INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 16, Issue: 2, Month: April, Year: 2021
Article Number: 4111, https://doi.org/10.15837/ijccc.2021.2.4111

CCC Publications

AGORA UNIVERSITY PRESS

# Deep Spatio-temporal Learning Model for Air Quality Forecasting

L. Zhang, D. Li, Q. Guo, J. Pan

**Lei Zhang*, Dong Li,Quansheng Guo, Jiaxing Pan**
Beijing Key Laboratory of Intelligent Processing for Building Big Data,
School of Electrical and Information Engineering
Beijing University of Civil Engineering and Architecture
1 Zhanlanguan Road
Beijing, 100044, China
*Corresponding author:lei.zhang@bucea.edu.cn
korolness@163.com
gqs1996722@163.com
zxcpjx@163.com

**Abstract**

In recent years, air pollution has seriously affected people's production and life, so the air prediction has become a research hotspot in recent years. When analyzing air data, it is found that this type of data has not only temporal correlation, but also spatial correlation. For these temporal and spatial characteristics, this paper studies deep spatio-temporal learning method to global prediction. The purpose is to learn the evolution rule behind the spatio-temporal sequence, and give an estimation for future state. To be specific, we propose two novel forecasting models based on video processing technology: Spatio-temporal Orthogonal Cube model (STOR-cube) and Spatio-temporal Dynamic Advection model (ST-DA), which effectively capture the spatio-temporal correlation and accurately predict the long-term air quality. STOR-cube contains three branches, i.e., a spatial branch for capturing moving objects, a temporal branch for processing motion, and an output branch for coupling the first two mutually orthogonal branches to generate a prediction frame. ST-DA constructs a spatio-temporal reasoning network to learn the characteristics of the spatio-temporal domain, and its impact on the future is explicitly modeled by pixel motion. Experiments results on the real-world datasets demonstrate our proposed approach significantly outperforms the state-of-the-art ones. Moreover, our model can be extended to other spatio-temporal data prediction tasks.

**Keywords:** spatio-temporal data mining, global prediction, 3D convolution, dynamic neural advection.

## 1 Introduction

Spatio-temporal forecasting has many important applications such as weather prediction, behavior prediction [8], transportation planning. With the development of urban industrialization, automobile exhaust emission and heavy metal diffusion also become part of air pollution [10]. So, predicting air pollution is a vital task, which can highly impact our daily life and guide government decisions. Due to the significance, it has received great attention both in academic and industrial community [21].

Air quality data is recorded by the fixed ground monitoring stations and other monitoring systems [1, 16] at regular time intervals, which cannot reflect the dynamical effects from the adjacent areas with the changing time [27]. Unlike clustering tasks [19], the long-term prediction still lacks a satisfactory progress in the literature, mainly due to the complex non-linear characteristics in spatio-temporal dimensions. The correlations of air pollution change significantly over time, and may fluctuate tremendously and suddenly (e.g., an accident). How to analyze the complex and dynamic relationships to develop a novel spatio-temporal data mining model and accurately predicting is a challenging issue.

Although some influential studies have been applied the spatio-temporal data on weather forecasting, most of these methods deal with continuous temporal attributes or limit independent spatial characteristics. More recently, numerous models [11, 14] have been proposed containing some external factors, however, they did not consider the evolution of spatio-temporal dimensions. Later, data-driven models replaced knowledge-driven ones, that is, using neural networks to predict air pollution. Nevertheless, neural networks are a shallow learning method which cannot fit spatio-temporal charactersand handle high-dimension dynamic attributes. Deep learning model is a new data mining tool due to its strong feature expression and high-dimension non-linear mapping relationships. Recent studies use image processing technology to analyze the spatio-temporal data [15, 26]. However, there existed some limitations in the time domain, and they did not consider the continuous motion information.

The above concerns motivate us to research deep learning model based on video processing, that is, forecasting the next frame from the previous consecutive ones. According to the dynamic spatio-temporal attributes and the complex nonlinear spatio-temporal coupling relationship, we construct spatio-temporal reasoning algorithm to fuse spatio-temporal features, and propose a generative model for global prediction. Specifically, firstly, we design a matrix representation of spatio-temporal data (image-like) and a matrix block expression (video-like). Then we propose two deep learning networks based on video processing: Spatio-temporal Orthogonal Cube model (STOR-cube) and Spatio-temporal Dynamic Advection model (ST-DA). Experiments on real-world datasets confirm that our models are superior to the most advanced performances.

The paper is divided as follows: we first present the background and related work in section II. Section III elaborates on our proposed approach. In Section IV, we discuss the datasets use and present the results and analysis. Finally, we conclude the paper in Section V.

## 2 Related Work

In this section, we describe the infrastructure for air quality forecasting and video processing.

### 2.1 Air Quality Forecasting

Air quality prediction has been extensively studied in past decades. A spatio-temporal support vector regression (ST-SVR) was proposed to establish a local support vector regression model with spatial auto correlation variables [23]. Huang et al. [6] integrated the time effect into Geographically Weighted Regression (GWR), that is, the Geographically and Temporally Weighted Regression (GTWR), to capture the heterogeneity of space and time, and then applied on a quantitative correlation model between Aerosol Optical Depth (AOD) and PM2.5 [4]. These regression models cannot express the complex non-linear spatio-temporal relationship and bio-dimension dynamics. Meanwhile, these machine learning methods have great difficulties in over-fitting and local minima, and their generalization ability is insufficient.

Generalizing deep learning algorithm to spatio-temporal data mining is an emerging topic in many fields, i.e. weather prediction. Yi et al. [24] proposed DeepAir, which was a distributed architecture with multi-source heterogeneous data to capture multiple impacts. However, it does not consider the influence of time dimension variation. Therefore, it is difficult to describe the spatial and temporal distribution of air pollution. Based on [26], spatio-temporal residual network (ST-ResNet) is proposed to predict air quality [25]. ST-ResNet use convolutional neural network (CNN) to model the correlation between space grids, and limit the time attributes to three attributes: periodicity, trend, and proximity.

It neglects the low-level temporal motion which is more important. Graph neural network (GNN) [20] has been popular for spatial relationship of nodes, however, its understanding of space primitives in space-time ontology is that space is limited to point-based and discrete. In actual situation, the air data is distributed in whole space, and GNN ignore the correlation among adjacent areas.

### 2.2 Video Processing

Some scholars have proposed using CNN for intelligent video behavior recognition [12], but compared with processing static images and video recognition, making predictions from dynamic videos is a huge challenge. Traditional 2D CNN does not consider the motion information encoded in multiple consecutive frames, so 3D convolution [7, 18] has been proposed to extract features from the spatial and temporal dimensions for capturing the motion information. Due to the expensive training cost of 3D convolution, many scholars paid much attention on more efficient variants. In related work, Qiu et al. [13] proposed a separable convolution instead of 3D convolution, in which spatial convolution (S-Conv) was first performed in form of 2D CNN, and temporal convolution (T-Conv) was applied on 1D CNN. Pseudo 3D convolution [22] decoupled 3D convolution to reduce computational cost and achieved good results.

Furthermore, some researchers have proposed to use recurrent networks for video prediction. Srivastav et al. [17] presented a fully connected LSTM (FC-LSTM) to construct an encoder-decoder for future frames prediction. Traditional LSTM units learn from one-dimensional vectors and lose spatial information, so FC-LSTM avoids directly predicting future video frames at the image level. Convolution LSTM (ConvLSTM) [15] takes a three-dimensional tensor as input, and impose convolution operation into gating unit to modify FC-LSTM. However, ConvLSTM cannot achieve better results due to it is almost impossible to consider motion information and object information at the same time. Fan et al. [3] proposed Cubic Long Short-Term Memory (CubicLSTM) for video prediction. CubicLSTM contains three branches to capture moving objects, processing motion, and generate the output of predicted frame. But CubicLSTM is also not enough to extract a larger range of spatial information. The above studies motivate us to propose STOR-cube and ST-DA.

## 3 Research Method

In this section, we describe our proposed model in detail.

### 3.1 Dimension Conversion

Figure 1 shows the procedure of dimension conversion. Based on the previous research [25], we transform the air quality data into a pixel matrix $A$, $A \in R^{N \times N}$ through spatial conversion, where $N \times N$ means that the city is divided into $N \times N$ grids according to longitude and latitude. The performance of dynamic system in the space-time region is stacked into a three-dimensional array B, $B \in R^{N \times N \times T}$, $T = \{1, 2 \quad \ldots, F\}$is time interval set. We define $C = \{B_1, \ldots, \quad B_{\text{step}}\}$along the time step (that is the number of three-dimensional array B). Therefore, given historical observation $C$, the prediction is defined as $A_{step*F+1}$.

### 3.2 Pseudo-3D convolution

Given a three-dimensional array of $N \times N \times T$ size, where $N$ and $T$ represent the length, height, and width of each frame. 3D convolution can simultaneously model spatial information like 2D convolution, and construct temporal connections across frames. Pseudo-3D [? ] decouples 3D convolution based on Kronecker product into a low-complexity convolution kernel:$\mathbf{K} = \mathbf{K}_{x,y} \otimes \mathbf{K}_t$, where $\mathbf{K}_{x,y}$ and $\mathbf{K}_t$ denote S-Conv in space domain and T-Conv in time domain, respectively. This means that one $3 \times 3 \times 3$ 3D convolution kernel can be decoupled into a $1 \times 3 \times 3$ kernel equivalent to 2D CNN on spatial domain and $3 \times 1 \times 1$ kernel like 1D CNN in temporal domain. It not only significantly reduces the model size, but also gives P3D CNN more ability to learn spatio-temporal data.

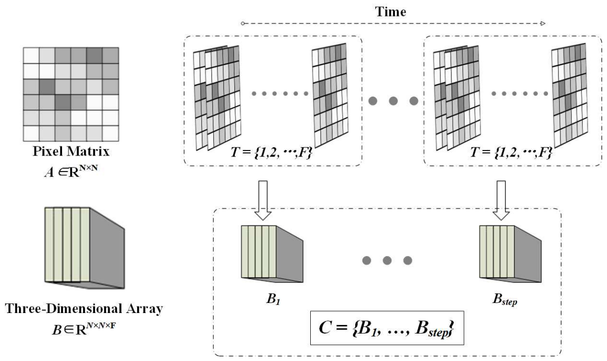


Figure 1: Procedure of dimension conversion

### 3.3 ConvLSTM

ConvLSTM [15] changes the fully connected weights in LSTM to the convolution kernel, so it can perform temporal modeling local features of 2D array with spatiotemporal characteristics. Traditional 3D convolution is affected by the limitation of convolution kernel, and ignores the long-term periodicity and other factors, while ConvLSTM can learn long-term time dependence. The input of ConvLSTM is the generated model from 3D convolutional network, giving full play to the time sequence memory advantages of the long short-term memory network, and realizing the fusion of spatio-temporal features. It can be expressed as follows:

$$f_t = \sigma\left(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf}{}^{\circ}\mathcal{C}_{t-1} + b_f\right) \tag{1}$$

$$\mathcal{C}_t = f_t{}^{\circ}\mathcal{C}_{t-1} + i_t{}^{\circ}\tanh\left(W_{xc} * \mathcal{X}_t + W_{hc} * \mathcal{H}_{t-1} + b_c\right) \tag{2}$$

$$o_t = \sigma\left(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1} + W_{co}{}^{\circ}\mathcal{C}_t + b_o\right) \tag{3}$$

$$\mathcal{H}_t = o_t{}^{\circ}\tanh\left(\mathcal{C}_t\right) \tag{4}$$

### 3.4 STOR-cube

Figure 2 illustrates the framework of STOR-cube. It contains three branches: time branch (x-axis), space branch (y-axis) and output branch (z-axis), therein, time branch and space branch are orthogonal. Enter the input Bi into the network: x-axis uses T-Conv to obtain the motion information$H_i^L$; y-axis conducts S-Conv to capture the moving object information$H_i'^L$; z-axis combines motion information and moving object information to generate final prediction, where $L$ is the network depth.

Cube-block (blue box) uses P3D to learn the spatio-temporal representation. The $L-th$ hidden layer in P3D CNN is defined:

$$H_i^L = \mathrm{f}\left(W_T{}^L \otimes H_i^{L-1} + b^k\right) \tag{5}$$

$$H_i'^L = \mathrm{f}\left(W_S{}^L \otimes H_i'^{L-1} + b'^k\right) \tag{6}$$

where $\otimes$ denotes 3D convolution operation. $\mathbf{f}(\cdot)$ is the activation function, here, we use rectifier function, that is, $\mathbf{f}(z) = \max(0,\ z)$. $W_T{}^L$ and $W_S{}^L$ are T-Conv kernel parameters and S-Conv kernel parameters in the $L-th$ layers, and the sizes are $1 \times 1 \times 3$ and $3 \times 3 \times 1$, respectively. After obtaining $H_i^L$ and $H_i'^L$, we use ConvLSTM to store and associate motion information with visual information along x-axis and y-axis in each time step. The input of ConvLSTM is the previous unit status $\mathcal{C}_{i-1}$ and the current information $H_i^L$, then it generates new $\mathcal{C}_i$. The updates are formulated as Equation 7.

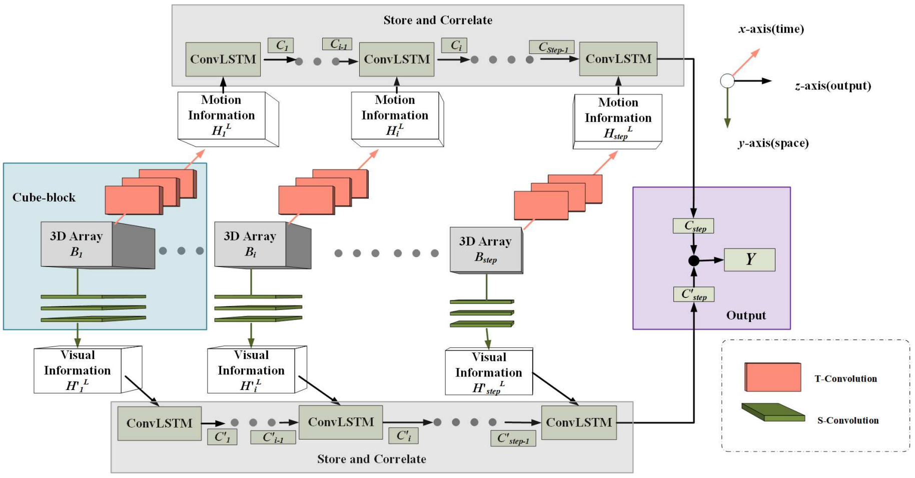


Figure 2: Framework of STOR-cube

$$\begin{aligned}&\text{temporal branch}\ :\\ &\mathcal{C}_i = \text{LSTM}\left(\left(\mathcal{C}_{i-1}, H_{i-1}^L\right); \mathcal{W}, b; *\right)\\ &\text{spatial branch}\ :\\ &\mathcal{C}'_i = \text{LSTM}\left(\left(\mathcal{C}'_{i-1}, H'^L_{i-1}\right); \mathcal{W}', b'; *\right)\end{aligned} \tag{7}$$

where $*$ represents 2D convolution operation. Cube-blocks are stacked along z-axis and the final prediction is defined as follows:

$$\text{outputbranch} : \text{Y} = \mathcal{W}'' * \left[\mathcal{C}_{step}, \mathcal{C}'_{step}\right] + b'' \tag{8}$$

### 3.5 ST-DA

Figure 3 describes the framework of ST-DA. Like STOR-cube, ST-DA also adopts P3D and ConvLSTM as spatio-temporal reasoning part (blue box), while it is a unified form of S-Conv and T-Conv for capturing the short-term spatio-temporal features, then use the memory network to learn long-term characters.

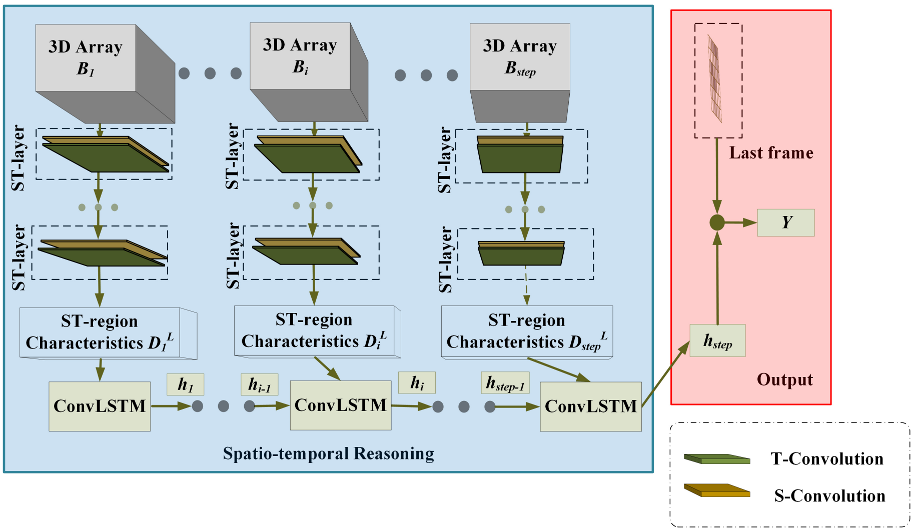


Figure 3: Figure 3: Framework of ST-DA

As shown in Figure 3, we cascade S-Conv and T-Conv as ST-layer:

$$\mathbf{ST}\left(D_i^{L-1}\right) = \mathbf{T}\left(\mathbf{S}\left(D_i^{L-1}\right)\right) = D_i^L \tag{9}$$

$\mathbf{T}$ and $\mathbf{S}$ represent the convolution in time domain and space domain; $D_i^L$ and $D_i^{L+1}$ are the input and output of the L-th hidden layer. For example, the input size is $d \times w \times h \times c$, $d$ is the length of time domain, $w \times h$ is the size of spatial domain, and $c$ is the number of channels. Therefore, the calculation cost of 3D-convolution (kernel size is $d_k \times k \times k$) is $d \times w \times h \times c \times d_k \times k \times k$, and the calculation cost of decoupled 3D-convolution is $d \times w \times h \times c \times (d_k + k \times k)$.

ConvLSTM adopts the state-to-state calculation by connecting the spatio-temporal characteristics in each time step, and its parameters are calculated as follows:

$$[\mathbf{o}_i, \mathbf{f}_i, \mathbf{i}_i, \mathbf{g}_i] = \sigma\left(\mathbf{K}^{ss} * h_{i-1} + \mathbf{K}^{is} * D_i\right) \tag{10}$$

$$\mathbf{c}_i = \mathbf{f}_i * \mathbf{c}_{i-1} + \mathbf{i}_i * \mathbf{g}_i \tag{11}$$

$$h_i = \mathbf{o}_i * \tanh\left(\mathbf{c}_i\right) \tag{12}$$

$\mathbf{K}^{ss}$ denotes the state-to-state weight, $\mathbf{K}^{is}$ represents the input-to-state weight; output gate, forget gate, input gate and content gate are respectively expressed as:$\mathbf{o}_i, \mathbf{f}_i, \mathbf{i}_i, \mathbf{g}_i$.

Based on Dynamic Neural Advection (DNA) [2], the output part (red box) applies the learned spatio-temporal motion influence on the latest state. In this approach, we predict a distribution over locations from the previous frame for each pixel in the new frame. The predicted pixel value is computed as an expectation under this distribution.

As shown in Figure 3, the spatio-temporal motion features obtained from the Spatio-temporal Reasoning are regarded as the constraint $\hat{J}_t^{(c)}$ of the discrete distribution of pixels in the next frame, where $c$ is the number of channels to generate predictions.

$$h_{step} = \hat{J}_t^{(c)} \tag{13}$$

$$\mathcal{Y} = \sum_c \hat{J}_t^{(c)} + \hat{I}_{t-1} \tag{14}$$

### 3.6 Loss Function

The loss function is Log-Cosh.

$$L\left(y, \hat{y}_i\right) = \sum_{i=1}^{n} \log\left(\cosh\left(\hat{y}_i - y_i\right)\right) \tag{15}$$

Log-Cosh is mostly the same as the mean square error (MSE), but it will not be strongly affected by the occasional extreme incorrect prediction. It has all the advantages of Huber Loss. The difference from Huber Loss is that it can be guided twice. For deep learning which use Newton's method to find the optimal solution, the second-order derivable function is more advantageous.

## 4 Experiments and Results

### 4.1 Datasets and Preprocessing

We used real datasets collected by 42 official monitoring stations in Beijing from Jan 2014 to March 2020. It contains various data, such as PM2.5, PM10, O3 and AQI. Here, we select two datasets, that are AQI and O3.

The real data is converted into a pixel matrix through space conversion, and the 2D pixel matrix is stacked to form a 3D pixel matrix block (the size of pixel matrix is 32×32, and the number of consecutive frames is 10). In other words, the shape of 3D matrix block is 32×32×10, so that it can fit the requirements of models.

## 4.2 Settings

All experiments are conducted on a 64-bit Linux server with NVIDIA Titan GPU and Keras programming environment (TensorFlow).

ST-DA adopts a pseudo 3D convolution, so each defined spatio-temporal convolutional layer of the reasoning network contains two layers, namely S-Conv layer (kernel size is 3×3×1) and T-Conv layer (kernel size is 1×1×3). The number of kernels in each layer is 32, and the depth of spatio-temporal layer is 3. The number of ConvLSTM units is 8, the number of hidden layers is 64, and all the input-to-state and state-to-state kernel sizes are 3×3. Relu() is the activation function. Batch normalization is used for local convolution components. In our experiments, the batch size is set to 64, the max epoch for training is set to 200. Adadelta is an optimization algorithm. The initial learning rate and input size of the network are 0.01and 8×10×32×32×1, respectively. The output of P3D convolution is converted into 8×32×32×90 by reshape layer, and then input into ConvLSTM.

We use 80% of the data for training, 20% for validation. The method of initialize the convolutional kernel is 'keras.initializers. glorot_uniform'.

STOR-cube use the similar settings as ST-DA.

## 4.3 Baselines

We compare our models with the following baseline methods:

**LSTM**:Long-short term memory (LSTM), due to its unique design structure, it is suitable for processing and predicting important events with long intervals and delays in time series.

**ARIMA**:Auto-regressive integrated moving average (ARIMA) [9], is an important method for studying time series, based on autoregressive model (referred to as AR model) and moving average model (referred to as MA model).

**SARIMA**:Seasonal Auto-regressive integrated moving average (SARIMA) [5], adds seasonal factors to ARIMA, which is one of the time series forecasting methods.

**MLP**:Multi-layer perceptron (MLP), its number of hidden units are 256.

**ConvLSTM**:Convolutional LSTM is a deep learning approach, it rewrites the gating unit of LSTM into CNN to process image time series data.

**ST-ResNet**:Deep learning network that originally proposed to predict traffic [26] and air quality [25].

## 4.4 Evaluation Metrics

We apply two widely used metrics to evaluate the performance of our models, i.e., Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), which are defined as follows:

$$\mathrm{RMSE}(y,\hat{y}_i)=\sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}(\hat{y}_i-y_i)^2} \tag{16}$$

$$\mathrm{MAE}(y,\hat{y}_i)=\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}|\hat{y}_i-y_i| \tag{17}$$

where$\hat{y}_i$ and $y_i$ mean the prediction value and real value in $i$ time stamp, and $n$ is the total number of cases.

## 4.5 Results and Discussion

### 4.5.1 Results

Our models are verified on two real-world datasets. In Part 1, we compare our models with some baseline models. In Part 2, we conduct hyperparameter evaluation of these two models.

**Comparison of our model with baselines**

Table 1 gives the compared results of each model on two datasets. Obviously, the training MAE are better than those of validation. This is because the regularization in Keras is closed during verifying. The prediction results of traditional time series methods (eg. LSTM, ARIMA, SARIMA and MLP) are not ideal. Their MAE on AQI is 25 35, and RMSE is 4.1 4.5. The MAE of O3 is 15 20, and RMSE is 2.9 3.1.

Table 1: Comparison of our model with baselines

| Model | Beijing AQI | | | | Beijing O3 | | | |
|---|---|---|---|---|---|---|---|---|
| | Training MAE | Validation MAE | Training RMSE | Validation RMSE | Training MAE | Validation MAE | Training RMSE | Validation RMSE |
| LSTM | 36.78 | 26.93 | 4.78 | 4.18 | 26.04 | 15.56 | 3.56 | 2.98 |
| ARIMA | 36.74 | 30.53 | 4.65 | 4.13 | 26.92 | 17.54 | 3.67 | 3.02 |
| SARIMA | 37.63 | 27.49 | 4.75 | 4.20 | 27.53 | 17.49 | 3.69 | 3.04 |
| MLP | 42.07 | 32.77 | 4.95 | 4.35 | 26.68 | 16.54 | 3.61 | 2.91 |
| ConvLSTM | 30.77 | 22.47 | 4.33 | 4.09 | 25.73 | 15.58 | 3.57 | 2.93 |
| ST-ResNet | 27.76 | 20.75 | 4.12 | 3.87 | 21.42 | 12.98 | 3.26 | 2.59 |
| STOR-cube | 30.57 | 21.03 | 4.44 | 4.07 | 20.32 | 12.95 | 3.20 | 2.55 |
| ST-DA | 25.76 | 18.36 | 4.13 | 3.99 | 17.46 | 10.50 | 2.92 | 2.48 |

On the contrary, deep learning models (eg. ConvLSTM, ST-ResNet, STOR-cube and ST-DA) considering temporal and spatial characteristics can get better results. Their MAE on AQI is 18 23, and RMSE is 3.8 4.1; The MAE of O3 is 10 16, and RMSE is 2.4 2.9. In deep learning model, ST-ResNet performs better on MAE and RMSE of AQI dataset than STOR-cube, but worse on O3. According to all the evaluation indexes, ST-DA is the best. ST-DA has advantages in describing spatio-temporal characteristics, while STOR-cube has several gaps, due to it orthogonally couples the temporal and spatial reasoning. Experiments show that STOR-cube focus on moving objects instead of physics, so it performs poorly when facing invisible objects (i.e. air quality). ST-DA does better promote to invisible objects.

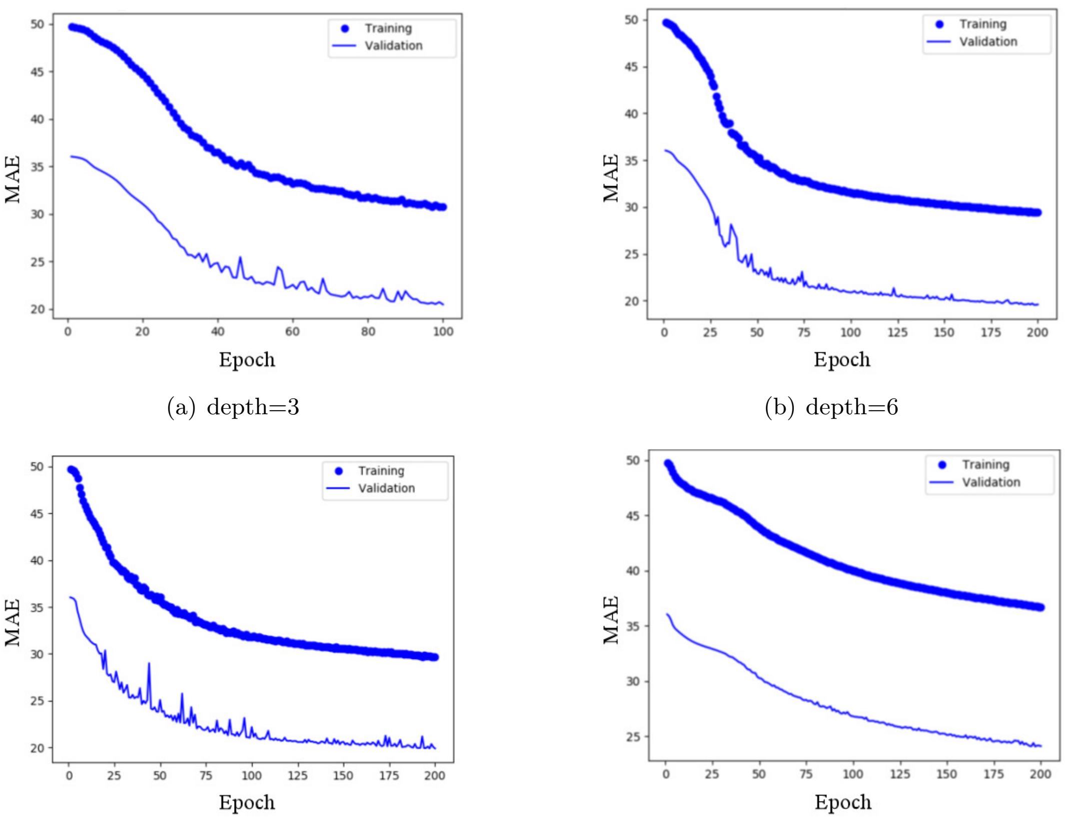
**Comparison of our model with baselines**

Figure 4 shows the prediction performance of STOR-cube with different network depths (the dot is Training MAE, and the line is Validation MAE). The prediction performance improves as the network depth increases, and achieve better performance at depth=6 and depth=9, then the performance decreases. Figure 5 and Figure 6 show the prediction performance of different depth and sequential step in ST-DA. Figure 5 describes the performance decreases as the depth increases, and when the depth reaches 6 and 12, there is even a chaotic situation, here Validation MAE outperforms Training MAE. Figure 6 shows that the best performance when step=8, and the increase in step will also lead to a decreasing performance. A potential reason is that more parameters need to be learned considering longer term dependence. The more steps, the time dependence among data will become more complicated, as a result, training process becomes more difficult.

### 4.5.2 Discussion

It can be seen from Part1 that the regression models such as LSTM, ARIMA, SARIMA and MLP which regard air pollution data as time series, so the result are the worst. The extraction of spatio-temporal features needs understanding the attributes of spatio-temporal data, that is, temporal ontology and spatial ontology. The spatio-temporal ontology in this paper is that the air occupies a continuous space (not discrete), and it is also related with time. Based on this, the spatio-temporal reasoning modules are designed based on CNN (deal with continuous spatial information) and memory network (deal with continuous temporal features). The traditional regression models, such as LSTM, ARIMA, Sarima and MLP, simply regard spatio-temporal data as time series, ignoring the complex correlation between spatial structure and spatio-temporal coupling, so their effect is the worst.

The deep learning networks, such as ConvLSTM and ST-ResNet, construct and analyze the interactive relationship of spatio-temporal data. ConvLSTM replaces LSTM's gating unit with CNN to capture the time connection in spatial level. However, its shortcoming is limiting the convolution depth of gating unit, and lack of high-dimensional feature extraction ability.ST-ResNet limits the time attribute to periodicity, trend and nearest proximity, and loses the low-level motion in time dimension, so their prediction accuracy is not good.



(a) depth=3

(b) depth=6

(c) depth=9

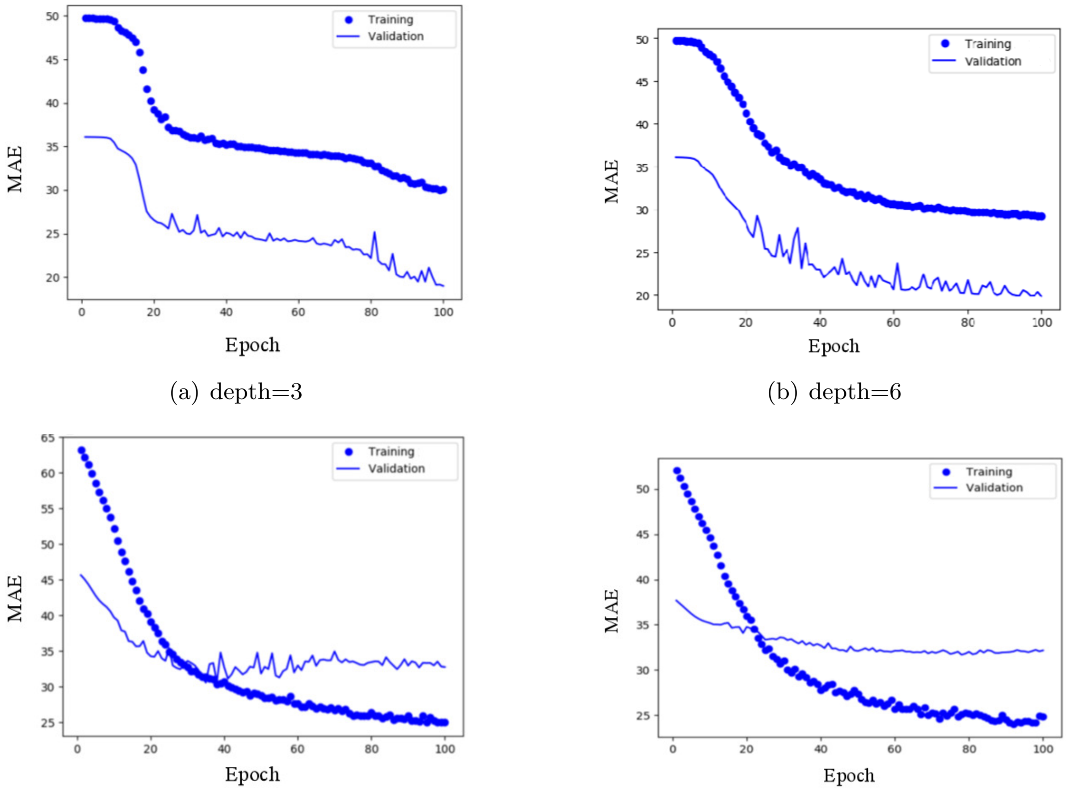
(d) depth=12

Figure 4: Influence of depth for STOR-cube We changed the network depth of the spatial dimension and the temporal dimension (S-Conv and T-Conv layer depth) for comparison experiments.

ST-DA builds a spatio-temporal overlay model, that is, to capture spatio-temporal features by using the cascaded ST-layer, adopting ConvLSTM as the memory network connect context, and then to build a generator for predictions based on dynamic neural advection method. The results of ST-DA are the best, because ST-DA is more focused on studying the dynamic system of time and space, and can better predict the invisible objects (such as air quality, etc.).

From Part 2, we can see that deep network usually has better results, because it can not only capture the dependencies between neighboring regions and recent time, but also obtain the dependencies between distant regions and long time. However, the deeper the network, the more difficult it is to train.

Some works have proved that decoupling 3D convolution kernel into a combination of 1×3×3 S-Conv and 3×1×1 T-Conv can capture the spatio-temporal features, and reduce the training parameters. Based on the above, we adopt S-Conv and T-Conv in ST-DA. The combination of S-Conv and T-Conv is parallel or cascade structure instead of 3×3×3 convolution kernel. We process spatio-temporal ontology on STOR-cube, that is orthogonal combination of time module and space module. ST-DA uses another cascade mode, that is constructing a unified spatio-temporal model (adding a time support module on the basis of the existing spatial one.



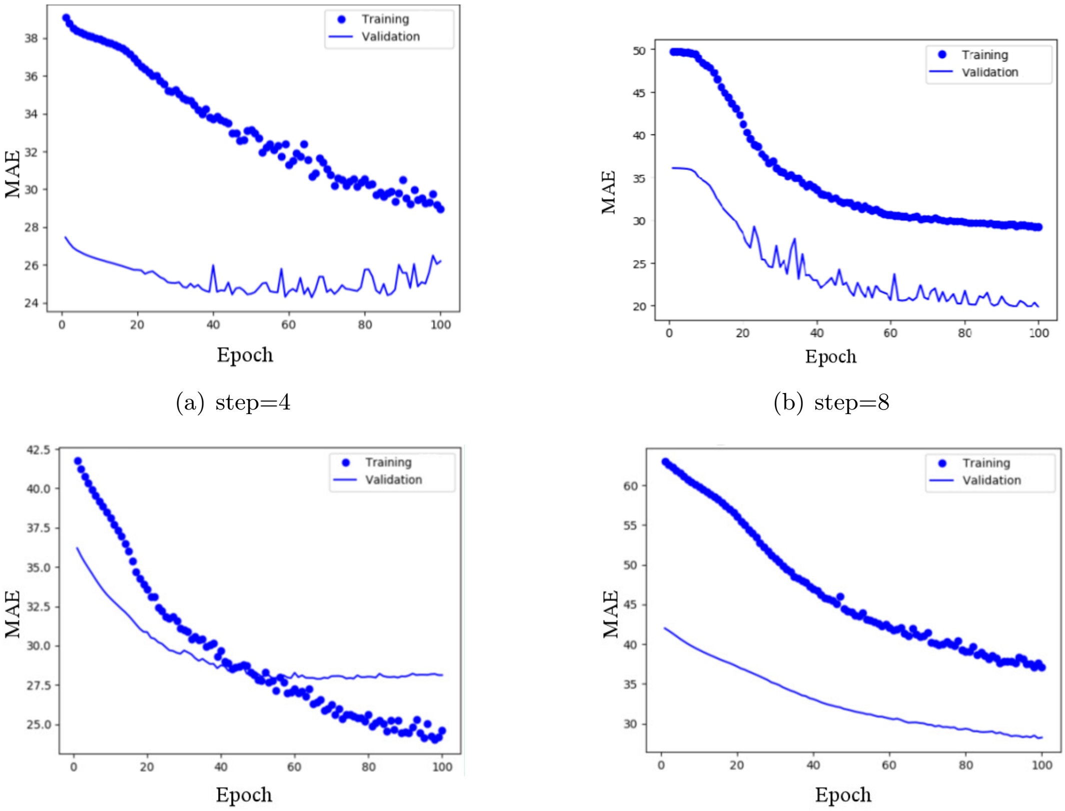
(a) depth=3

(b) depth=6

(c) depth=9

(d) depth=12

Figure 5: Influence of depth for ST-DA We changed the network depth of the spatio-temporal reasoning module for comparison experiments.



(a) step=4

(b) step=8

(c) step=11

(d) step=20

Figure 6: Influence of sequential step for ST-DA We changed the sequence step of the long-term connection part (ConvLSTM) for comparison experiments.

## 5 Conclusions

In this paper, we proposed deep spatio-temporal learning mechanisms to predict air quality. Based on video processing technology, two innovative spatio-temporal data mining models were designed, namely STOR-cube and ST-DA.

- STOR-cube consists of three branches, i.e., spatial branch for capturing moving objects, temporal branch for processing motion, and output branch for coupling the first two mutually orthogonal branches to generate a prediction frame.
- ST-DA designs a spatio-temporal reasoning network to learn the characteristics of spatio-temporal domain, and its impact on future is explicitly modeled by pixel motion. It is good at processing the complex and non-linear conditions.

The main contributions of our work can be summarized as follows:

- We propose deep spatio-temporal learning mechanisms by extending two-dimension spatio-temporal data into three-dimension to model the dynamic spatial and non-linear temporal correlations based on video processing technology.
- STOR-cube and ST-DA modules are presented to extract different spatio-temporal features, and evaluated on real-world datasets to achieve global predictions.
- Quantitative comparison results against existing and state-of-the-art models.

In future work, we plan to conduct more experiments with different spatio-temporal prediction tasks, such as water consumption prediction. Furthermore, we will enrich this research by considering multi-source heterogeneous spatio-temporal big data and various external factors.

### Funding

This work was supported by grants from the Social Science Planning Foundation of Beijing (20GLC059), National Natural Science Foundation of China (61871020), High Level Innovation Team Construction Project of Beijing Municipal Universities (IDHT20190506), Scientific Research Project of Beijing Municipal Education Commission (KM202010016011), BUCEA Post Graduate Innovation Project (PG2020046), and Ministry of Housing & Urban Construction Science and Technology Project of China (2016-K2-034).

### Author contributions

The authors contributed equally to this work.

### Conflict of interest

The authors declare no conflict of interest.

## References

[1] Dutta, P.K. ; Banerjee, S. (2019). Monitoring of Aerosol and Other Particulate Matter in Air Using Aerial Monitored Sensors and Real Time Data Monitoring and Processing, *Journal of System and Management Sciences*, 9(2), 104-113, 2019.

[2] Finn, C. ; Goodfellow, I. ; Levine, S. (2016). Unsupervised learning for physical interaction through video prediction, *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 64-72, 2016.

[3] Fan, H. ; Zhu, L. ; Yang, Y. (2019) Cubic LSTMs for Video Prediction, *33rd AAAI Conference on Artificial Intelligence*, 33(01), 8263-8270, 2019.

[4] He, Q. ; Huang, B. (2018). Satellite-based mapping of daily high-resolution ground PM2. 5 in China via space-time regression modeling, *Remote Sensing of Environment*, 206, 72-83, 2018.

[5] Hillmer, S. C. ; Tiao, G. C. (2012). An ARIMA-Model-Based Approach to Seasonal Adjustment, *Journal of the American Statistical Association*, 377(77), 63-70, 2012.

[6] Huang, B. ; Wu, B. ; Barry, M. (2010). Geographically and Temporally Weighted Regression for Modeling Spatio-temporal Variation in House Prices, *International Journal of Geographical Information Science*, 24(3), 383-401, 2010.

[7] Ji, S. ; Xu, W. ; Yang, M. ; Yu, K. (2012). 3D convolutional neural networks for human action recognition, *IEEE transactions on pattern analysis and machine intelligence*, 35(1), 221-231, 2012.

[8] Li, J ; Pan, S.X. ; Huang, L. ; Zhu, X. (2019). A Machine Learning Based Method for Customer Behavior Prediction, *Tehnicki vjesnik-Technical Gazette*, 26(6), 1670-1676, 2019.

[9] Moreira-Matias, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; Damas, L. (2013). Predicting Taxi-passenger demand using streaming data, *IEEE Transactions on Intelligent Transportation Systems*, 14(3), 1393-1402, 2013.

[10] Pirbadali-Somarin, A. ; Peyghambarzadeh, S. (2020). Air pollution by heavy metals from petrochemical incinerators: measurement and dispersion modelling, *Environmental Engineering and Management Journal*, 19, 379-390, 2020.

[11] Prasad, K.; Gorai, A. K.; Goyal, P. (2016). Corrigendum to ×Development of ANFIS Model for Air Quality Forecasting and Input Optimization for Reducing the Computational Cost and Time×, *Atmospheric Environment*, 246-262, 2016.

[12] Qin, L. L.; Yu, N. W.; Zhao, D. H. (2018). Applying the Convolutional Neural Network Deep Learning Technology to Behavioural Recognition in Intelligent Video, *Tehnicki vjesnik-Technical Gazette*, 25(1), 528-535, 2018.

[13] Qiu, Z.; Yao, T. ; Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks, *Proceedings of IEEE International Conference on Computer Vision (CVPR)*, 5533-5541, 2020.

[14] Saide, P. E.; Mena-Carrasco, M.; Tolvett, S.; Hernandez, P.; Carmichael, G. R. (2016). Air quality forecasting for Winter-time PM2. 5 episodes occurring in multiple cities in central and southern Chile, *Journal of Geophysical Research: Atmospheres*, 121(1), 558-575, 2016.

[15] Shi, X.; Chen, Z.; Wang, H.; Yeung, D. Y.; Wong, W. K.; Woo, W. C.(2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting, *Advances in neural information processing systems*, 28, 802-810, 2015.

[16] Simo, A.; Dzitac, S.; Frigura-Iliasa, F. M.; Musuroi, S.; Andea, P.; Meianu, D. (2020). Technical Solution for a Real-Time Air Quality Monitoring System, *Internation journal of computers computers communications & control*, 15(4), 2020.

[17] Srivastava, N.; Mansimov, E.; Salakhudinov, R. (2015). Unsupervised learning of video representations using lstms, *International conference on machine learning*, 843-852, 2015.

[18] Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks, *In Proceedings of the IEEE international conference on computer vision*, 4489-4497, 2015.

[19] Wang, L.; Hao, Z.; Han, X.; Zhou, R. (2018). Gravity Theory-Based Affinity Propagation Clustering Algorithm and Its Applications, *Tehnički vjesnik*, 25(4), 1125-1135, 2018.

[20] Wang, S.; Li, Y.; Zhang, J.; Meng, Q.; Meng, L.; Gao, F. (2020) . PM2. 5-GNN: A Domain Knowledge Enhanced Graph Neural Network For PM2. 5 Forecasting, *26th International Conference on Knowledge Discovery and Data Mining*, 2020.

[21] Wu, H.; Tsai, A.; Wu, H. (2019) . A hybrid multi-criteria decision analysis approach for environmental performance evaluation: an example of the TFT-LCD manufacturers in Taiwan, *Environmental Engineering and Management Journal*, 18, 597-616, 2019.

[22] Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. (2018) . Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification, *In Proceedings of the European Conference on Computer Vision (ECCV)*, 305-321, 2018.

[23] Yang, W.; Deng, M.; Xu, F.; Wang, H. (2018) . PM2. 5-GNN: A Domain Knowledge Enhanced Graph Neural Network For PM2. 5 Forecasting, *Atmospheric Environment*, 181, 12-19, 2018.

[24] Yi, X.; Zhang, J.; Wang, Z.; Li, T.; Zheng, Y. (2018) . Deep distributed fusion network for air quality prediction, *In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 965-973, 2018.

[25] Zhang, L.; Li, D.; Guo, Q. (2020). Deep Learning from Spatio-temporal Data using Orthogonal Regularizaion Residual CNN for Air Prediction, *IEEE Access*, 8, 66037-66047, 2020.

[26] Zhang, J.; Zheng, Y.; Qi, D.; Li, R.; Yi, X.; Li, T. (2018). Predicting Citywide Crowd Flows Using Deep Spatio-Temporal Residual Networks, *Artificial Intelligence*, 259, 147-166, 2018.

[27] Zheng, Y.; Capra, L.; Wolfson, O.; Yang, H. (2014) . Urban computing: concepts, methodologies, and applications, *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3), 1-55, 2014.

[28] Zong, M.; Wang, R.; Chen, Z.; Wang, M.; Wang, X.; Potgieter, J. (2020) . Multi-cue-based 3D residual network for action recognition, *Neural Computing and Applications*, 1-15, 2020.

*Cite this paper as:*

Zhang, L.; Li, D.; Guo, Q.; Pan, J. (2021). Deep Spatio-temporal Learning Model for Air Quality Forecasting, *International Journal of Computers Communications & Control*, 16(1), 4111, 2021.
https://doi.org/10.15837/ijccc.2021.2.4111