ACCURATE AIR QUALITY PREDICTION: A PHYSICAL-TEMPORAL COLLECTION MODEL CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Air quality is closely related to public health. Several known health issues such as cardiovascular and respiratory diseases are in connection with long exposure in highly polluted environment. Accurate air quality forecast becomes extremely important, particularly for people living in less developed countries. To estimate the variation and growth of several air pollution concentrations, previous researchers proposed either model-based or data-driven approaches, such as Community Multiscale Air Quality model (CMAQ) and neural networks. These methods either require manual adjustment based on experience or purely learn from existing data and thus struggle to catch the dynamics of air pollutants. This paper proposes a physical-temporal collection (PTC) model to integrate both the model-based and data-driven strategies, aiming to reduce the systematic error and improve forecast accuracy. To validate the proposed method, we collected over two years of air pollutants record in Chengdu, one of the most polluted cities in China, covering a period from January 1st, 2016 to December 31st, 2017. Experiment results show that the proposed PTC achieves excellent performance with the average accuracy improved by 6.9% comparing with the baseline CMAQ model. Furthermore, the model can learn the long and short term effects from historical data, and thus facilitate deployment in the areas that lack trained experts and help vulnerable people to make informed planning for upcoming air pollution.

1 INTRODUCTION

Air particle pollution is becoming a global issue in recent years with studies reveal that various human chronic diseases are caused by these molecular contaminants including SO_2 (sulfur dioxide), NO_2 (nitrogen dioxide) and NO (nitric oxide) (Lelieveld et al., 2015). To develop effective air quality forecast methods is gaining increased attention, especially for the developing areas where environment protection is often sacrificed for economic development. In the last two decades, the Community Multiscale Air Quality model (CMAQ) (Appel et al., 2007; Lightstone et al., 2017; Foley et al., 2010) has been widely applied to predict the spatial distribution of pollutants in various scales. Because pollutants are characterised by a combination of temporal and spatial distribution, the CMAQ model is less effective and shows systematic bias for some measurements (Lightstone et al., 2017). In addition, the CMAQ model is limited by its grid prediction feature (Queen & Zhang, 2008) which is unable to forecast the air conditions in fine spatial resolution. Nevertheless, the CMAQ model has achieved wide applications and provides a solid baseline for further research.

Most recently, various data-driven approaches have achieved great advances in constructing linear and non-linear models to predict air pollutants. To address the time-varying effect, the artificial neural networks (ANN) (Fernando et al., 2012; Fu et al., 2015; Niska et al., 2004; Yu et al., 2017) have drawn overwhelming attention to characterise and forecast the complicated distribution of air pollutants. The internal relationship between different air pollutants is also exploited for achieving more accurate predicts (Kumar et al., 2017). Existing ANN based methods often pay less attention to the long-term effect of air quality forecast, and thus recurrent neural networks (RNN) with long short term memory (LSTM) that excel on time series have been explored. Apart from deep-learning tools, non-linear machine learning algorithms such as the extreme gradient boosting (XGBoost)

(Chen & Guestrin, 2016; Zamani Joharestani et al., 2019) which aims to measure the importance of input factors are applied to predict pollutant concentrations.

Motivated by the existing research work, this paper proposes a physical-temporal collection (PTC) model to reduce the bias in CMAQ predictor by exploring the short and long-term effects learned from historical records. This ensemble model includes a cascaded LSTM (C-LSTM) (Hochreiter & Schmidhuber, 1997; Li et al., 2017; Khan et al., 2020; Greff et al., 2016; Zhao et al., 2017) for time series prediction, and exploits the auxiliary information including meteorological data and the associations between air pollutants and seasonal factors to further improve the forecast. To better understand the importance and contribution of temporal and weather factors, we study the temporal pattern and weather pattern of the input sequences by XGBoost, which are used to prune the less important features in order to reduce noise. Our main contributions are summarised as follows:

1) We develop a physical temporal collection (PTC) model that integrates deep neural networks (NN) and LSTM methods, trained on historical pollutant concentrations and meteorological data with the CMAQ model, to provide accurate air quality prediction. Comparing with previous baseline models including CMAQ, the proposed PTC model improves the forecast accuracy and reduces the wrong extreme predictions.
2) We adopt XGBoost to extract both temporal and weather impacts to estimate the prominent factors of input variables and remove the insignificant contributions.
3) We evaluate the robustness of our proposed model by applying it to a London station, which has long-time records of air quality measurements, verifying the correctness of the data-driven and model-driven components and the overall PTC model. Results confirm it achieves improved performance by overcoming the drawbacks of each component.

The rest part of the paper is organized as follows. In Section 2, we propose the PTC model and describe the data. Section 3 compares our results with some previous methods, such as traditional neural networks and linear regression models. Future improvement and conclusion are covered in Section 4 and Section 5, respectively.

2 METHODS AND DATA

This section introduces the PTC model. The overall architecture is given in Figure 1. The main method includes two LSTM layers cascaded with the CMAQ model, and an Extreme Gradient Boosting (XGBoost) model cascaded with a DNN. To extract the temporal patterns at various scales, the XGBoost model first takes the following inputs: $D_1, D_2, ..., D_{24}$, 24-h CMAQ, 48-h CMAQ, 72-h CMAQ, where the first 24 inputs represent true records of the air pollutants, and the last three represent CMAQ values predicted for the next 24 hour, 48 hour and 72 hour. Then, the output is sent to the two cascaded LSTM layers. For each LSTM layer, we introduce a dropout layer to avoid over-fitting. The first LSTM layer has 50 nodes, with no returning sequences. The second one has 100 nodes and produces sequences. As a general setting, the drop rates are both set to 20%.

A key component of the PTC model is the bias correction module (the blue blocks in Figure 1), which takes meteorological data as the input to a XGBoost model and pass the output of XGBoost to a DNN and uses them for extracting the impact of weather. With the help of this meteorological feature extracting model, some important features dismissed by CMAQ model or LSTM can gain more weights, and therefore, improve the prediction accuracy. The correction module calculates the weight metrics using a neural network and determines the mutual influence of these values. The prediction value of this bias correction model is combined with the output of CMAQ cascaded two LSTM layers to give the final output. The XGBoost model uses tree booster with 0.05 learning rate and 0.07 gamma value. A dense layer using linear activation function is used for producing the output values. This model is trained using the ADAM optimizer with 0.01 learning rate, the batch size is set to 128, and the mean square error (MSE) is used as the loss function.

To train the PTC model, the CMAQ cascaded LSTM model need to be trained first. Then, this trained part will be compared with the real value and produce an error array, which will be used as the target output of the bias correction module. We choose the following meteorological measurements including temperature, humidity, wind speed, and air pressure given their direct impact on the concentration and dispersion of air pollutants.

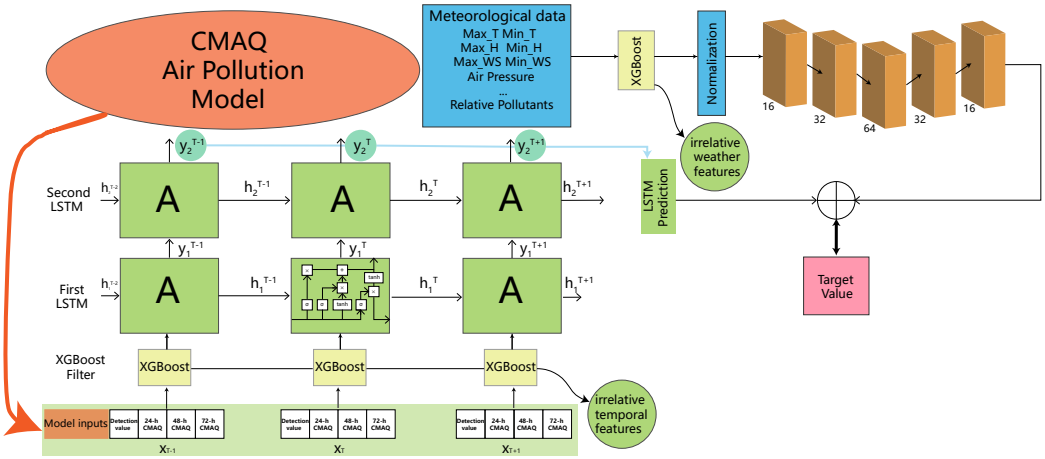

Figure 1: The PTC model.

2.1 Area and Data

This project collected over two years of air pollutants data in Chengdu, one of the most polluted cities in China. According to the statistics from the World Health Organization, the air quality index (AQI) of Chengdu has increased steadily from 2010 to 2018. Besides, the population of the city reached 16.33 million in 2018. Such a big number of citizens has high possibility to be exposed to air pollutants. Therefore, constructing an urban-wise and long-term air pollution system becomes significant to better control the pollutants and protect people's health. The Figure 2 indicates the distribution of air pollution monitoring stations and meteorological stations taken part in this study, which can cover the main urban area in Chengdu[1]. The 5 air quality monitoring stations provide a good coverage of the pollution conditions in Chengdu, and their names are "Jingquanlianghe", "Liangjiaxiang", "Shilidian", "Sanwayao", and "Shahepu". To be brief, we encode them based on their locations as follows in the rest of this paper: "Jinquanlianghe" refers to A1, , "Shahepu" refers to A2, "Liangjiaxiang" refers to A3, "Shilidian" refers to A4 , and "Sanwayao" refers to A5.

All air quality data are provided by the Environmental Protection and Research Institute of Chengdu[2]. The meteorological data are provided by the National Meteorological Information Center (NMICC), China[3]. The data collected and used in this paper is under the influence of subtropical humid monsoon climate, with high frequency static wind and significant urban heat island effect in summer and stable atmospheric stratification in winter. These characteristics are typical in low speed perennial wind or in basin terrain. The pollution type of Chengdu is coal-smoke air pollution (Wu et al., 2014), and the monitoring stations are situated mostly in the urban area, which suggest this dataset can represent an industrial metropolitan city model. The pollutants are most common ones including CO, NO_2, SO_2, ozone (one-hour average, $O_3 1h$), ozone (eight-hour average, $O_3 8h$). The meteorological data include the following items: maximum temperature (max_T), minimum temperature (min_T), maximum humidity (max_H), minimum humidity (min_H), maximum wind speed (Max_WS), minimum wind speed (Min_WS) and air pressure from 24 hours ago. We use meteorological data from 1 January 2016 to 30 June 2017 as the training data, and data after July 2017 as the testing and validation data. In addition, for each pollutant, we develop and train an individual model.

[1]The main urban area of Chengdu is approximately defined by its First Belt Highway.

[2]https://www.cmascenter.org/cmaq/

[3]http://data.cma.cn/

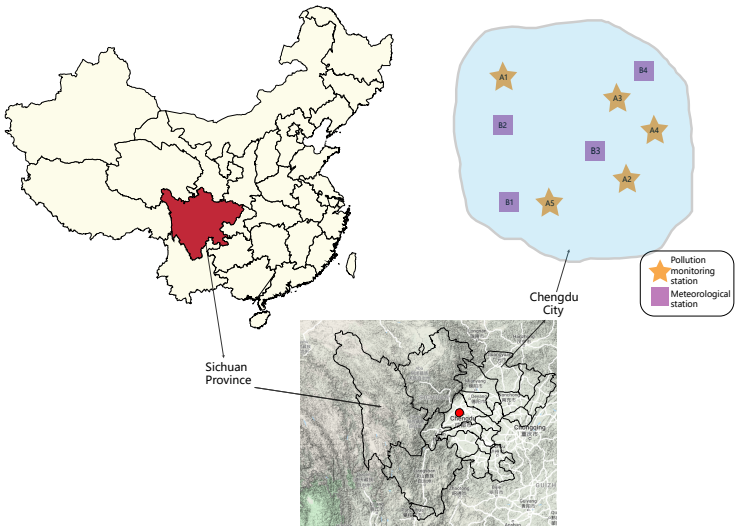

Figure 2: Locations of the air quality monitoring stations and meteorological stations in the urban area of Chengdu, Sichuan, China. Stations A1 to A5 are five main pollution monitoring stations distributed in Chengdu. B1 to B4 are 4 ground meteorological stations distributed in Chengdu.

3 Results and Evaluation

3.1 Model Performance Evaluation

We validate our model by measuring the Euclidean distance (ϵ_{base}) between CMAQ's 24 hours' prediction (the estimated future 24 hours' air pollutant indexes) and the true values, together with the distance (ϵ_{model}) between our model's predictions and the true values:

$$\varepsilon_{\text{base}} = \frac{1}{L}\sum_{t=1}^{\text{m}} \|Y_{\text{CMAQ24h}} - Y_{\text{true}}\|_2^2 \tag{1}$$

$$\varepsilon_{\text{model}} = \frac{1}{L}\sum_{t=1}^{\text{m}} \|Y_{\text{model}} - Y_{\text{true}}\|_2^2 \tag{2}$$

where L stands for the number of time point in the test set, $\text{Y}_{\text{CMAQ24h}}$ stands for the CMAQ 24 hour prediction value vector of CMAQ output, Y_{true} stands for the real value vecteor, Y_{model} stands for every model prediction vector and $||\boldsymbol{x}||$ is the L^2 distance of two vectors. To measure the accuracy, the following value is used:

$$Accuracy = \left(\varepsilon_{\text{base}} - \varepsilon_{\text{model}}\right)/\varepsilon_{\text{base}} \times 100\% \tag{3}$$

We use 24-h CMAQ prediction as our baseline instead of 48 or 72 hours for the reason that 24-h CMAQ prediction has the least Euclidean distance to the real value normally, and thus provides the most accurate estimation comparing the rest two.

3.2 Results

In this section, we evaluate our proposed PTC model against several methods including an integrated LSTM neural networks (LSTM+NN), Gated recurrent units with XGBoost (GRU+XGBoost), and DNN with XGBoost(DNN+XGBoost).

In Figure 3, the predicted daily PM2.5 ($\mu g/m^3$)and SO_2 (mg/m^3) concentrations (green dash line) are compared with the baseline CMAQ values (blue dotted line) and real concentrations (red line) at the five stations in Chengdu. Since the accurate air quality prediction is hard to obtain in summer caused by its dynamic change, we presented the forecast results from July 1st to August 5th, to evaluate the robustness of our model. As shown in Figure 3, the proposed PTC model with the L2

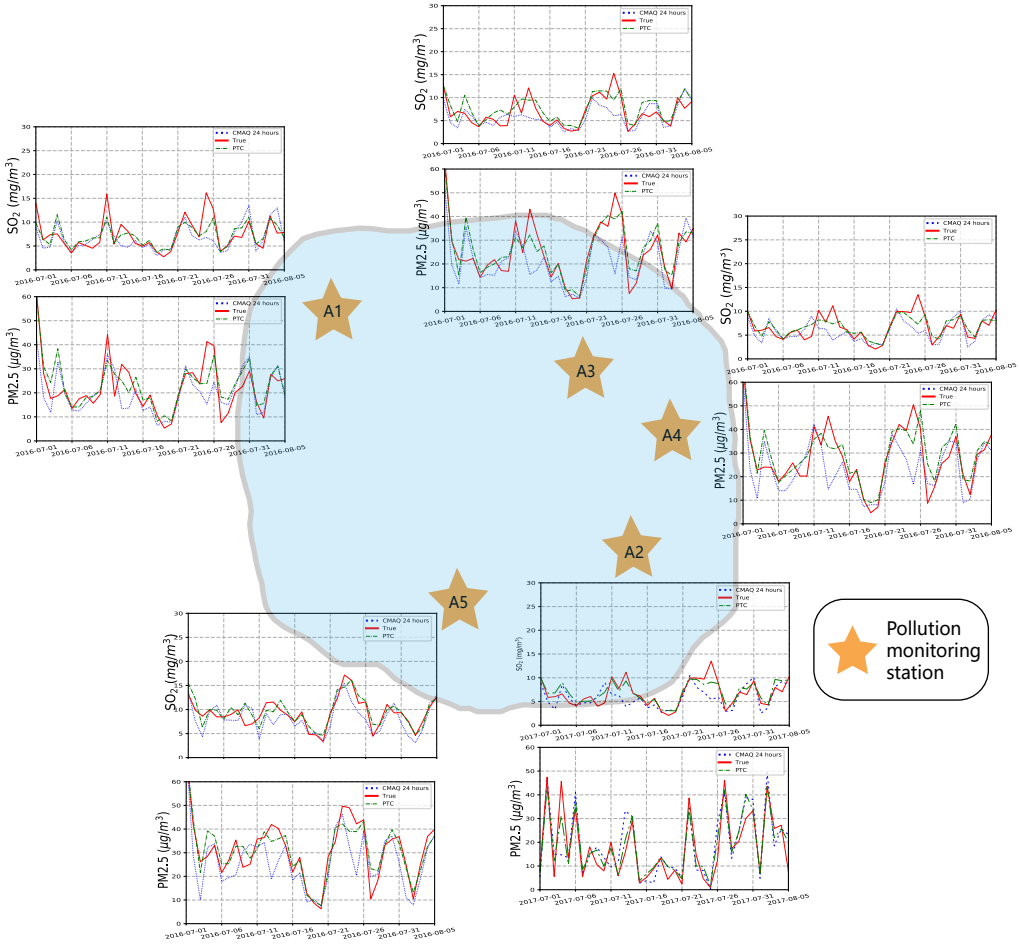

Figure 3: Prediction of PM2.5 and SO_2 in urban-wise area of Chengdu city

loss is able to suppress extreme values wrongly estimated by the CMAQ model and the predicted values are more close to the true measurements.

Figure 4 presents statistics of the five stations on the accuracy improvement against the baseline CMAQ model, as measured by how close the predicted values to the true measurements. We compare the four models and 4 main air pollutants include CO, NO_2, SO_2 and PM2.5. The final values are calculated using Equations (1), (2) and (3). The results reveal that our PTC model has least errors for CO, NO_2 and SO_2, where accuracy improvements are by 4% to 9% compared with the CMAQ prediction. For PM2.5, both the PTC model and CAMQ have relatively lower accuracy, and thus the improvement is not significant (around 1% or lower).

To further verify the robustness of our proposed model, we test it using data of air pollutant concentrations collected at London in 2018, through the UK-AIR (Air Information Resource[4]). Only the prediction of PM2.5 is illustrated due to its importance. We choose Bexley's records since it provides a more complete dataset. It can be observed in Figure 5 that PTC successfully suppresses extreme prediction values and the results more closely march the actual data in comparison with CMAQ model. In quantitative analysis, the proposed model improves the forecasting accuracy by 29.54%, showing the improved accuracy of the proposed model. We recognise the importance of carrying out extensive experiments on more stations and areas, as well as evaluating the performance on other air pollutants, which will be provided in an extended version. The dataset is accessible under request.

[4] https://www.cmascenter.org/cmaq/

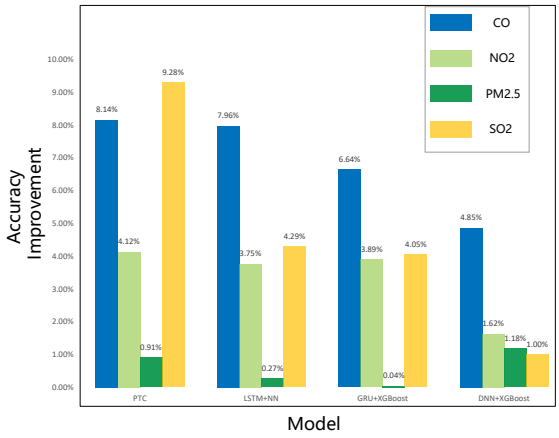

Figure 4: Accuracy improvement comparison

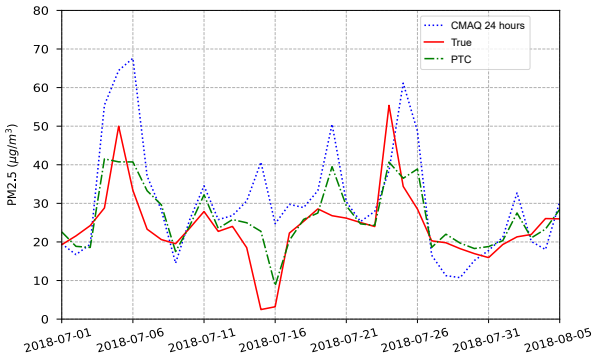

Figure 5: The prediction of daily PM2.5 concentration in Bexley, London

4 Conclusions

In this paper, we propose a physical temporal collection for accurately predicting air pollutants based on traditional air quality forecast models. The proposed model utilizes the traditional model data, historical records, and meteorological data to give a more accurate prediction on air pollutant concentrations. By observing the trend of real air quality and extracting meteorological impacts, the CMAQ model prediction accuracy is further improved. In addition, the feature importance, which includes meteorological data and public holidays, are examined to see how much these features are influencing on the performance of CMAQ model.

The assessment on different models show that our PTC model produces more accurate results comparing with traditional models. It is worth noting that, with the feature importance mechanism via XGBoost, our model can utilize parameters more flexibly and efficiently. In the future, this PTC model can contribute to the original ensemble process for producing more accurate air quality forecast.

References

K Wyat Appel, Alice B Gilliland, Golam Sarwar, and Robert C Gilliam. Evaluation of the community multiscale air quality (cmaq) model version 4.5: sensitivities impacting model performance: part i—ozone. *Atmospheric Environment*, 41(40):9603–9615, 2007.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. ACM, 2016.

H JS Fernando, MC Mammarella, G Grandoni, P Fedele, R Di Marco, R Dimitrova, and P Hyde. Forecasting pm10 in metropolitan areas: Efficacy of neural networks. *Environmental pollution*, 163:62–67, 2012.

KM Foley, SJ Roselle, KW Appel, PV Bhave, JE Pleim, TL Otte, R Mathur, G Sarwar, JO Young, RC Gilliam, et al. Incremental testing of the community multiscale air quality (cmaq) modeling system version 4.7. *Geoscientific Model Development*, 3(1):205–226, 2010.

Minglei Fu, Weiwen Wang, Zichun Le, and Mahdi Safaei Khorram. Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model. *Neural computing and applications*, 26(8):1789–1797, 2015.

Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10): 2222–2232, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

A. H. Khan, X. Cao, S. Li, and C. Luo. Using social behavior of beetles to establish a computational model for operational management. *IEEE Transactions on Computational Social Systems*, pp. 1–11, 2020. ISSN 2373-7476. doi: 10.1109/TCSS.2019.2958522.

Navneet Kumar, Anirban Middey, and Padma S Rao. Prediction and examination of seasonal variation of ozone with meteorological parameter through artificial neural network at neeri, nagpur, india. *Urban Climate*, 20:148–167, 2017.

Jos Lelieveld, John S Evans, Mohammed Fnais, Despina Giannadaki, and Andrea Pozzer. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569):367, 2015.

Xiang Li, Ling Peng, Xiaojing Yao, Shaolong Cui, Yuan Hu, Chengzeng You, and Tianhe Chi. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental pollution*, 231:997–1004, 2017.

Samuel Lightstone, Fred Moshary, and Barry Gross. Comparing cmaq forecasts with a neural network forecast model for pm2. 5 in new york. *Atmosphere*, 8(9):161, 2017.

Harri Niska, Teri Hiltunen, Ari Karppinen, Juhani Ruuskanen, and Mikko Kolehmainen. Evolving the neural network model for forecasting air pollution time series. *Engineering Applications of Artificial Intelligence*, 17(2):159–167, 2004.

Ashley Queen and Yang Zhang. Examining the sensitivity of mm5–cmaq predictions to explicit microphysics schemes and horizontal grid resolutions, part iii—the impact of horizontal grid resolution. *Atmospheric Environment*, 42(16):3869–3881, 2008.

Libin Wu, Shuhua Zhou, Changjian Ni, Peichuan Liu, and Kun Liu. Study on spatial and temporal variation of haze in chengdu and surrounding areas. *Plateau and Mountain Meteorology Research*, 2:63–67, 2014.

Xingrui Yu, Xiaomin Wu, Chunbo Luo, and Peng Ren. Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sensing*, 54(5):741–758, 2017. doi: 10.1080/15481603.2017.1323377. URL `https://doi.org/10.1080/15481603.2017.1323377`.

Mehdi Zamani Joharestani, Chunxiang Cao, Xiliang Ni, Barjeece Bashir, and Somayeh Talebiesfandarani. Pm2. 5 prediction based on random forest, xgboost, and deep learning using multisource remote sensing data. *Atmosphere*, 10(7):373, 2019.

Zheng Zhao, Weihai Chen, Xingming Wu, Peter CY Chen, and Jingmeng Liu. Lstm network: a deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 11(2): 68–75, 2017.