



**HAL**  
open science

# Graph convolutional networks for traffic forecasting with missing values

Jingwei Zuo, Karine Zeitouni, Yéhia Taher, Sandra Garcia-Rodriguez

► **To cite this version:**

Jingwei Zuo, Karine Zeitouni, Yéhia Taher, Sandra Garcia-Rodriguez. Graph convolutional networks for traffic forecasting with missing values. *Data Mining and Knowledge Discovery*, 2023, 37 (2), pp.913-947. 10.1007/s10618-022-00903-7. hal-04069257

**HAL Id: hal-04069257**

**<https://hal.science/hal-04069257v1>**

Submitted on 6 Jun 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



# Graph convolutional networks for traffic forecasting with missing values

Jingwei Zuo<sup>1</sup> · Karine Zeitouni<sup>2</sup> · Yehia Taher<sup>2</sup> · Sandra Garcia-Rodriguez<sup>3</sup>

Received: 11 December 2021 / Accepted: 25 November 2022 / Published online: 16 December 2022  
© The Author(s) 2022

## Abstract

Traffic forecasting has attracted widespread attention recently. In reality, traffic data usually contains missing values due to sensor or communication errors. The Spatio-temporal feature in traffic data brings more challenges for processing such missing values, for which the classic techniques (e.g., data imputations) are limited: (1) in temporal axis, the values can be randomly or consecutively missing; (2) in spatial axis, the missing values can happen on one single sensor or on multiple sensors simultaneously. Recent models powered by Graph Neural Networks achieved satisfying performance on traffic forecasting tasks. However, few of them are applicable to such a complex missing-value context. To this end, we propose GCN-M, a Graph Convolutional Network model with the ability to handle the complex missing values in the Spatio-temporal context. Particularly, we jointly model the missing value processing and traffic forecasting tasks, considering both local Spatio-temporal features and global historical patterns in an attention-based memory network. We propose as

---

Responsible editor: Albrecht Zimmermann and Peggy Cellier.

---

This research has been developed for the most part in the context of the main author's Ph.D. at DAVID Lab, UVSQ, Université Paris-Saclay.

---

✉ Jingwei Zuo  
jingwei.zuo@tii.ae  
Karine Zeitouni  
karine.zeitouni@uvsq.fr  
Yehia Taher  
yehia.taher@uvsq.fr  
Sandra Garcia-Rodriguez  
sandra.garciaRodriguez@cea.fr

<sup>1</sup> Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, UAE

<sup>2</sup> DAVID Lab, UVSQ, Université Paris-Saclay, Versailles, France

<sup>3</sup> Data Analysis and Systems Intelligence Laboratory, CEA, LIST, Gif Sur Yvette, France

well a dynamic graph learning module based on the learned local-global features. The experimental results on real-life datasets show the reliability of our proposed method.

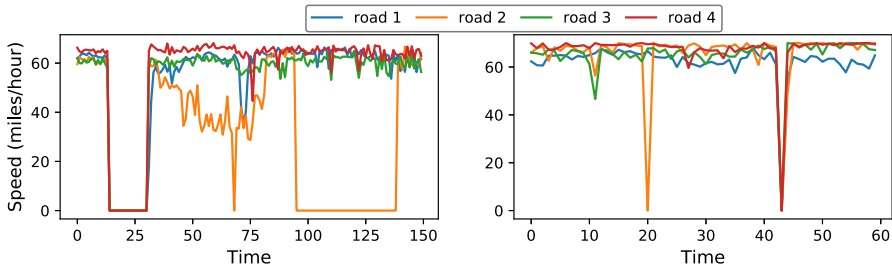
**Keywords** Traffic forecasting · Missing values · Graph convolutional networks · Memory networks · Neural networks · Deep learning

## 1 Introduction

Traffic forecasting has played a critical role in intelligent transportation systems, which helps the transportation department better manage and control traffic congestion. Generally represented by geo-located Multivariate Time Series (MTS), traffic data not only shows the typical characteristics of MTS, i.e., temporal dependency (Zuo et al. 2021), but also integrates the spatial information of the traffic network, i.e., the spatial dependency between the sensor traffic nodes over the road network.

In recent years, by leveraging the spatial-temporal patterns in traffic data, many deep learning models based on recurrent neural network (RNN) (Li et al. 2018), temporal convolutional network (TCN) (Wu et al. 2019), graph convolutional networks (GCN) (Li et al. 2021), etc., have been applied in traffic forecasting tasks and achieved state-of-the-art performance. They all have a strong assumption that the data is complete or has been well-preprocessed (Yu et al. 2018). However, since the traffic data is generally collected from geo-located sensors, sensor failures or communication errors will result in missing values in the collected data, thus deteriorating the performance of the forecasting model. We should remark that the missing measures are usually marked as *zero* in traffic data (Li et al. 2021), which should be distinguished from the non-missing measures but with *zero* values. A typical example (Tian et al. 2018) comes from the traffic flow data: no vehicles are detected during the night, then the traffic measures are marked as *zero* instead of being considered as missing. This can be commonly observed from real-life traffic data for which the missing rate evolves periodically during the day (Lopez 2018).

The missing values can either be ignored in the learning model when calculating the loss function (Wang et al. 2020) or be considered before or during the training process (Cui et al. 2020b). Ignoring the missing values, especially when the missing ratio is high (Cui et al. 2020b), hinders the model from benefiting from the rich data information for better performance. When considering the missing values in traffic data, most work (Cirstea et al. 2019) conducts data imputation during the preprocessing step, then imports the completed data into the training step, i.e., *two-step processing*. Recent work tends to jointly consider the missing values and the forecasting modeling during the training step (i.e., *one-step processing*) and declared better performance than the two-step processing (Che et al. 2018; Cui et al. 2020a, b; Tian et al. 2018; Tang et al. 2020). However, the above-mentioned work suffers from three major issues. First, the missing and zero values are usually considered to be the same, leading to unnecessary, even harmful data imputations, thus contradicting the raw data information. Second, most of the work (Che et al. 2018; Cui et al. 2020a; Tian et al. 2018; Tang et al. 2020) considers missing values from the temporal aspect, ignoring the rich information from the spatial perspective. Third, they are generally designed for processing the missing



**Fig. 1** Missing measures of traffic speed data from METR-LA dataset (Li et al. 2018). *left*) long-range missing on entire network (i.e., Spatio-temporal block) and partial network (i.e., temporal block); *right*) short-range random missing on partial network (i.e., temporal values) and entire network (i.e., Spatio-temporal vectors)

values in some basic scenarios, such as random missing values or temporal block missing values, but lack power for the complex scenarios as shown in Fig. 1. In the real world, the missing values in traffic data occur in both long-range (e.g., *device power-off*) and short-range (e.g. *device errors*) settings, in partial (e.g., *local sensor errors*) and entire transportation network (e.g., *control center errors*). Therefore, a holistic approach is required for handling various types of missing values together in complex scenarios.

To handle both the Spatio-temporal patterns and complex missing-value scenarios in traffic data, we propose **Graph Convolutional Networks for Traffic Forecasting with Missing Values (GCN-M)**. The graph neural network-based structure allows jointly modeling the Spatio-temporal patterns and the missing values in a one-step process. We construct local statistical features from spatial and temporal perspectives for handling short-range missing values. This is further enhanced by a memory module to extract global historical features for processing long-range missing blocks. The combined local-global features allow not only for identifying the missing measures from the inherent zero values but also for enriching the traffic embeddings, thus generating dynamic traffic graphs to model the dynamic spatial interactions between traffic nodes. The missing values on a partial and entire network can then be considered from spatial and temporal perspectives.

We summarize the paper's main contributions as follows:

- **Complex missing value modeling:** We study the complex scenario where missing traffic values occur on both short & long ranges and on partial & entire transportation networks.
- **Spatio-temporal memory module:** We propose a memory module that can be used by GCN-M to learn both local Spatio-temporal features and global historical patterns in traffic data for handling the complex missing values.
- **Dynamic graph modeling:** We propose a dynamic graph convolution module that models the dynamic spatial interactions. The dynamic graph is characterized by the learned local-global features at each timestamp, which not only offset the missing values' impact but also help learn the graph.
- **Joint model optimization:** We jointly model the Spatio-temporal patterns and missing values in one-step processing, which allows processing missing values

specifically for traffic forecasting tasks, thus bringing better model performance than two-step processing.

- **Extensive experiments on real-life data:** The experiments are carried out on two real-life traffic datasets. We provide detailed evaluations with 12 baselines, which show the effectiveness of GCN-M over state-of-the-art.

The rest of this paper starts with a review of the most related work. Then, we formulate the problems of the paper. Later, we present in detail our proposal GCN-M, followed by the experiments on real-life datasets and the conclusion.

## 2 Related works

We start with defining the notions used in the paper:

**Definition 1** (One-step processing). For one-step processing models, the missing values and the traffic forecasting are jointly modeled in one single step.

**Definition 2** (Two-step processing). The two-step processing models first handle the missing values in a preprocessing step, then apply a forecasting model on the completed data.

### 2.1 Graph convolutional networks for traffic forecasting

Graph Convolutional Network (GCN) is a special kind of Convolutional Neural Network (CNN) generalized for graph-structured data. Most of the GCN-related work focuses on graph representation, which learns node embedding by integrating the features from the node's local neighbors based on the given graph structure, i.e., adjacency matrix. The traffic data shows strong dependencies between the spatial nodes, for which GCN can be naturally suitable. Various work (Li et al. 2018; Wu et al. 2020; Yu et al. 2018; Wang et al. 2020) empowered by GCN achieved remarkable performance when doing traffic forecasting tasks, relying on spatial and temporal completion of the data, or calculating loss function for non-zero entries, i.e., only calculating the loss on entries that contain valid sensor readings. However, these techniques may introduce derivations when modeling the Spatio-temporal relations between the sensor nodes. In other words, where non-missing measures are required to characterize the dynamic graph at each timestamp, missing values may hinder the traffic graph learning (Li et al. 2021), especially dynamic graph learning (Guo et al. 2021).

### 2.2 Missing value processing

The simplest solution for processing missing values in MTS would be data imputation, such as statistic imputation (e.g., mean, median), EM-based imputation (García-Laencina et al. 2010), K-nearest neighborhood (Batista et al. 2002), and matrix factorization (Dong et al. 2022). It's generally believed that those methods fail to model temporal dynamics in a time series (Tang et al. 2020). In other words, they

are not applicable for handling long-range missing values. Recent generative models (Yoon et al. 2019; Dong et al. 2022) show reliable performance for long-range time series imputation. However, isolating the imputation model from the forecasting model leads to two-step processing, and may generate sub-optimal results (Cirstea et al. 2019; Wells et al. 2013; Che et al. 2018). To handle this issue, recent studies (Che et al. 2018; Tang et al. 2020; Wang et al. 2021; Zhong et al. 2021) jointly model the missing values and forecasting task in one-step processing. For instance, GRU-D (Che et al. 2018) considers the nearby temporal statistical features to do imputations inside GRUs, whereas LSTM-I (Cui et al. 2020a) infers missing values at the current time step from preceding LSTM cell states and hidden states, and SGMN (Cui et al. 2020b) improved the state transition process via a Graph Markov Process. Limited to short-period missing context, those methods are further enhanced by LGnet (Tang et al. 2020) with the global temporal dynamics to handle the long-range missing issue, and by LSTM-M (Tian et al. 2018) with multi-scale modeling to better explore historical information. However, the above-mentioned models handle missing values by focusing on the temporal aspect without considering the complex Spatio-temporal features in traffic data. Specifically, the strong spatial connections between the sensor nodes should provide us with more information to handle the missing values. Moreover, one-step processing models are generally designed for single-step forecasting without considering the multi-step settings. Table 1 shows the method comparison for traffic forecasting with missing values.

### 3 Problem formulation

We aim to predict future traffic data by leveraging historical traffic data. Traffic data can be represented as a multivariate time series on a traffic network. Let the traffic network  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ , where  $\mathcal{V} = \{v_1, \dots, v_N\}$  is a set of  $N$  traffic sensor nodes and  $\mathcal{E} = \{e_1, \dots, e_E\}$  is a set of  $E$  edges connecting the nodes. Each node contains  $F$  features representing traffic flow, speed, occupancy, etc. We use  $\mathcal{X} = \{\mathbf{X}_t\}_{t=1}^\tau \in \mathcal{R}^{N \times F \times \tau}$  to denote all the feature values of all the nodes over  $\tau$  time slices,  $\mathbf{X}_t = (\mathbf{x}_t^1, \dots, \mathbf{x}_t^N) \in \mathcal{R}^{N \times F}$  denotes the observations at time  $t$ , where  $\mathbf{x}_t^i \in \mathcal{R}^F$  is the  $i$ -th variable of  $\mathbf{X}_t$ . We define a mask sequence  $\mathcal{M} = \{\mathbf{M}_t\}_{t=1}^\tau \in \mathcal{R}^{N \times F \times \tau}$ ,  $\mathbf{M}_t = (\mathbf{m}_t^1, \dots, \mathbf{m}_t^N) \in \mathcal{R}^{N \times F}$ .  $\mathbf{m}_t^i \in \{0, 1\}^F$  denotes the features' missing status for the  $i$ -th variable. To simplify, we adopt  $x_t^i \in \mathcal{R}$  and  $m_t^i \in \mathcal{R}$  to denote respectively the observation and mask value of one single feature for the  $i$ -th variable of  $\mathbf{X}_t$ . We take  $m_t^i = 0$  if  $x_t^i$  is missing, otherwise  $m_t^i = 1$ .

We aim to build a model  $f$ , which can take an incomplete traffic sequence  $\{\mathcal{X}, \mathcal{M}\}$  and the traffic network  $\mathcal{G}$  as input, to predict the traffic data for the next  $T_p$  time steps  $\mathbf{Y} = \{y_{\tau+1}, \dots, y_{\tau+T_p}\} \in \mathcal{R}^{N \times T_p}$ .

### 4 Proposal: GCN-M

Traffic data is collected under complex urban conditions. Apart from the Spatio-temporal patterns in the traffic data, we also consider the scenarios of complex missing

**Table 1** Existing methods for traffic forecasting with missing values

	Imputation-based	GRU-D	LSTM-I	LSTM-M	LGnet	SGMN	GCN-M
Short-range missing	✓	✓	✓	✓	✓	✓	✓
Long-range missing	-	-	-	✓	✓	-	✓
Multi-scale modeling	-	-	-	✓	-	-	✓
Spatial modeling	-	-	-	-	-	✓	✓
One-step processing	-	✓	✓	✓	✓	✓	✓
Multi-step forecasting	✓	-	-	-	✓	-	✓

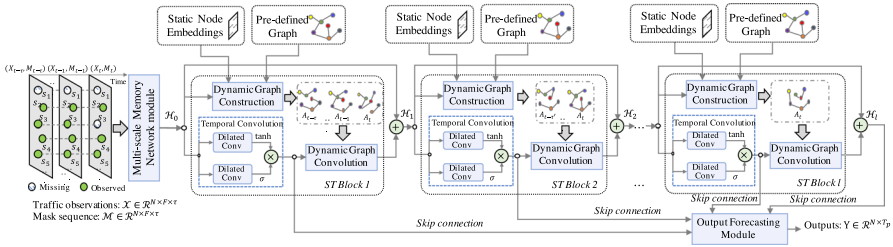


Fig. 2 Main architecture of GCN-M

values. We design a solution that models the local Spatio-temporal features and global historical patterns in a dynamic manner. The complex missing values are considered when building the forecasting model, i.e., one-step processing.

### 4.1 Model architecture

The global structure of GCN-M is shown in Fig. 2, integrating a Multi-scale Memory Network module, an Output Forecasting module, and  $l$  Spatio-Temporal (ST) blocks. Each ST block integrates three key components: Temporal Convolution, Dynamic Graph Construction, and Dynamic Graph Convolution. The input traffic observations  $\mathcal{X} \in \mathbb{R}^{N \times F \times \tau}$  and the mask sequence  $\mathcal{M} \in \mathbb{R}^{N \times F \times \tau}$  are fed into the multi-scale memory network to extract the local statistic features and global historical patterns thus enriching the traffic embeddings. On the one hand, the enriched embeddings  $\mathcal{H}_i$  on each ST block are used to mark the dynamic traffic status, thus generating dynamic graphs by combining both static node embeddings and predefined graph information. On the other hand, the learned dynamic graphs are combined with the temporal convolution module via a dynamic graph convolution to capture temporal and spatial dependencies in the traffic embeddings. We adopt residual connections between the input and output of each ST block to avoid the gradient vanishing problem. The output forecasting module takes the skip connections on the output of the final ST block and the hidden states after each temporal convolution for final prediction.

### 4.2 Multi-scale memory network

To extract the local statistic features and global historical patterns then form an enriched embedding, we adopt the concept of memory network, which was firstly proposed in (Weston et al. 2015) with primary application in Question-Answer (QA) systems. As shown in Fig. 3, the main idea of our memory network is to learn from historical memory components which conserve the long-range multi-scale patterns, i.e., recent, daily-periodic, and weekly-periodic dependencies. The scale range depends on the data characteristics. Specifically, we first extract local Spatio-temporal features as keys to query the memory components; the weighted historical long-range patterns will be cooperated with the local statistic features to eliminate the side effect from the missing values. Then, the local-global features will be output as the enriched traffic embeddings.



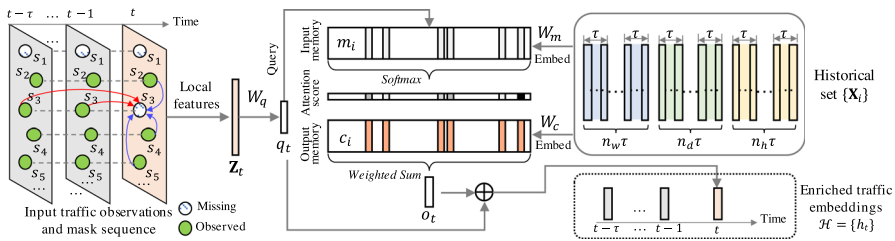


Fig. 3 Memory module enriches traffic embeddings with multi-scale global features

### 4.2.1 Local spatio-temporal features

We first extract the Spatio-temporal features using the contextual information from observed parts of the time series. Unlike prior studies (Che et al. 2018), we consider both temporal and spatial aspects for generating the following statistic features of every timestamp:

*Empirical Temporal Mean:* The mean of previous observations reflects the recent traffic state and serves as a contextual knowledge of  $x_t^i$ . Therefore, for a missing value  $x_t^i \in \mathcal{R}$ , we construct its temporal mean using  $L$  past samples  $x_{*}^i$  before time  $t$ :

$$\bar{x}_t^i = \sum_{l=t-L}^{t-1} m_l^i x_l^i / \sum_{l=t-L}^{t-1} m_l^i \tag{1}$$

*Last Temporal Observation:* We adopt the assumption in (Che et al. 2018) that any missing value inherits more or less the information from the last non-missing observation. In other words, the temporal neighbor stays close to the current missing value. We use  $\dot{x}_t^i$  to denote the last temporal observation of  $x_t^i$ , their temporal distance is defined as  $\delta_t^i$ .

*Empirical Spatial Mean:* Another contextual knowledge of  $x_t^i$  is from the nearby nodes, which reflects the current local traffic situation. For each missing value  $x_t^i$ , we construct its empirical spatial mean using  $S$  nearby samples  $x_t^{*}$  of the sensor node  $i$ :

$$\bar{\bar{x}}_t^i = \sum_{s=1}^S m_t^s x_t^s / \sum_{s=1}^S m_t^s \tag{2}$$

*Nearest Spatial Observation:* Typically, the state of a graph node remains relatively similar to its neighbors, especially in a traffic graph where the nearby nodes share similar traffic situations. We define  $\ddot{x}_t^i$  as the nearest spatial observation of  $x_t^i$ , their spatial distance is denoted as  $\delta_t^i$ .

Generally, when  $\delta_t^i$  or  $\ddot{\delta}_t^i$  is smaller, we tend to trust  $\dot{x}_t^i$  or  $\ddot{x}_t^i$  more. When the spatial/temporal distance becomes larger, the spatial/temporal mean would be more representative. Under this assumption, we model the temporal and spatial decay rate

$\gamma$  as

$$\gamma_t(\delta_t^i) = \exp\{-\max(0, w^i \delta_t^i + b^i)\} \tag{3}$$

$$\gamma_s(\delta_t^i) = \exp\{-\max(0, w_t \delta_t^i + b_t)\} \tag{4}$$

where  $w^i, w_t, b^i$  and  $b_t$  are model parameters that we train jointly with other parameters of the traffic forecasting model. We chose the exponentiated negative rectifier (Che et al. 2018) so that the decay rates  $\gamma_t$  and  $\gamma_s$  decrease monotonically in the range between 0 and 1. Considering the trainable decays, our proposed model incorporates the spatial/temporal estimations to define the local features of  $x_t^i$ :

$$z_t^i = m_t^i x_t^i + (1 - m_t^i)(\gamma_t \check{x}_t^i + \gamma_s \ddot{x}_t^i + (1 - \gamma_t) \bar{x}_t^i + (1 - \gamma_s) \bar{\bar{x}}_t^i) \tag{5}$$

Therefore, for  $\mathbf{X}_t \in \mathcal{R}^{N \times F}$ , we can get its local features  $\mathbf{Z}_t \in \mathcal{R}^{N \times F}$ .

### 4.2.2 Multi-scale memory construction

Global historical patterns play a critical role in building an enriched traffic embedding. The historical observations in multiple scales (e.g., hourly, daily, weekly) can be embedded into memory as complement information for the local features  $\mathbf{Z}_t \in \mathcal{R}^{N \times F}$ . The main idea is to adopt local features to query similar historical patterns in the memory and output a weighted feature representation for the current timestamp. In this manner, the enriched multi-scale historical and local features allow not only eliminating the side effect of missing values but also improving the current feature embeddings. At time  $t$ , the query  $q_t$  of  $\mathbf{X}_t$  can be embedded from the local features  $\mathbf{Z}_t \in \mathcal{R}^{N \times F}$ :

$$q_t = \mathbf{Z}_t W_q + b_q \in \mathcal{R}^{N \times d} \tag{6}$$

where  $W_q \in \mathcal{R}^{F \times d}, b_q \in \mathcal{R}^{N \times d}$  are parameters,  $d$  is the embedding dimension.

The input memory components are the temporal segments of multiple scales:

- The recent (e.g., hourly) segment is:  $X_h = \{\mathbf{X}_i\}_{i=t-\tau}^{t-1} \in \mathcal{R}^{N \times F \times n_h \tau}$ , with  $n_h$  recent periods (e.g., hours) before  $t$ , each period contains  $\tau$  observations.
- The daily-periodic segment is:  $X_d = \{\mathbf{X}_i\} \in \mathcal{R}^{N \times F \times n_d \tau}$  with  $i \in [t - n_d T_d - \tau/2 : t - n_d T_d + \tau/2] \parallel [t - (n_d - 1)T_d - \tau/2 : t - (n_d - 1)T_d + \tau/2] \parallel \dots \parallel [t - T_d - \tau/2 : t - T_d + \tau/2]$ , we store  $\tau$  samples around time  $t$  for each of the past  $n_d$  days.  $T_d$  denotes the sample number during one day, and  $\parallel$  indicates the concatenation operation.
- The weekly-periodic segment is:  $X_w = \{\mathbf{X}_i\} \in \mathcal{R}^{N \times F \times n_w \tau}$  with  $i \in [t - n_w T_w - \tau/2 : t - n_w T_w + \tau/2] \parallel [t - (n_w - 1)T_w - \tau/2 : t - (n_w - 1)T_w + \tau/2] \parallel \dots \parallel [t - T_w - \tau/2 : t - T_w + \tau/2]$ , we store  $\tau$  samples around time  $t$  for each of the past  $n_w$  weeks.  $T_w$  denotes the sample number during one week, and  $\parallel$  indicates the concatenation operation.

The input set of  $\{\mathbf{X}_i\} = [X_h \parallel X_d \parallel X_w] \in \mathcal{R}^{N \times F \times (n_d + n_w + n_h) \tau}$  are embedded into the input memory vectors  $\{m_i\}$  and output memory vectors  $\{c_i\}$ :

$$m_i = \mathbf{X}_i W_m + b_m \in \mathcal{R}^{N \times d} \tag{7}$$

$$c_i = \mathbf{X}_i W_c + b_c \in \mathcal{R}^{N \times d} \quad (8)$$

where  $W_m, W_c \in \mathcal{R}^{F \times d}$ ,  $b_m, b_c \in \mathcal{R}^{N \times d}$  are parameters.

In the embedding space, we compute the attention score between the query  $q_t$  and each memory  $m_i$  by taking the inner product followed by a Softmax:

$$p_{t,i} = \text{Softmax}(q_t^T m_i) \quad (9)$$

The attention score represents the similarity of each historical observation to the query. Any pattern with a higher attention score is more similar to the context of targeting missing values. As shown in Fig. 3, the response vector from memory is then a sum over the output memory vectors, weighted by the attention score from the input:

$$o_t = \sum_{i=1}^{(n_d+n_w+n_h)\tau} c_i p_{t,i} \in \mathcal{R}^{N \times d} \quad (10)$$

We can finally integrate both local Spatio-temporal and global multi-scale features and output the enriched traffic embeddings:

$$h_t = (q_t \parallel o_t) W_h + b_h \in \mathcal{R}^{N \times d} \quad (11)$$

where  $W_h \in \mathcal{R}^{2d \times d}$ ,  $b_h \in \mathcal{R}^d$  are parameters, and  $\parallel$  denotes the concatenation operation. Therefore, for input  $\mathcal{X} = \{\mathbf{X}_t\}_{t=1}^\tau \in \mathcal{R}^{N \times F \times \tau}$ , we can get its enriched traffic embeddings  $\mathcal{H} = \{h_t\}_{t=1}^\tau \in \mathcal{R}^{N \times d \times \tau}$ .

### 4.3 Dynamic graph construction

A predefined graph is usually constructed with the distance or the connectivity between the spatial nodes. However, recent studies (Wang et al. 2020; Wu et al. 2020; Li et al. 2021) show that the cross-region dependence does exist for those nodes which are not physically connected but share similar patterns. Learning dynamic graphs should show better performance than learning static graphs or adopting the predefined graphs. Considering the missing values in traffic data, instead of using the raw traffic observations to mark the dynamic traffic status (Li et al. 2021; Han et al. 2021b), we construct dynamic graphs (i.e., adjacency matrix) with the enriched traffic embeddings  $\mathcal{H}_i$  at each ST block, which integrates both local and global multi-scale patterns at each time step. This allows capturing the spatial relationship between traffic nodes robustly. As shown in Fig. 4, the main idea here is to generate dynamic filters from the predefined graphs  $\mathcal{G}$  and the traffic embeddings  $\mathcal{H}_i \in \mathcal{R}^{N \times d \times \tau_i}$  ( $\tau_i$  is the sequence length at the  $i$ -th ST block), which are applied on the randomly initialized static node embeddings to construct dynamic adjacency matrices. In more detail, the core steps in Fig. 4 are illustrated as follows:

**[Dynamic Filter Generation]** Given  $\mathcal{H}_i = \{h_t\} \in \mathcal{R}^{N \times d \times \tau_i}$ , the traffic embedding  $h_t$  at time  $t$  is firstly combined with the predefined adjacency matrix  $A_G \in \mathcal{R}^{N \times N}$  to generate dynamic graph filters via a diffusion convolution layer as proposed in Li

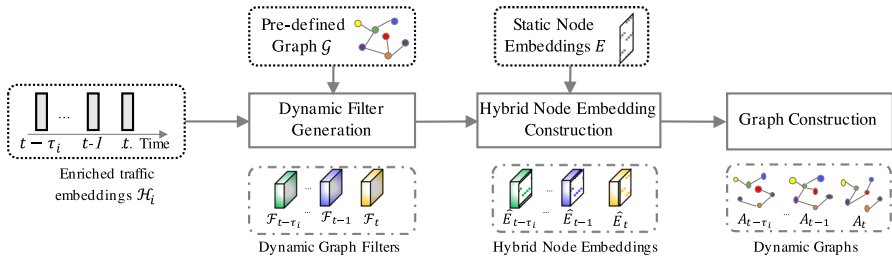


Fig. 4 Dynamic graph construction from the enriched traffic embeddings

et al. (2018):

$$\mathcal{F}_t = \sum_{k=0}^K P_k h_t W_k \in \mathcal{R}^{N \times d} \tag{12}$$

where  $K$  denotes the diffusion step,  $P_k = A_G / \text{rowsum}(A_G)$  represents the power series of the transition matrix (Wu et al. 2019), and  $W_k \in \mathcal{R}^{d \times d}$  is the model parameter matrix.

**[Hybrid Node Embedding Construction]** Considering both the source and target traffic node, we initialize two random node embeddings  $E^1, E^2 \in \mathcal{R}^{N \times d}$ , representing the static node features (Wang et al. 2020) which are not reflected in the observations but learnable during training. Thus, two dynamic filters are applied over the static node embeddings:

$$\begin{aligned} \hat{E}_t^1 &= \tanh(\alpha(\mathcal{F}_t^1 \odot E^1)) \in \mathcal{R}^{N \times d} \\ \hat{E}_t^2 &= \tanh(\alpha(\mathcal{F}_t^1 \odot E^2)) \in \mathcal{R}^{N \times d} \end{aligned} \tag{13}$$

where  $\odot$  denotes the Hadamard product (Wu et al. 2019).  $\hat{E}_t^1$  and  $\hat{E}_t^1$  are hybrid node embeddings combining both static and dynamic settings of the traffic data.

**[Graph Construction]** As mentioned in the previous study (Wu et al. 2020), in multivariate time series forecasting, we expect that the change of a node’s condition causes the change of another node’s condition such as traffic flow. Therefore the learned relationship is supposed to be uni-directional. We construct the graph by extracting uni-directional relationships between traffic nodes. The dynamic adjacency matrix is constructed from the hybrid embeddings:

$$A_t = \text{ReLU} \left( \tanh(\alpha(\hat{E}_t^1 \hat{E}_t^{2T} - \hat{E}_t^2 \hat{E}_t^{1T})) \right) \in \mathcal{R}^{N \times N} \tag{14}$$

Therefore, we can construct the dynamic graphs  $A_{\mathcal{D}_i} = \{A_t\} \in \mathcal{R}^{N \times N \times \tau_i}$  for the enriched traffic embeddings  $\mathcal{H}_i \in \mathcal{R}^{N \times d \times \tau_i}$  at the  $i$ -th ST block. As the computation and memory cost grows quadratically with the increase of graph size, in practice, it is possible to adopt a sampling approach (Wu et al. 2020), which only calculates pairwise relationships among a subset of nodes.

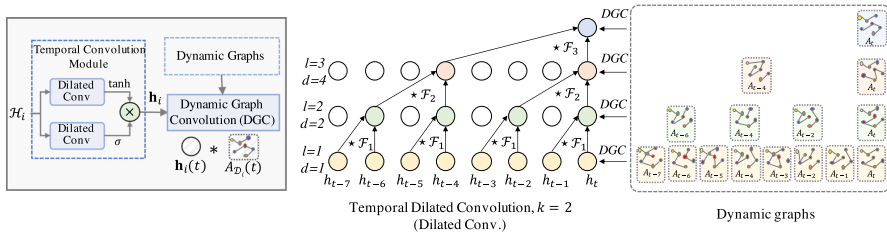


Fig. 5 Temporal convolution module with dynamic graph convolution

### 4.4 Temporal convolution module

The temporal convolution network (TCN) (Lea et al. 2017) consists of multiple dilated convolution layers, which allows extracting high-level temporal trends. Compared to RNN-based approaches, dilated causal convolution networks are capable of handling long-range sequences in a parallel manner. The output of the last layer is a representation that captures temporal dynamics in history. As shown in Fig. 5, considering the temporal dynamics in traffic data, we adopt the temporal convolution module (Wu et al. 2019) with the consideration of the gating mechanism over the enriched traffic embeddings  $\mathcal{H}_i$ . One dilated convolution block is followed by a tangent hyperbolic activation function to output the temporal features. The other block is followed by a sigmoid activation function as a gate to determine the ratio of information that can pass to the next module. In particular, the sigmoid gate controls which input of the current states is relevant for discovering compositional structure and dynamic variances in time series. Applying the sigmoid nonlinearity on the input states differs from other well-known architectures (e.g., LSTM or GRU), which ignore the compositional structure features in time series (Yu et al. 2018).

Given the enriched traffic embeddings  $\mathcal{H}_i = \{h_t\} \in \mathcal{R}^{N \times d \times \tau_i}$ , a filter  $\mathcal{F} \in \mathcal{R}^{1 \times K}$ ,  $K$  is the temporal filter size,  $K = 2$  by default. The dilated causal convolution operation of  $\mathcal{H}_i$  with  $\mathcal{F}$  at time  $t$  is represented as:

$$\mathcal{H}_i \star \mathcal{F}_i(t) = \sum_{s=0}^K \mathcal{F}_i(s) \mathcal{H}_i(t - \mathbf{d} \times s) \in \mathcal{R}^{N \times d \times \tau_{i+1}} \tag{15}$$

where  $\star$  is the convolution operator,  $\mathbf{d}$  is the dilation factor,  $d$  is the embedding dimension size,  $\tau_{i+1}$  is the new sequence length after the convolution operation, which equals to one on the last layer. Figure 5 shows a three-layer dilated convolution block with  $K = 2$ ,  $\mathbf{d} \in [1, 2, 4]$ . Considering the gating mechanism, we define the output of the temporal convolution module:

$$\mathbf{h}_i = \tanh(W_{\mathcal{F}1} \star \mathcal{H}_i) \odot \sigma(W_{\mathcal{F}2} \star \mathcal{H}_i) \in \mathcal{R}^{N \times d \times \tau_{i+1}} \tag{16}$$

where  $W_{\mathcal{F}1}$ ,  $W_{\mathcal{F}2}$  are learnable parameters of convolution filters,  $\odot$  denotes the element-wise multiplication operator,  $\sigma(\cdot)$  is the sigmoid function.

A classic temporal convolution module stacks the temporal features at each time step  $t$ . Therefore, the upper layer contains richer information than the lower layer. The gating mechanism allows filtering the temporal features on the lower layers by

weighting features on different time steps without considering the spatial node interactions at each time step. Moreover, the spatial interactions in traffic data always show a dynamic nature (Wu et al. 2020). To this end, the gating mechanism from a dynamic spatial aspect is envisaged to better capture the Spatio-temporal patterns.

### 4.5 Dynamic graph convolution

Spatial interactions between the traffic nodes could be used to improve traffic forecasting performance. The dynamic spatial interaction leads to considering a dynamic version of graph convolution to conduct it on different graphs at different timestamps. Different from previous work (Li et al. 2021) which uses raw traffic observations to mark the dynamic traffic status, we adopt the enriched traffic embeddings, which consider the missing-value issues to generate robust dynamic graphs.

As shown in Fig. 5, we apply the dynamic graph convolution on  $\mathbf{h}_i$ , i.e., the output of the temporal convolution module, to further select the features at each time step from the spatial perspective. As mentioned in Sect. 4.3, the dynamic graphs  $A_{\mathcal{D}_i} \in \mathcal{R}^{N \times N \times \tau_i}$  are generated from the enriched traffic embeddings  $\mathcal{H}_i \in \mathcal{R}^{N \times d \times \tau_i}$  at the  $i$ -th ST block.  $A_t$  reflects the spatial relationships between nodes at time  $t$ . The temporal features  $\mathbf{h}_i(t)$  aggregate spatial information according to the adjacency matrix  $A_t$ . Inspired by DCRNN (Li et al. 2018), we consider the traffic situation as the diffusion procedure on the graph. The graph convolution will generate the aggregated spatial information at each time step:

$$\mathcal{H}'_i(t) = \sum_{k=0}^K (A_{\mathcal{D}_i}(t))^k \mathbf{h}_i(t) W_k \in \mathcal{R}^{N \times d} \tag{17}$$

where  $K$  denotes the diffusion step, and  $W_k$  is the learnable parameter matrix. We adopt the residual connection (He et al. 2016) between the input and output of each ST block to avoid the gradient vanishing issue in the model’s training. Therefore, the input of the  $(i + 1)^{th}$  ST block is defined as:

$$\mathcal{H}_{i+1}(t) = \mathcal{H}_i(t) + \mathcal{H}'_i(t) \tag{18}$$

### 4.6 Output forecasting module

The outputs  $\mathbf{h}_i \in \mathcal{R}^{N \times d \times \tau_{i+1}}$  of the middle temporal convolution modules and  $\mathcal{H}_l \in \mathcal{R}^{N \times d \times 1}$  of the last ST block are considered for the final prediction, which represent the hidden states at various Spatio-temporal levels. We add skip connections on each of the hidden states which are essentially  $1 \times \tau_{i+1}$  standard convolutions ( $\tau_{i+1}$  denotes the sequence length at the output of the  $i$ -th ST block). The concatenated output features are defined as follows:

$$O = (\mathbf{h}_0 W_s^0 + b_s^0) \parallel \dots \parallel (\mathbf{h}_i W_s^i + b_s^i) \parallel \dots \parallel (\mathbf{h}_{l-1} W_s^{l-1} + b_s^{l-1}) \parallel (\mathcal{H}_l W_s^l + b_s^l) \tag{19}$$

where  $O \in \mathcal{R}^{N \times (l+1)d}$ ,  $W_s^i, b_s^i$  are learnable parameters for the convolutions. Two fully-connected layers are added to project the concatenated features into the desired output dimension:

$$\hat{\mathbf{Y}} = (\text{ReLU}(OW_{fc}^1 + b_{fc}^1))W_{fc}^2 + b_{fc}^2 \in \mathcal{R}^{N \times T_p} \quad (20)$$

where  $W_{fc}^1, W_{fc}^2, b_{fc}^1, b_{fc}^2$  are learnable parameters for the fully-connected layers,  $N$  is the node number,  $T_p$  denotes the forecasting steps.

Given the ground truth  $\mathbf{Y} \in \mathcal{R}^{N \times T_p}$  and the predictions  $\hat{\mathbf{Y}} \in \mathcal{R}^{N \times T_p}$ , we use mean absolute error (MAE) as our model's loss function for training:

$$L = \frac{1}{NT_p} \sum_{n=1}^N \sum_{t=1}^{T_p} |\hat{\mathbf{Y}}_t^n - \mathbf{Y}_t^n| \quad (21)$$

## 5 Experiments

In this section, we demonstrate the effectiveness of GCN-M<sup>1</sup> with real-life traffic datasets. The experiments were designed to answer the following research questions (RQs):

- RQ 1 *Performance on raw benchmark datasets*: How well does our model perform on traffic datasets with a few missing values or without?
- RQ 2 *Complex scenarios of missing values*: How successful is our model at forecasting traffic data considering the complex missing values scenarios?
- RQ 3 *Dynamic graph modeling*: How does our method perform on dynamic graph modeling considering the missing values?
- RQ 4 *One-step processing VS two-step processing*: How will our method perform when adopting distinct missing-value processing strategies?

### 5.1 Experimental settings

**[Datasets]** We base our experiments on the public traffic datasets: PEMS-BAY and METR-LA released by Li et al. (2018), which are widely used in the literature. PEMS-BAY records six months of traffic speed on 325 sensors in the California Bay Area. METR-LA records four months of traffic flow on 207 sensors on the highways of Los Angeles County. Both datasets contain some zero and/or missing values, though PEMS-BAY has been pre-processed by the domain experts from the data provider (Caltrans 2015) to interpolate most of the missing values. Following Li et al. (2018), the datasets are split with 70% for training, 10% for validation, and 20% for testing. In order to validate the model in complex scenarios of missing values, we introduce complex missing values in the datasets (see details in Sect. 5.4). In practice, the model should forecast future values from the input data with missing values. Therefore, in the testing set, we mask out the observations from the input sequence  $\mathcal{X}$  (i.e., inject missing

<sup>1</sup> The source code is publicly available in <https://github.com/JingweiZuo/GCN-M>

values) but maintain the complete information for the target  $\mathbf{Y}$ . We use recent  $\tau = 12$  timestamps as input to predict the next  $T_p$  timestamp. Considering that the missing values are marked as zeros, we scale the input by dividing it with the max speed of the training set instead of applying Z-score normalization. This avoids changing the zero values and facilitates the computation process. Table 2 shows the summary statistics of the datasets.

**[Evaluation metrics]** The forecasting accuracy of all tested models is evaluated by three metrics: mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE).

$$\begin{aligned}
 MAE(\mathbf{Y}, \hat{\mathbf{Y}}) &= \frac{1}{NT_p} \sum_{n=1}^N \sum_{t=1}^{T_p} |\hat{\mathbf{Y}}_t^n - \mathbf{Y}_t^n| \\
 RMSE(\mathbf{Y}, \hat{\mathbf{Y}}) &= \sqrt{\frac{1}{NT_p} \sum_{n=1}^N \sum_{t=1}^{T_p} |\hat{\mathbf{Y}}_t^n - \mathbf{Y}_t^n|^2} \\
 MAPE(\mathbf{Y}, \hat{\mathbf{Y}}) &= \frac{1}{NT_p} \sum_{n=1}^N \sum_{t=1}^{T_p} \left| \frac{\hat{\mathbf{Y}}_t^n - \mathbf{Y}_t^n}{\mathbf{Y}_t^n} \right|
 \end{aligned} \tag{22}$$

and  $N$  denotes the node numbers,  $T_p$  represents the forecasting steps.

When evaluating each model's performance on the testing set, we mask out the inherent zero values in the prediction targets when computing the metrics.

We conduct statistical tests to assess the statistical significance of the differences between the models. In order to compare the forecasters over multiple datasets (Ismail Fawaz et al. 2019), we adopted the critical difference diagrams recommended by Demšar (2006) and used the Friedman test (Friedman 1940) to reject the null hypothesis (i.e., check whether there are significant differences at all). We followed the pairwise post-hoc analysis recommended by Benavoli et al. (2016) and adapted the critical difference diagrams (Demšar 2006) with the change that all forecasters are compared with pairwise Wilcoxon signed-rank test (Wilcoxon 1992). Additionally, we formed cliques using Holm's alpha (5%) correction (Holm 1979) rather than the post-hoc Nemenyi test originally used in Demšar (2006).

**[Execution and Parameter Settings]** The proposed model is implemented by PyTorch 1.6.0 and is trained using the Adam optimizer with a learning rate of 0.001. All the models are tested on a single Tesla V100 GPU of 32 Go memory. In the multi-scale memory module,  $L$ ,  $S$  are set to 12 and 5.  $n_h$ ,  $n_d$ ,  $n_w$  are all set to 2. We apply four ST blocks in which the Temporal Convolution module contains two dilated layers with dilation factor  $\mathbf{d} \in [1, 2]$ . The embedding dimension  $d$  is set to 32.

## 5.2 Baseline approaches

We only compare with the baseline models whose source code is publicly available. We follow the default parameter settings described in each paper for training each model. According to the strategy for handling missing values, the baseline models can be organized into two categories:



**Table 2** Summary statistics of PEMS-BAY and METR-LA

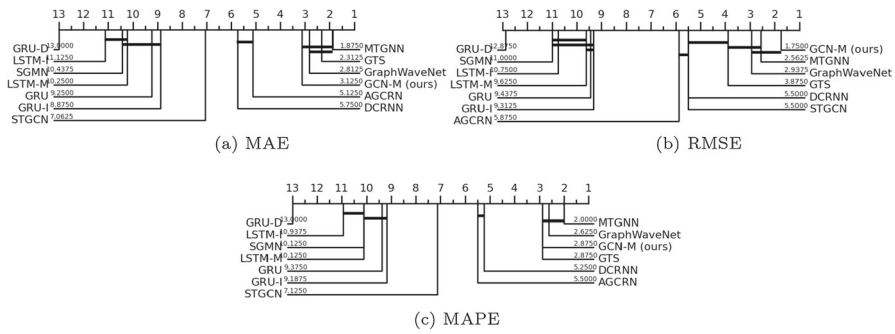
Data	#Nodes	#Edges	Length	Sample rate (mins)	Observations	Zero ratio (%)
PEMS-BAY	325	2369	52 116	5	16 937 179	0.0031
METR-LA	207	1515	34 272	5	6 519 002	8.11

1. Ignore the missing values when optimizing the model, i.e., consider missing values in the input sequence as actual zero values and mask out the missing values when computing the loss error:
  - DCRNN (Li et al. 2018): Based on the predefined graphs, DCRNN integrates GRU with dual directional diffusion convolution.
  - STGCN (Yu et al. 2018): Based on the predefined graphs, STGCN combines graph convolution into 1D convolutions.
  - Graph WaveNet (Wu et al. 2019): Graph WaveNet learns an adaptive graph and integrates diffusion graph convolutions with temporal convolutions.
  - MTGNN (Wu et al. 2020): MTGNN learns an adaptive graph and integrates mix-hop propagation layers in the graph convolution module. Moreover, it designed the dilated inception layers in temporal convolutions.
  - AGCRN (Bai et al. 2020): AGCRN learns an adaptive graph and integrates with recurrent graph convolutions with node adaptive parameter learning.
  - GTS (Shang et al. 2020): GTS learns a probabilistic graph which is combined with the recurrent graph convolutions to do traffic forecasting.

Note: In practice, any model can ignore the missing values in their optimization process. We list here some classic models and the most recent models designed specifically for traffic forecasting.
2. Jointly model the missing values and forecasting task, i.e., one-step processing models:
  - GRU (Chung et al. 2014): Gated Recurrent Unit (GRU) can be considered as a basic structure for traffic forecasting.
  - GRU-I (Che et al. 2018): A variation of GRU, which infers the missing values with the predictions from previous steps.
  - GRU-D (Che et al. 2018): Based on GRU, GRU-D helps improve the prediction performance by incorporating the missing patterns, including the masking information and time intervals between missing and observed values.
  - LSTM-I (Cui et al. 2020b): Based on LSTM, LSTM-I is similar to GRU-I for inferring the missing values.
  - LSTM-M (Tian et al. 2018): Based on LSTM, LSTM-M is designed for traffic forecasting on data with short-period and long-period missing values.
  - SGMN (Cui et al. 2020b): Based on the graph Markov process, SGMN does traffic forecasting on data with random missing values by incorporating a spectral graph convolution.

### 5.3 RQ 1: performance on raw benchmark datasets

Recently, a lot of traffic forecasting models (Jiang and Luo 2022) have been proposed, achieving remarkable performance on the benchmark datasets PEMS-BAY and METR-LA. Our objective is not to beat all the models in terms of forecasting accuracy, but to validate our proposal for jointly modeling missing values and forecasting. Therefore, it's essential to know how GCN-M performs in a primary setting, i.e., on the original datasets with a few missing values or without. We pick three classic



**Fig. 6** Critical difference diagrams showing rankings by various evaluation metrics and the statistical difference comparison of 13 forecasting models on PEMS-BAY and METR-LA datasets under multiple forecasting horizons

models (DCRNN (Li et al. 2018), STGCN (Yu et al. 2018) and Graph WaveNet (Wu et al. 2019)) and the three most recent models (MTGNN (Wu et al. 2020), AGCRN (Bai et al. 2020) and GTS (Shang et al. 2020)), which focus on the Spatio-temporal modeling of traffic data, and generally ignore the missing values when training the model. Additionally, we consider the group of works (Che et al. 2018; Cui et al. 2020b; Tian et al. 2018) which are specifically designed for modeling the missing values in the forecasting model, i.e., one-step processing models.

Tables 3 and 4 show the performance comparison on the raw PEMS-BAY and METR-LA datasets, respectively. It should be noted that the original datasets already contain missing values (0.0031% missed in PEMS-BAY, 8.11% missed in METR-LA). We train the models for single-step (horizon=1) and multi-step (horizon=3,6,12) forecasting. We report the evaluation errors on each horizon step. We observe from the results that no model achieved evident better performance than the others. However, the first group of works (e.g., DCRNN) performs better than the one-step processing models, which is not surprising as they incorporate the advanced graph models (e.g., *mix-hop propagation* (Wu et al. 2020)) and training techniques (e.g., *curriculum learning* (Wu et al. 2020)) to improve the Spatio-temporal forecasting performance. Surprisingly, among the one-step processing models, GRU-D (Che et al. 2018) shows much worse performance than the others, probably due to the fact that it has been designed for health care applications, whose data is more stable than dynamic traffic data. LSTM-M (Tian et al. 2018) and SGMN (Cui et al. 2020b), designed for traffic forecasting with missing values, show relatively good performance in PEMS-BAY especially on single-step forecasting. However, they did not show a clear advantage over the first group of works. The one-step processing models are generally designed for single-step forecasting; their performance gap with the first group of works becomes larger under a multi-step forecasting setting.

We present in Fig. 6 the critical difference diagrams (Demšar 2006) which show the average rankings and visualize the statistical difference between the forecasting models, where a thick horizontal line shows a group (i.e., clique) of models that are not significantly different in terms of evaluation metrics. From Fig. 6, we observe that even though GCN-M belongs to the one-step processing models, its performance

**Table 3** Performance comparison on raw PEMS-BAY dataset

PEMS-BAY Models	Horizon = 1 (5 mins)			Horizon = 3 (15 mins)			Horizon = 6 (30 mins)			Horizon = 12 (60 mins)		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
DCRNN	0.96	1.63	1.81	1.38	2.95	2.90	1.74	3.97	3.90	2.07	4.74	4.90
STGCN	0.98	1.84	1.98	1.44	2.88	3.16	1.85	3.82	4.20	2.21	4.52	5.09
GraphWaveNet	0.91	<b>1.56</b>	1.72	<b>1.31</b>	2.75	<b>2.73</b>	1.65	3.75	3.74	1.99	4.62	4.78
MTGNN	<b>0.87</b>	1.57	<b>1.70</b>	1.33	2.80	2.80	1.65	3.75	3.70	1.95	4.49	4.56
AGCRN	0.95	1.81	1.94	1.37	2.92	2.94	1.69	3.87	3.82	1.99	4.61	4.62
GTS	0.91	1.64	1.77	1.32	2.80	2.75	1.63	3.74	<b>3.63</b>	<b>1.90</b>	<b>4.40</b>	<b>4.44</b>
GRU	1.29	2.46	2.54	1.89	3.53	3.98	2.27	4.24	5.02	2.65	4.90	5.92
GRU-I	1.30	2.57	2.57	1.89	3.52	3.99	2.26	4.22	4.99	2.62	4.89	5.87
GRU-D	5.40	9.25	13.83	5.34	9.25	13.76	5.42	9.26	13.85	5.41	9.27	13.85
LSTM-I	1.71	2.69	2.80	1.97	3.45	4.08	2.57	5.52	5.62	2.74	5.00	6.21
LSTM-M	1.35	2.31	2.71	1.87	3.39	3.95	2.33	4.33	5.17	3.45	8.32	7.29
SGMN	0.98	1.85	1.88	1.63	3.40	3.32	2.29	4.91	4.88	3.31	6.86	7.32
GCN-M (ours)	0.91	1.57	1.75	1.33	<b>2.72</b>	2.76	<b>1.62</b>	<b>3.64</b>	3.64	1.95	<b>4.40</b>	4.61

The results show the forecasting errors at each forecasting step (i.e., horizon). The bold values represent the best results among all the models

**Table 4** Performance comparison on raw METR-LA dataset

METR-LA Models	Horizon = 1 (5 mins)			Horizon = 3 (15 mins)			Horizon = 6 (30 mins)			Horizon = 12 (60 mins)		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
DCRNN	2.45	4.21	5.99	2.77	5.38	7.30	3.15	6.45	8.80	3.60	7.60	10.50
STGCN	2.58	4.32	6.22	3.04	5.48	8.00	3.60	6.51	9.97	4.21	7.37	11.61
GraphWaveNet	2.41	4.29	5.93	<b>2.68</b>	<b>5.14</b>	6.87	3.06	6.14	8.23	3.52	7.25	<b>9.77</b>
MTGNN	<b>2.24</b>	3.92	<b>5.39</b>	<b>2.68</b>	5.16	<b>6.86</b>	<b>3.05</b>	6.16	8.19	<b>3.50</b>	<b>7.24</b>	9.83
AGCRN	2.41	4.27	6.08	2.86	5.54	7.66	3.22	6.55	8.92	3.58	7.45	10.24
GTS	2.32	4.15	6.12	2.72	5.42	7.11	3.11	6.47	<b>7.49</b>	3.52	7.49	10.07
GRU	2.83	4.56	6.78	3.48	5.80	9.02	3.97	6.74	10.72	4.65	7.86	13.00
GRU-I	2.80	4.52	6.70	3.49	5.83	9.05	3.97	6.74	10.75	4.60	7.88	12.80
GRU-D	7.46	11.82	24.55	7.43	11.85	24.62	7.45	11.84	24.62	7.47	11.86	24.68
LSTM-I	2.86	4.57	6.77	3.57	5.88	9.05	4.10	6.85	10.94	4.78	8.13	13.34
LSTM-M	3.15	5.58	7.03	3.46	5.74	8.75	4.08	6.86	10.89	4.63	7.83	12.92
SGMN	3.11	6.02	7.01	4.23	8.54	9.89	5.46	10.88	13.01	7.37	13.78	17.81
GCN-M (ours)	2.34	<b>3.89</b>	5.88	2.74	5.21	6.94	3.12	6.18	8.25	3.54	<b>7.12</b>	10.01

The results show the forecasting errors at each forecasting step (i.e., horizon). The bold values represent the best results among all the models

remains close to the first group of works. GCN-M is not significantly different to MTGNN, GTS, and GraphWaveNet on all three evaluation metrics. Moreover, the advanced graph models and training techniques in recent work (Wu et al. 2020) can be considered to have improved the performance of GCN-M further.

#### 5.4 RQ 2: complex scenarios of missing values

In this section, we demonstrate the power of GCN-M in handling complex scenarios of missing values for the purpose of traffic forecasting.

As mentioned previously in Fig. 1, there are several scenarios of missing values in real-life traffic datasets (e.g., METR-LA): short-range or long-range missing; partial or entire network missing. The results in Tables 3 and 4 did not show the superiority of GCN-M over other models on the original datasets with a low missing rate. To test the model's capability of handling complex missing values, we have designed three scenarios with various missing rates (10%, 20%, and 40%), and removed the observations from the datasets accordingly. We use  $\hat{x}_i \in \mathcal{R}^{n \times f \times t}$  to represent each of the observation tensors to be removed from  $\mathcal{X} \in \mathcal{R}^{N \times F \times \tau}$ , therefore a local mask tensor  $\hat{m}_i \in \mathcal{R}^{n \times f \times t}$  can be defined accordingly. This annotates the locations of missing values in the original dataset. All the local mask tensors constitute the global mask sequence  $\mathcal{M} = \{\hat{m}_i\}$  which allows injecting missing values with a given missing rate. Then, we designed the scenarios of:

- Short-range missing: we randomly set  $n \in [1, \dots, N]$ ,  $f = F$ ,  $t = 1$
- Long-range missing: we randomly set  $n \in [1, \dots, N]$ ,  $f = F$ ,  $t = \tau$
- Mix-range missing: we randomly set  $n \in [1, \dots, N]$ ,  $f = F$ ,  $t \in [1, \dots, \tau]$

In Tables 5 and 6, we show the performance comparison on the PEMS-BAY and METR-LA datasets under various missing value scenarios. We highlight the best results among the one-step processing models (underlined values) and all the models (bold values). Globally, GCN-M shows the best performance under all the settings when compared with other one-step processing models. The graph-based model SGMN (Cui et al. 2020b) performs much worse than other one-step processing models under long-range and mix-range missing settings, indicating that it applies only to simple missing scenarios, i.e., short-range random missing. GCN-M does not always show superiority compared with the first group of works, especially in the short-range missing scenario, where MTGNN and GTS usually show good performances. MTGNN typically performs better than GCN-M when the missing rate is low (10%), except under the mix-range missing scenario of PEMS-BAY. We can draw a conclusion from this observation: a robust Spatio-temporal forecasting model can offset the impact of the missing values to some extent, as it allows exploring the information thoroughly from the observed measures. GCN-M becomes the best forecasting model when the missing rate gets higher, as the missing values become a more critical factor that impacts the forecasting model than Spatio-temporal pattern modeling.

Compared to the short-range missing scenario, GCN-M shows a more robust performance under long-range and mix-range missing scenarios, where the recent temporal values and the nearby nodes' values are not always observed. The multi-scale memory block in GCN-M allows enriching the traffic embedding at each timestamp, thus mak-

**Table 5** Performance comparison on the **incomplete** PEMS-BAY dataset with various random settings on missing values

Models	Missing rate = 10%			Missing rate = 20%			Missing rate = 40%		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
<i>Short-range missing</i>									
DCRNN	1.76	3.94	3.94	1.82	3.96	4.01	1.85	4.26	4.04
STGCN	1.82	4.11	4.25	1.91	4.18	4.41	1.97	4.33	4.42
GraphWaveNet	1.69	3.79	3.81	1.74	3.75	3.75	1.79	3.87	3.90
MTGNN	<b>1.58</b>	<b>3.42</b>	<b>3.33</b>	1.72	3.78	3.83	1.83	4.03	3.94
AGCRN	1.65	3.81	3.78	1.66	3.81	3.79	1.72	3.96	3.95
GTS	1.65	3.86	3.74	1.65	3.86	3.76	<b>1.69</b>	3.92	<b>3.86</b>
GRU	2.60	4.64	5.75	2.67	4.78	5.90	2.86	5.10	6.37
GRU-I	2.29	4.28	5.06	2.31	4.31	5.09	2.41	4.47	5.38
GRU-D	5.38	9.29	13.84	5.46	9.36	13.96	7.20	11.58	16.91
LSTM-I	2.35	4.33	5.22	2.82	6.63	6.05	3.06	7.47	6.56
LSTM-M	2.47	4.55	5.50	2.56	4.70	5.74	3.34	7.09	7.68
SGMN	2.32	4.96	4.94	2.34	5.01	5.00	2.45	5.20	5.23
GCN-M (ours)	<u>1.62</u>	<u>3.67</u>	<u>3.60</u>	<b>1.63</b>	<b>3.73</b>	<b>3.68</b>	<u>1.75</u>	<b>3.81</b>	<u>3.90</u>
<i>Long-range missing</i>									
DCRNN	1.83	4.07	4.22	1.96	4.22	4.42	2.07	4.45	4.67
STGCN	1.92	4.22	4.42	2.03	4.37	4.72	2.14	4.52	4.76
GraphWaveNet	1.74	3.96	4.03	1.87	4.09	4.18	1.94	4.21	4.33
MTGNN	<b>1.65</b>	<b>3.68</b>	<b>3.72</b>	1.89	4.01	4.17	2.01	4.42	4.61
AGCRN	1.72	3.78	3.94	1.84	4.11	4.13	1.90	4.18	4.31
GTS	1.68	3.86	3.91	1.78	4.12	4.97	1.88	4.17	4.22
GRU	2.93	5.12	6.32	3.06	5.31	6.63	3.35	5.78	7.03

Table 5 continued

Models	Missing rate = 10%			Missing rate = 20%			Missing rate = 40%		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
GRU-I	2.52	4.51	5.33	2.53	4.57	5.73	2.71	4.82	5.51
GRU-D	9.33	14.51	22.31	9.89	13.94	22.86	11.07	15.88	23.13
LSTM-I	2.65	4.65	5.88	3.13	6.35	6.82	3.62	9.53	7.12
LSTM-M	3.93	7.25	9.17	5.45	10.06	13.67	5.57	10.12	14.59
SGMN	8.86	12.57	14.54	11.45	14.56	18.31	14.62	17.23	23.13
GCN-M (ours)	<u>1.70</u>	<u>3.75</u>	<u>3.74</u>	<u>1.73</u>	<u>3.88</u>	<u>3.92</u>	<u>1.79</u>	<u>4.07</u>	<u>4.14</u>
<i>Mix-range missing</i>									
DCRNN	1.81	4.01	4.15	1.91	4.16	4.31	2.02	4.36	4.52
STGCN	1.85	4.13	4.21	1.98	4.31	4.56	2.11	4.43	4.68
GraphWaveNet	1.72	3.92	3.96	1.83	4.06	4.14	1.89	4.11	4.21
MTGNN	1.69	3.77	3.78	1.86	4.03	4.11	1.98	4.32	4.44
AGCRN	1.67	3.85	3.88	1.72	3.95	3.99	1.80	4.10	4.13
GTS	1.70	3.96	3.92	1.75	3.98	3.89	1.79	4.09	4.09
GRU	2.71	4.88	6.03	2.82	5.08	6.28	3.05	5.43	6.82
GRU-I	2.31	4.30	5.11	2.34	4.39	5.18	2.40	4.50	5.37
GRU-D	8.90	13.71	20.03	9.46	14.50	21.04	10.21	15.19	22.44
LSTM-I	2.46	4.51	5.49	2.75	5.85	6.02	3.39	9.15	6.88
LSTM-M	3.86	7.06	8.93	5.19	9.71	13.15	5.27	9.74	13.29
SGMN	7.41	10.91	13.47	9.95	13.49	17.56	13.10	16.96	22.58
GCN-M (ours)	<u>1.65</u>	<u>3.67</u>	<u>3.69</u>	<u>1.66</u>	<u>3.72</u>	<u>3.62</u>	<u>1.69</u>	<u>3.79</u>	<u>3.83</u>

The results show a one-hour (12-step) average of the forecasting errors. The underlined values represent the best results among the one-step processing models, the bold values represent the best results among all the models



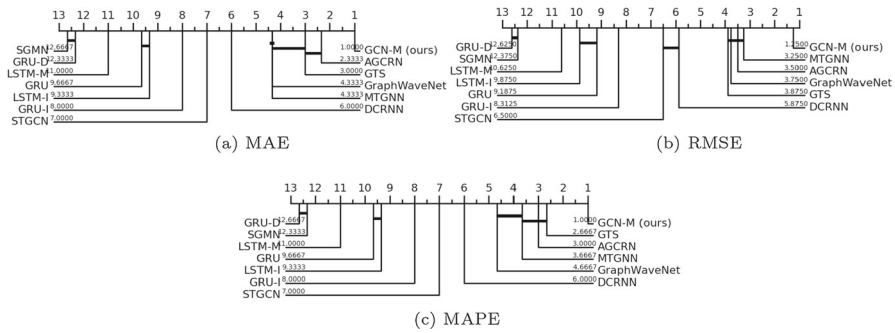
**Table 6** Performance comparison on the **incomplete** METR-LA dataset with various random settings on missing values

Models	Missing Rate = 10%			Missing Rate = 20%			Missing Rate = 40%		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPEy (%)	MAE	RMSE	MAPE (%)
<i>Short-range missing</i>									
DCRNN	3.31	6.61	9.47	3.44	6.80	9.57	3.50	6.90	9.68
STGCN	3.53	7.08	9.73	3.59	7.25	10.25	3.66	7.45	10.41
GraphWaveNet	3.28	6.60	9.11	3.36	6.74	9.50	3.45	6.81	9.57
MTGNN	<b>2.98</b>	<b>6.03</b>	<b>8.35</b>	3.19	<b>6.44</b>	8.69	3.26	6.59	9.07
AGCRN	3.19	6.47	8.81	3.24	6.60	9.01	3.25	6.61	9.19
GTS	3.08	6.40	8.59	<b>3.14</b>	6.52	<b>7.58</b>	<b>3.12</b>	6.56	<b>8.61</b>
GRU	4.20	7.09	11.27	4.27	7.16	11.42	4.45	7.41	11.94
GRU-I	4.02	6.83	10.89	4.03	6.88	10.83	4.09	6.91	10.88
GRU-D	7.50	11.87	24.69	7.45	11.86	24.66	7.53	11.91	24.76
LSTM-I	4.12	6.89	11.04	4.18	6.98	11.10	4.21	7.08	11.26
LSTM-M	4.10	6.92	10.91	4.15	6.98	11.03	4.26	7.18	11.44
SGMN	5.54	10.93	13.17	5.61	10.99	13.35	5.81	11.18	13.83
GCN-M (ours)	<u>3.17</u>	<u>6.33</u>	<u>8.72</u>	<u>3.23</u>	<u>6.47</u>	<u>8.99</u>	<u>3.26</u>	<u>6.35</u>	<u>8.98</u>
<i>Long-range missing</i>									
DCRNN	3.46	6.78	9.62	3.54	6.96	9.75	3.62	7.02	9.89
STGCN	3.71	7.20	9.91	3.76	7.39	10.42	3.88	7.66	10.67
GraphWaveNet	3.43	6.64	9.07	3.57	6.92	9.62	3.61	7.03	10.71
MTGNN	3.19	<b>6.32</b>	<b>8.48</b>	3.39	6.85	9.21	3.50	6.95	9.74
AGCRN	3.31	6.54	8.94	3.33	6.68	8.98	3.33	6.78	9.45
GTS	3.25	6.61	8.93	3.29	6.74	8.85	3.46	6.86	9.37
GRU	4.37	7.28	12.54	4.47	7.44	11.72	4.76	7.81	12.71

Table 6 continued

Models	Missing Rate = 10%			Missing Rate = 20%			Missing Rate = 40%		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPEy (%)	MAE	RMSE	MAPE (%)
GRU-I	4.20	6.91	11.78	4.09	6.97	15.42	4.21	7.03	11.43
GRU-D	7.59	11.94	24.72	7.82	12.46	25.67	7.96	12.45	26.12
LSTM-I	4.20	7.09	11.08	4.25	7.12	11.32	4.36	7.32	11.56
LSTM-M	4.53	7.47	11.88	5.21	8.84	15.34	6.08	10.02	18.13
SGMN	9.47	14.30	20.72	11.49	16.01	24.55	13.97	18.24	29.10
GCN-M (ours)	<b>3.18</b>	<u>6.39</u>	<u>8.71</u>	<b>3.23</b>	<b>6.56</b>	<b>8.78</b>	<b>3.27</b>	<b>6.68</b>	<b>9.12</b>
<i>Mix-range missing</i>									
DCRNN	3.33	6.69	9.53	3.47	6.85	9.64	3.56	6.95	9.78
STGCN	3.56	7.12	9.81	3.64	7.28	10.33	3.73	7.51	10.62
GraphWaveNet	3.28	6.51	9.02	3.43	6.78	9.52	3.51	6.94	9.62
MTGNN	<b>3.04</b>	<b>6.18</b>	<b>7.84</b>	3.14	6.72	9.07	3.44	6.82	9.12
AGCRN	3.19	6.49	8.77	3.21	6.56	8.95	3.26	6.65	8.98
GTS	3.12	6.51	8.61	3.22	6.61	8.84	3.34	6.72	8.86
GRU	4.30	7.14	11.47	4.35	7.31	11.68	4.65	7.72	12.58
GRU-I	4.05	6.83	10.94	4.01	6.86	10.83	4.11	6.97	10.98
GRU-D	7.53	11.89	24.74	7.71	12.32	25.43	7.89	12.34	25.63
LSTM-I	4.15	6.94	11.06	4.19	7.01	11.18	4.30	7.18	11.35
LSTM-M	4.40	7.38	11.91	5.14	8.77	14.92	6.02	9.92	17.88
SGMN	9.33	14.16	20.47	11.42	15.87	24.40	13.84	18.13	28.97
GCN-M (ours)	<u>3.08</u>	<u>6.34</u>	<u>8.59</u>	<b>3.12</b>	<b>6.42</b>	<b>8.71</b>	<b>3.23</b>	<b>6.50</b>	<b>8.76</b>

The results show a one-hour (12-step) average of the forecasting errors. The underlined values represent the best results among the one-step processing models, the bold values represent the best results among all the models



**Fig. 7** Critical difference diagrams showing rankings by various evaluation metrics and the statistical difference comparison of 13 forecasting models on PEMS-BAY and METR-LA datasets with various missing rates under mix-range missing scenario

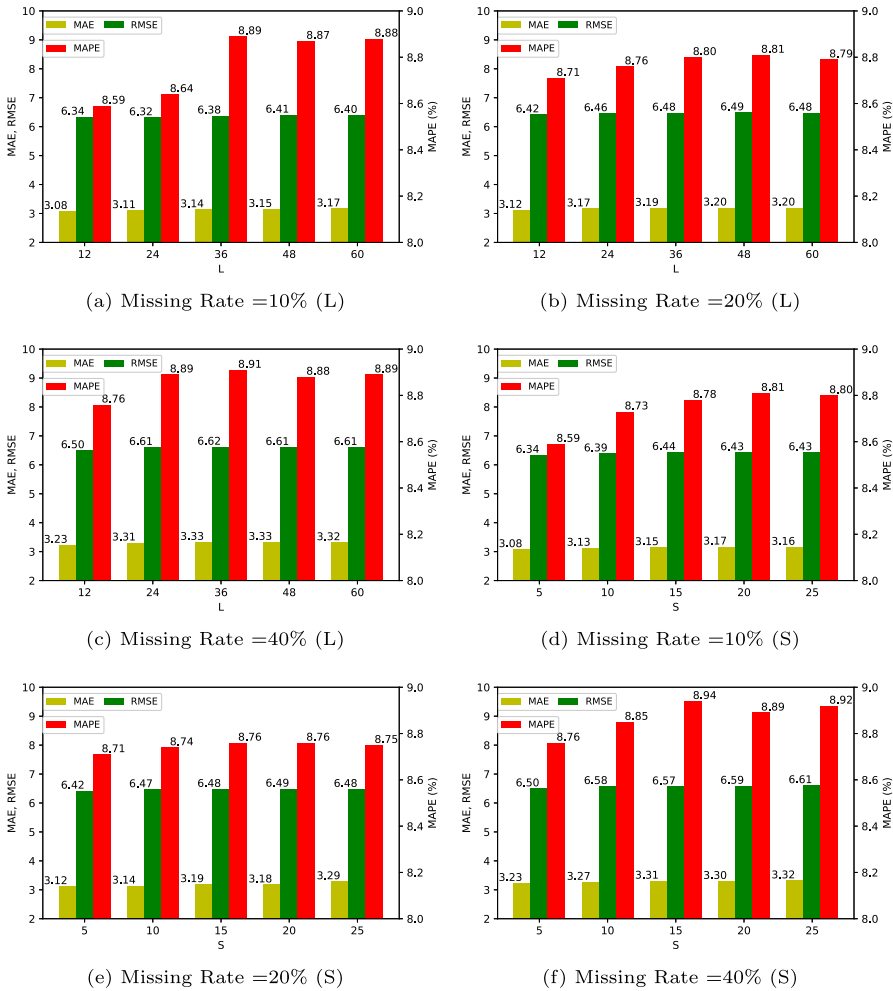
ing the model robust in the two complex scenarios. The memory block searches for the periodic global patterns from historical data and the valuable local features from nearby nodes or recent observations at each timestamp. When nearby node values are unobserved, GCN-M favors more recent observations and vice versa. As the zero values usually show periodicity while missing values show contingency (Caltrans 2015), the memory module with the periodic historical patterns can distinguish the inherent zero values from the missing values. The current node readings combined with the historical patterns will eliminate the effect of missing values but conserve that of zero values.

In Fig. 7, we show the critical difference diagrams (Demšar 2006) on PEMS-BAY and METR-LA datasets under the mix-range missing scenario. A thick horizontal line shows a group of models that are not significantly different in terms of different evaluation metrics. We can also observe that GCN-M is significantly different from other model groups on all the evaluation metrics, which validates the model’s performance for processing missing values under complex scenarios.

In Fig. 8, we show the effects of the memory module’s parameters  $L$  and  $S$  on the model’s performance. The two parameters represent the searching range of the local temporal and spatial features, respectively. We report the model’s evaluation errors with various missing rates. From the results in Fig. 8, we observe that when the searching range becomes more extensive, the model’s performance decreases more. This can be explained by the mean value of a larger space and the less recent observations will lead to a weaker information dependency with the current timestamp, thus affecting the information enrichment of the traffic embedding. In real-life datasets, we can set the parameters from a small value, such as considering local features during the last one hour ( $L=12$ ) with five nearest sensor nodes ( $S=5$ ).

**5.5 RQ 3: dynamic graph modeling**

In the dynamic traffic system, the spatial dependency can be considered as a dynamic system status, which evolves over time (Han et al. 2021b). The traffic observations at



**Fig. 8** Parameters effects: we report the model errors of GCN-M on METR-LA dataset considering (a-c)  $L$  observed samples before current timestamp for constructing the *Empirical Temporal Mean* in equation 1; (d-f)  $S$  observed samples nearby current node for constructing the *Empirical Spatial Mean* in equation 2

each timestamp are always adopted to characterize the dynamic traffic status and help learn the dynamic graphs (Li et al. 2021). However, due to the missed observations, the traffic status at certain timestamps can not be characterized, thus affecting the dynamic graph learning process.

This issue can be handled by the enriched traffic embeddings proposed in GCN-M. It allows considering the local static features and global historical patterns, which avoids the deviation introduced by the missing values and helps learn the dynamic graphs. To validate the performance of the learned dynamic graphs, we designed the following variants of our GCN-M model:

- GCN-M-obs: instead of using the enriched traffic embeddings, the raw traffic observations (Li et al. 2021) are adopted to construct dynamic graphs.
- GCN-M-adp: instead of learning dynamic graphs and applying dynamic convolution, an adaptive static graph (Wu et al. 2019) is learned to do the graph convolution.
- GCN-M-pre: instead of learning graphs from the traffic embedding or observations, the predefined graphs (Li et al. 2018) calculated with the directed distances between traffic nodes are adopted for doing graph convolution.
- GCN-M-com: combine both predefined and learned static graphs (Wu et al. 2019) to do the graph convolution.

We show in Table 7 the performance comparison on various model variants of the spatial graph modeling. We report the model errors on multiple horizons. We consider the complex scenario of mix-range missing values with a missing rate of 40% on both PEMS-BAY and METR-LA datasets. The results in Table 7 suggest that the dynamic graphs learned from the enriched traffic embeddings perform the best when compared to other variants. In contrast, the model obtains the worst performance when learning the dynamic graphs from the raw observations, which is mainly due to the missing values hindering the graph learning process in inferring the dynamic traffic status. GCN-M-obs performs even worse than GCN-M-adp in which the static graph is learned from the entire observations, eliminating the effect from local missing values.

## 5.6 RQ 4: one-step VS two-step processing

In this section, we show the performance comparison when adopting different missing-value processing strategies for traffic forecasting. As a one-step processing model, GCN-M jointly models the Spatio-temporal patterns and missing values for traffic forecasting. The two-step processing models handle the missing values in a preprocessing step, then apply a forecasting model to the completed data.

To compare the processing strategies fairly, we use GCN-M as a base model to test the two-step processing approaches. A preprocessing step is adopted to replace the GCN-M memory module designed to handle missing values.

We consider the following imputation methods to fill in missing values and apply GCN-M to the completed data.

- MEAN (García-Laencina et al. 2010): This approach replaces missing values with the mean of observed measures based on the respective feature of the input sequence.
- KNN (Batista et al. 2002): This method replaces missing values with the mean of  $k$ -nearest temporal neighbors. We linearly interpolate the missing values with  $k = 2$ , considering the previous and next non-empty values.
- MICE (Van Buuren and Groothuis-Oudshoorn 2011): MICE is the multiple imputation method that fills the missing values from the conditional distributions by Markov chain Monte Carlo (MCMC) techniques.

We show in Table 8 the performance comparison between one-step processing (i.e., GCN-M) and two-step processing (i.e., GCN-M variants) models. We conducted the experiments under the complex scenario of missing values (i.e., mix-range missing)

**Table 7** Performance comparison on graph module in mix-range missing value scenario with missing rate = 40%

Models	Horizon = 3 (15 mins)			Horizon = 6 (30 mins)			Horizon = 12 (60 mins)		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
<i>PEMS-BAY</i>									
GCN-M-obs	1.63	3.48	3.53	1.93	4.16	4.31	2.25	4.81	5.10
GCN-M-adp	1.53	3.11	3.14	1.82	3.93	4.03	2.14	4.72	4.92
GCN-M-pre	1.61	3.27	3.21	1.87	4.05	4.11	2.18	4.74	5.03
GCN-M-com	1.54	3.13	3.11	1.79	3.92	3.97	2.11	4.62	3.91
GCN-M	<b>1.45</b>	<b>3.03</b>	<b>3.09</b>	<b>1.70</b>	<b>3.81</b>	<b>3.89</b>	<b>2.06</b>	<b>4.64</b>	<b>4.86</b>
<i>METR-LA</i>									
GCN-M-obs	2.97	5.68	7.71	3.31	6.57	8.78	3.71	7.54	10.07
GCN-M-adp	2.84	5.51	7.44	3.19	6.41	8.59	3.68	7.34	9.96
GCN-M-pre	2.92	5.56	7.64	3.23	6.42	8.63	3.72	7.42	10.04
GCN-M-com	2.84	5.52	7.45	3.17	6.4	8.63	3.68	7.41	9.97
GCN-M	<b>2.82</b>	<b>5.47</b>	<b>7.42</b>	<b>3.16</b>	<b>6.38</b>	<b>8.55</b>	<b>3.58</b>	<b>7.31</b>	<b>9.92</b>

The results show the forecasting errors at each forecasting step (i.e., horizon). The bold values represent the best results among all the models

with missing rates varying from 10 to 40%. Table 8 shows that the preprocessing step with different imputation techniques leads to worse performance than the one-step setting. Even though authors in Shleifer et al. (2019) justified that the imputation techniques in the preprocessing step improve the model's performance under simple scenarios of missing values (e.g., random short-range missing), they are not applicable for complex scenarios of missing values with the results obtained from Table 8. On the one hand, the complex missing values (e.g., on short & long ranges, partial & entire variables) challenges the imputation tasks in considering local and global Spatio-temporal patterns; on the other hand, the patterns of missing values modeled in the preprocessing step are totally isolated from forecasting models, which may not be valuable for the forecasting tasks.

## 5.7 Discussions

Our approach has several advantages. First, starting from the real-world data, GCN-M considered the complex scenarios of missing values in traffic data. Different from the previous work (Che et al. 2018; Cui et al. 2020b; Tian et al. 2018) which consider the missing value from a part of the real-life scenarios: either under the short-range or long-range missing settings, under partial or entire networking missing settings. GCN-M considers the complex mix-missing value context covering various real-life scenarios for missing values.

Second, GCN-M is capable of handling such complex missing value scenarios with a multi-scale memory module. This combines local Spatio-temporal features (short-range missing, partial and entire network missing) and global historical patterns (long-range missing) to generate the enriched traffic embeddings. The embeddings allow distinguishing the inherent zero values from the missing values. In this way, GCN-M jointly models the Spatio-temporal patterns and missing values in one-step processing, which generally allows a better model performance than two-step processing (Cui et al. 2020b).

Third, GCN-M allows generating reliable dynamic graphs from the enriched traffic embeddings, which opens a path for learning robust dynamic graphs under missing value settings. Moreover, the generated dynamic graphs can cooperate with various advanced graph convolution modules (Wu et al. 2020) to improve the model's performance further.

Last but not least, even though GCN-M is designed for traffic forecasting, it is applicable to wider application domains sharing similar Spatio-temporal characteristics and missing-value scenarios, such as crowd flow forecasting (Xie et al. 2020), weather and air pollution forecasting (Han et al. 2021a; El Hafyani et al. 2022; Abboud et al. 2021), etc. The Spatio-temporal patterns in those data and the missing values caused by the sensor issues or control center errors form similar research problems to this paper.

However, GCN-M does have a limitation in terms of computational efficiency. Table 9 shows the per epoch training time comparison on the full datasets between GCN-M and the baseline models. The one-step processing baseline models are much more efficient than other models. This is basically because of their simple structure without

**Table 8** Performance comparison between one-step and two-step processing models under mix-range missing value scenario with various missing rates

Models	Missing Rate = 10%			Missing Rate = 20%			Missing Rate = 40%		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
<i>PEMS-BAY</i>									
MEAN-GCN-M	1.72	3.88	3.90	1.82	4.01	4.08	1.97	4.35	4.45
KNN-GCN-M	1.74	3.86	3.93	1.84	3.99	4.06	2.06	4.39	4.55
MICE-GCN-M	1.75	3.91	3.98	1.82	3.94	4.04	1.95	4.28	4.39
GCN-M	<b>1.65</b>	<b>3.67</b>	<b>3.69</b>	<b>1.66</b>	<b>3.72</b>	<b>3.62</b>	<b>1.69</b>	<b>3.79</b>	<b>3.83</b>
<i>METR-LA</i>									
MEAN-GCN-M	3.24	6.61	9.42	3.58	7.12	9.83	3.76	7.46	10.21
KNN-GCN-M	3.19	6.63	9.48	3.62	7.26	9.91	3.81	7.52	10.38
MICE-GCN-M	3.14	6.52	8.92	3.52	6.87	9.68	3.72	7.54	10.12
GCN-M	<b>3.08</b>	<b>6.34</b>	<b>8.59</b>	<b>3.12</b>	<b>6.42</b>	<b>8.71</b>	<b>3.23</b>	<b>6.50</b>	<b>8.76</b>

The results show a one-hour (12-step) average of the forecasting errors. The bold values represent the best results among all the models



**Table 9** Model efficiency: training time per epoch (s)

Models	PEMS-BAY	METR-LA	Models	PEMS-BAY	METR-LA
DCRNN	468.22	178.23	GRU	3.65	2.45
STGCN	55.32	27.70	GRU-I	4.22	3.67
GraphWaveNet	118.77	48.16	GRU-D	7.82	5.43
MTGNN	86.20	38.70	LSTM-I	4.32	4.64
AGCRN	67.40	32.9	LSTM-M	8.12	5.76
GTS	191.4	62.3	SGMN	3.45	2.38
GCN-M (ours)	241.69	118.65	–	–	–

integrating the costly graph convolution modules. GCN-M still performs better than DCRNN, but worse than other forecasting models. This is mainly caused by two factors: 1) generating the enriched traffic embeddings requires a huge computation cost on the attention score's calculation in the memory module; 2) generating the dynamic graphs for graph convolution requires learning a large number of parameters, thus increasing computation cost. Possible solutions might be to reduce the time complexity for calculating the attention with *ProbSparse Attention* proposed in Zhong et al. (2021), and to apply more efficient dynamic graph convolution such as graph tensor decomposition (Han et al. 2021b) and node sampling when generating the graphs (Wu et al. 2020).

## 6 Conclusion

In this paper, we propose GCN-M, a graph convolutional network-based model for handling complex missing values in traffic forecasting. We studied the complex scenario where missing traffic values occur on both short & long ranges and on partial & entire transportation networks. The enriched traffic embeddings learned by a Spatio-temporal memory module allow handling the complex missing values and constructing dynamic traffic graphs to improve the model's performance. A joint model optimization is applied to consider missing values and traffic forecasting in one-step processing. We compare GCN-M with the one-step processing models, which are specifically designed for processing incomplete traffic data and the recent advanced traffic forecasting models. The extensive experiments on two benchmark traffic datasets with 12 baselines demonstrate that GCN-M shows a clear advantage under various scenarios of complex missing values, as compared to the advanced traffic forecasting models, while at the same time maintaining comparable performance on complete traffic datasets. These experiments also provide an up-to-date comparison of the traffic forecasting models would it be with or without missing values. In future work, we will explore the aforementioned optimizations to reduce computational costs. From a longer-term perspective, one can consider noisy data or external events that may impact the predictions.

**Acknowledgements** This research was supported by DATAIA convergence institute as part of the *Programme d'Investissement d'Avenir*, (ANR-17-CONV-0003) operated by DAVID Lab, University of Versailles Saint-Quentin, and the MASTER project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie-Sklodowska Curie grant agreement N. 777695. The authors would like to thank as well the publication support from the Technology Innovation Institute.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abboud M, El Hafyani H, Zuo J, et al (2021) Micro-environment recognition in the context of environmental crowdsensing. In: Workshops of the EDBT/ICDT joint conference, EDBT/ICDT-WS
- Bai L, Yao L, Li C, et al (2020) Adaptive graph convolutional recurrent network for traffic forecasting. *Adv Neural Inf Process Syst (NeurIPS)* 33
- Batista GE, Monard MC et al (2002) A study of k-nearest neighbour as an imputation method. *His* 87(251–260):48
- Benavoli A, Corani G, Mangili F (2016) Should we really use post-hoc tests based on mean-ranks? *J Mach Learn Res (JMLR)* 17(1):152–161
- Caltrans (2015) An introduction to the caltrans performance measurement system (pems). [https://pems.dot.ca.gov/PeMS\\_Intro\\_User\\_Guide\\_v5.pdf](https://pems.dot.ca.gov/PeMS_Intro_User_Guide_v5.pdf)
- Che Z, Purushotham S, Cho K et al (2018) Recurrent neural networks for multivariate time series with missing values. *Sci Rep* 8(1):1–12
- Chung J, Gulcehre C, Cho K, et al (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 workshop on deep learning
- Cirstea RG, Yang B, Guo C (2019) Graph attention recurrent neural networks for correlated time series forecasting. *MileTS19@ KDD*
- Cui Z, Ke R, Pu Z et al (2020a) Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values. *Transp Res Part C Emerg Technol* 118(102):674
- Cui Z, Lin L, Pu Z et al (2020b) Graph markov network for traffic forecasting with missing data. *Transp Res Part C Emerg Technol* 117(102):671
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res (JMLR)* 7(1):1–30
- Dong H, Ding F, Tan H et al (2022) Laplacian integration of graph convolutional network with tensor completion for traffic prediction with missing data in inter-city highway network. *Physica A* 586(126):474
- El Hafyani H, Abboud M, Zuo J, et al (2022) Learning the micro-environment from rich trajectories in the context of mobile crowd sensing. *Geoinformatica*. <https://doi.org/10.1007/s10707-022-00471-4>
- Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. *Ann Math Stat* 11(1):86–92
- García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR (2010) Pattern classification with missing data: a review. *Neural Comput Appl* 19(2):263–282
- Guo S, Lin Y, Wan H, et al (2021) Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Trans Knowl Data Eng (TKDE)*
- Han J, Liu H, Zhu H, et al (2021a) Joint air quality and weather prediction based on multi-adversarial spatiotemporal networks. In: Proceedings of the 35th AAAI conference on artificial intelligence (AAAI)

- Han L, Du B, Sun L, et al (2021b) Dynamic and multi-faceted spatio-temporal deep learning for traffic speed forecasting. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp 547–555
- He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
- Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 65–70
- Ismail Fawaz H, Forestier G, Weber J et al (2019) Deep learning for time series classification: a review. *Data Min Knowl Discov* 33(4):917–963
- Jiang W, Luo J (2022) Graph neural network for traffic forecasting: a survey. *Expert Syst Appl* 117921
- Lea C, Flynn MD, Vidal R, et al (2017) Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 156–165
- Li Y, Yu R, Shahabi C, et al (2018) Diffusion convolutional recurrent neural network: data-driven traffic forecasting. In: International conference on learning representations (ICLR)
- Li F, Feng J, Yan H, et al (2021) Dynamic graph convolutional recurrent network for traffic prediction: benchmark and solution. *ACM Trans Knowl Discov Data (TKDD)*
- Lopez AL (2018) Traffic state estimation and prediction in freeways and urban networks. Ph.D. thesis, Université Grenoble Alpes
- Shang C, Chen J, Bi J (2020) Discrete graph structure learning for forecasting multiple time series. In: International conference on learning representations (ICLR)
- Shleifer S, McCreery C, Chitters V (2019) Incrementally improving graph wavenet performance on traffic prediction. arXiv preprint [arXiv:1912.07390](https://arxiv.org/abs/1912.07390)
- Tang X, Yao H, Sun Y, et al (2020) Joint modeling of local and global temporal dynamics for multivariate time series forecasting with missing values. In: Proceedings of the 34th AAAI conference on artificial intelligence (AAAI), pp 5956–5963
- Tian Y, Zhang K, Li J et al (2018) LSTM-based traffic flow prediction with missing data. *Neurocomputing* 318:297–305
- Van Buuren S, Groothuis-Oudshoorn K (2011) mice: multivariate imputation by chained equations in R. *J Stat Softw* 45:1–67
- Wang X, Ma Y, Wang Y et al (2020) Traffic flow prediction via spatial temporal graph neural network. *Proc Web Conf 2020*:1082–1092
- Wang S, Gao M, Wang Z et al (2021) Fine-grained spatial-temporal representation learning with missing data completion for traffic flow prediction. In: International conference on collaborative computing: networking, Applications and Worksharing, pp 138–155
- Wells BJ, Chagin KM, Nowacki AS, et al (2013) Strategies for handling missing data in electronic health record derived data. *EGEMS* 1(3)
- Weston J, Chopra S, Bordes A (2015) Memory networks. In: International conference on learning representations (ICLR)
- Wilcoxon F (1992) Individual comparisons by ranking methods. In: *Breakthroughs in statistics*. Springer, pp 196–202
- Wu Z, Pan S, Long G, et al (2019) Graph wavenet for deep spatial-temporal graph modeling. In: Proceedings of the 28th international joint conference on artificial intelligence (IJCAI), pp 1907–1913
- Wu Z, Pan S, Long G, et al (2020) Connecting the dots: multivariate time series forecasting with graph neural networks. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp 753–763
- Xie P, Li T, Liu J et al (2020) Urban flow prediction from spatiotemporal data using machine learning: a survey. *Inf Fus* 59:1–12
- Yoon J, Jarrett D, Van der Schaar M (2019) Time-series generative adversarial networks. *Adv Neural Inf Process Syst (NeurIPS)* 32
- Yu B, Yin H, Zhu Z (2018) Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the 27th international joint conference on artificial intelligence (IJCAI)
- Zhong W, Suo Q, Jia X, et al (2021) Heterogeneous spatio-temporal graph convolution network for traffic forecasting with missing values. In: 2021 IEEE 41st international conference on distributed computing systems (ICDCS), IEEE, pp 707–717

- Zhou H, Zhang S, Peng J, et al (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the 35th AAAI conference on artificial intelligence (AAAI), pp 11,106–11,115
- Zuo J, Zeitouni K, Taher Y (2021) Smate: Semi-supervised spatio-temporal representation learning on multivariate time series. In: 2021 IEEE international conference on data mining (ICDM), IEEE, pp 1565–1570

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.