International Journal of Automotive Technology, Vol. 24, No. 4, pp. 1013-1023 (2023)
DOI 10.1007/s12239–023–0083–9

pISSN 1229–9138/ eISSN 1976–3832

TRAFFIC FLOW FORECASTING OF GRAPH CONVOLUTIONAL NETWORK BASED ON SPATIO-TEMPORAL ATTENTION MECHANISM

Hong Zhang*, Linlong Chen, Jie Cao, Xijun Zhang, Sunan Kan and Tianxin Zhao

College of Computer & Communication, Lanzhou University of Technology, Lanzhou 730050, China

(Received 16 May 2022; Revised 7 February 2023; Accepted 2 March 2023)

ABSTRACT–Accurate traffic flow forecasting is a prerequisite guarantee for the realization of intelligent transportation. Due to the complex time and space features of traffic flow, its forecasting has always been a research hotspot in this field. Aiming at the difficulty of capturing and modelling the temporal and spatial correlation and dynamic features of traffic flow, this paper proposes a novel graph convolutional network traffic flow forecasting model (STAGCN) based on the temporal and spatial attention mechanism. STAGCN model is mainly composed of three modules: Spatio-temporal Attention (STA-Block), Graph Convolutional Network (GCN) and Standard Convolutional Network (CN), model the periodicity, spatial correlation and time dependence of traffic flow respectively. STA-Block module models the spatio-temporal correlation between different time steps through the spatio-temporal attention mechanism and gating fusion mechanism, and uses GCN and CN to capture the spatial and temporal features of traffic flow respectively. Finally, the output of the three components is predicted through a gated fusion mechanism. A large number of experiments have been conducted on two data sets of PeMS. The experimental results demonstrate that compared with the baseline method, the STAGCN model proposed in this paper has better forecasting performance.

KEY WORDS : Traffic flow forecasting, Spatio-temporal attention mechanism, Graph convolutional network, Spatio-temporal correlation, Gated fusion mechanism

NOMENCLATURE

G	: traffic road network
V	: group of road nodes
N	: number of nodes
E	: set of edges
A	: adjacency matrix
τ	: time slices
Y	: future traffic flow sequence
i	: road node
t	: time, min
f	: traffic flow sequence
D	: value of the features of all nodes
X_t^i	: eigenvalues of node i at time t
X_t	: feature values of nodes at the moment of t
y_t^i	: traffic flow of node i at time t
T_h	: input of the recent component
T_d	: input of the daily-period component

1. INTRODUCTION

Traffic flow forecasting is an important part of intelligent traffic system (ITS), which can provide a scientific basis for the management and planning of urban traffic systems (Cui *et al*., 2019). Accurate and timely traffic flow forecasting based on historical observations will help road users to make better travel plans, alleviate traffic congestion, and improve traffic operation efficiency (Lv *et al*., 2015).

Traffic flow forecasting is a widely studied problem. Early traffic flow forecasting methods are usually statistical methods based on time series (Kumar and Vanajakshi, 2015) or simple machine learning methods (Kumar *et al*., 2013). In practice, these methods are difficult to deal with non-linear traffic flow data, and it is hard to consider the spatio-temporal correlation of high-dimensional traffic flow data at the same time. In recent years, traffic flow forecasting methods based on deep learning have been extensively studied (Polson and Sokolov, 2017). Some researchers (Zhang *et al*., 2016a) models the traffic network as a grid and use a convolutional neural network (CNN) to capture the spatial correlation. However, due to the irregularity of the road, the use of grid for modeling will lose the topological information in the transportation network. In response to this problem, researchers integrated graph neural networks (GNN) that can effectively capture non-Euclidean distances into recurrent neural networks (RNN) (Chen *et al*., 2020) or CNN (Yu *et al*., 2017), capture the temporal and spatial features of traffic flow by gathering information from neighboring nodes.

Although the use of deep learning models in traffic flow

**Corresponding author*. e-mail: zhanghong@lut.edu.cn forecasting considers spatial correlation and temporal dependence, the existing methods have two main limitations. On the one hand, the spatial correlation between different locations only depends on the similarity of historical traffic flow (Zhang *et al*., 2016a) and the static spatial correlation of model learning. However, the time dependence between different locations will change over time. On the other hand, many existing studies ignore the long-term cyclical dependence. Traffic flow data show strong periodicity, and this periodic feature has a great effect on forecasting. However, traffic flow data are not strictly periodic. For example, the peak time of the working day usually occurs in the afternoon, but on different days, the peak time may vary from afternoon to evening. Regardless of the fact that studies (Zhang *et al*., 2016b) have considered periodicity, they have not considered the dynamic randomness of the sequence.

In order to capture the complex spatio-temporal correlation of traffic flow, this paper proposes a traffic flow forecasting model based on spatio-temporal attention mechanism graph convolutional network to predict traffic flow. The model is composed of spatio-temporal attention module and spatio-temporal convolutional network module. The spatio-temporal attention module is composed of a spatio-temporal attention mechanism and a gating fusion mechanism. The spatial attention mechanism is used to capture the spatial correlation between different sensors, and the temporal attention mechanism is used to capture the dynamic time dependence between different time. The gated fusion mechanism is used to adaptively fuse information extracted through the spatio-temporal attention mechanism. Then, the spatio-temporal convolutional network is used to capture the spatio-temporal correlation of the traffic flow. Finally, the output of the three components is predicted through a gated fusion mechanism.

The main contributions of this article are summarized as follows:

1) This paper proposes a novel spatial attention mechanism and time attention mechanism to learn the dynamic spatial correlation and nonlinear time dependence of traffic flow data respectively. In addition, this paper designs a gated fusion mechanism to adaptively fuse the information extracted through the spatio-temporal attention mechanism to reduce the propagation of errors in the forecasting process.

2) This paper proposes a novel spatio-temporal convolutional network to capture the spatio-temporal correlation of traffic flow. The network consists of a graph convolutional network and a standard convolutional network. The graph convolutional network is used to the capture spatial correlation, and the standard convolutional network is used to capture temporal dependence.

3) This paper conducts a large number of comparative experiments on two sets of traffic data. The experimental results demonstrate that compared with the existing baseline methods, the model in this paper has achieved excellent forecasting performance on different data sets.

2. RELATED WORKS

After countless research and practice, traffic flow forecasting has made numerous research results in the past few decades. In the field of time series, autoregressive integrated moving average model (ARIMA) (Lippi *et al*., 2013) and Kalman filter model (Shekhar and Williams, 2007) have been extensively used in the field of traffic flow forecasting. These early methods separately studied the time series of traffic flow at each location. In recent years, some studies have begun to consider the spatial information of traffic flow, such as similar traffic road networks in different locations (Deng *et al*., 2016) and external environmental information (Tong *et al*., 2017), such as Weather conditions, traffic incidents, etc. (Wu *et al*., 2016). However, these methods are still based on traditional time series models or machine learning models, and cannot well capture complex nonlinear spatio-temporal dependencies.

In recent years, deep learning has achieved great success in many challenging learning tasks (Lecun *et al*., 2015). For example, a number of studies (Zhang *et al*., 2016a) models the city's traffic flow as a heat map image and use CNN to model nonlinear spatial correlation. In order to model nonlinear time dependence, Cui *et al*. (2018) proposed a traffic flow forecasting model based on a recurrent neural network (RNN). Yao *et al*. (2018) further proposed a model combining CNN and long-term short-term memory (LSTM) (Wang *et al*., 2019) to model the temporal and spatial correlation of traffic flow.

Some research work uses temporal convolutional networks (Sen *et al*., 2019) to enable the model to process longer sequences in less time. However, these studies did not explicitly model the interdependence between different time series. A recent study (Xu *et al*., 2020) uses the transformer architecture for traffic flow forecasting. Due to the large number of trainable parameters, such work usually requires a large number of training samples (Zhang *et al*., 2019).

At present, deep learning methods have been widely used in various traffic tasks. DCRNN (Li *et al*., 2017) defines the spatial correlation of traffic flow data as a diffusion process, and extends the previous GCN (Defferrard *et al*., 2016) to a directed graph. Graph WaveNet (Wu *et al*., 2019) combines GCN with expanded causal convolutional network, and proposes an adaptive adjacency matrix as a supplement to the predefined adjacency matrix to capture spatial correlation. Huang *et al*. (2014) used deep belief network (DBN) to learn the effective features of traffic flow forecasting in an unsupervised manner. Li *et al*. (2017) proposed a hybrid model based on GCN, which captures the spatial correlation of random walks on the traffic network and uses LSTM to capture the time dependence of traffic flow. Yu *et al*. (2017) proposed a spatio-temporal graph convolutional network (STGCN), which also uses a convolutional structure to extract the spatio-temporal features of traffic flow. In recent years, some studies, such as STSGCN (Song *et al*., 2020) and

GMAN (Zheng *et al*., 2019), have also added more complex spatial and temporal attention mechanisms with GCN to capture the temporal and spatial features of traffic flow. However, these methods can only capture the sharing patterns between traffic flow sequences and rely on predefined spatial connection graphs.

In addition, many researchers use deep learning methods to process high-dimensional traffic flow data, that is, RNN is used to effectively extract the time features of traffic flow data, and GCN is used to extract the spatial features of road network graph data. However, these methods are still unable to simultaneously model the temporal and spatial features and dynamic correlation of traffic flow data.

In recent years, due to its high efficiency and flexibility in dependency modeling, the attention mechanism has been widely used in various fields (Shen *et al*., 2018). The core idea of the attention mechanism is to adaptively focus on the most relevant features according to the input data (Cheng *et al*., 2018). For example, Liang *et al*. (2018) proposed a multi-level attention network to adaptively adjust the correlation between multiple sensor time series. Xu *et al*. (2015) proposed two attention mechanisms in the image description task, and used a visual method to intuitively demonstrate the effect of the attention mechanism. Researchers apply the attention mechanism to graph structured data (Veličković *et al*., 2017) to model the spatial correlation of graph classification.

In summary, there are still huge challenges in traffic flow forecasting, mainly as follows:

1) Dynamic spatial correlation: The correlation of traffic conditions between sensors in a road network changes over time, such as during peak and off-peak periods. How to dynamically select relevant sensor data to predict the traffic conditions of long-term target sensors is a challenging problem.

2) Non-linear time dependence: The traffic situation where the sensor is located may suddenly change to affect the correlation between different points in time, such as a traffic accident. When the forecasting time point is longer, how to model the nonlinear time dependence is still a huge challenge.

3) Sensitivity to error propagation: In long-term traffic flow forecasting, the errors in each forecasting step may be superimposed in further forecasts. This error propagation makes future forecasts more challenging.

Driven by the above research, considering the topological structure of the traffic network and the dynamic spatio-temporal pattern of traffic data, this paper uses graph convolutional network and attention mechanism to model the traffic flow data of the network structure at the same time. The temporal and spatial attention mechanism is used to reduce the sensitivity to error propagation, and the graph convolutional network and standard convolutional network are used to model the dynamic correlation and nonlinear time dependence of traffic flow.

3. MODEL BUILDING

In this research, the transportation network is defined as an undirected graph $G = (V, E, A)$, where V is a group of road nodes, $V=\{v_1, v_2, ..., v_N\}$, N is the number of nodes, and E is a set of edges, $A \in R^{N\times N}$ represents the adjacency matrix of graph G. In undirected graph G, the adjacency matrix represents the matrix of adjacency relations between vertices, which is symmetrical, the main diagonal must be zero, and the degree of any vertex i is the number of all non-zero elements in the i column (or the i row). Each node on the transportation network G detects F measured values at the same sampling frequency, that is, each node generates a feature vector of length F on each time slice, as shown in Figure 1.

Suppose the f-th time series on each node in the traffic network G is the traffic flow sequence, $f \in (1, ..., F)$. $x_t^{c,i} \in R$ represents the c-th eigenvalue of node i at time t, and $X_t^i \in R^F$ represents all the eigenvalues of node i at time t. $X_t = \left(X_t^1, X_t^2, ..., X_t^N\right)^T \in R^{N\times N}$ represents all the feature values of all nodes at the moment of t. $D = \left(X_1, X_2, ..., X_\tau\right)^T \in R^{N\times F\times\tau}$ represents the value of all the features of all nodes in τ time slices. Let $y_t^i = x_t^{f,i} \in R$ represent the traffic flow of node i at time t.

Through the historical measurement values of all nodes on the road network in the past τ time slices, the future traffic flow sequence $Y = \left(y^1, y^2, ..., y^N\right)^T \in R^{N\times T_p}$ of all nodes

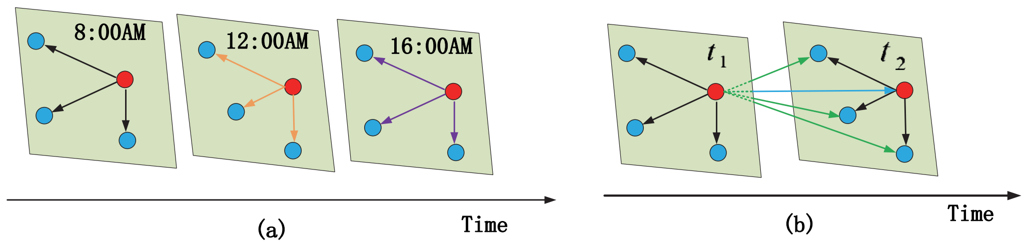


Figure 1. (a) Spatiotemporal structure of traffic data; (b) Influence of red nodes in the spatiotemporal network.

on the entire traffic network on the next time slice T_p can be predicted, where $Y=\left(y^1,y^2,...,y^N\right)^T\in R^{N\times T_p}$ represents the traffic flow from $\tau+1$ to node *i*.

Figure 2 shows in the overall framework of the STAGCN model proposed in this paper. The model is composed of three independent parts with the same structure, respectively modeling the recent, daily-period and weekly-period dependencies of historical traffic flow data.

Assuming that the sampling frequency of the detector is *q* times a day, the current time and forecasting window size are t_0 and T_p, respectively. In this paper, three time series segments with lengths of T_h, T_d and T_w are respectively intercepted in chronological order and used as the input of the recent, daily-period and weekly-period components respectively, where T_h, T_d and T_w are all integer multiples of T_p. The details of the three time series segments are as follows.

(1) The recent segment:
The recent segment is a historical time series directly adjacent to the forecast period. Since the formation of traffic congestion is a gradual process, historical traffic flow will inevitably have an impact on future traffic flow.

$$D_h=\left(X_{t_0-T_h+1},X_{t_0-T_h+2},...,X_{t_0}\right)\in R^{N\times F\times T_h} \tag{1}$$

(2) The daily-periodic segment:
It is composed of the same time period as the forecast period in the historical time series. Due to daily routines, traffic data may show repetitive patterns, such as daily morning rush hours and evening rush hours. The day time period part is to model the daily periodicity of traffic flow data.

$$D_d=\Big(X_{t_0-\left(\frac{T_d}{T_p}\right)*q+1},...,X_{t_0-\left(\frac{T_d}{T_p}\right)*q+T_p},X_{t_0-\left(\frac{T_d}{T_p}-1\right)*q+1},...,$$
$$X_{t_0-\left(\frac{T_d}{T_p}-1\right)*q+T_p},...,X_{t_0-q+1},...,X_{t_0-q+T_p}\Big)\in R^{N\times F\times T_d} \tag{2}$$

(3) The weekly-periodic segment:
It is composed of the time periods of the past few weeks, and they have the same weekly period attributes and time intervals as the forecast period.

$$D_w=\Big(X_{t_0-7*\left(\frac{T_w}{T_p}\right)*q+1},...,X_{t_0-7*\left(\frac{T_w}{T_p}\right)*q+T_p},X_{t_0-7*\left(\frac{T_w}{T_p}-1\right)*q+1},...,$$
$$X_{t_0-7*\left(\frac{T_w}{T_p}-1\right)*q+T_p},...,X_{t_0-7*q+1},...,X_{t_0-7*q+T_p}\Big)\in R^{N\times F\times T_w} \tag{3}$$

These three parts have the same network structure, and each part is composed of multiple STA-Block, GCN, CN and a fully connected layer stacked together. Each STA-Block has a spatial attention mechanism module, a temporal attention mechanism module and a gated fusion mechanism module. Among them, the data after the spatio-temporal attention mechanism dynamically assigns different weights is input to GCN and CN to capture the spatio-temporal features of traffic flow, and the output of GCN and CN is input to FC to ensure that each output has the same size and shape as the predicted target. Finally, based on the parameter matrix, the output of the three partial components is fused through a gated fusion mechanism to obtain the final forecasting result.

3.1. Description of Optimisation Problem

In the spatial dimension, the traffic condition on one road will be affected by other different roads, and this influence is very dynamic. To model these attributes, this paper designs a spatial attention mechanism to adaptively capture the correlation between different geographic locations in the road network. The spatial attention mechanism aims to dynamically assign different weights to different vertices (such as sensors) at different time points.

Take the spatial attention mechanism of the recent component as an example:

$$S=V_s\cdot sigmoid\left(\left(D_h^{(r-1)}W_1\right)W_2\left(W_3D_h^{(r-1)}\right)^T+b_s\right) \tag{4}$$

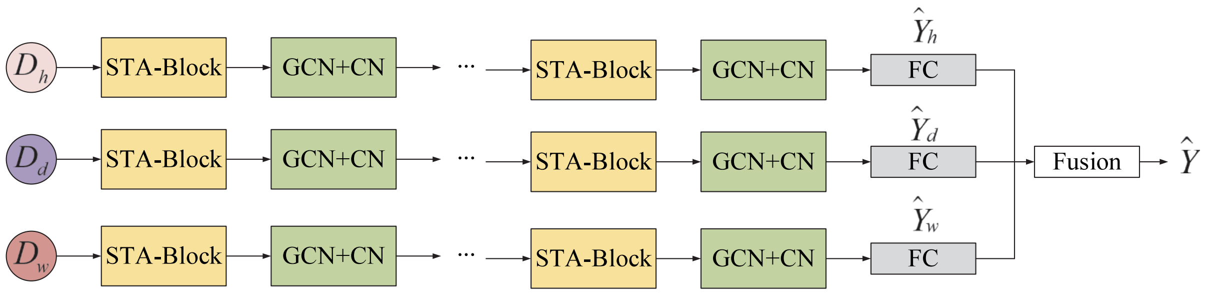


Figure 2. Frame diagram of STAGCN model: STA-Block, Spatio-temporal attention block; GCN, Graph convolutional network, CN: Standard convolutional network, FC: Fully connected layer.

$$S_{i,j}^{'} = \frac{exp\left(S_{i,j}\right)}{\sum_{j=1}^{N} exp\left(S_{i,j}\right)} \tag{5}$$

Among them, $D_h^{(r-1)} = \left(X_1, X_2, ..., X_{T_{r-1}}\right) \in R^{N \times C_{r-1} \times T_{r-1}}$ is the input of the r-th STA-Block, C_{r-1} is the channels of the input data in the r-th layer. When $r = 1$ and $C_0 = F$, T_{r-1} is the length of the time sequence in the rth layer. When $r = 1$, in the recent component $T_0 = T_h$, V_s, $b_s \in R^{N \times N}$, $W_1 \in R^{T_{r-1}}$, $W_2 \in R^{C_{r-1} \times T_{r-1}}$, $W_3 \in R^{C_{r-1}}$ is a learnable parameter, and the softmax function is used to constrain the total attention weight of the node to 1. The spatial attention matrix S is dynamically calculated from the current input of the layer. In the spatial attention matrix S, $S_{i,j}$ represents the strength of the correlation between node i and node j.

3.2. Temporal Attention Mechanism

The traffic situation at a certain time has a correlation with its previous traffic situation, and this correlation shows a non-linear change as the time step increases. To model these attributes, this paper designs a temporal attention mechanism to adaptively model the nonlinear correlation between different time points.

Take the recent component's time attention mechanism as an example:

$$E = V_e \cdot sigmoid\left(\left(\left(D_h^{(r-1)}\right)^T U_1\right) U_2 \left(U_3 D_h^{(r-1)}\right) + b_e\right) \tag{6}$$

$$E_{i,j}^{'} = \frac{exp\left(E_{i,j}\right)}{\sum_{j=1}^{T_{r-1}} exp\left(E_{i,j}\right)} \tag{7}$$

Among them, V_e, $b_e \in R^{T_{r-1} \times T_{r-1}}$, $U_1 \in R^N$, $U_2 \in R^{C_{r-1} \times N}$, $U_3 \in R^{C_{r-1}}$ is a learnable parameter. The time attention matrix E is determined by the input. In the time attention of matrix S, $S_{i,j}$ represents the strength of the dependence between time i and j. The attention matrix E is normalized by the softmax function. Taking the normalized time attention matrix E as input, we get $\widehat{D}_h^{(r-1)} = (\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_{T_{r-1}}) = (X_1, X_2, \ldots, X_{T_{r-1}}) E^{'} \in R^{N \times C_{r-1} \times T_{r-1}}$, dynamically adjust the input by fusing relevant information.

3.3. Gating Fusion Mechanism

The traffic condition of a road at a specific point in time is correlated with its previous measurement values and the traffic conditions of other roads. As showed in Figure 3, this paper designs a gated fusion mechanism to adaptively merge the spatial attention mechanism and the temporal attention mechanism.

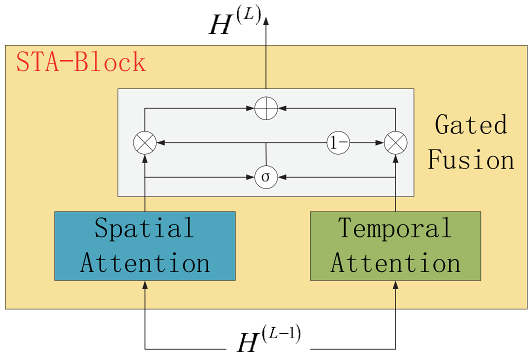


Figure 3. STA-Block structure diagram: STA-Block integrates the spatial attention mechanism and the temporal attention mechanism through a gated fusion mechanism.

In the L-th STA-Block, the output of the spatial attention mechanism and the temporal attention mechanism are expressed as $H_S^{(L)}$ and $H_T^{(L)}$, respectively. $H_S^{(L)}$ and $H_T^{(L)}$ are fused by Equation (8):

$$H^{(L)} = z \cdot H_S^{(L)} + (1 - z) \cdot H_T^{(L)} \tag{8}$$

$$z = sigmoid\left(H_S^{(L)} W_{z,1} + H_T^{(L)} W_{z,2} + b_z\right) \tag{9}$$

Among them, $W_{z,1} \in R^{D \times D}$, $W_{z,2} \in R^{D \times D}$, $b_z \in R^D$ is a learnable parameter, z is the gate. The fusion mechanism adaptively controls the spatial correlation and time dependence of the traffic flow in each node and the time step.

3.4. Graph Convolutional Networks and Convolutional Networks

STA-Block makes the network adaptively pay more attention to valuable information. The output after the fusion of the gated fusion mechanism is input to the GCN and CN modules, the spatial correlation of the neighborhood is captured by GCN, and the temporal dependence of the neighboring time is captured by CN.

To make full use of the topological features of the transportation network, this paper uses the graph convolutional network based on the spectrogram theory to directly process the signal on each time slice. The spectrum method converts the transportation network graph into algebraic form to analyze the topological properties of the graph. Graph convolution is a convolution operation realized by using a linear operator, which replaces the classic convolution operator by diagonalization in the Fourier domain. Therefore, the signal x on the graph G is filtered by a kernel $g\theta$:

$$g\theta * Gx = g\theta(L)x = g\theta\left(U \Lambda U^T\right)x = Ug\theta(\Lambda)U^T x \tag{10}$$

Among them, *G is the graph convolution operation, where the graph Fourier basis $U \in R^{n\times n}$ is the normalized graph Laplacian $L = I_N - D^{-(1/2)}WD^{-(1/2)} = U\Lambda U^T \in R^{n\times n}$ eigenvector matrix, I_N is an identity matrix, $D \in R^{n\times n}$ is a pair angle matrix, where $D_{ii} = \sum_j W_{ij}$; $\Lambda \in R^{n\times n}$ is a diagonal matrix of the eigenvalues of L, and filter $g\theta(\Lambda)$ is also a diagonal matrix.

However, when the scale of the graph is large, it is difficult to directly perform eigenvalue decomposition on the Laplacian matrix. Therefore, this paper uses Chebyshev polynomial approximation to effectively solve this problem (Simonovsky and Komodakis, 2017):

$$g\theta * Gx = g\theta\left(L\right)x = \sum_{k=0}^{K-1}\theta_k T_k\left(\tilde{L}\right)x \tag{11}$$

Among them, parameter $\theta \in R^K$ is a vector of polynomial coefficients, $\tilde{L} = (2/\lambda_{max})L - I_N$, λ_{max} is the largest eigenvalue of the Laplace matrix. The recursive definition of Chebyshev polynomial is $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, $T_0(x) = 1$, $T_1(x) = x$. The graph convolution module uses the $ReLU$ function as the final activation function.

To dynamically adjust the correlation between nodes, for each term of the Chebyshev polynomial, this paper multiplies $T_k(\tilde{L})$ with the spatial attention matrix $S' \in R^{N\times N}$, and then obtains $T_k(\tilde{L})*S'$, where * means the Hadamard product. Therefore, Equation (11) can be rewritten as Equation (12):

$$g\theta * Gx = g\theta\left(L\right)x = \sum_{k=0}^{K-1}\theta_k \left(T_k\left(\tilde{L}\right) * S'\right)x \tag{12}$$

After the graph convolution operation captures the neighboring information of each node in the graph, the standard convolutional layer is further stacked to update the information of the node by fusing the information on adjacent time slices. Take the operation on the r-th layer in the recent component as an example:

$$D_h^{(r)} = ReLU\left(\mu * \left(ReLU\left(g\theta * G\widehat{D}_h^{(r-1)}\right)\right)\right) \in R^{C_r\times N\times T_r} \tag{13}$$

Among them, μ is the parameter of the standard convolution kernel, and $ReLU$ is the activation function.

When fusing the output of different components, the influence weights of the three components of each node are different. To improve the accuracy of forecasting, they should be learned from historical data. Therefore, the final forecasting result after fusion is:

$$\hat{Y} = W_h * \hat{Y}_h + W_d * \hat{Y}_d + W_w * \hat{Y}_w \tag{14}$$

Among them, W_h, W_d and W_w are learnable parameters, reflecting the degree of influence of the three time dimension components on the forecasting target.

In summary, the spatio-temporal attention mechanism and the gating fusion mechanism form the STA-Block, and the GCN and CN modules can well capture the spatial and temporal features of traffic flow data. Multiple STA-Block, GCN and CN modules are superimposed to further extract a larger range of dynamic spatio-temporal correlation. Finally, through FC, and using ReLU as the activation function, to ensure that the output of each component has the exact same size and shape as the predicted target.

3.5. Loss Function

The purpose of training is to minimize the error between the actual traffic speed and the predicted traffic speed in the road network. In this paper, the mean square error (MSE) is used as the loss function. The actual traffic speed and predicted traffic speed of different road sections are represented by y_i and $\hat{y}_i$ respectively, and n is the number of samples.

Therefore, the loss function of the STAGCN model is shown in Equation (15):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{15}$$

4. EXPERIMENT

4.1. Data Description

This paper verifies the forecasting performance of the STAGCN model proposed in this paper on two California highway traffic data sets PeMSD04 and PeMSD08, which are collected in real time every 30 seconds by the Caltrans Performance Measurement System (PeMS) (Chao *et al.*, 2000). The experimental traffic data set contains different attributes, such as location, date, period, speed, and traffic volume. The detailed information of the experimental data set is shown in Table 1.

Table 1. Description of experimental data set.

Datasets	Number of sensors	Time range
PeMSD04	307	1/1/2018-2/28/2018
PeMSD08	170	7/1/2016-8/31/2016

PeMSD04: This data set contains 3848 detectors on 29 roads. The time span of this data set is from January to February 2018. This article selects the first 50 days of data as the training set, and the rest as the test set.

PeMSD08: This data set contains 1979 detectors on 8 roads. The time span of this data set is from July to August 2016. This article selects the first 50 days of data as the training set, and the last 12 days of data as the test set.

In this paper, some redundant detectors are deleted. Among them, there are 307 detectors in PeMSD04 and 170 detectors in PeMSD08. Traffic data are aggregated every 5 minutes, so each detector contains 288 data points per day. For the missing data in the data set, this paper uses linear interpolation method (Blu *et al*., 2004) to fill in the missing values. Before entering the data into the predictive model, this article uses the Z-score standardization method to process the data so that the average value is 0. The standardized formula is:

$$Z = \frac{x-\mu}{\sigma} \tag{16}$$

Where μ represents the mean of all sample data, σ represents the standard deviation of all sample data, and Z represents the standardized input data.

4.2. Experimental Environment and Parameter Settings

This experiment is compiled and run on a Linux server (CPU: Intel(R)Xeon(R)CPU E5-2620 v4 @ 2.10GHz, GPU: NVIDIA GeForce GTX 1080). Based on the MXNET deep learning framework, the traffic flow forecasting model training is completed in the PyCharm development environment.

In the model in this article, the Adam Optimizer is used to train the model. According to Kipf's research (Kipf and Welling, 2016), the term number $K \in \{1,2,3\}$ of Chebyshev polynomial, with the increase of *K*, the prediction performance is slightly improved. The same is true of the kernel size in the time dimension. Considering the computational efficiency and prediction performance, this paper sets $K = 3$, and the kernel size along the time dimension is 3. In this paper, the hyperparameter is optimized by optimizer, learning rate, epoch, batch size and so on. In the training phase, the initial learning rate is manually set to 0.001 and the batch size is 64. All image convolutional layers and standard convolutional layers use 64 convolution kernels, and the size of the forecasting window T_p is 12. This article uses one-hour historical data to predict the traffic flow of the next hour, that is, past 12 continuous time steps are used to predict the future 12 continuous time steps.

- Baseline Methods

(1) LSTM (Yang *et al*., 2019): Long and short-term memory network for time series forecasting.

(2) STGCN (Yu *et al*., 2017): Spatio-temporal graph convolutional network uses ChebNet and 2D convolutional network to capture spatial correlation and temporal dependence respectively.

(3) DCRNN (Li *et al*., 2017): Diffusion convolutional recurrent neural network, which uses two-way random walks to model spatial correlation, and uses encoder-decoder architecture to model time dependence for traffic flow forecasting.

(4) Graph WaveNet (Wu *et al*., 2019): Graph WaveNet combines graph convolutional networks and extended causal convolutional networks to capture the temporal and spatial correlation of traffic flow.

(5) STSGCN (Song *et al*., 2020): Spatio-temporal synchronization graph convolutional network, which uses the local spatio-temporal subgraph module to independently model the local spatio-temporal correlation.

4.3. Model Evaluations

To better analyze the experimental results and evaluate the forecasting performance of the model, this paper evaluates the error between the actual traffic flow speed and the predicted results based on the following indicators:

(1) Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{17}$$

(2) Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{18}$$

(3) Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{19}$$

In the formula, y_i and $\hat{y}_i$ represent actual traffic speed and predicted traffic speed, respectively. n is the number of observations. Use MAE, RMSE and MAPE to measure the forecasting error. The smaller the forecasting value, the better the forecasting effect.

4.4. Experimental Results and Analysis

In this paper, the STAGCN model is tested on the PeMSD04 and PeMSD08 data sets, and compared with eight baseline methods. Table 2 shows the forecasting performance of the STAGCN model and different baseline models on the PeMSD04 and PeMSD08 data sets. It can be seen from Table 2 that the STAGCN model in this paper show the best predictive performance on the two sets of data.

For example, on the PeMSD08 data set, compared with the SVR model, the MAE of the STAGCN model, DCRNN, STGCN, STSGCN, and Graph WaveNet model are reduced by about 33.76 %, 30.01 %, 28.08 %, 33.52 %, 21.08 %, and RMSE respectively. It is about 30.28 %, 29.77 %, 29.07 %, 29.31 %, 16.25 %. Compared with the LSTM model, the MAPE of the STAGCN model, DCRNN, STGCN, STSGCN and Graph WaveNet model are reduced by about 3.01 %, 3.31 %, 3.29 %, 3.76 %, 1.54 %, respectively. As HA, SVR, VAR and LSTM models only consider time dependence,

Table 2. Performance comparison of different traffic flow prediction models in two datasets.

Baseline methods	Datasets	PeMSD04 (60 min)			PeMSD08 (60 min)		
	Metrics	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE
HA		38.03	27.88	59.24	34.86	24.07	52.04
SVR		28.71	19.21	44.57	23.26	14.75	36.18
VAR		24.54	17.24	38.61	23.46	15.42	36.33
LSTM		27.34	18.60	41.80	22.38	14.79	34.38
STGCN		23.34	14.80	36.30	18.16	11.50	28.03
DCRNN		24.92	17.49	38.38	17.89	11.48	27.88
Graph WaveNet		25.48	17.53	39.74	19.21	13.25	31.12
STSGCN		21.29	13.95	33.85	17.42	11.03	27.98
STAGCN		21.07	14.56	33.62	17.39	11.78	27.77

Table 3. Forecasting performance of the STAGCN model and the two variant models at different time points.

Baseline methods	Datasets	PeMSD04 (15/30/60 min)			PeMSD08 (15/30/60 min)		
	Metrics	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE
Without attention		26.57/26.65/26.77	18.95/18.95/19.04	42.70/42.57/42.93	22.12/22.19/23.20	14.51/14.63/14.78	33.25/33.39/34.79
Without gated fusion		23.51/24.44/25.63	16.02/16.71/17.59	36.52/38.00/40.07	17.32/18.02/18.86	11.44/12.15/12.76	27.70/27.91/29.18
STAGCN		20.43/20.80/21.07	14.07/14.41/14.56	32.72/32.92/33.62	15.87/16.48/17.39	10.84/11.21/11.78	25.95/26.58/27.77

ignoring the spatial correlation of the transportation network. STGCN, DCRNN, Graph WaveNet, STSGCN, and the STAGCN model in this paper all consider spatial correlation, so they have better forecasting performance than methods that are only used for time series forecasting.

On the PeMSD04 data set, compared with the STGCN, DCRNN, Graph WaveNet and STSGCN models, the MAE of the STAGCN model in this paper is reduced by about 10.77 %, 18.27 %, 20.93 %, 1.04 %, respectively. RMSE was reduced by about 7.97 %, 14.16 %, 18.20 % and 0.68 % respectively. Because DCRNN, STGCN and Graph WaveNet use two modules to model the spatial correlation and time dependence, respectively, while ignoring the time dependence and periodic changes in the traffic flow data. The model in this paper simultaneously captures the temporal and spatial correlation in traffic flow data, and considers time dependence and periodic changes. Therefore, the method in this paper has better predictive performance than the baseline model.

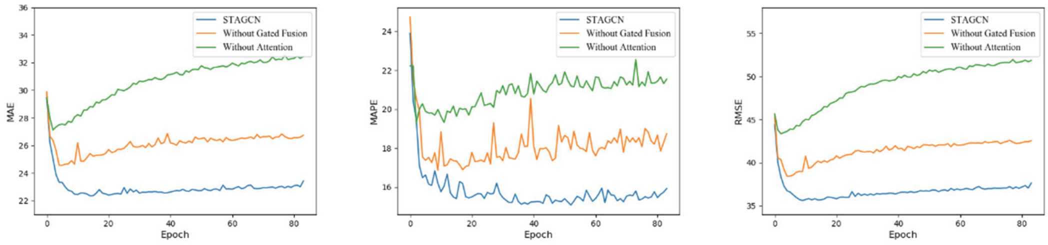
To further study the performance of the different modules of STAGCN, this paper designs two variants of the STAGCN model, studies the influence of the attention mechanism and the gating fusion mechanism on the model performance, and compares these two variants with the STAGCN model in PeMSD04 and PeMSD08 for comparison on the data set, 15-minute, 30-minute, and 60-minute traffic flow forecasting are performed, as shown in Table 3. The differences between these two variant models and the STAGCN model are:

Without Attention: This model has no attention mechanism and gated fusion mechanism modules.

Without Gated Fusion: This model has an attention mechanism module and not gated fusion mechanism.

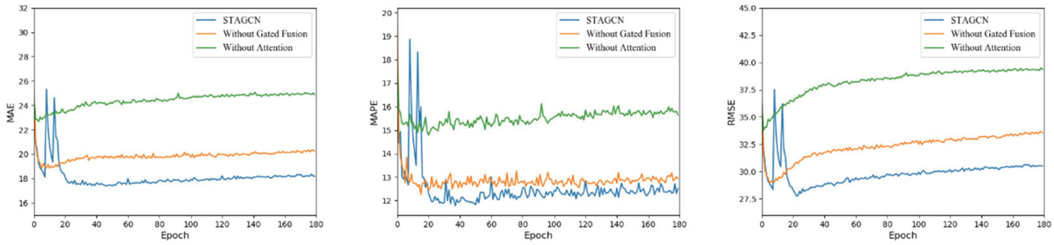
In 15 minutes, compared with the Without Attention and Without Gated Fusion models, the MAE of the STAGCN model on the PeMSD08 dataset was reduced by about 39.38 % and 9.14 %, and the RMSE was reduced by about 28.13 % and 6.74 %, respectively. In 30 minutes, MAE was reduced by approximately 34.65 % and 9.35 %, respectively, and RMSE was reduced by 25.62 % and 5.01 %, respectively. In 60 minutes, the MAE was reduced by about 33.41 % and 8.45 %, and the RMSE was reduced by about 25.28 % and 5.08 %, respectively. In the same way, the STAGCN model also achieved better forecasting performance on the PeMSD08 data set. In addition, according to Table 3, the STAGCN model achieved the best forecasting performance at different time points.

To better explain the STAGCN model, the training results of the STAGCN model in the PeMSD04 and PeMSD08 data sets are visualized, as showed in Figure 4 and Figure 5.



(a) MAE (b) MAPE (c) RMSE

Figure 4. 60-minute training performance comparison in the PeMSD04 data set.



(a) MAE (b) MAPE (c) RMSE

Figure 5. 60-minute training performance comparison in the PeMSD08 data set.

In general, as the forecasting time step increases, the corresponding forecasting difficulty becomes greater and greater, so the forecasting error will also increase. Because this model is a sub-model of the attention mechanism model, the fitting performance of this sub-model has some defects. When the attention mechanism is not included, the training MAE and RMSE will rise. It can be seen from Figures 4 and 5 that the Without Attention model does not show good forecasting performance compared with the Without Gated Fusion model and the STAGCN model in this paper, regardless of the time step. As the forecasting time step increases, Without the MAPE, MAE, and RMSE of the Gated Fusion model are becoming larger and larger. This is because the Without Attention model does not consider the temporal and spatial correlation of traffic flow at the same time. As the forecasting time step increases, the MAPE, MAE, and RMSE of the Without Gated Fusion model increase slowly. In contrast, the STAGCN model proposed in this paper has achieved better forecasting performance in almost all time steps. It shows that the strategy of combining spatio-temporal attention mechanism and gating fusion mechanism can better mine the spatio-temporal correlation in traffic flow data.

5. CONCLUSION

This paper proposes a novel graph convolutional network traffic flow forecasting model based on the spatio-temporal attention mechanism. On this basis, a gated fusion mechanism is further proposed adapting the spatial attraction mechanism and the temporal attention mechanism. This model combines a spatio-temporal attention mechanism and a spatio-temporal convolutional network. The spatio-temporal convolutional network includes a GCN in the spatial dimension and a standard convolutional network in the time dimension to simultaneously capture the spatio-temporal features of traffic flow data. The performance of different modules of STAGCN is further studied. This article designs two variants of the STAGCN model, and studies the influence of the attention mechanism and the gating fusion mechanism on the performance of the model. The two variants are compared with the STAGCN model on two datasets. The experimental results show that the combination of spatio-temporal attention mechanism and gated fusion mechanism can better mine the spatio-temporal correlation in traffic flow data. At the same time, the forecasting performance of the STAGCN model was verified on two traffic data sets and compared with the baseline model, The experimental results show that the forecasting accuracy of the STAGCN model is better than the baseline model in different forecasting time periods, which proves the accuracy of the model in traffic flow forecasting.

In fact, the traffic flow of expressways will be affected by many external factors, such as weather and holidays. In future work, we will further consider combining some external factors to further improve the accuracy of forecasts.

ACKNOWLEDGEMENT–This work was supported by Natural Science Foundation of Gansu Province, China (Grant No. 20JR 5RA450); National Key Research and Development Plan (Grant No.2019YFB1707303).

REFERENCES

Blu, T., Thevenaz, P. and Unser, M. (2004). Linear interpolation revitalized. *IEEE Trans. Image Processing* **13**, **5**, 710–719.

Chao, C., Petty, K. and Skabardonis, A. (2000). Freeway performance measurement: Mining loop detector data. *Transportation Research Record J. Transportation Research Board* **1748**, **1**, 96–102.

Chen, W., Chen, L., Xie, Y., Cao, W., Gao, Y. and Feng, X. (2020). Multi-range attentive bicomponent graph convolutional network for traffic forecasting. *34th AAAI Conf. Artificial Intelligence (AAAI)*, New York, USA.

Cheng, W., Shen, Y., Zhu, Y. and Huang, L. (2018). A neural attention model for urban air quality inference: Learning the weights of monitoring stations. *32nd AAAI Conf. Artificial Intelligence (AAAI)*, New Orleans, Louisiana, USA.

Cui, Z., Henrickson, K., Ke, R. and Wang, Y. (2019). Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Trans. Intelligent Transportation Systems* **21**, **11**, 4883–4894.

Cui, Z., Ke, R., Pu, Z. and Wang, Y. (2018). Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. *arXiv*. 1801.02143.

Defferrard, M., Bresson, X. and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *30th Int. Conf. Neural Information Processing Systems (NIPS)*, Barcelona, Spain.

Deng, D., Shahabi, C., Demiryurek, U., Zhu, L., Yu, R. and Liu, Y. (2016). Latent space model for road networks to predict time-varying traffic. *22nd ACM SIGKDD Int. Conf. knowledge Discovery and Data Mining (KDD)*, San Francisco, California, USA.

Huang, W., Song, G., Hong, H. and Xie, K. (2014). Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Trans. Intelligent Transportation Systems* **15**, **5**, 2191–2201.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv*: 1609.02907.

Kumar, K., Parida, M. and Katiyar, V. K. (2013). Short term traffic flow prediction for a non urban highway using artificial neural network. *Procedia-Social and Behavioral Sciences*, **104**, 755–764.

Kumar, S. V. and Vanajakshi, L. (2015). Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review* **7**, **3**, 1–9.

Lecun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature* **521**, **7553**, 436–444.

Li, Y., Yu, R., Shahabi, C. and Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv*. 1707.01926.

Liang, Y., Ke, S., Zhang, J., Yi, X. and Zheng, Y. (2018). GeoMAN: Multi-level attention networks for geo-sensory time series prediction. *27th Int. Joint Conf. Artificial Intelligence (IJCAI)*, Stockholm, Sweden.

Lippi, M., Bertini, M. and Frasconi, P. (2013). Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Trans. Intelligent Transportation Systems* **14**, **2**, 871–882.

Lv, Y., Duan, Y., Kang, W., Li, Z. and Wang, F. Y. (2015). Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intelligent Transportation Systems* **16**, **2**, 865–873.

Polson, N. G. and Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, **79**, 1–17.

Sen, R., Yu, H. F. and Dhillon, I. S. (2019). Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. *arXiv*. 1905.03806.

Shekhar, S. and Williams, B. M. (2007). Adaptive seasonal time series models for forecasting short-term traffic flow. *Transportation Research Record* **2024**, **1**, 116–125.

Shen, T., Zhou, T., Long, G., Jiang, J. Pan, S. and Zhang, C. (2018). DiSAN: Directional self-attention network for RNN/CNN-free language understanding. *32nd AAAI Conf. Artificial Intelligence (AAAI)*, New Orleans, Louisiana, USA.

Simonovsky, M. and Komodakis, N. (2017). Dynamic edge-conditioned filters in convolutional neural networks on graphs. *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA.

Song, C., Lin, Y., Guo, S. and Wan, H. (2020). Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. *34th AAAI Conf. Artificial Intelligence (AAAI)*, New York, USA.

Tong, Y., Chen, Y., Zhou, Z., Chen, L., Wang, J., Yang, Q., Ye, J. and Lv, W. (2017). The simpler the better: A unified approach to predicting original taxi demands based on large-scale online platforms. 23rd *ACM SIGKDD Int. Conf. Knowledge Discovery & Data Mining (KDD)*, Halifax, Nova Scotia, Canada.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. and Bengio, Y. (2017). Graph attention networks. *arXiv*. 1710.10903.

Wang, J., Cao, Y., Du, Y. and Li, L. (2019). DST: A deep urban traffic flow prediction framework based on spatial-temporal features. *12th Int. Conf. Knowledge Science, Engineering and Management (KSEM)*, Athens, Greece.

Wu, F., Wang, H. and Li, Z. (2016). Interpreting traffic dynamics using ubiquitous urban data. *24th ACM SIGSPATIAL Int. Conf. Advances in Geographic Information Systems (GIS)*, San Francisco, California, USA.

Wu, Z., Pan, S., Long, G., Jiang, J. and Zhang, C. (2019). Graph wavenet for deep spatial-temporal graph modeling. *arXiv*. 1906.00121.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *32nd Int. Conf. Machine Learning (ICML)*, Lille, France.

Xu, M., Dai, W., Liu, C., Gao, X., Lin, W., Qi, G. J. and Xiong, H. (2020). Spatial-temporal transformer networks for traffic flow forecasting. *arXiv*. 2001.02908.

Yang, B., Sun, S., Li, J., Lin, X. and Tian, Y. (2019). Traffic flow prediction using LSTM with feature enhancement. *Neurocomputing*, **332**, 320–327.

Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J. and Li, Z. (2018). Deep multi-view spatial-temporal network for taxi demand prediction. *32nd AAAI Conf. Artificial Intelligence (AAAI)*, New Orleans, Louisiana, USA.

Yu, B., Yin, H. and Zhu, Z. (2017) Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv*. 1709.04875.

Zhang, J., Yu, Z. and Qi, D. (2016a) Deep spatio-temporal residual networks for citywide crowd flows prediction. *arXiv*. 1610.00081.

Zhang, J., Zheng, Y., Qi, D., Li, R. and Yi, K. (2016b) DNN-based prediction model for spatio-temporal data. *24th ACM SIGSPATIAL Int. Conf. Advances in Geographic Information Systems (GIS)*, San Francisco, California, USA.

Zhang, Y., Jiang, Q., Li, S., Jin, X., Ma, X. and Yan, X. (2019). You may not need order in time series forecasting. *arXiv*. 1910.09620.

Zheng, C., Fan, X., Wang, C. and Qi, J. (2020). Gman: A graph multi-attention network for traffic prediction. *34th AAAI Conf. Artificial Intelligence (AAAI)*, New York, USA.

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