# Improving Intra-Urban Prediction of Atmospheric Fine Particles Using a Hybrid Deep Learning Approach

**Zhengyu Zhang [1], Jiuchun Ren [1] and Yunhua Chang [2,*]**

[1] Department of Communication Engineering, School of Information Science and Engineering, Fudan University, Shanghai 200433, China; 21210720078@m.fudan.edu.cn (Z.Z.)
[2] Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), NUIST Center on Atmospheric Environment, Nanjing University of Information Science and Technology, Nanjing 210044, China
* Correspondence: changy13@nuist.edu.cn

**Supplementary Information Text**

**Text S1.** Hourly averaged concentrations of air pollutants and meteorological information including 52705 instances and 15 attributes were collected at six monitoring stations distributed across the city of Pudong, Shanghai (Fig. S1). These stations are supervised by the Ministry of the Environment. Air pollutant data include the hourly concentration of PM2.5, PM10, SO2, NO2, CO, and O3. Hourly meteorological data include temperature, humidity, precipitation, wind direction, and wind speed. The average temperature in Shanghai is 27℃ in summer (DJF) and 15◦C in winter, and the average annual precipitation is 1178.2mm.

We used data during the entire 2020 year, with eight months used for training (adjusting the parameters of our models), and three months for the test (to check the performance of our models).

The data used in this study are shown in Table S1 and contain both meteorological data and air pollution data.

**Table S1.** Air quality and meteorological data used in the modeling study.

| | | Area | Unit |
|---|---|---|---|
| | **Input parameters** | | |
| 1 | $PM_{2.5}$ | Township level | μm/m3 |
| 2 | $PM_{10}$ | Township level | μm/m3 |
| 3 | $SO_2$ | Township level | μm/m3 |
| 4 | CO | Township level | μm/m3 |
| 5 | $NO_2$ | Township level | μm/m3 |
| 6 | $O_3$ | Township level | μm/m3 |
| 7 | Temperature | District level | ℃ |
| 8 | Wind direction | District level | ° |
| 9 | Wind speed | District level | m/s |
| 10 | Pressure | District level | hPa |
| 11 | Precipitation | District level | mm |
| 12 | Humidity | District level | - |
| | **Output parameter** | | |
| 1 | Next hour's $PM_{2.5}$ concentration | Township level | μm/m3 |

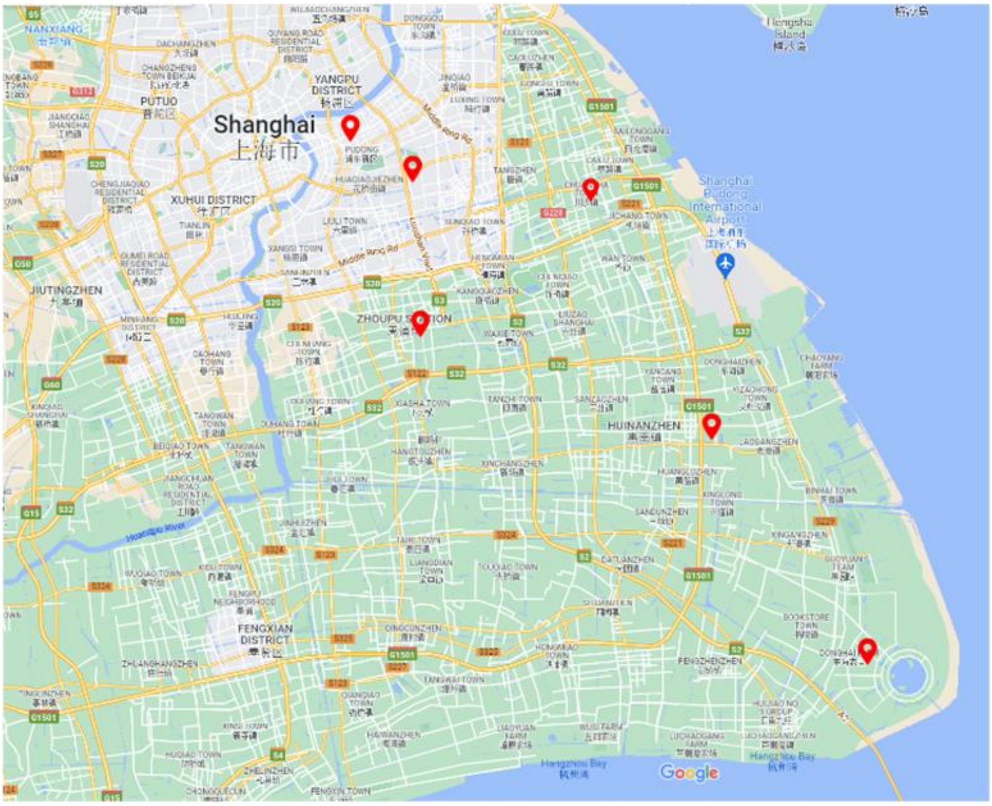


**Figure S1.** Map of Pudong, Shanghai, indicating the location of the city's air quality monitoring stations.

**Text S2.** Given the limitation from environment and monitor equipment, it is normal for meteorological and pollution data to have missing values. Therefore, it is necessary to have a method that addresses these gaps in the data. Because the amount of missing data is huge and thus does not allow a reliable daily reconstruction of the information, we apply an efficient data filling method based on linear spline imputation. This method can predict the missing value by penalizing the absolute error between the limited value and the predicted value calculated by a linear function, and it has been widely applied to interpolation problems.

PM2.5 concentrations can be highly influenced by other air pollutants and meteorological parameters. Therefore, studying the non-linear correlation between various pollutants not only helps us to have a general understanding of various pollutants from a statistical analysis point of view, but also proves the importance of the extraction of information in the cross-domain air quality data. By using Pearson's correlation coefficient, we calculated the correlation matrix between air pollutants and meteorological parameters, and some of the results are shown in Fig. S2. From the figure, we can see that PM2.5 concentration shows a strong positive correlation with pollutant data such as SO2 concentration and NOx concentration, a negative correlation with wind speed and humidity, and a weak positive correlation with air pressure. Also, we found various positive and negative correlations between and within the pollutant data and weather data. Analyzing and studying these correlations is the basis for our proper modeling and reasoning about the changes in PM2.5 concentrations.

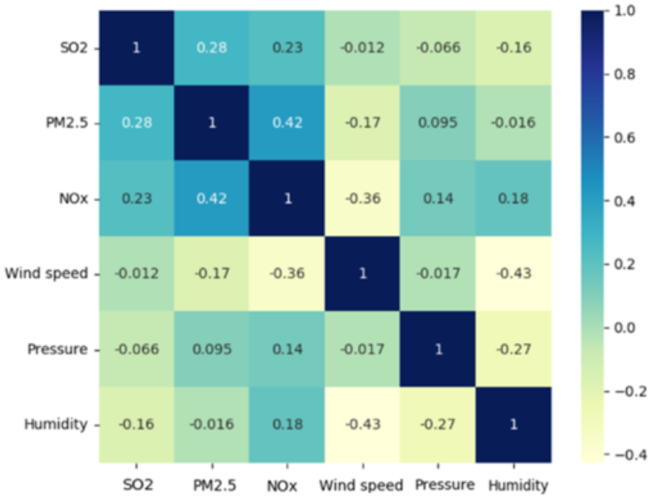


**Figure S2.** Correlation matrix for SO2, PM2.5, NOX, wind speed, pressure, and humidity. The Spearman correlation is represented as a color scale, with deep blue indicating R = 1 and yellow indicating R = -1.

**Text S3.** The hyperparameters in the proposed model are identified during the training process, which means that we get the best parameter setting based on the validation set through the RMSE. We used 128 filters in CNN, and the kernel size is 3. In LSTM the output dimension is set as 50, and the learning rate is fit in the learning process. Our model is built on Keras and Tensorflow, and we train all models using the SGD optimizer with the batch size of 40.

The results are compared with the baseline approaches such as CNN, LSTM, CNN-LSTM (Wu and Lin, 2019a), SVR (Murillo-Escobar et al., 2019), FDN-learning (Zou et al., 2021), and ARIMA. Among the baseline approaches, the CNN-LSTM, CNN, and FDN-learning were utilized for predicting the air pollutant concentration considering the meteorological parameter and data of other stations as auxiliary inputs, while the other methods are merely time series prediction models.

## References


1. Wu, Q.L.; Lin, H.X. A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. *Sci. Total Environ.* **2019**, *683*, 808–821. https://doi.org/10.1016/j.scitotenv.2019.05.288.
2. Murillo-Escobar, J.; Sepulveda-Suescun, J.P.; Correa, M.A.; Orrego-Metaute, D. Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburra Valley, Colombia. *Urban Clim.* **2019**, *29*, 100473. https://doi.org/10.1016/j.uclim.2019.100473.
3. Zou, G.J.; Zhang, B.; Yong, R.H.; Qin, D.M.; Zhao, Q. FDN-learning: Urban PM2.5-concentration Spatial Correlation Prediction Model Based on Fusion Deep Neural Network. *Big Data Res.* **2021**, *26*, 100269. https://doi.org/10.1016/j.bdr.2021.100269.