i, j, t)$ denote the predicted and actual passenger demand, respectively, from origin station i to destination station j at time interval t . N_s represents the number of stations, and T is the total number of time intervals in the test dataset.

5.3. Experimental Settings

Based on the constructed three-dimensional tensor, the training and test datasets are further organized into input–output sample pairs for the TNEKF model, as illustrated in Figure 4. Specifically, each input sample consists of OD demand matrices from 6 consecutive time intervals, which are reshaped into a matrix of size 1024×6 , where $1024 = 32 \times 32$ corresponds to all OD pairs. The corresponding output sample is defined as the OD demand matrix at the subsequent time interval, represented as a vector of size 1024×1 .

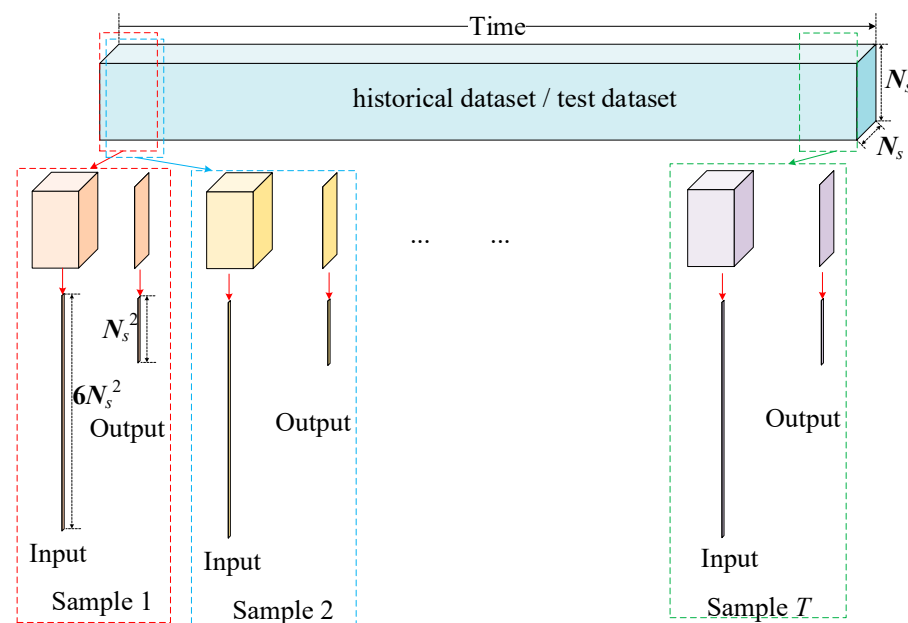


Figure 4. Three metro lines of Hangzhou. Conversion of historical and test datasets into input–output samples for the TNEKF model.

The configurations of all benchmark and proposed methods are summarized as follows.

ARIMA: An individual ARIMA model is developed for each OD pair. The model orders are selected within the range of 1–6 based on validation performance on the training dataset.

EKF: A standalone extended Kalman filter (EKF) is implemented for each OD pair independently. This setting does not account for spatial interactions among different OD flows.

STGCN [30]: A spatiotemporal graph convolutional network is employed to capture spatial dependencies through graph-based convolutions and temporal dynamics via sequential convolution operations.

LSTM-GCN [31]: This hybrid framework integrates graph convolutional networks with long short-term memory (LSTM) units to jointly model spatial and temporal correlations.

DBSTNet [32]: A deep bidirectional spatiotemporal network is adopted to characterize complex traffic patterns.

Minor adjustments to the hyperparameters and network architectures of STGCN, LSTM-GCN, and DBSTNet were conducted using the validation set to ensure fair and consistent performance comparison across all benchmark models.

TNEKF: The proposed TNEKF model is based on a Volterra system with memory length $p = 6$, input dimension $m = 1024$, output dimension $l = 1024$, and system order $d = 1$. To control rank growth in tensor operations, truncated singular value decomposition (SVD) with a threshold of 10^{-4} is applied. The effective input dimension is $N = m \times p + 1 = 6145$. In the absence of tensor network representation, a conventional EKF with a full covariance matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$ would incur a storage cost on the order of 1011 elements (approximately 400 billion), rendering it computationally infeasible.

The experimental procedure consists of three steps. First, AFC records are processed to construct sequential OD demand matrices and forecasting samples. Second, all benchmark models and the proposed TNEKF are trained and validated using identical data partitions. Finally, one-step-ahead prediction is conducted on the testing dataset, and performance is evaluated using MAE, RMSE, and wMAPE.

The selected baselines cover representative forecasting paradigms, including statistical methods (ARIMA), recursive Bayesian estimation (EKF), graph-based spatiotemporal models (STGCN and LSTM-GCN), and deep spatiotemporal networks (DBSTNet), enabling comprehensive evaluation of TNEKF against both model-driven and data-driven approaches.

5.4. Comparative Experiments Under Normal Operating Conditions

Figure 5 presents the OD passenger flow prediction results for several representative OD pairs, including Station 15–Station 8, Station 8–Station 19, Station 19–Station 15, and Station 15–Station 10. Although the four OD pairs exhibit different passenger demand patterns, similar observations can be made across all subfigures. The results show that all methods are able to capture the overall temporal variation patterns of passenger demand. However, compared with the conventional EKF and the deep learning-based DBSTNet model, the proposed TNEKF achieves consistently higher prediction accuracy, particularly during peak and low-demand periods where demand fluctuations are more pronounced. This improvement can be attributed to the fact that the conventional EKF treats each OD pair independently and lacks the ability to capture spatial dependencies, while DBSTNet, although capable of modeling spatiotemporal features, relies purely on data-driven learning. In contrast, TNEKF integrates tensor-based spatial representations with recursive state estimation, enabling a more effective characterization of the coupled spatiotemporal dynamics of metro OD demand, thereby enhancing both prediction accuracy and robustness.

To further evaluate the prediction performance of different models at the network level, Figures 6 and 7 illustrate the estimated OD demand distributions across all 32 stations during a representative peak period (time interval 105) and an off-peak period (time interval 196), respectively. In Figure 6, the four subfigures compare the ground truth and

the prediction results produced by different methods under the same peak-period demand scenario. A consistent trend can be observed across all prediction maps that TNEKF achieves consistently higher prediction accuracy. Similar observations can be obtained during the off-peak period, indicating improved accuracy and robustness.

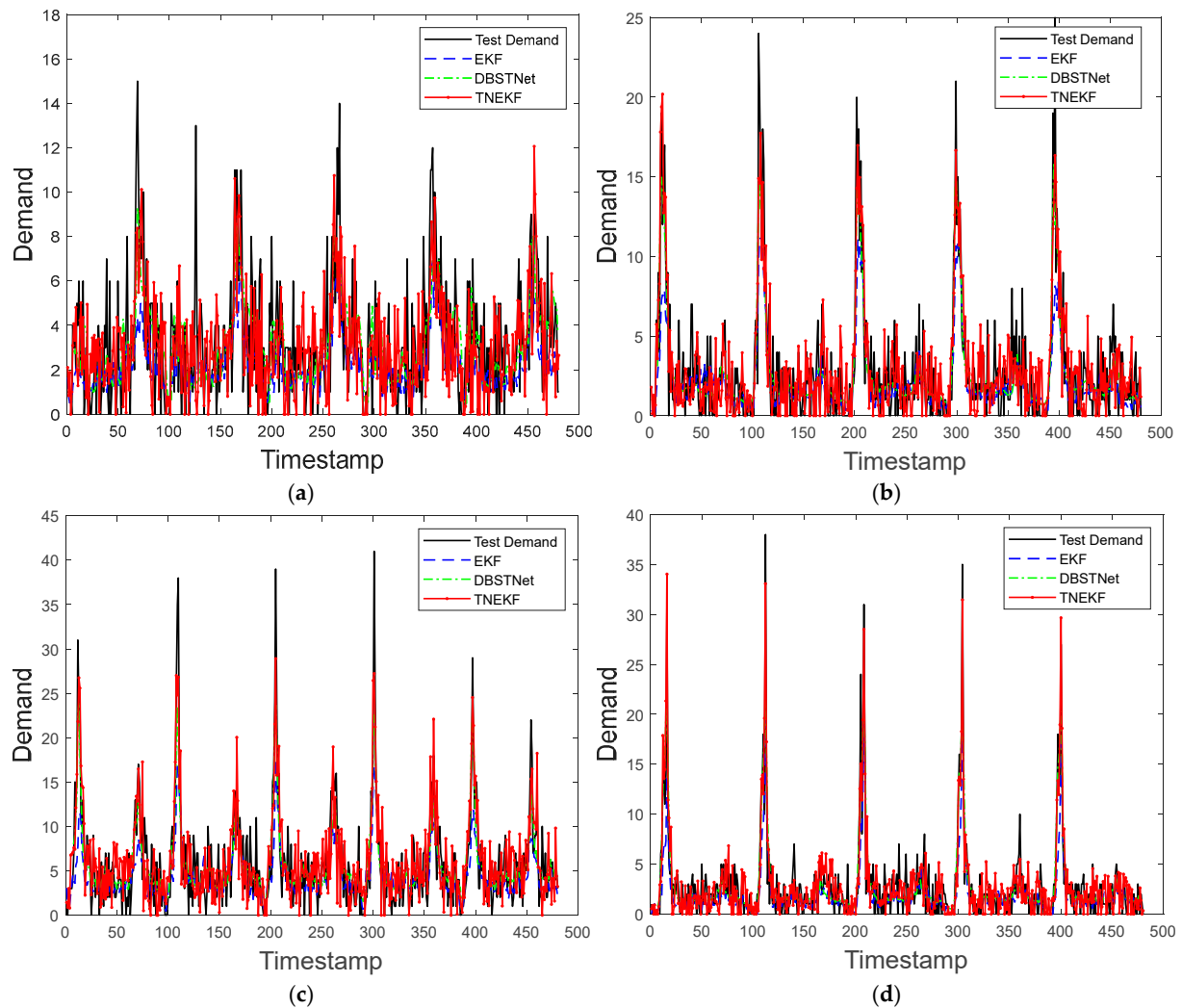
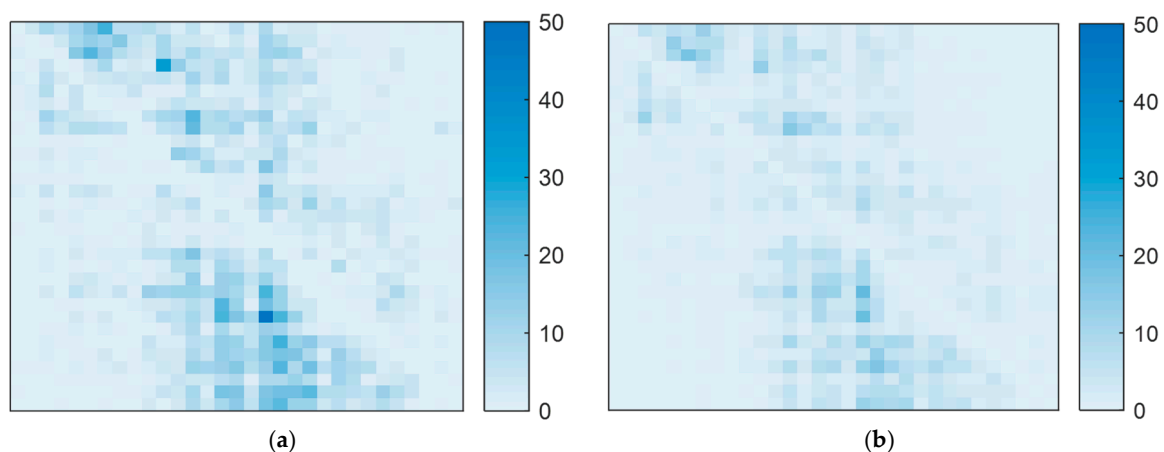


Figure 5. OD passenger demand forecasting results for representative OD pairs. (a) OD pair: Station 15-Station 8. (b) OD pair: Station 8-Station 19. (c) OD pair: Station 19-Station 15. (d) OD pair: Station 15-Station 10.



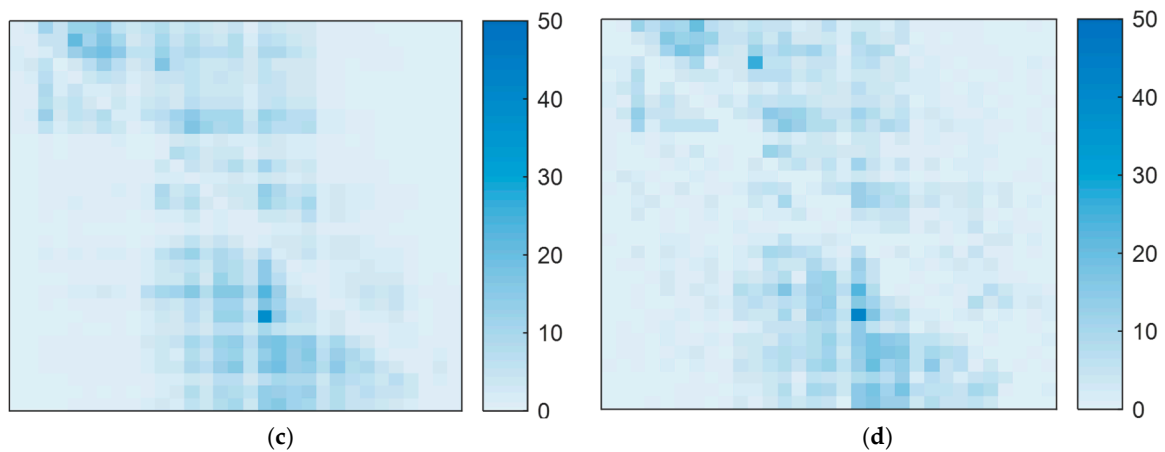


Figure 6. Comparison of OD demand matrix prediction results during a peak period (time interval 105). (a) Ground-truth demand distribution. (b) Prediction results of the EKF model. (c) Prediction results of the DBSTNet model. (d) Prediction results of the TNEKF model.

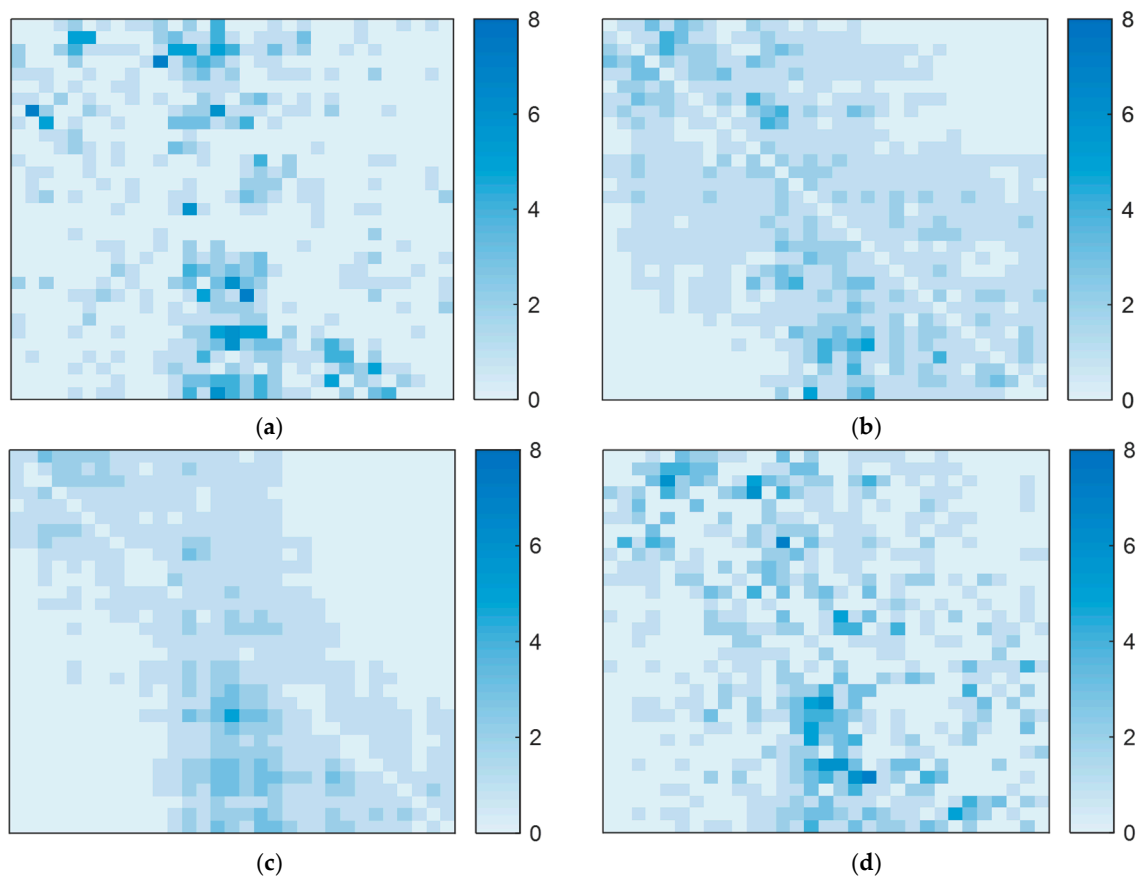


Figure 7. Comparison of OD demand matrix prediction results during an off-peak period (time interval 196). (a) Ground-truth demand distribution. (b) Prediction results of the EKF model. (c) Prediction results of the DBSTNet model. (d) Prediction results of the TNEKF model.

To further evaluate the prediction performance from the perspective of station-level passenger flows, the OD demand matrices were aggregated according to departure stations to obtain the total outbound passenger demand at each station. Figure 8 first compares the predicted total departure passenger flows at Stations 8, 16, 19, and 31. The results show that TNEKF achieves consistently higher prediction accuracy across different stations, although the stations exhibit different demand levels.

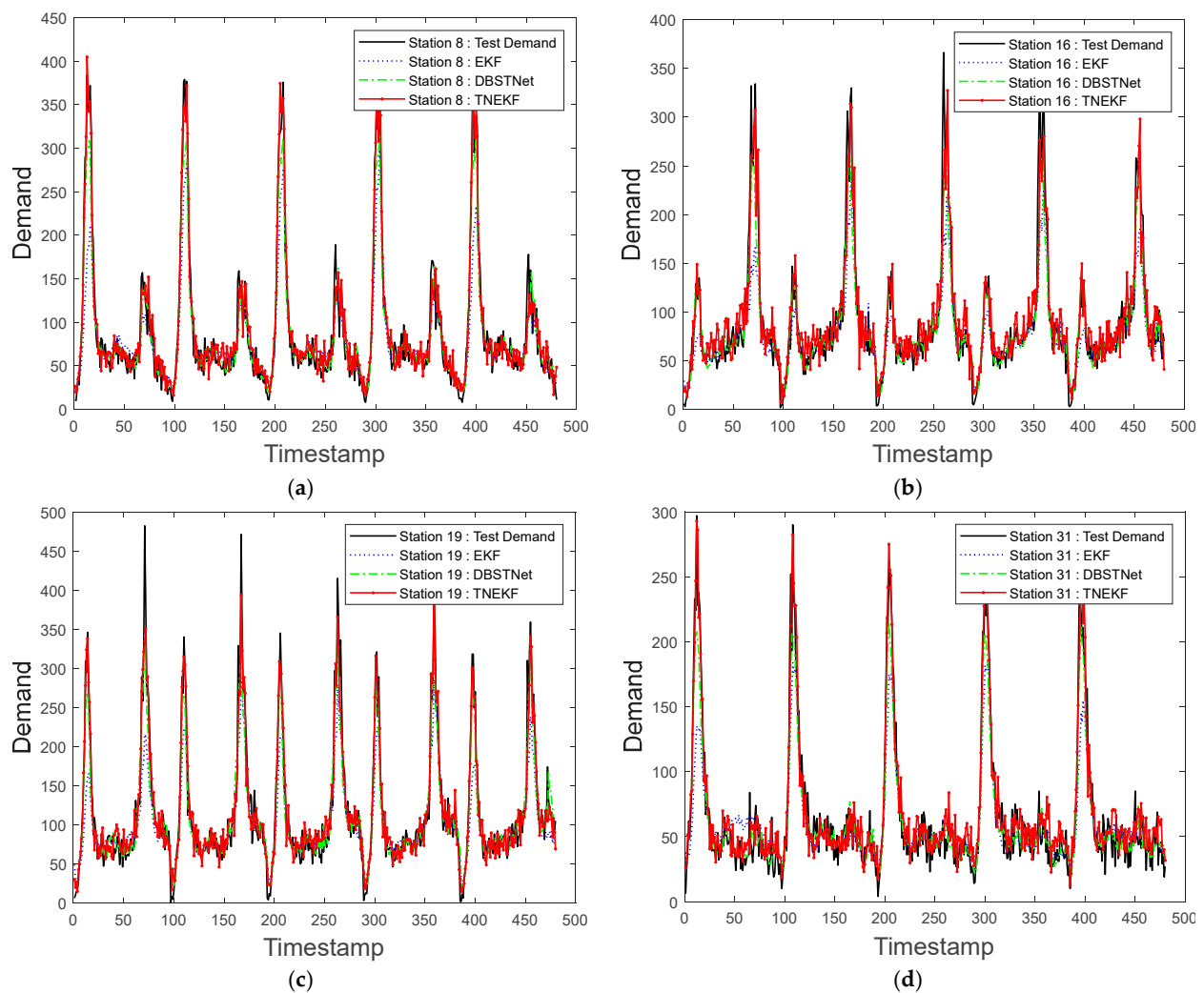


Figure 8. Prediction results of total departure passenger flows for representative stations. (a) Total departure passenger flow at Station 8. (b) Total departure passenger flow at Station 16. (c) Total departure passenger flow at Station 19. (d) Total departure passenger flow at Station 31.

Furthermore, Figure 9 presents the prediction results of total departure passenger flows for all stations over a one-day period. The results further verify that TNEKF not only performs well in local station-level forecasting but also demonstrates strong overall prediction accuracy and robustness at the network level. By jointly modeling spatial dependencies and temporal state evolution, TNEKF is better able to capture the complex spatio-temporal characteristics of metro passenger demand, thereby improving forecasting performance.

Table 2 reports the performance comparison of different short-term passenger flow forecasting methods under three evaluation scenarios: whole-day (W), peak-hour (P), and off-peak (N). The all-day period spans 6:00–22:00, with peak hours defined as 7:00–9:00 and 16:30–18:30, and the remaining intervals considered off-peak. The comparison includes both classical approaches (ARIMA, EKF) and advanced spatiotemporal models from the literature, such as STGCN, LSTM-GCN, and DBSTNet.



Figure 9. Prediction results of total departure passenger flows for all stations over a one-day period.

Table 2. Comparison of Performance of Different demand Forecasting Methods.

Scenarios and Methods		MAE	RMSE	MAPE
W	ARIMA	2.47	4.33	0.59
	EKF	2.12	3.76	0.48
	STGCN [30]	2.00	3.83	0.50
	LSTM-GCN [31]	1.96	3.81	0.52
	DBSTNet [32]	1.99	3.79	0.56
	TNEKF	1.39	3.39	0.45
P	ARIMA	2.76	5.32	0.58
	EKF	2.60	5.15	0.54
	STGCN	2.69	5.10	0.48
	LSTM-GCN	2.64	5.07	0.52
	DBSTNet	2.67	5.09	0.57
	TNEKF	1.83	3.95	0.47
N	ARIMA	1.95	3.96	0.59
	EKF	1.94	3.06	0.41
	STGCN	1.75	3.36	0.52
	LSTM-GCN	1.69	3.33	0.52
	DBSTNet	1.73	3.30	0.56
	TNEKF	1.27	2.95	0.43

As shown in Table 2, TNEKF consistently achieves lower MAE, RMSE, and MAPE values in most scenarios, indicating superior overall predictive accuracy. Compared with purely data-driven deep learning models, the filtering-based approaches demonstrate improved robustness and adaptability to dynamic demand variations. In particular, TNEKF exhibits more stable performance across different time periods, suggesting that the integration of tensor-based spatial representations with recursive state estimation effectively captures the coupled spatiotemporal dynamics of metro passenger demands.

5.5. Robustness Validation Under Disruption Scenarios

Finally, unplanned station closures are introduced to evaluate model robustness under uncertain conditions. Passenger redistribution is assumed to follow three typical behavioral responses: shifting to nearby stations, detouring to alternative stations within walking distance, and canceling or postponing trips. In this experiment, Station 14 is assumed to be closed for two consecutive days (Day 2 and Day 3 in the test set). Among the affected passengers, 20% are assumed to reroute via Station 13 as an alternative transfer point, 70% shift to Station 15, and the remaining 10% adopt alternative travel modes.

Figure 10 presents the prediction results for representative OD pairs under the disruption scenario. It can be observed that the prediction accuracy of data-driven deep learning models, including STGCN, LSTM-GCN, and DBSTNet, degrades when abrupt changes occur in passenger demand patterns. In contrast, the proposed TNEKF model maintains stable and accurate predictions under such disturbances.

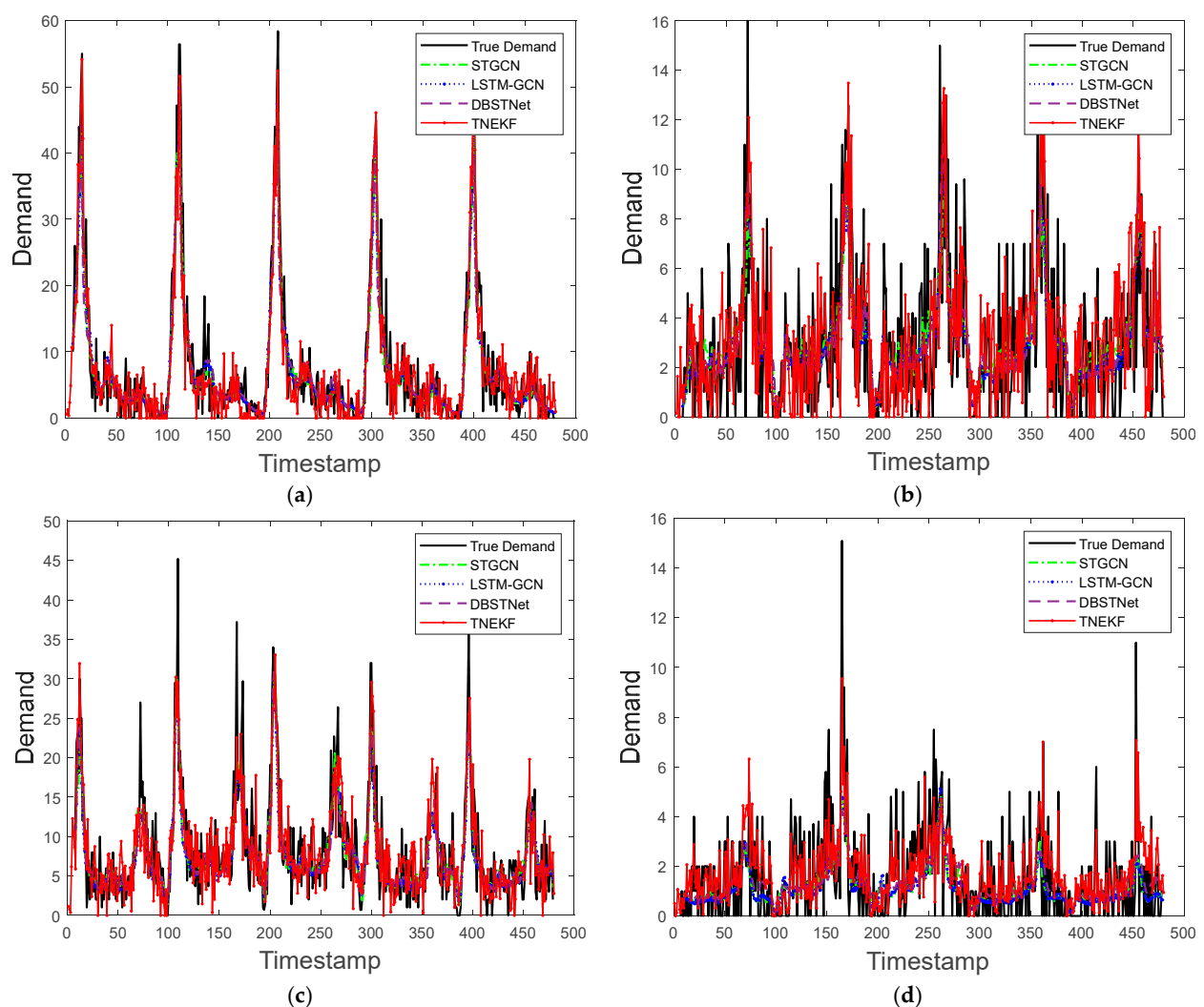


Figure 10. OD passenger flow forecasting results under an unplanned station closure scenario. (a) OD pair: Station 9-Station 13. (b) OD pair: Station 13-Station 7. (c) OD pair: Station 22-Station 15. (d) OD pair: Station 15-Station 2.

This advantage can be attributed to its ability to jointly capture temporal dynamics and spatial correlations within a unified tensor-based filtering framework, enabling effective adaptation to sudden demand variations. Overall, the results demonstrate that TNEKF not only achieves strong performance under normal operating conditions but also

exhibits superior robustness and adaptability in the presence of unexpected disruptions, making it a reliable tool for dynamic passenger flow prediction and management in complex metro systems.

Table 3 presents a comprehensive evaluation of all methods across these scenarios, focusing on OD flows related to Stations 13 and 15 over the disruption period. The performance is summarized using the mean and standard deviation of MAE, RMSE, and wMAPE.

Table 3. Comparison of prediction performance under unplanned station closure scenarios aggregated across passenger redistribution configurations (mean $\pm$ standard deviation).

Method	MAE	RMSE	wMAPE
STGCN	3.086 $\pm$ 0.021	4.323 $\pm$ 0.041	0.662 $\pm$ 0.005
LSTM-GCN	3.104 $\pm$ 0.027	4.387 $\pm$ 0.041	0.667 $\pm$ 0.005
DBSTNet	3.050 $\pm$ 0.031	4.327 $\pm$ 0.053	0.655 $\pm$ 0.007
TNEKF	1.926 $\pm$ 0.030	3.6537 $\pm$ 0.059	0.497 $\pm$ 0.004

The results indicate that TNEKF consistently outperforms all benchmark models in terms of prediction accuracy, while also demonstrating lower variability across different redistribution settings. In comparison with state-of-the-art spatiotemporal models such as STGCN, LSTM-GCN, and DBSTNet, the proposed method exhibits enhanced resilience to structural shifts in demand.

5.6. Practical Implications

Beyond improving prediction accuracy, the proposed TNEKF provides practical value for geo-information-enabled intelligent metro management. Unlike station-level forecasting, OD demand prediction explicitly captures spatial interactions between origins and destinations, preserving the spatiotemporal distribution of passenger movements across the metro network. Such detailed demand information supports refined operational decisions, including dynamic capacity allocation, train service adjustment, and passenger flow management, particularly for transfer hubs and congested corridors.

Moreover, the recursive Bayesian framework continuously updates latent demand characteristics as new AFC observations become available, enabling adaptation to evolving spatiotemporal travel patterns. This capability is valuable under dynamic operating conditions, such as temporary station closures or special events, where forecasting models without recursive updating may suffer from degraded performance. The disruption experiments further demonstrate the robustness of TNEKF under abnormal scenarios.

Finally, the low-rank tensor network implementation improves the scalability of recursive OD forecasting by exploiting correlations among origins, destinations, and temporal dimensions. By avoiding explicit computation and storage of high-dimensional covariance matrices, the proposed framework provides an efficient solution for large-scale geo-spatiotemporal transportation data analysis and facilitates its integration into intelligent transportation management systems.

6. Conclusions

This paper proposes a TNEKF framework for short-term metro OD demand forecasting. The proposed framework formulates metro OD forecasting as a recursive Bayesian state-space estimation problem and integrates a multi-input multi-output Volterra model with Extended Kalman Filtering to characterize the nonlinear spatiotemporal evolution of passenger demand. To address the computational challenges associated with large-scale OD systems, low-rank tensor-network representations are adopted for both the latent

state and covariance, allowing all recursive filtering operations to be performed directly in tensor-network form without explicitly constructing high-dimensional vectors and matrices. As a result, the proposed framework enables scalable recursive forecasting while preserving the nonlinear characteristics of metro passenger demand.

Experimental results on the Hangzhou metro dataset demonstrate that TNEKF consistently outperforms representative statistical, recursive Bayesian, and deep learning methods under whole-day, peak-hour, and off-peak conditions. Moreover, the proposed method maintains superior prediction accuracy under simulated station disruption scenarios, indicating strong robustness to abrupt demand variations. Beyond the quantitative improvements, the experimental results provide several important insights into large-scale metro OD forecasting. First, explicitly modeling passenger demand within a nonlinear recursive Bayesian framework improves forecasting accuracy while accounting for demand uncertainty. Second, recursive state updating enables continuous adaptation to evolving passenger demand patterns, resulting in improved robustness under both normal operating conditions and unexpected disruptions. Third, the low-rank tensor-network implementation substantially reduces the computational burden of recursive estimation by exploiting the inherent low-rank structure of high-dimensional OD demand, making network-wide recursive forecasting computationally feasible. These findings indicate that uncertainty-aware recursive estimation and tensor-network computation complement each other, providing an effective and scalable framework for forecasting complex spatiotemporal transportation systems.

Beyond its forecasting performance, the proposed framework also has important practical implications for intelligent urban rail transit management. Unlike station-level passenger flow forecasting, network-wide OD demand prediction explicitly preserves passenger movements between origins and destinations, providing richer spatiotemporal information for operational decision-making. Such information can support dynamic train scheduling, capacity allocation, passenger flow management, and disruption response, particularly under rapidly changing operating conditions. More broadly, the proposed recursive tensor-network framework is not limited to metro OD forecasting and provides a scalable computational paradigm for uncertainty-aware prediction in other large-scale spatiotemporal dynamic systems.

The present study still has several limitations. External factors, including weather conditions, special events, and operational control strategies, are not explicitly incorporated into the forecasting framework, and the proposed method has been validated using a single metro network. Future work will investigate heterogeneous data fusion, adaptive tensor-rank optimization, and more advanced nonlinear state-space models to further improve forecasting accuracy, robustness, and computational efficiency under more realistic operating conditions. In addition, the proposed recursive tensor-network framework will be extended to multimodal transportation systems and other large-scale geo-spatiotemporal prediction problems requiring efficient recursive Bayesian inference.

Author Contributions: Aijing Su: formal analysis, investigation, visualization, methodology, writing—original draft. Bing Wu: investigation, validation, supervision, funding acquisition, writing—review & editing. Xiaoxing Fang: conceptualization, funding acquisition, data curation, methodology, writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Nature Science Foundation of China under Grant No. 62201269 and 52172331.

Data Availability Statement: The data that support the findings of this study are available on request from the first author (1910910@tongji.edu.cn).

Acknowledgments: The authors would like to acknowledge the editors and reviewers of the ISPRS International Journal of Geo-Information for their constructive comments and suggestions. The authors are very grateful to Tianchi Data Sets for providing datasets (<https://tianchi.aliyun.com/dataset/21904#14>, accessed on 30 May 2026) for research use.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Shao, J.; Liu, G.; Yuan, H.; Song, Q.; Yang, M.; Luo, D.; Zhang, X.; Tan, Y.; Zhang, Y. Evaluation and scale forecast of underground space resources of historical and cultural cities in China. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 31.
2. Gu, Y.; Dou, M. Nonlinear and Threshold Effects on Station-Level Ridership: Insights from Disproportionate Weekday-to-Weekend Impacts. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 365.
3. Liu, B.; Xu, Y.; Guo, S.; Yu, M.; Lin, Z.; Yang, H. Examining the Nonlinear Impacts of Origin-Destination Built Environment on Metro Ridership at Station-to-Station Level. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 59.
4. Li, L.; Li, Y.; Li, Z. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transp. Res. Part C Emerg. Technol.* **2013**, *34*, 108–120.
5. Li, Y.; Li, Z.; Li, L. Missing traffic data: Comparison of imputation methods. *IET Intell. Transp. Syst.* **2014**, *8*, 51–57.
6. Tan, H.; Wu, Y.; Cheng, B.; Wang, W.; Ran, B. Robust missing traffic flow imputation considering nonnegativity and road capacity. *Math. Probl. Eng.* **2014**, *2014*, 763469.
7. Tan, H.; Wu, Y.; Shen, B.; Jin, P.J.; Ran, B. Short-term traffic prediction based on dynamic tensor completion. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2123–2133.
8. Tichavský, P.; Phan, A.-H.; Cichocki, A. Sensitivity in tensor decomposition. *IEEE Signal Process. Lett.* **2019**, *26*, 1653–1657.
9. Chang, S.Y.; Wu, H.-C.; Kuan, Y.-C.; Wu, Y. Tensor Levenberg–Marquardt algorithm for multi-relational traffic prediction. *IEEE Trans. Veh. Technol.* **2023**, *72*, 11275–11290.
10. Shekhar, S.; Williams, B.M. Adaptive seasonal time series models for forecasting short-term traffic flow. *Transp. Res. Rec.* **2007**, *2024*, 116–125.
11. Moreira-Matias, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; Damas, L. Predicting taxi-passenger demand using streaming data. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 1393–1402.
12. Aljebreen, M.; Alamro, H.; Al-Mutiri, F.; Othman, K.M.; Alsumayt, A.; Alazwari, S.; Hamza, M.A.; Mohammed, G.P. Enhancing traffic flow prediction in intelligent cyber-physical systems: A novel Bi-LSTM-based approach with Kalman filter integration. *IEEE Trans. Consum. Electron.* **2024**, *70*, 1889–1902.
13. Wang, Y.; Papageorgiou, M. Real-time freeway traffic state estimation based on extended Kalman filter: A general approach. *Transp. Res. Part B Methodol.* **2005**, *39*, 141–167.
14. Wang, Y.; Papageorgiou, M.; Messmer, A. RENAISSANCE—A unified macroscopic model-based approach to real-time freeway network traffic surveillance. *Transp. Res. Part C Emerg. Technol.* **2006**, *14*, 190–212.
15. Tampere, C.M.J.; Immers, L.H. An extended Kalman filter application for traffic state estimation using CTM with implicit mode switching and dynamic parameters. In Proceedings of the IEEE Intelligent Transportation Systems Conference, Seattle, WA, USA, 30 September–3 October 2007; pp. 209–216.
16. van Hinsbergen, C.P.; Schreiter, T.; Zuurbier, F.S.; van Lint, J.W.C.; van Zuylen, H.J. Localized extended Kalman filter for scalable realtime traffic state estimation. *IEEE Trans. Intell. Transp. Syst.* **2012**, *13*, 385–394.
17. Liu, G.; Neupane, P.; Wu, H.-C.; Xiang, W.; Pu, L.; Chang, S.Y. Novel robust indoor device-free moving-object localization and tracking using machine learning with Kalman filter and smoother. *IEEE Syst. J.* **2022**, *16*, 6253–6264.
18. Chang, S.Y.; Wu, H.-C.; Yan, K. Variable-bit-rate video framesize prediction by the extended Kalman filter using Levenberg–Marquardt algorithm. *IEEE Trans. Broadcast.* **2023**, *69*, 75–84.
19. Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.-Y. Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 865–873.
20. Ke, J.; Yang, H.; Zheng, H.; Chen, X.; Jia, Y.; Gong, P.; Ye, J. Hexagon-based convolutional neural network for supply-demand forecasting of ride-sourcing services. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 4160–4173.
21. 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.

22. Ke, J.; Zheng, H.; Yang, H.; Chen, X. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transp. Res. Part C Emerg. Technol.* **2017**, *85*, 591–608.
23. Noursalehi, P.; Koutsopoulos, H.N.; Zhao, J. Dynamic origin-destination prediction in urban rail systems: A multi-resolution spatio-temporal deep learning approach. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 5106–5115.
24. Batselier, K.; Chen, Z.; Wong, N. A Tensor Network Kalman filter with an application in recursive MIMO Volterra system identification. *Automatica* **2017**, *84*, 17–25.
25. Batselier, K.; Wong, N. Matrix output extension of the tensor network Kalman filter with an application in MIMO Volterra system identification. *Automatica* **2018**, *95*, 413–418.
26. Ma, X.; Zhang, J.; Du, B.; Ding, C.; Sun, L. Parallel architecture of convolutional bi-directional LSTM neural networks for network-wide metro ridership prediction. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 2278–2288.
27. Wei, L.; Guo, D.; Chen, Z.; Yang, J.; Feng, T. Forecasting Short-Term Passenger Flow of Subway Stations Based on the Temporal Pattern Attention Mechanism and the Long Short-Term Memory Network. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 25.
28. Fu, R.; Zhang, Z.; Li, L. Using LSTM and GRU neural network methods for traffic flow prediction. In Proceedings of the 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, China, 11–13 November 2016; pp. 324–328.
29. Han, Y.; Wang, S.; Ren, Y.; Wang, C.; Gao, P.; Chen, G. Predicting station-level short-term passenger flow in a citywide metro network using spatiotemporal graph convolutional neural networks. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 243.
30. 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.
31. Chen, P.; Fu, X.; Wang, X. A graph convolutional stacked bidirectional unidirectional-LSTM neural network for metro ridership prediction. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 6950–6962.
32. Rong, J.; Xu, W.; Wen, Y. A spatiotemporal model for urban taxi origin–destination prediction based on multi-hop GCN and hierarchical LSTM. *Alex. Eng. J.* **2025**, *128*, 905–917.

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.