# 基于 GAM-CMAQ 过滤机制的 CNN-LSTM 和 CNN-GRU 的广东省某地区每日臭氧浓度预测

苏拥英[1]，邵键帆[2]，王国长[2]，徐世荣[3]
1.广州科技贸易职业学院通识教育学院，广东 广州 511442
2.暨南大学经济学院，广东 广州 510006
3.UCLA Statistics & Data Science，USA

**摘　　要：** 随着全球对臭氧污染关注度的提升和科学研究的深入，建立准确的臭氧预测模型对于减轻其对人类健康和经济的影响至关重要。现有模型或仅考虑气候和气态污染物，或整合 CMAQ 数据，虽取得一定成效，但未能充分利用 CMAQ 数据或减少噪声积累等问题。本文提出了一种新方法，即通过广义加性模型（GAM）预处理 CMAQ 数据，并结合前馈神经网络（FNN）、CNN-LSTM 和 CNN-GRU 模型进行臭氧浓度预测。利用 2018 年 4 月 26 日至 2020 年 7 月 31 日广东省某地区监测站的数据，我们发现，相较于未经 GAM 处理的数据，经 GAM 处理后的数据在均方误差上分别降低了 34.82%、37.66% 和 29.89%，显示出 GAM 处理对提高预测精度的显著效果。此外，考虑到多日 CMAQ 数据可能导致噪声累积，影响预测准确性。本研究引入卷积神经网络（CNN）以提取数据的局部特征，进一步降低了 CNN-LSTM 和 CNN-GRU 模型的均方误差，分别达到 36.6% 和 33.20% 的降低率。通过与六个模型的预测性能比较，CNN-LSTM 和 CNN-GRU 显示出最佳的预测性能。因此，本文建议使用 GAM 预处理后的 CMAQ 数据结合 CNN-LSTM 或 CNN-GRU 模型进行臭氧数据分析。

**关 键 词：** GAM-CMAQ；臭氧；FNN；CNN；CNN-LSTM；CNN-GRU

# Daily Ozone Concentration Prediction based on CNN-LSTM and CNN-GRU Model Coupled with GAM-CMAQ Filtering

Su Yongying[1], Shao Jianfan[2], Wang Guochang[2], Xu Shirong[3]
1.College of General Studies，Guangzhou Vocational College of Technology &Business, Guangzhou, Guangdong 511442
2.School of Economics, Jinan University, Guangzhou, Guangdong 510632
3. UCLA Statistics & Data Science, USA

**Abstract:** With the growing global concern and in-depth scientific research on ozone pollution, establishing an accurate ozone prediction model is crucial for mitigating its impacts on health and the economy. Existing models, which either only consider climate and gaseous pollutants or integrate CMAQ data, have achieved certain results but fail to fully utilize CMAQ data or suffer from noise accumulation issues. This paper proposes a new method that preprocesses CMAQ data through the Generalized Additive Model (GAM) and combines it with Feedforward Neural Network (FNN), CNN-LSTM, and CNN-GRU models for ozone concentration prediction. Utilizing data from a monitoring station in a certain area of Guangdong Province, China, from April 26, 2018, to July 31, 2020, we found that compared to the CMAQ data not filtered through GAM, the mean square error of the CMAQ data filtered through GAM was reduced by 34.82%, 37.66%, and 29.89%, respectively, demonstrating the significant effect of GAM processing in improving prediction accuracy. Furthermore, considering the noise accumulation from multiple days of CMAQ data that can negatively impact prediction accuracy, this study introduces Convolutional Neural Networks (CNN) to extract local features from the input data, further reducing the mean square error of the CNN-LSTM and CNN-GRU models by 36.6% and 33.20%, respectively. Finally, comparing the predictive performance of the proposed models with six other models over a three-day period on the CMAQ dataset, the results show that the CNN-LSTM and CNN-GRU models have the best predictive performance. Therefore, this paper recommends using the CNN-LSTM or CNN-GRU model based on GAM-filtered CMAQ data for analyzing ozone data.

**Keywords:** GAM-CMAQ; Ozone; FNN; CNN; CNN-LSTM; CNN-GRU

## 引言

长期暴露于空气污染物，如一氧化碳 (CO)、二氧化硫 ( $SO_2$)、氮氧化物 (NOx)、挥发性有机化合物 (VOCs)、臭氧 ($O_3$)、重金属和可吸入颗粒物 (PM2.5和 PM10)，已被广泛证实对人类健康造成急性或慢性影响 ( Kampa和 Castanas，2008 ) [1]。这些健康问题不仅影响个体生活质量，还对社会经济造成重大负担。以北京为例，空气污染导致的过早死亡每年造成经济损失约5.8302亿人民币 (占 GDP的0.03%) ( Zhao，X，2016 ) [2]。在中国，由于南北气候差异，北方地区更关注 PM2.5问题，而南方则因臭氧污染问题日益严重，成为全球关注的焦点 ( Lu等，2018 ) [3]。中国南方的臭氧污染事件与自然因素，尤其是太阳辐射周期密切相关 ( Chen等，2020 ) [4]。

因此，建立准确的臭氧预测模型对于提供可靠的空气质量预警系统至关重要，并协助政府关闭一些高污染企业，提醒人们注意空气污染，从而减少健康风险。尽管已有多种模型用于预测空气污染物浓度，包括确定性模型、统计模型和机器学习模型 ( Chemel等，2010；Grell等，2005 ) [5-6]，这些模型在基本假设和方法论上存在差异，且往往未能充分利用所有可用数据，如 CMAQ模型数据。此外，这些模型在处理多天数据时可能会累积噪声，影响预测准确性。因此，本研究提出了一种 GAM过滤与深度学习模型相结合的新方法，即通过广义加性模型 ( GAM ) 过滤 CMAQ数据，并结合卷积神经网络 ( CNN ) 和长短期记忆网络 ( LSTM ) 或门控循环单元 ( GRU ) 来提高臭氧浓度预测的准确性。这方法不仅充分利用了 CMAQ数据，还通过 GAM过滤减少噪声的影响，并通过 CNN提取输入数据的局部特征，增强模型对时间序列数据之间依赖关系的捕捉能力。

## 一、数据和评估指标

### ( 一 ) 数据预处理

本研究的每日臭氧浓度数据采集自广东省的某一空气质量监测站，收集时间为2018年4月26日至2020年7月31日。除了臭氧浓度数据，本文还收集了气象辅助变量，包括每日最低大气压、每日最高大气压、每日平均大气压、最低温度、每日最高温度、每日平均温度、每日平均露点温度、每日最低相对湿度、每日最高相对湿度、每日平均相对湿度、最大十分钟平均风向、每日最大十分钟平均风速、每日最大瞬时风向、每日最大瞬时风速等。这些辅助变量已被证明可以提高空气质量预测模型的准确性 ( Bai等，2016 ) [7]。所有数据均来自广东省最近的气象监测站，以确保数据的相关性和准确性。

为了避免辅助数据维度不同带来的预测错误和收敛缓慢的问题，对辅助气象数据进行标准归一化。标准归一化公式如下：

$$x' = \frac{x-\mu}{\sigma}$$

其中 $\mu$ 和 $\sigma$ 分别表示变量 $x$ 的均值和标准差。作为预测臭氧浓度的辅助数据 CMAQ数据，其原始尺度对于模型训练是必要的，因此无需进行标准化处理。

此外，风向和风速数据被转换为风矢量，以便更好地反映风的特性。风矢量的转换公式如下：

$$W_r = W_d \times (\frac{\pi}{180})$$
$$W_x = W_v \times (\cos W_r)$$
$$W_y = W_v \times (\sin W_r)$$
$$\max W_r = \max W_d \times (\frac{\pi}{180})$$
$$\max W_x = \max W_v \times (\cos \max W_r)$$
$$\max W_y = \max W_v \times (\sin \max W_r)$$

其中 $(W_x, W_y)$ 和 $(\max W_x, \max W_y)$ 分别表示最大十分钟平均向量和日最大瞬时风向量，$W_d, W_v, \max W_d$ 和 $\max W_v$ 分别表示日最大平均风向、日最大十分钟平均风速、日最大瞬时风向和日最大瞬时风速。

### ( 二 ) CMAQ数据过滤

CMAQ模型产生的预报数据包含了基于物理和化学过程的复杂模拟，这些数据对于臭氧浓度的预测至关重要。然而，由于模拟真实大气环境的复杂性，CMAQ数据不可避免地存在预测误差和噪声。为了减少这些噪声对模型性能的影响，本文采用了广义加性模型 ( GAM ) 对 CMAQ数据进行过滤，GAM模型如下：

$$y = f_0 + \sum_{j=1}^{p} f_j(x_j) + \varepsilon$$

其中 $f_0$ 表示为常数项，$f_1, \cdots f_p$ 表示未知函数，$\varepsilon$ 是随机误差。为了识别模型，假设 $E(f_j(x_j)) = 0$ 。

为了实现变量选择并估计 GAM模型，本人需在给定的基上增加非参数函数 $f_j(x_j)$ ，如样条基、傅立基和小波基。本文使用三次自然样条基，并通过最小化以下方程来获得估计值。

$$PRSS(f_0, f_1, f_2, \cdots, f_p) = \sum_{i=1}^{N}(y_i - f_0 - \sum_{i=1}^{N} f_j(x_{ij}))^2 + \sum_{j=1}^{p} \beta \int (f_j''(x_j))^2 dt$$

其中 $\beta$ 是惩罚参数，$f_j''(x_j)$ 表示 $f_j(x_j)$ 的二阶导数。

利用反拟合算法来估计广义可加模型，具体的估计算法如下：

Algorithm1　The Backfitting Algorithm for Additive Models

1.Initialize: $f_0 = E(Y), f_1^1(\cdot) \equiv f_2^1(\cdot) \equiv \cdots f_p^1(\cdot) \equiv 0, m = 0$

2.Iterate: $m = m + 1$

*for*　$j = 1$　*to*　$p$　:

$$R_j = Y - f_0 - \sum_{k=1}^{j-1} f_k^m(\chi_k) - \sum_{k=j+1}^{p} f_k^m(\chi_k)$$

$$f_j^m(\chi_j) = E(R_j | \chi_j)$$

Untill: $RSS = E(Y - f_0 - \sum_{j=1}^{p} f_j^m(\chi_j))^2$ fails to decrease

与此同时，根据 GCV准则选择惩罚参数。

### ( 三 ) 评估标准

模型的预测效果将由三个评估指标来评价，例如：均方误差 ( MSE ) 、标准差 ( RMSE ) 和平均绝对误差 ( MAE )。具体定义如下：

$$MSE = \frac{\sum_i (\hat{y}_i - y_i)^2}{n}$$
$$RMSE = \sqrt{\frac{\sum_i (\hat{y}_i - y_i)^2}{n}}$$
$$MAE = \frac{\sum_i |\hat{y}_i - y_i|}{n}$$

其中 $\hat{y}_i$ 表示预测的臭氧浓度，$y_i$ 表示真实的臭氧浓度，$n$ 是样本容量。数据集包含799样本，分为三部分：训练集（50%）、验证集（30%）和测试集（20%)。

## 二、方法论

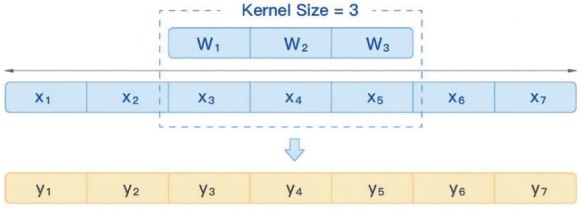
本研究采用了三种深度学习模型：1D卷积神经网络（CNN）、前馈神经网络（FNN）和结合了CNN的长短期记忆网络（CNN-LSTM）与门控循环单元（CNN-GRU），以预测广东省某地区的每日臭氧浓度。

### （一）用于特征提取的1D卷积神经网络

卷积神经网络CNN是一种深度学习模型，特别适合于处理具有网格结构的数据，如图像分类、人脸识别、物体识别、图像分割等。图1给出1维卷积神经网的框架。本文利用1维卷积神经网络从时间序列数据中提取局部特征。对于时间序列数据，局部特征由于相邻数据的依赖性而受到更多关注。1维卷积神经网络可以提取累计的延迟的信息。1维卷积神经网络机制如下所示：

$$y_t = \sum_{k=1}^{K} w_k x_{t-k+1}$$

其中，$K$ 表示滤波器的长度，$w_1, w_2, \cdots$ 表示滤波器，$x_1, x_2, \cdots$ 表示时间序列。




> 图1. 1维卷积的框架

在本研究中，CNN通过滑动窗口捕捉臭氧浓度数据的局部模式，为后续的LSTM和GRU模型提供更丰富的输入特征。我们选择适当的卷积核大小和步长，以最优地捕捉时间序列数据中的局部依赖关系。

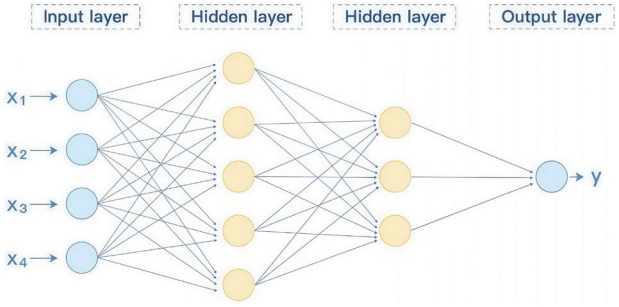
### （二）用于基准模型和时间序列预测的前馈神经网络

前馈神经网络（FNN）是一种基本的神经网络结构，它展示了逼近复杂函数的出色能力 (Bebis and Georgiopoulos，1994)[8]。FNN由输入层、隐藏层和输出层组成。通过逐层的信息传递，FNN根据一般的近似原理完成了非线性关系的建模 (Sanger等， 1989)[9]。在FNN内部，隐藏层中的线性全连接和非线性激活的组合是整个模型核心部分。图2给出了FNN的框架。FNN通过线性和非线性过程的迭代将输入数据转换为输出目标。该的方程如下所示：

$$z^{(i)} = W^{(i)} a^{(i)} + b^{(i)}$$
$$a^{(i)} = f_i(z^{(i)})$$
$$a^{(0)} = 0$$

其中，$z^{(i)}$ 表示净激活，$a^{(i)}$ 表示激活，$W^{(i)}$ 表示权重，$b^{(i)}$ 偏置。反向传播算法被用于参数更新，包括权重和偏置。对于神经网络模型，适当的参数可以提高预测效果并节省计算。参数调整包括层的数量、神经元的数量、激活类型、批处理规模等。

由于FNN的通用性和有效性，常被用作基准模型来测试其他现有或新提出的模型在空气污染中的预测效果。在本研究中，FNN也用于与CNN-LSTM和CNN-GRU模型的性能比较，以验证卷积和递归网络结构的有效性。




> 图2. FNN的框架

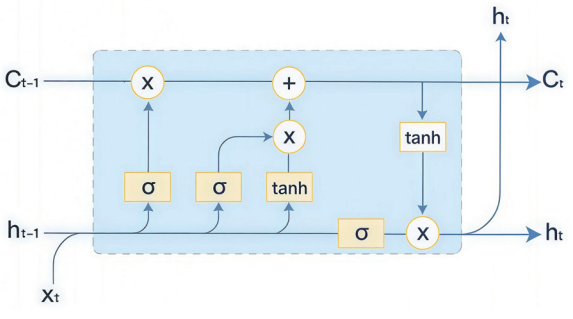
### （三）用于时间序列预测的长短期记忆网络和门控循环单元

循环神经网络（RNN）作为一种具有短期记忆能力的神经网络，已经被证明在处理时间序列数据方面表现出色。RNN基于时间反向传播算法构建了具有循环的网络结构，使得RNN具有自反馈能力。但输入序列过长时，会出现长距离依赖问题（Bengio等，1994）[10]。为了解决此问题，本文引入了长短期记忆网络（LSTM）和门控循环单元（GRU）。

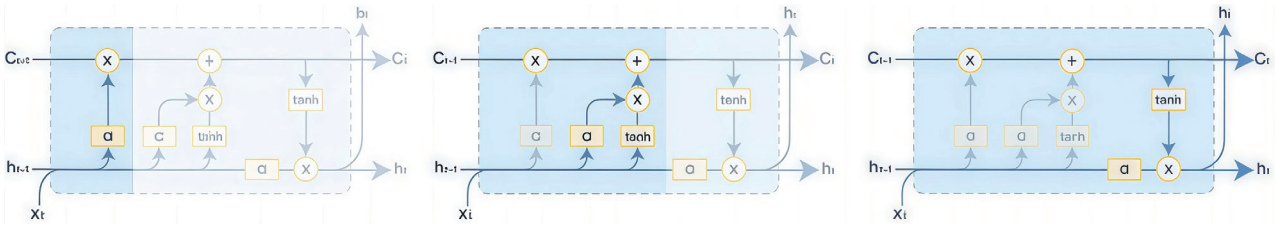
门控模型具有控制信息积累速度的能力，包括添加新信息或有选择地忘记之前的信息。图3给出了LSTM的框架示意图。LSTM作为一种循环神经网络的变形，通过输入门（$i_t$）、遗忘门（$f_t$）和输出门（$o_t$）来控制信息的传输路径（Greff等，2016）[11]。LSTM的操作机制方程如下（具体的门控机制如图4所示）

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t$$
$$h_t = o_t \otimes \tanh(c_t)$$
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

其中，$c_t$ 表示内部状态，$\tilde{c}_t$ 表示候选状态，$h_t, i_t, f_t, o_t$ 分别表示隐藏状态输入门、遗忘门和输出门，$\sigma(\cdot)$ 表示逻辑斯蒂克函数，$x_t$ 表示输入。




> 图3. LSTM框架

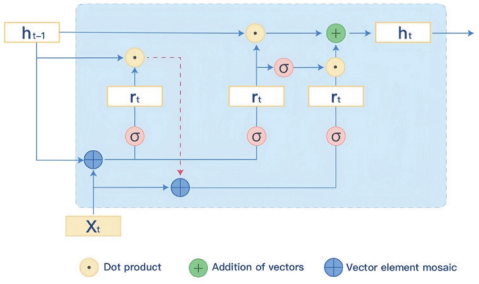


> 图4. LSTM门控机制

与LSTM相比，GRU以更简洁的方式控制信息(Yang等，2020)[12]。GRU引入了更新门（$z_t$）以确定哪些信息可以保留，还使用了重置门以指定候选状态是否依赖于前一个候选状态。图5展示了GRU的框架。GRU运作机制的方程如下：

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$
$$\tilde{h}_t = \sigma(W_h x_t + U_h (r_t \otimes h_{t-1}) + b_h)$$
$$h_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes \tilde{h}_t$$

其中，$x_t$表示输入，$h_{t-1}$表示上一次的输出，$z_t$和$r_t$分别表示更新门和重置，$\sigma(\cdot)$表示激活函数。




> 图5. GRU框架

LSTM和GRU是两种处理时间序列数据的循环神经网络，能够捕捉长期依赖关系。在本文中将被用于处理臭氧浓度数据的时间序列特性，并与CNN相结合，构建CNN-LSTM和CNN-GRU模型，以提高预测准确性。

## 三、实证分析

在本文中，整个数据集分为训练集（50%）、验证集（30%）和测试集（20%），目标为预测未来3天的臭氧浓度。考虑7种不同的数据处理方式，如表1所示，考虑三种不同模型训练数据；FNN、CNN-LSTM和CNN-GRU。下文将讨论如何提高模型预测准确度。在这项研究中，输入臭氧浓度的最大滞后期为10天，以下图表仅显示了最佳滞后天数，以便简洁地呈现了公平的比较标准。

### （一）GAM在变量选择中的表现

为了验证基于GAM的变量选择是否可提高预测的准确性，本文使用FNN、LSTM和GRU三种模型来比较T2（经GAM过滤）与T3数据（未经GAM过滤）的预测准确性。结果如表2所示。从表2中可知，对于T2而言，三种模型的预测准确性都要高于T3。以FNN模型为例，对于利用一天前的预测，T2的RMSE比T3低约8，MAE比T3低约4；对于两天前的预测，T2的RMSE比T3低约3，MAE比T3低约2。对于总体而言，一天前预测中GRU的性能最好，而两天前预测中FNN的性能最好。

表1：数据集模式缩写及其对应详情

| |
|---|
| T1: Previous ozone concentration only |
| T2 : Previous ozone concentration plus CMAQ auxiliary data(with GAM filtration) |
| T3 : Previous ozone concentration plus CMAQ auxiliary data (No GAM filtration) |
| T4 : Previous ozone concentration plus Meteorological auxiliary data |
| T5: Previous ozone concentration plus standard Meteorological auxiliary data |
| T6: Previous ozone concentration plus CMAQ auxiliary data and Meteorological auxiliary data |
| T7 : Previous ozone concentration plus CMAQ auxiliary data and Meteorological auxiliary data (standardized) |

表2: 相同模型下T2和T3数据预测性能比较，预测时间为一天、两天或三天前

| | | FNN | | | CNN-LSTM | | | CNN-GRU | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Predict day | MSE | RMSE | MAE | MSE | RMSE | MAE | MSE | RMSE | MAE |
| T2 | 1 | 1076.58 | 32.81 | 24.63 | 884.56 | 29.74 | 21.98 | 854.22 | 29.23 | 21.89 |
| | 2 | 1033.57 | 32.15 | 23.67 | 1063.54 | 32.61 | 24.25 | 965.65 | 31.07 | 24.06 |
| | 3 | 1364.38 | 36.94 | 26.04 | 1372.50 | 37.05 | 26.14 | 1278.67 | 35.76 | 24.96 |
| T3 | 1 | 1651.61 | 40.64 | 28.40 | 1418.97 | 37.67 | 26.03 | 1218.37 | 34.91 | 24.67 |
| | 2 | 1286.50 | 35.87 | 25.20 | 1408.64 | 37.53 | 26.10 | 1227.39 | 35.03 | 24.70 |
| | 3 | 1428.21 | 37.79 | 26.19 | 1414.66 | 37.61 | 26.39 | 1314.88 | 36.26 | 25.17 |

### （二）利用CNN特征提取的性能

本小节将讨论CNN在臭氧浓度预测中提取局部信息的性能。已有论文验证循环神经网络和CNN相结合能够提高空气污染物的预测性能（Qin等，2019；Yan等，2021）[13-14]。在本文将考虑把卷积层嵌入到原始LSTM和GRU的顶部。通过顶层卷积层，输入数据捕捉臭氧浓度的局部信息得到更好的处理。从上一节可知T2数据集预测性能比T3更好，因此本文考虑分别用四种模型：LSTM、GRU、CNN-LSTM和CNN-GRU利用除T3之外的六种数据预测一天后的臭氧浓度。结果见表3。从表3中可知，无论是CNN-LSTM还是CNN-GRU，其性能都比LSTM和GRU要好。以T6为例，与LSTM相比较，CNN-LSTM的MSE、RMSE和MAE分别降低了381.07、6.58、3.2。同样地，与GRU相比较，CNN-GRU的MSE、RMSE和MAE分别降低了337.38、5.2、4.39。由表3可知，CNN-GRU模型在T6数据集中

的 MAE是24种情况中最小的。

表3：纯循环模型和 CNN混合模型的性能比较

| | | LSTM | | | | CNN-LSTM | | |
|---|---|---|---|---|---|---|---|---|
| Dataset pattern | MSE | RMSE | MAE | lag period | MSE | RMSE | MAE | lag period |
| T1 | 1251.00 | 34.86 | 24.67 | 1 | 1270..68 | 35.65 | 27.06 | 2 |
| T2 | 1096.83 | 33.12 | 23.99 | 2 | 884.56 | 29.74 | 21.98 | 1 |
| T4 | 1505.57 | 38.80 | 29.29 | 10 | 1217.43 | 34.89 | 25.26 | 4 |
| T5 | 1555.98 | 38.16 | 30.01 | 5 | 1269.18 | 35.63 | 26.30 | 1 |
| T6 | 1040.13 | 32.25 | 23.23 | 3 | 659.07 | 25.67 | 19.94 | 7 |
| T7 | 1352.39 | 36.77 | 28.25 | 4 | 820.34 | 28.64 | 22.10 | 1 |
| | | GRU | | | | CNN-GRU | | |
| Dataset pattern | MSE | RMSE | MAE | lag period | MSE | RMSE | MAE | lag period |
| T1 | 1251.00 | 35.37 | 24.80 | 1 | 1204.36 | 34.70 | 24.78 | 1 |
| T2 | 1094.68 | 33.09 | 23.17 | 1 | 854.22 | 29.23 | 21.89 | 1 |
| T4 | 1392.75 | 37.32 | 27.68 | 9 | 1203.31 | 34.69 | 24.14 | 8 |
| T5 | 1582.04 | 39.77 | 28.91 | 4 | 1316.89 | 36.29 | 25.79 | 1 |
| T6 | 1016.26 | 31.88 | 23.63 | 6 | 678.88 | 26.06 | 19.24 | 1 |
| T7 | 1445.21 | 38.02 | 29.36 | 2 | 806.39 | 28.40 | 20.37 | 1 |

**（三）数据集类型的确定**

由上两个小节结论可知，选择 CMAQ数据来减少累计噪声，并且 FNN、GRU应结合循环神经网络输入数据处理来提高预测准确性。本节将考虑使用哪种数据来训练 FNN、CNN-LSTM和 CNN-GRU。为了简化，只考虑一天前的臭氧浓度预测，把六种数据集 T1、T2、T4-T7分别应用至这三种模型。结果如表4所示，FNN、CNN-LSTM和 CNN-GRU在 T6上都有最小的 MSE和 MAE，T7次之，但优于其他数据集。因此，本文将 T6作为最终的数据进一步研究。

表4： 不同数据集在三个模型预测效果比较

| | | FNN | | | | CNN-LSTM | | | | CNN-GRU | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset pattern | MSE | RMSE | MAE | lag period | MSE | RMSE | MAE | lag period | MSE | RMSE | MAE | lag period |
| T1 | 1269.71 | 35.63 | 25.23 | 1 | 1270.68 | 35.65 | 27.06 | 2 | 1204.36 | 34.70 | 24.78 | 1 |
| T2 | 1076.58 | 32.81 | 24.63 | 4 | 884.56 | 29.74 | 21.98 | 1 | 854.22 | 29.23 | 21.89 | 1 |
| T4 | 1332.16 | 36.50 | 24.86 | 4 | 1217.43 | 34.89 | 25.26 | 4 | 1203.31 | 34.69 | 24.14 | 8 |
| T5 | 1261.76 | 35.52 | 25.47 | 1 | 1269.18 | 35.63 | 26.30 | 1 | 1316.89 | 36.29 | 25.79 | 1 |
| T6 | 730.10 | 27.02 | 20.02 | 3 | 659.07 | 25.67 | 19.94 | 7 | 678.88 | 26.06 | 19.24 | 1 |
| T7 | 777.24 | 27.88 | 20.67 | 1 | 820.34 | 28.64 | 22.10 | 1 | 806.39 | 28.40 | 20.37 | 1 |

**（四）细分网络参数**

现在需要确定 FNN、CNN-LSTM和 CNN-GRU的调整参数。 FNN的参数包括全层的数量、每层的节点数、激活类型、优化器类型、正则化过程、训练周期和批量大小。由于增加了卷积层循环机制，CNN-LSTM和 CNN-GRU的优化参数包括卷积滤波器的数量、步幅、LSTM/GRU层的数量、连接层的数量、每层的节点数、激活类型、优化器类型、正则化过程、训练周期和批量大小。详细的调整参数见表5，由表5可知，CNN-LSTM和 CNN-GRU的计算成本远高于 CNN。对于 CNN-LSTM，训练以获得优化则需要更大的批量大小和 dropout比率。

表5: 细分三个模型的网络结构及内部参数

| | | FNN | | CNN-LSTM | | | | CNN-GRU | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | 1 | Time lag 2 | 3 | 1 | Time lag 2 | 3 | 1 | Time lag 2 | 3 |
| Number of iterations (epochs) | 200 | 200 | 200 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| Lag period | 3 | 5 | 4 | 7 | 1 | 5 | 1 | 1 | 2 |
| Non-linear activation | RELU | RELU | RELU | RELU | RELU | RELU | RELU | RELU | RELU |
| Batch size | 32 | 32 | 32 | 32 | 100 | 100 | 32 | 32 | 32 |
| Dropout ratio | 0.2 | 0.2 | 0.1 | 0.2 | 0.5 | 0.4 | 0.2 | 0.2 | 0.4 |
| Optimizer | Adam | Adam | Adam | Adam | Adam | Adam | Adam | Adam | Adam |

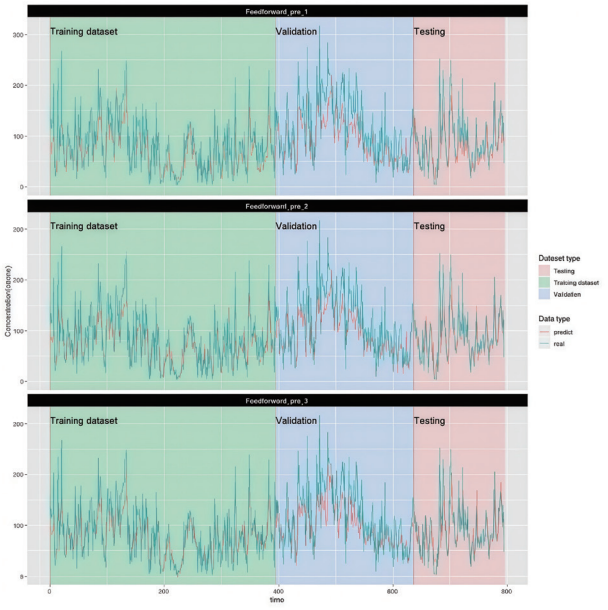
**（五）六个模型的预测性能比较**

为了检验所提出模型在 CMAQ数据集上三天内预测性能，考虑机器学习模型 XGBoost、DTR和 SVR与本文的模型 FNN、CNN-LSTM、CNN-GRU做比较。模型预测性能结果如表6所示，从表6可知，CNN-LSTM在六个模型中预测性能最好，其 MSE和 RMSE最低，MAE与 CNN-GRU相似。FNN的性能比

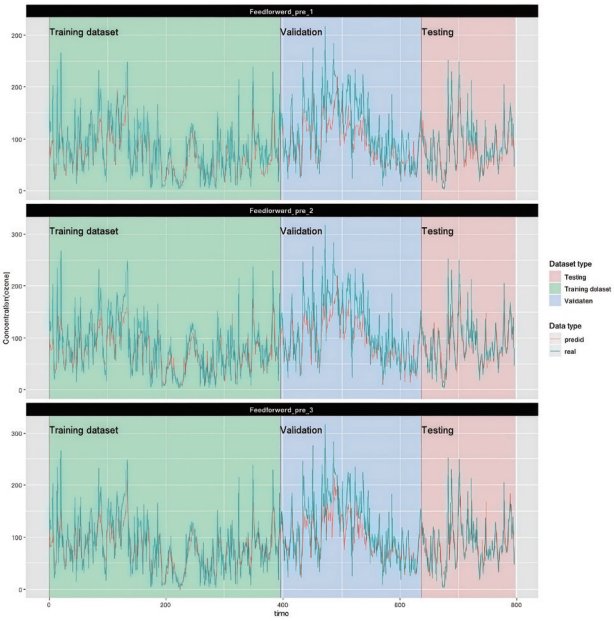
CNN-LSTM和CNN-GRU差，但比XGBoost、DTR和SVR好。图6-8中绘制FNN、CNN-LSTM和CNN-GRU的预测值与真实观测值。从图7-8中可知，CNN-LSTM和CNN-GRU的预测性能相似。

表6：六个模型的预测性能比较

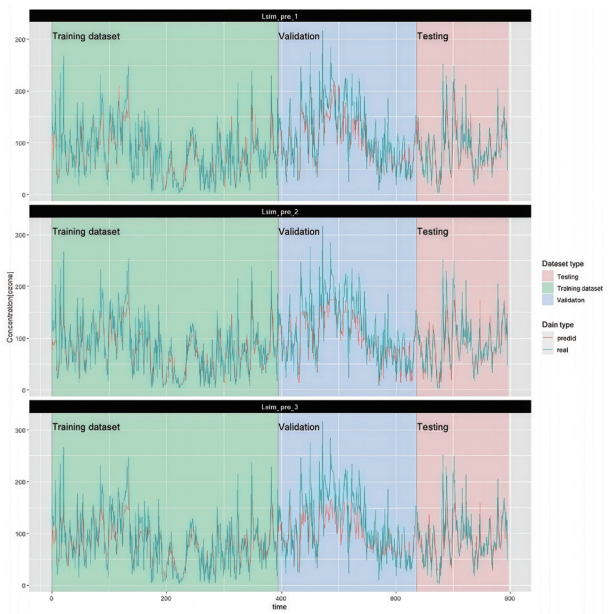
| | 1 DAY | | | 2 DAY | | | 3 DAY | | |
|---|---|---|---|---|---|---|---|---|---|
| | MSE | RMSE | MAE | MSE | RMSE | MAE | MSE | RMSE | MAE |
| XGBoost | 1740.05 | 41.71 | 32.01 | 1890.44 | 43.48 | 34.18 | 1948.06 | 44.14 | 33.00 |
| SVR | 1732.15 | 41.62 | 30.38 | 1854.67 | 43.07 | 31.22 | 1908.46 | 43.69 | 31.53 |
| DTR | 1283.35 | 35.82 | 25.63 | 1270.63 | 35.65 | 26.16 | 1422.89 | 37.72 | 28.19 |
| FNN | 730.10 | 27.02 | 20.02 | 937.00 | 30.61 | 22.26 | 1152.35 | 33.95 | 24.40 |
| CNN-GRU | 678.88 | 26.06 | 19.24 | 897.51 | 29.96 | 22.61 | 1071.97 | 32.74 | 23.80 |
| CNN-LSTM | 659.07 | 25.67 | 19.94 | 891.74 | 29.86 | 22.62 | 1069.56 | 32.70 | 23.43 |




> 图6. FNN的预测值与真实观测值比较




> 图8. CNN-GRU的预测值与真实观测值比较




> 图7. CNN-LSTM的预测值与真实观测值比较

## 四、结论

本文使用三种不同的神经网络模型：FNN、CNN-LSTM、CNN-GRU用于预测日臭氧浓度。得益于处理时间序列数据时能够捕捉到长期依赖关系的能力，CNN-LSTM和CNN-GRU具有最低的MSE、RMSE和MAE，表现出了优越的预测性能。在输入数据处理方面：引入气象辅助变量和CMAQ模型预测变量与历史臭氧浓度数据，这两个辅助变量为FNN、CNN-LSTM和CNN-GRU提供了更全面的数据支持，提高了模型的预测效果。利用GAM过滤CMAQ辅助变量，充分发挥了GAM的非线性拟合特性，达到了预期的模型效果。从表2中的提前一天预测臭氧浓度的结果可知，T3（未过滤）与T2（已过滤）相比，FNN的MSE从1651.61降至1076.58 (4.82%)，CNN-LSTM从1418.97降至884.56(37.6%)，CNN-GRU从1218.37降至965.65(29.89%)。在模型框架方面，在LSTM和GRU之上嵌入卷积层，则混合的CNN-RNN型能够更深入地挖掘数据的局部特征。在

CNN(LSTM和GRU)顶部添加卷积层后，由表3可得，CNN-LSTM的MSE从1040.13降 至659.07(36.64%)，CNN-GRU的MSE从1016.26 降至678.88 (33.20%)。由表4可得，六个数据集在不同的模型中比较，T6数据集在相同的模型架构下达到了最佳的预测效果：FNN时，MSE为730.10、CNN-LSTM时为659.07、CNN-GRU时为678.88。与其他最新的机器学习模型相比，本文提出的CNN-GRU和CNN-LSTM在提前预测三天臭氧浓度方面效果最好。

## *参考文献*


[1]Kampa M, Castanas E. Human health effects of air pollution [J]. Environmental pollution, 2008, 151(2): 362-367.

[2]Zhao X, Yu X, Wang Y, et al. Economic evaluation of health losses from air pollution in Beijing, China [J]. Environmental Science and Pollution Research, 2016, 23(12): 11716-11728.

[3]LuX, Hong J, Zhang L, et al. Severe surface ozone pollution in China: aglobal perspective [J]. Environmental Science & Technology Letters, 2018, 5(8): 487-494.

[4]Chen X, Zhong B, Huang F, et al. The role of natural factors in constraining long-term tropospheric ozone trends over Southern China [J]. Atmospheric Environment, 2020, 220:117060.

[5]Chemel C, Sokhi R S, Yu Y, et al. Evaluation of a CMAQ simulation at high resolution over the UK for the calendar year 2003 [J]. Atmospheric Environment, 2010, 44(24): 2927-2939.

[6]Grell G A, Peckham S E, Schmitz R, et al. Fully coupled "online" chemistry within the WRF model [J]. Atmospheric Environment, 2005, 39(37): 6957-6975

[7]Bai Y, Li Y, Wang X, et al. Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions [J]. Atmospheric pollution research, 2016, 7(3): 557-566.

[8]Bebis G, Georgiopoulos M. Feed-forward neural networks [J]. IEEE Potentials, 1994, 13(4): 27-31.

[9]Sanger T D. Optimal unsupervised learning in a single-layer linear feedforward neural network [J]. Neural networks, 1989, 2(6): 459-473.

[10]Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult [J]. IEEE transactions on neural networks, 1994, 5(2): 157-166.

[11] Greff K, Srivastava R K, Koutn í k J, et al. LSTM: A search space odyssey [J]. IEEE transactions on neural networks and learning systems, 2016, 28(10): 2222-2232.

[12] Yang S, Yu X, Zhou Y. LSTM and GRU neural network performance comparison study: Taking Yelp review dataset as an example [C] //2020 International workshop on electronic communication and artificial intelligence (IWECAI). IEEE, 2020: 98-101.

[13]Qin D, Yu J, ZouG, et al. A novel combined prediction scheme based on CNN and LSTM for urban PM 2.5 concentration [J]. IEEE Access, 2019, 7: 20050-20059. =

[14]Yan R, Liao J, Yang J, et al. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering [J]. Expert Systems with Applications, 2021, 169: 114513