COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

EAEE E4220 – Machine Learning Applications
for Environmental Engineering and Sciences

Integrated Air Quality Predictor by Recurrent Neural Network with Gated Recurrent Unit

Final Project

Submitted by:

Tianxiao Shen(ts3326)

Faculty Advisor:

Prof. Pierre Gentine

https://github.com/Stx980212/Integrated-Air-Quality-Predictor-by-RNN-with-GRU

December 2021

Columbia University in the City of New York

New York, New York

1. Introduction

In recent years, the environmental issue has become a heated topic, especially about air quality. As the report from WHO and PICC indicates, people worry that, with the further expansion of industrialization and overexploitation of fossil fuels taking place all around the world, the quality of air would deteriorate and finally pose a serious threat to public health. In order to derive the current and future situation of air pollution and inform the public about the ongoing pollution event, a reliable and accurate prediction method of air quality in both long term and short term is essential.

With the rapid development of machine learning, a study of computer algorithms that can improve automatically through experience and by the use of data, nowadays people are trying to apply the methods of machine learning to predict the Air Quality Index (AQI). Previous study has used neural networks to make a prediction on the AQI of the following days, based on the historical AQI and weather conditions data.

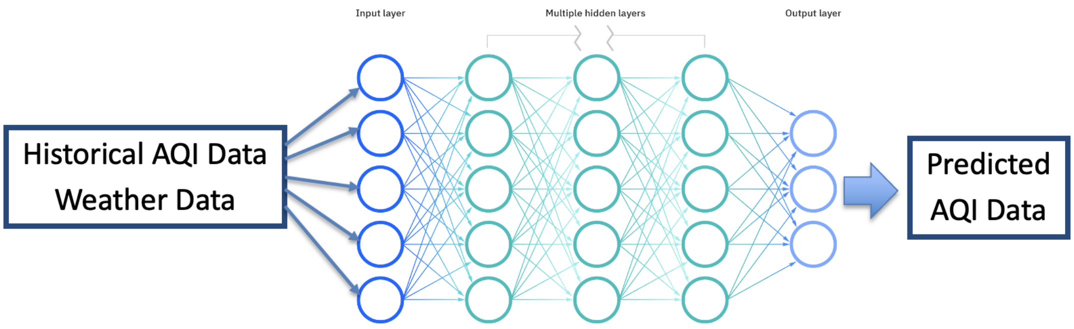


Figure 1. Diagram of previous approach to predict AQI data with machine learning.

Figure 1 shows the diagram of past machine learning approaches to predict AQI data, where layers of simple feedforward neural networks were used to process the past AQI and weather data and make predictions for the future. But this prediction method can hardly reveal the real factor that drives AQI value up and down, thus not being able to produce reliable results.

Actually, AQI is not an index that is derived from completely numerical calculations with the concentrations of pollutants in the atmosphere, instead it involves the comparison between the severity of different major pollutants in the atmosphere. The pollutant which falls in the most serious concentration category is chosen as the principal pollutant of that time, and its concentration is then converted into the value of AQI. So in order to accurately predict the AQI, the concentration of AQI pollutants, including Particulate Matters (mainly $PM_{2.5}$), Sulphur dioxide (SO_2), and Nitrogen Oxides (NO_x), also in some cases, ozone, should be correctly predicted first.

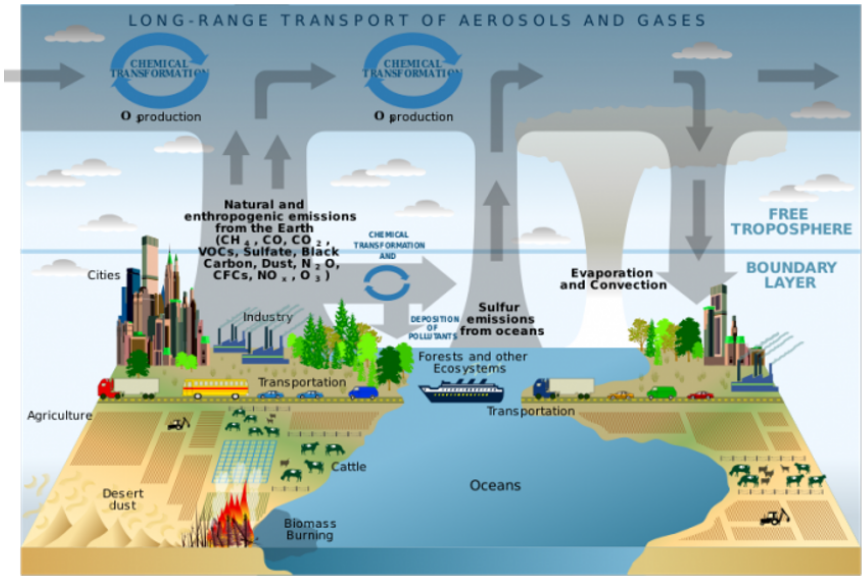

LONG-RANGE TRANSPORT OF AEROSOLS AND GASES
CHEMICAL TRANSFORMATION
O production
CHEMICAL TRANSFORMATION
O production
Natural and enthropogenic emissions from the Earth (CH 4 , CO, CO 2 , VOCs, Sulfate, Black Carbon, Dust, N 2 O, CFCs, NO x , O 3)
CHEMICAL TRANSFORMATION AND
DEPOSITION OF POLLUTANTS
Sulfur emissions from oceans
Evaporation and Convection
FREE TROPOSPHERE
BOUNDARY LAYER
Cities
Industry
Forests and other Ecosystems
Transportation
Transportation
Agriculture
Cattle
Oceans
Desert dust
Biomass Burning

Figure 2. Diagram of atmospheric processes affecting the pollutants.

As is shown in Figure 2, the changing of atmospheric pollutants could be attributed to a sophisticated diagram of various atmospheric mechanisms, mainly including emission, diffusion, photochemical reaction, and transportation across aeras. To take those mechanisms into account, more sources of data other than merely historical AQI and weather data need to be listed in the input data of the prediction.

The aim of this project is to propose an integrated air quality prediction model with machine learning, which is able to produce a reliable and accurate prediction for the concentration of major AQI pollutants. The potential method and algorithm is discussed in the later part of the report.

2. Method and Model

2.1. Universal Differential Equations

Proposed recently, Universal Differential Equations (UDEs) is a combination of the concept of the Universal Approximator and the Physical-Informed Neural Network, which is designed to derive all kinds of scientific relationships in the form of differential equations from the observation data.

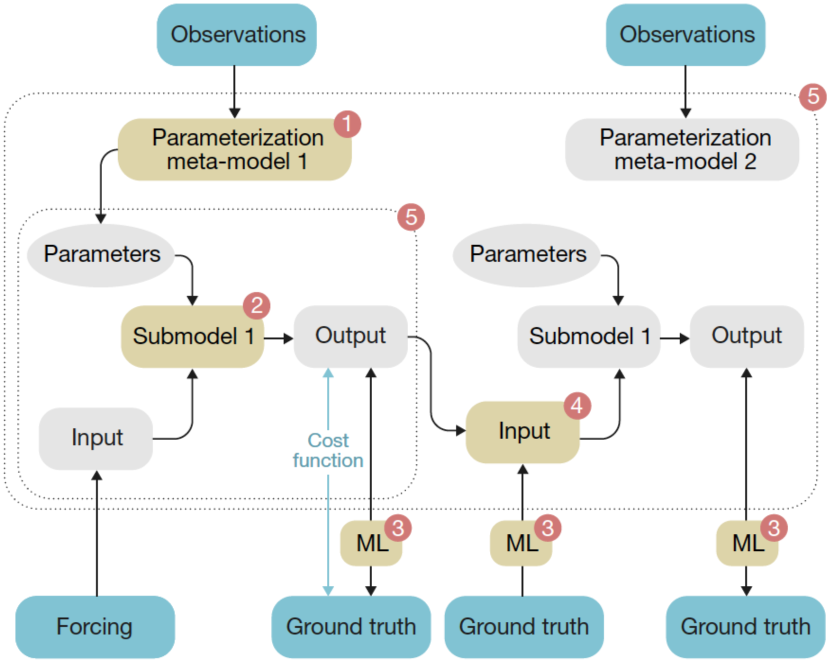


Figure 3. Diagram of algorithm of UDEs.

As is shown in the Figure 3, though the structure of UDEs is normally very complicated, the differentiable programming applied in its calculation and the ability of incorporating any prior knowledge into the model endow UDEs with high data efficiency and computing speed.

$$\frac{\partial C_{Pollutants}}{\partial t} = \underbrace{\left(Q_{Emission} + Q_{Transportation} + Q_{Reaction}\right) * E_{Diffusion}}_{\text{Effect of sources and sinks}} + \underbrace{\frac{C_{Pollutants}}{E_{Diffusion}} * \frac{\partial E_{Diffusion}}{\partial t}}_{\text{Effect of changing diffusion condition}}$$

$Q_{Emission}$

A neural network representing emission of pollutants

Input data:

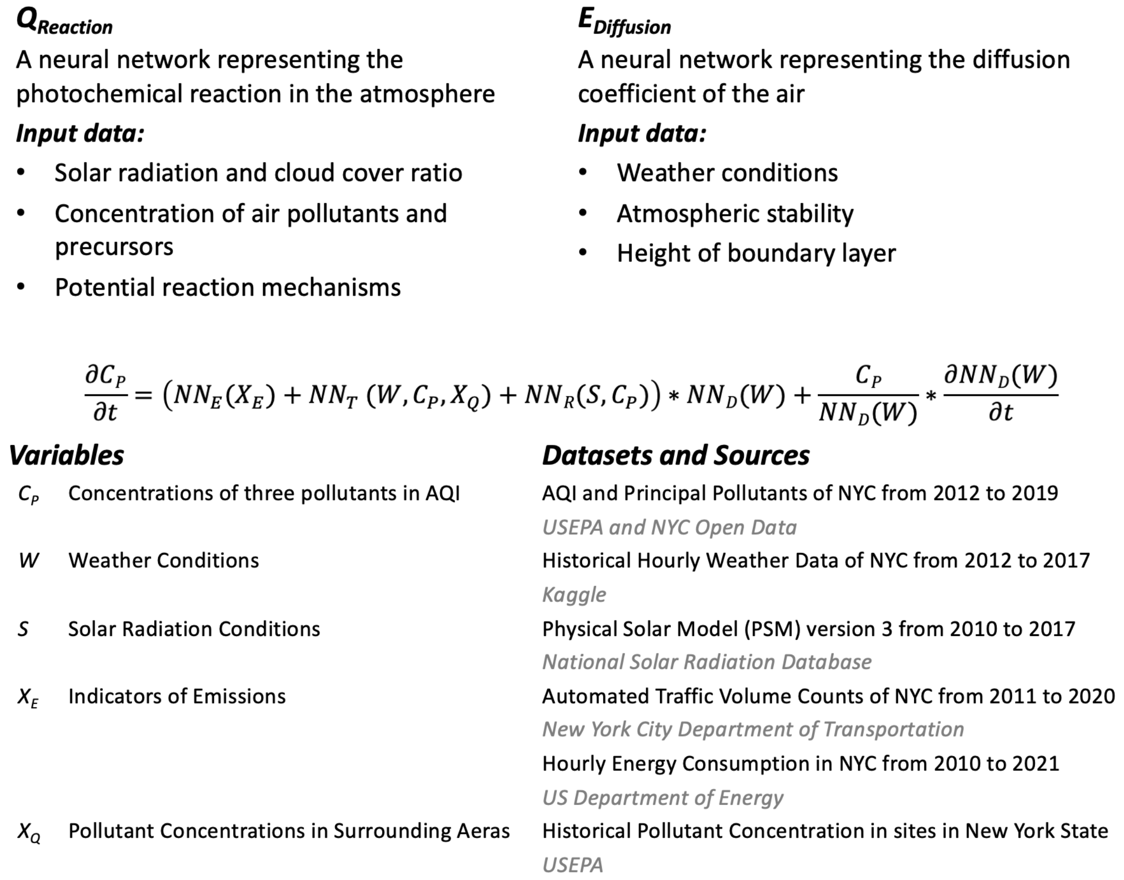
- Automated Traffic Volume Counts
- Emission factor of factories
- Electricity and gas consumption

$Q_{Transportation}$

A neural network representing importation and exportation of pollutants

Input data:

- Concentration of air pollutants
- Wind speed and wind direction
- AQI, direction and distance of surrounding aeras



$Q_{Reaction}$

A neural network representing the photochemical reaction in the atmosphere

Input data:

- Solar radiation and cloud cover ratio
- Concentration of air pollutants and precursors
- Potential reaction mechanisms

$E_{Diffusion}$

A neural network representing the diffusion coefficient of the air

Input data:

- Weather conditions
- Atmospheric stability
- Height of boundary layer

$$\frac{\partial C_P}{\partial t} = \left(NN_E(X_E) + NN_T\,(W, C_P, X_Q) + NN_R(S, C_P)\right) * NN_D(W) + \frac{C_P}{NN_D(W)} * \frac{\partial NN_D(W)}{\partial t}$$

Variables		***Datasets and Sources***
C_P	Concentrations of three pollutants in AQI	AQI and Principal Pollutants of NYC from 2012 to 2019 *USEPA and NYC Open Data*
W	Weather Conditions	Historical Hourly Weather Data of NYC from 2012 to 2017 *Kaggle*
S	Solar Radiation Conditions	Physical Solar Model (PSM) version 3 from 2010 to 2017 *National Solar Radiation Database*
X_E	Indicators of Emissions	Automated Traffic Volume Counts of NYC from 2011 to 2020 *New York City Department of Transportation* Hourly Energy Consumption in NYC from 2010 to 2021 *US Department of Energy*
X_Q	Pollutant Concentrations in Surrounding Aeras	Historical Pollutant Concentration in sites in New York State *USEPA*

Figure 4. Designed UDEs structure and formula for AQI prediction.

Figure 4 shows the preliminary design of the structure, formula, and input data for the Integrated Air Quality Predictor, which was my initial plan for this project.

<u>The reason why UDEs is not working in this project is</u>, the differentiable programming is designed to explore the gradients of all variables at the same time, which means, inside a UDEs predictor of the AQI data, the gradient of all the variables, including weather, solar, and emission conditions, will be derived from other input variables with a given formula or a neural network as an approximator. And the rest of UDEs are made up of known and unknown parameters, which does not change with the time and could not be defined by a dataset over a time series. But obviously, there is actually no reliable method to derive the gradients of these input variables with themselves. So, as long as the current UDEs does not support an input dataset without deriving its gradient, a reliable prediction of AQI data could not be expected with UDEs at the time being.

2.2. RNN

Compared to feedforward neural networks, Recurrent Neural Networks is more suitable for the data in a consecutive time series.

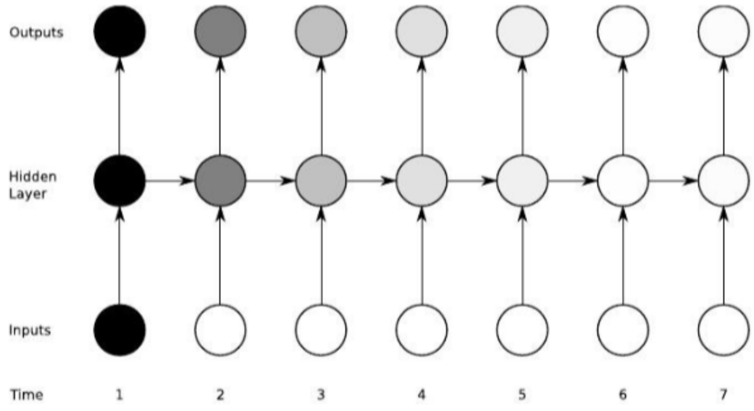


Figure 5. Diagram of Recurrent Neural Network.

Figure 5 shows the diagram of RNN, where the state of the hidden layer in the previous time would be the input of the hidden layer in the next time step, so that the hidden layer could represent the current state of certain variables inside a system over changing time. Designed for language and picture processing, RNN is now widely used in the prediction over a time series, which is suitable for the AQI prediction in this case.

2.3. LSTM and GRU

Vanilla RNNs, or RNNs with normal hidden layers, are trained by back propagation, while due to the rounding error of the computing, its long-term gradients during back propagation usually tend to be zero, which is also known as *vanish gradient problem*, indicates the inability for vanilla RNNs to make reliable predictions over a long-term factor.

Using Long-Short Term Memory (LSTM) as the gating mechanism of the hidden layer of RNN could avoid such vanish gradient problems, thus providing a better prediction which could correctly calculate the long-term gradients.

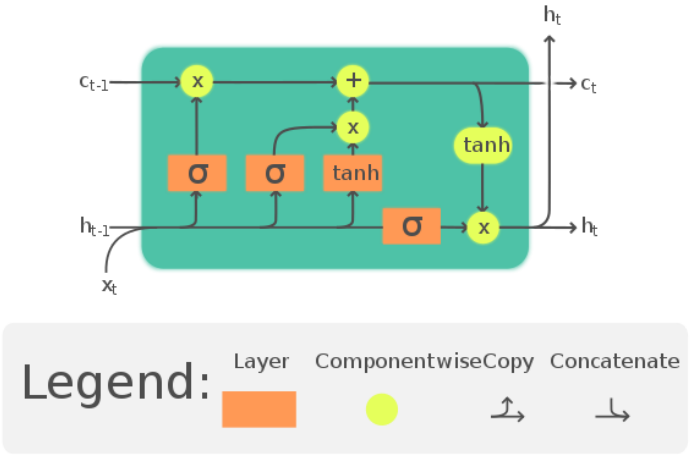


Figure 6. Diagram of an ordinary LSTM.

Figure 6 shows the diagram of a common LSTM unit, which is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

Recently, another gating mechanism called Gate Recurrent Unit (GRU) was proposed. Similar to LSTM, GRU is also designed to solve the problem of vanishing gradient in the back propagation, though GRU does not include an output gate and has fewer parameters than LSTM.

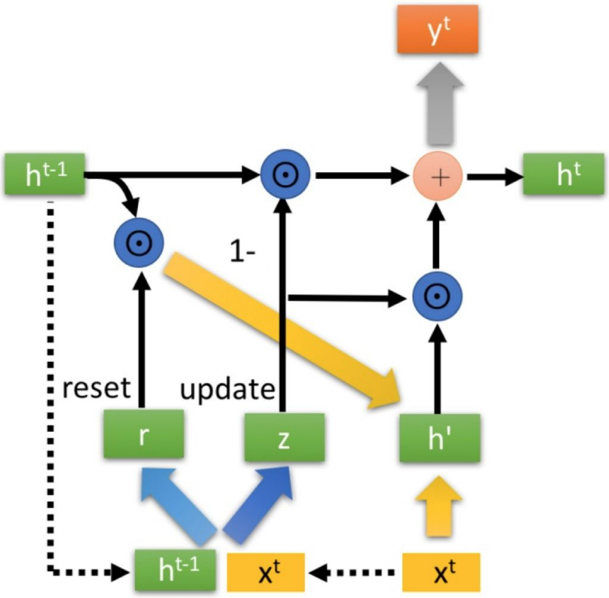


Figure 7. Diagram of GRU.

Figure 7 shows the diagram of GRU units. As is reported in some research, compared with LSTM, the use of GRU can achieve comparable results and is easier to train than LSTM, which can largely improve the training efficiency, so it is often preferred to use GRU.

2.4. Problems and Plans

For the purpose of selecting and proposing a reliable and accurate method for the air quality prediction with RNN, some questions should be answered: 1) How much data is sufficient to make the best possible prediction? In other words, would the introduction of more sources of data further improve the accuracy of prediction? 2) Which gating mechanism of RNN, LSTM or GRU, would give a better performance in this case? 3) What structure of the neural network would produce the optimal learning process and the lowest prediction error? In order to explore the answer to these questions, a comprehensive training and analysis of these models is conducted in this project.

In the first part of this project, a simple RNN autoregression with LSTM on the historical concentration of AQI pollutants in the past 24 hours is conducted as the base case of this project. With only one LSTM layer, this simple model is applied to investigate the autoregression ability of these concentration data, which means the feasibility of predicting their future behavior based on their past trends.

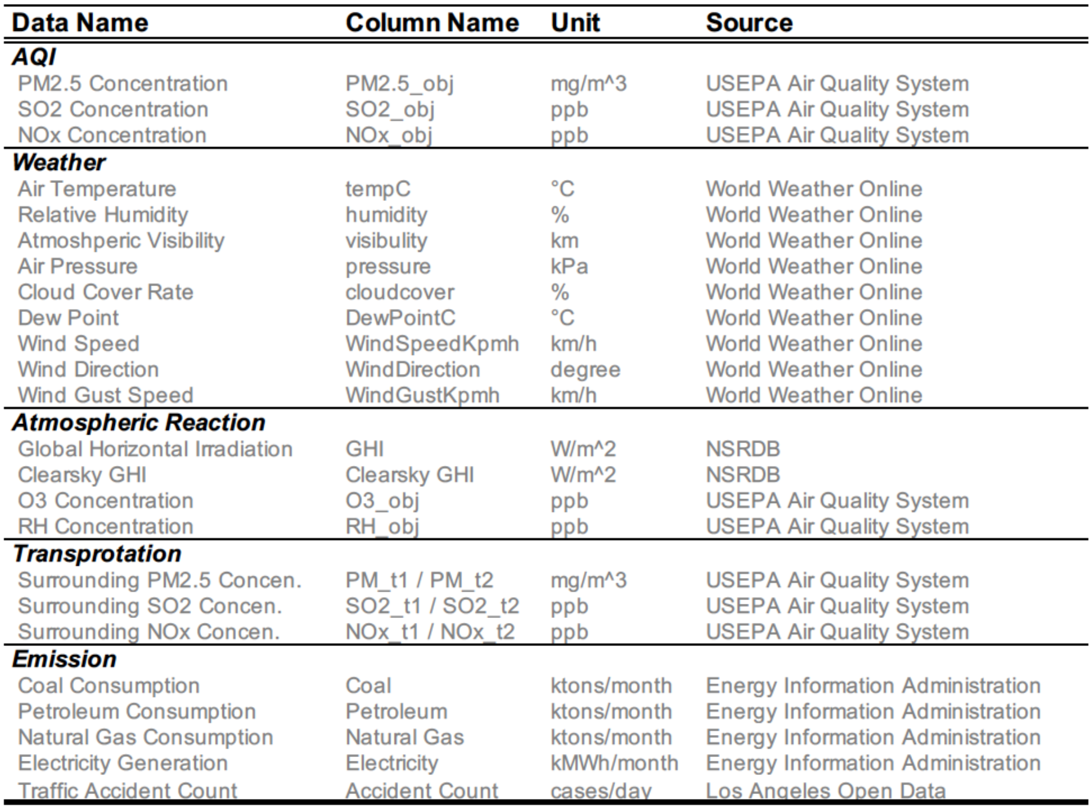
The second part of this project involves a RNN prediction model on the historical weather and AQI data, which is designed as an improvement to the past approach of using feedforward neural networks to predict the air quality. Compared to the simple autoregression model, this prediction model will have a more complicated structure to integrate the weather condition into the changing trend of the concentration of pollutants. Also, two different gating mechanisms in the hidden layer, LSTM and GRU, will be trained and analyzed respectively to make a comparison between them.

For the last part of this project, the Integrated Air Quality Predictor will be proposed. Established with the most sophisticated structure among other models in this project, the aim of this model is to integrate various kinds of data from different sources to derive multiple mechanisms including emission, diffusion, photochemical reaction, and transportation, then make a complete analysis across different factors. Similar to the previous RNN prediction model, the comparison between LSTM and GRU will also be presented, and the one with better performance will be chosen as the recommended method.

3. Data

3.1. Datasets, Categories and Sources

Table 1. Summary of obtained datasets by category



Data Name	Column Name	Unit	Source
AQI			
PM2.5 Concentration	PM2.5_obj	mg/m^3	USEPA Air Quality System
SO2 Concentration	SO2_obj	ppb	USEPA Air Quality System
NOx Concentration	NOx_obj	ppb	USEPA Air Quality System
Weather			
Air Temperature	tempC	°C	World Weather Online
Relative Humidity	humidity	%	World Weather Online
Atmoshperic Visibility	visibulity	km	World Weather Online
Air Pressure	pressure	kPa	World Weather Online
Cloud Cover Rate	cloudcover	%	World Weather Online
Dew Point	DewPointC	°C	World Weather Online
Wind Speed	WindSpeedKpmh	km/h	World Weather Online
Wind Direction	WindDirection	degree	World Weather Online
Wind Gust Speed	WindGustKpmh	km/h	World Weather Online
Atmospheric Reaction			
Global Horizontal Irradiation	GHI	W/m^2	NSRDB
Clearsky GHI	Clearsky GHI	W/m^2	NSRDB
O3 Concentration	O3_obj	ppb	USEPA Air Quality System
RH Concentration	RH_obj	ppb	USEPA Air Quality System
Transprotation			
Surrounding PM2.5 Concen.	PM_t1 / PM_t2	mg/m^3	USEPA Air Quality System
Surrounding SO2 Concen.	SO2_t1 / SO2_t2	ppb	USEPA Air Quality System
Surrounding NOx Concen.	NOx_t1 / NOx_t2	ppb	USEPA Air Quality System
Emission			
Coal Consumption	Coal	ktons/month	Energy Information Administration
Petroleum Consumption	Petroleum	ktons/month	Energy Information Administration
Natural Gas Consumption	Natural Gas	ktons/month	Energy Information Administration
Electricity Generation	Electricity	kMWh/month	Energy Information Administration
Traffic Accident Count	Accident Count	cases/day	Los Angeles Open Data

Table 1 shows all the obtained data from various sources, grouped by the categories of their potential mechanisms to affect the concentration of AQI pollutants. Except the energy consumption data from US EIA, where only monthly data in whole states are available, all the datasets contain hourly data of 17544 hours from 2016 to 2017 in local Los Angeles. All the objective AQI pollutant concentrations are the raw hourly sampling result of observation site 4008 and 4004 in Los Angeles, California set by USEPA, and surrounding AQI pollutant concentrations are the result of site 5005 and 1103, which are added to the machine learning process to show the transportation of pollutants from surrounding areas.

3.2. Data Preprocessing

3.2.1. Wind Speed and Wind Direction

In the data obtained from the source, the property of wind is presented by two variables, wind speed in km/h and wind direction in degree. This representation of wind property has two fatal disadvantages. Firstly, for the wind direction, 0 degree and 359 degree are actually adjacent directions, but they are the lowest and the highest value presented by wind direction. Then, even if the wind speed is very close to zero, the wind direction still serves as an independent variable with a unique weight factor.

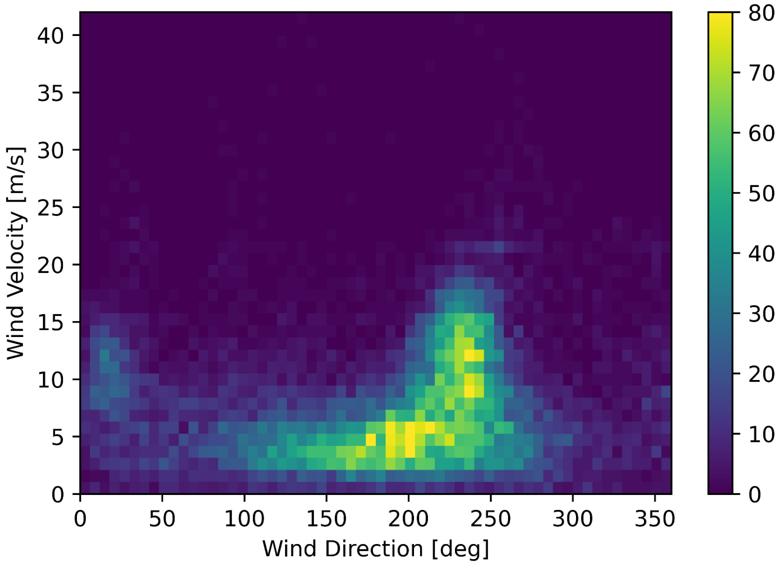


Figure 8. Frequency distribution of wind speed and wind direction.

Figure 8 shows the haphazard nature of the frequency distribution of wind speed and wind direction. Represented in this method, none of the two variables resemble a normal distribution, and similar wind could be defined as very different. So in order to improve the quality of the data, the wind speed and direction are split into the partial speed in the directions of x-axis and y-axis.

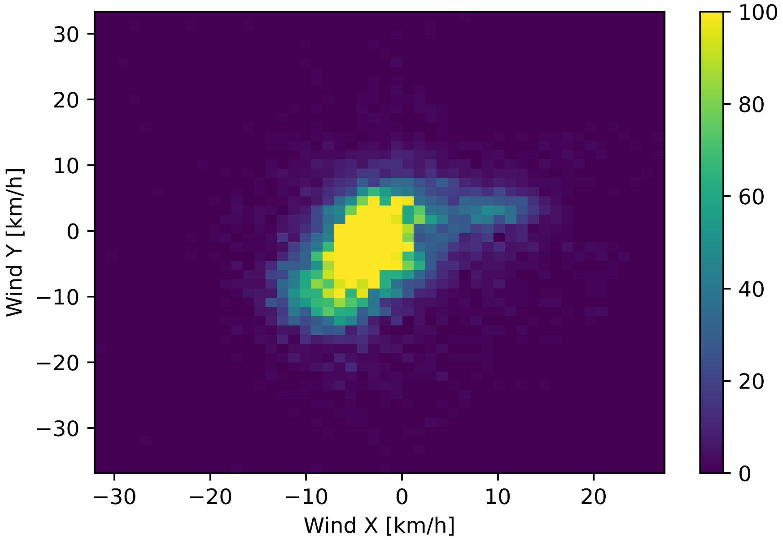


Figure 9. Frequency distribution of processed wind data.

The frequency distribution of processed wind data is shown in Figure 9, where the wind speed and direction is represented by the partial speed of two directions. The distribution of processed data resembles a normal distribution, and best avoids the disadvantages of representing in speed and direction.

3.2.2. Daily and Seasonal Period

The time in the obtained data includes numeric values of year, month, day, and hour for each row of the data, whose linear distribution is not able to derive the seasonal cycle. So instead of using those numeric values to represent different time and seasons, sine and cosine of hours and months are used to describe the daily and seasonal period respectively.

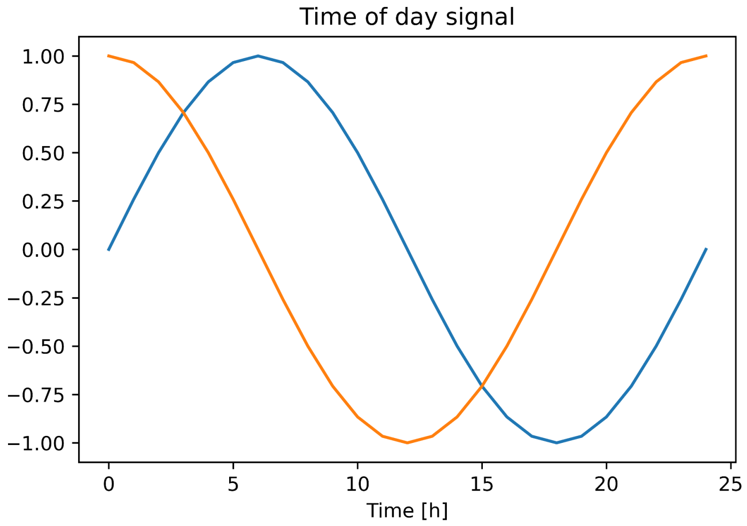


Figure 10. Daily period represented in sine and cosine of hours.

Figure 10 shows the processed hour data, and how it represents the daily cycle. Similar process is applied to the month data to represent the seasonal cycle.

3.3. Data Preview

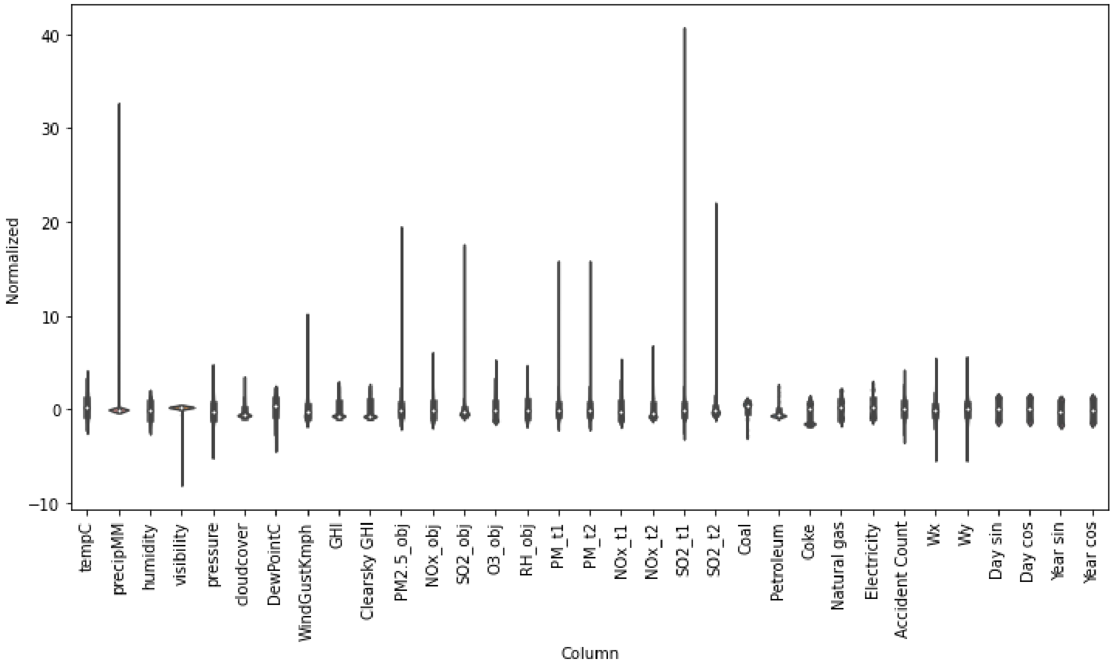


Figure 11. Distribution of normalized data of this project.

Figure 11 shows the distribution of normalized data involved in this project, which would be converted into 24-hour windows to form the input data of the project, in order to make predictions for the next hour based on the past 24-hour data.

4. Results

4.1. Simple RNN Autoregression with LSTM

To begin with, a simple autoregression through LSTM is conducted on the historical 24-hour concentration data of the three AQI pollutants as the base case of the whole project.

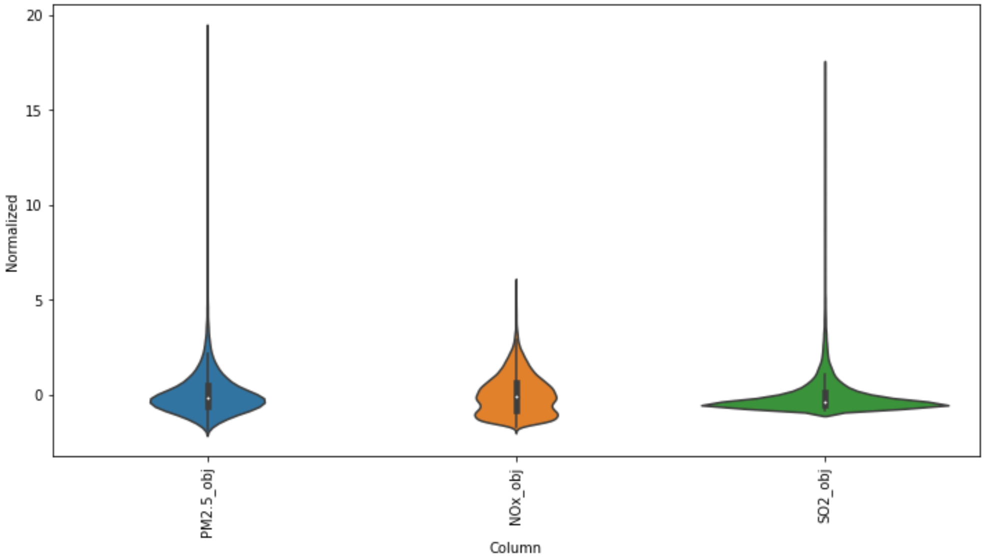
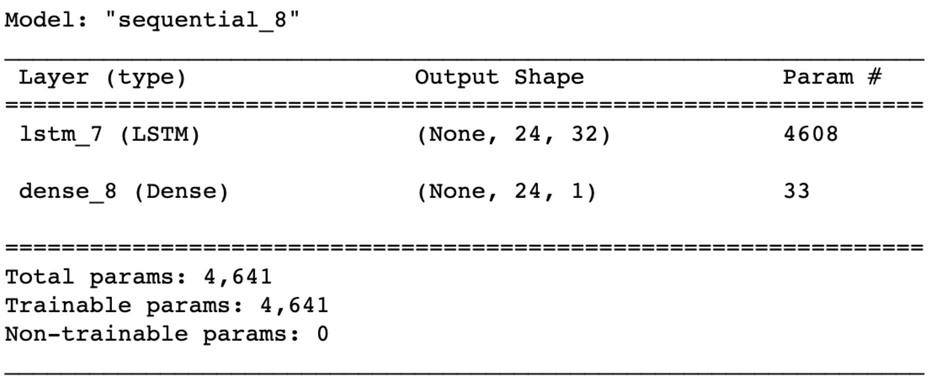


Figure 12. Distribution of the normalized concentration of AQI pollutants.

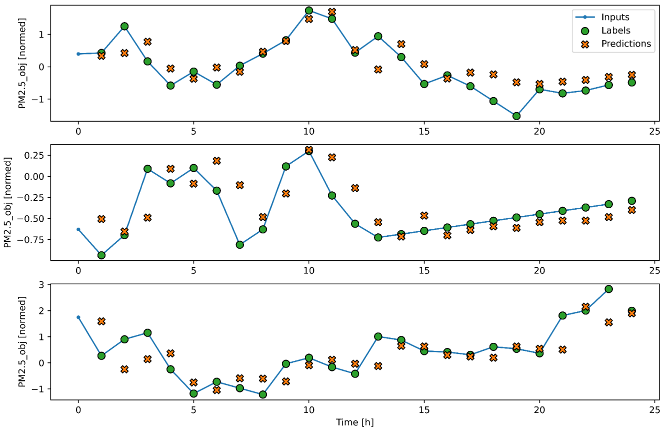
Figure 12 shows the distribution of normalized concentration of AQI pollutants, which would be the input data of this RNN autoregression model. All of the three concentrations are generally symmetrically distributed around the average, while the presence of occasional extreme high data would have an impact on the difficulty of prediction.

Table 2. Summary of simple autoregression LSTM.

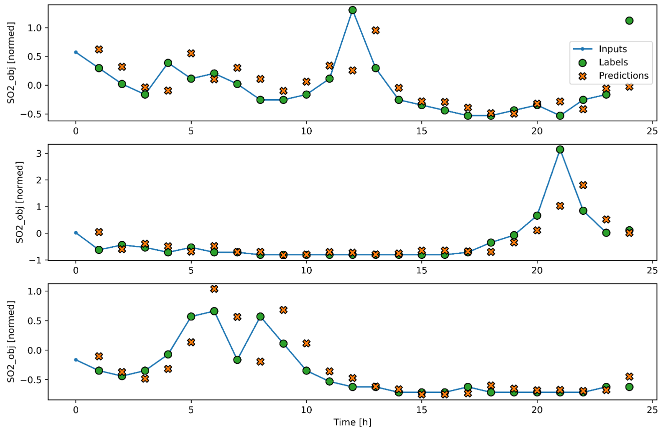


```
Model: "sequential_8"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_7 (LSTM)               (None, 24, 32)            4608

 dense_8 (Dense)             (None, 24, 1)             33

=================================================================
Total params: 4,641
Trainable params: 4,641
Non-trainable params: 0
_________________________________________________________________
```

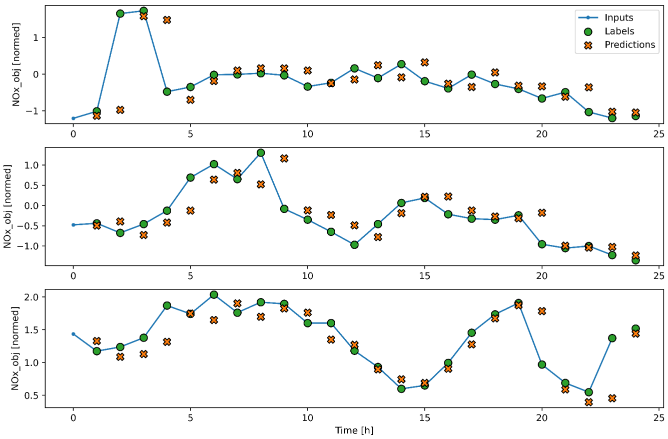
Table 2 shows the structure of the simple RNN autoregression with LSTM, with a single layer of LSTM with 32 neurons and a single output layer with one resulting neuron for each concentration of AQI pollutants. It is set to predict the concentration in the following hour with the data in the past 24 hours.



(a)



(b)



(c)

Figure 13. Prediction Result of simple RNN autocorrelation with LSTM on (a) PM2.5 concentration, (b) SO2 concentration, (c) NOx concentration

Figure 13 shows the result of using simple RNN autocorrelation with LSTM on the concentration of AQI pollutants in the past 24 hours. The prediction is basically correct when the concentration is stable, but during the time when the concentration is fluctuating, the model always gives a lagged prediction of the trend of fluctuation. The reason for the lag is very obvious - the autocorrelation will not try to change the trend of its prediction unless at least one of its input historical data has shown an obviously different trend. And since the concentration of the last hour would for sure play an important role in predicting the concentration in the next hour, the prediction is likely to lag for an hour. So, in order to predict the trend ahead of time and improve the accuracy of prediction, more information than merely historical concentration should be taken into account.

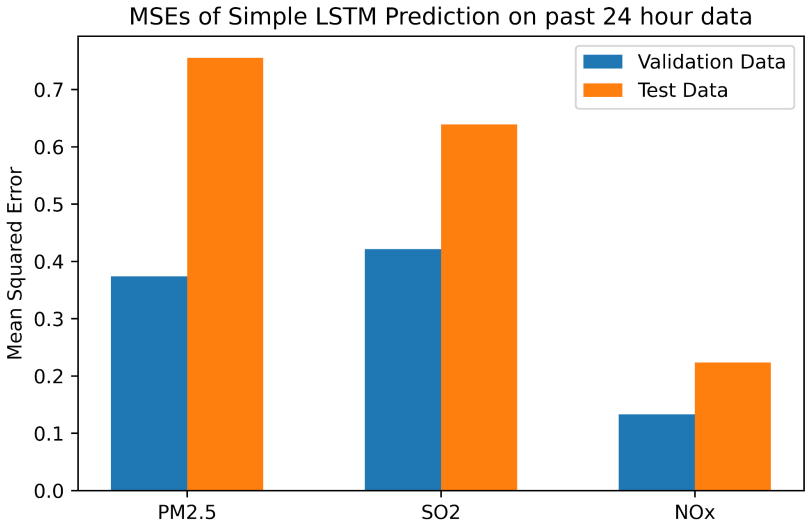


Figure 14. Distribution of the normalized concentration of AQI pollutants.

The prediction quality of simple RNN autoregression is shown in Figure 14, where the mean square error of the prediction on validation dataset and test dataset is calculated by different kinds of pollutants. Among these pollutants, NOx shows a better autocorrelation performance than other pollutants, indicating the fact that the concentration of NOx could be more accurately predicted by its historical data, while the change in concentration of PM2.5 and SO2 have more complicated mechanisms and should be predicted with other sources of data.

Also, the huge difference between MSEs on validation data and test data should be noticed. This phenomenon is reasonable, because the test data and validation data actually come from different part of a consecutive time series - test data comes from the last 10% of 17544 hours in year 2016 ~ 2017, which should be the winter of 2017, while validation data comes from the 60% ~ 90% percentile, representing the middle months of 2017. But this huge difference in MSE still indicates a different nature of changing trend of pollutants' concentration, which add to the importance of using different sources of data to make accurate predictions.

4.2. RNN Prediction Model with LSTM or GRU on weather and AQI data

Then, the RNN prediction model is established and trained to find out the potential of predicting the concentration of AQI pollutants with historical 24-hour weather and AQI data, same as the previous study has input into the feedforward neural network. Both RNN models with LSTM and GRU layer are trained, to find out the performance of these two different gating mechanisms in their internal state.

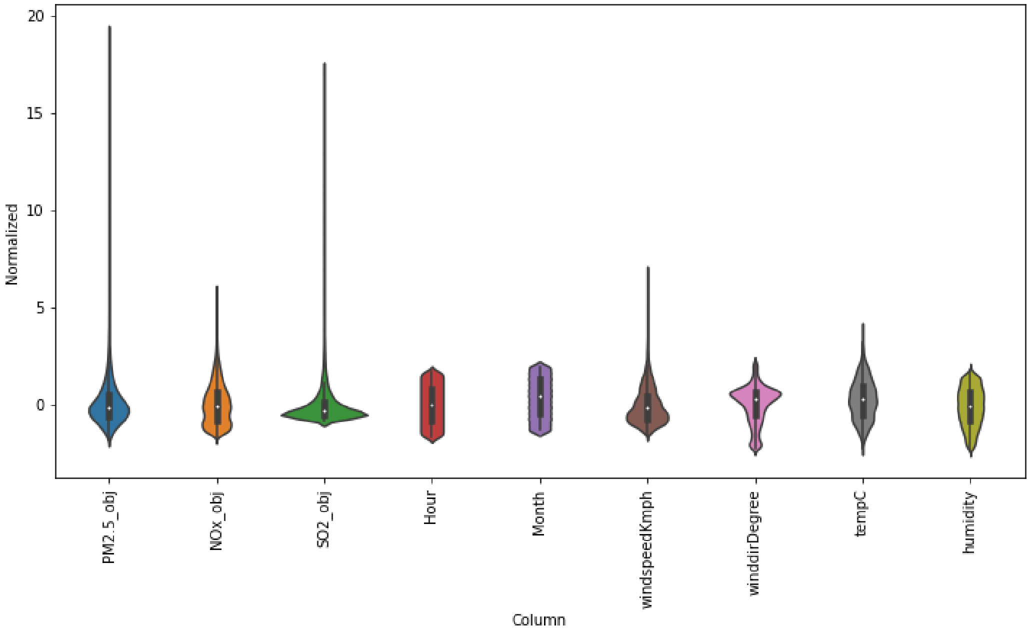
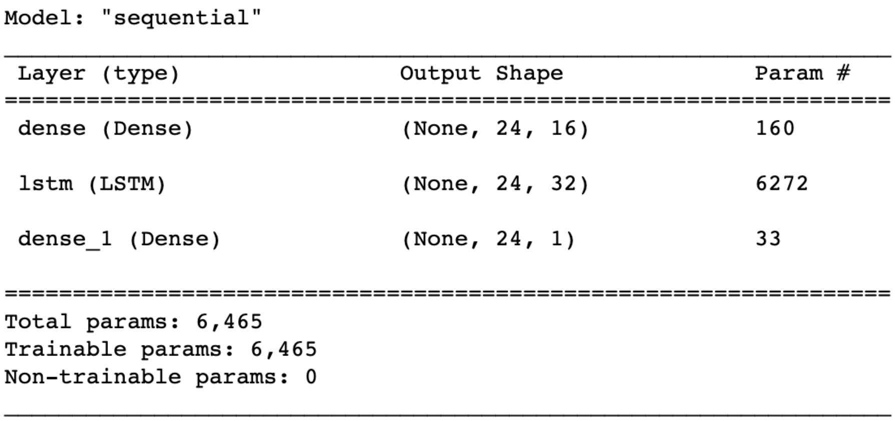


Figure 15. Distribution of the normalized input data.

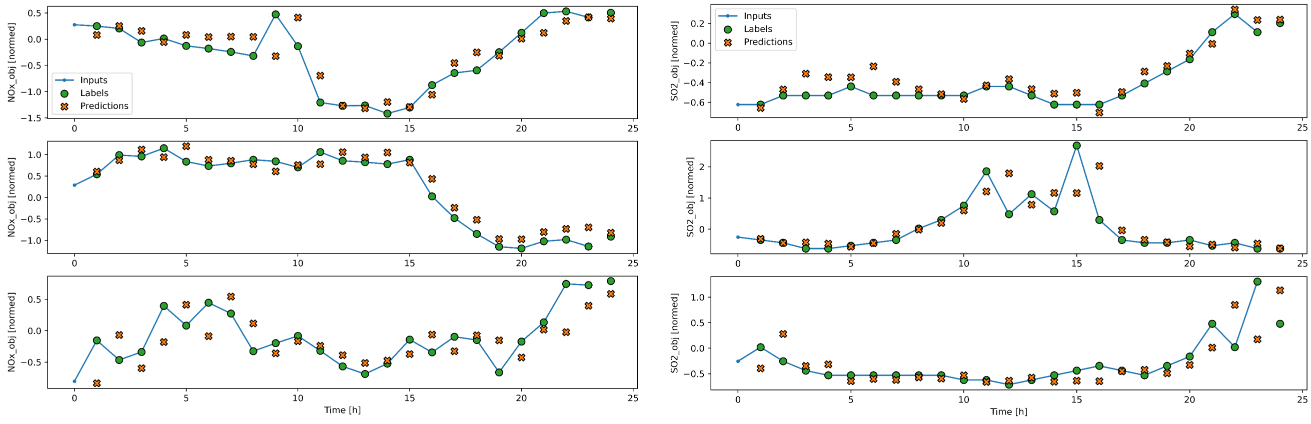
As is shown in Figure 15, apart from the concentration of three AQI pollutants, time data including hour of a day and month of a year representing daily and seasonal fluctuations, as well as the wind speed, wind degree, temperature and humidity is included in the input data of the model. Note that there is no proven linear relationship between the change of pollutants' concentration and weather conditions, indicating the necessity for a more complicated model.

Table 3. Summary of RNN Prediction model with LSTM on weather and AQI data



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 24, 16)            160

 lstm (LSTM)                 (None, 24, 32)            6272

 dense_1 (Dense)             (None, 24, 1)             33

=================================================================
Total params: 6,465
Trainable params: 6,465
Non-trainable params: 0
_________________________________________________________________
```

As is shown in Table 3, in order to explain the rather complicated relationship than the simple linear relationship between pollutants' concentration and weather data, the prediction model contains an extra layer with 16 neurons and ReLu activation function.



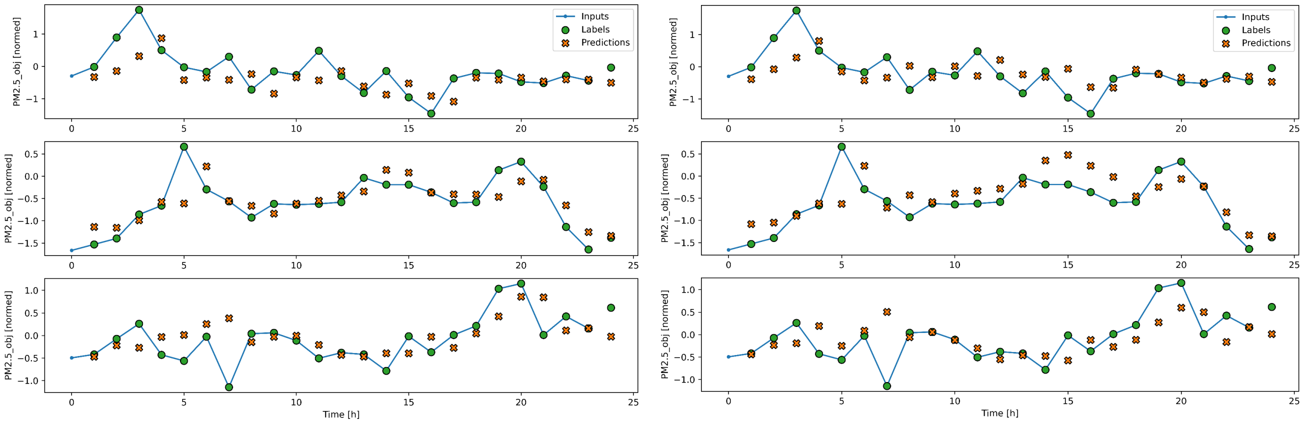
(a)

(b)

Figure 16. Prediction Result of RNN Prediction Model with LSTM on

(a) NOx concentration, (b) SO2 concentration

Figure 16 shows an example of the prediction result of the LSTM-RNN model on NOx and SO2 concentration. Compared to the prediction result of simple LSTM-RNN autoregression in section 4.1, the introduction of weather condition data and time data has slightly improved the accuracy of prediction, presenting a better performance in catching the changing trend of pollutants' concentration.



(a) (b)

Figure 17. Prediction Result of RNN Prediction Model of PM2.5 with

(a) LSTM, (b) GRU

Figure 17 shows a comparison between the prediction result of RNN prediction model with LSTM and GRU. The result of two different gating mechanisms are quite similar, with only a slight difference in several points.

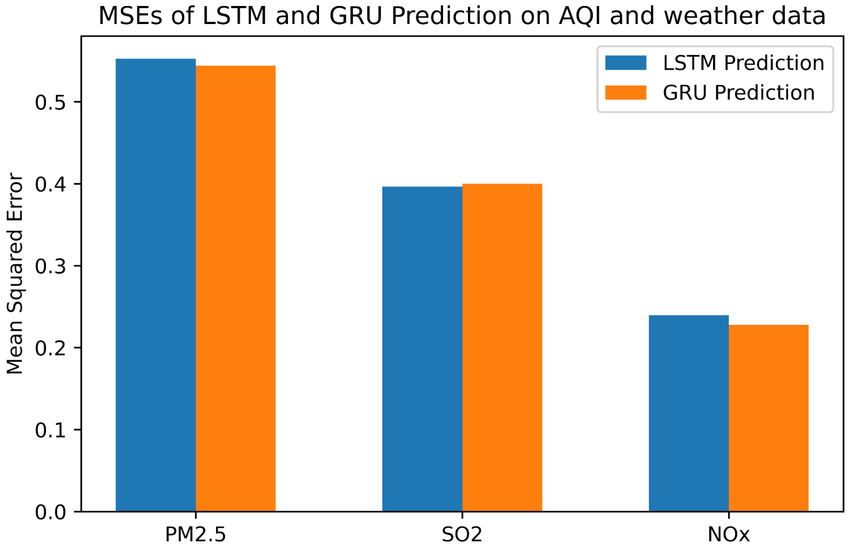


Figure 18. Mean Square Error of RNN prediction on AQI and weather data

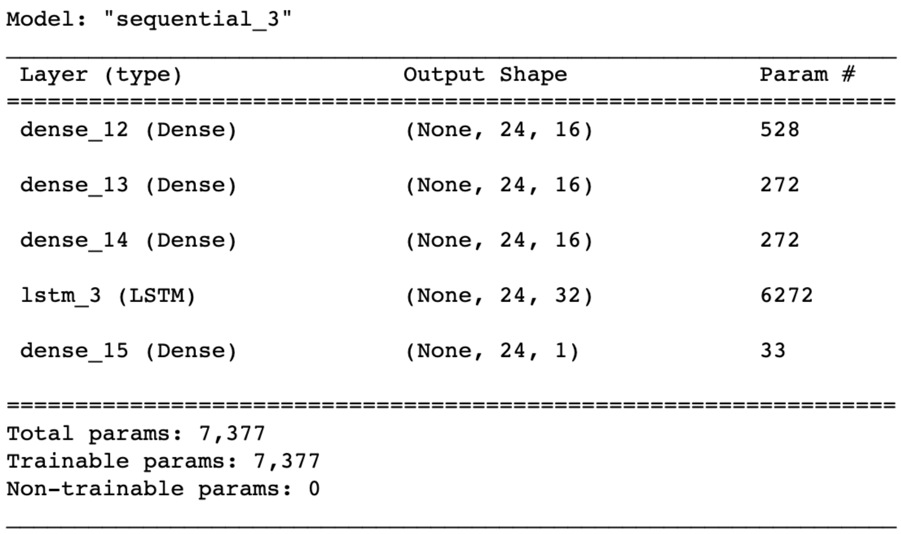
Figure 18 further shows a more comprehensive and quantified comparison between LSTM and GRU by calculating their mean squared error on the test dataset, which indicates very similar prediction qualities in this case through LSTM and GRU on only AQI and weather data.

4.3. Integrated Air Quality Predictor by RNN with GRU

Finally, the Integrated Air Quality Predictor by RNN with GRU is proposed and trained, in order to derive the effect of all other possible mechanisms to change the concentration of AQI pollutants. Similarly, two RNN models are established with LSTM and GRU gating mechanisms respectively, and their performance will be evaluated and compared.

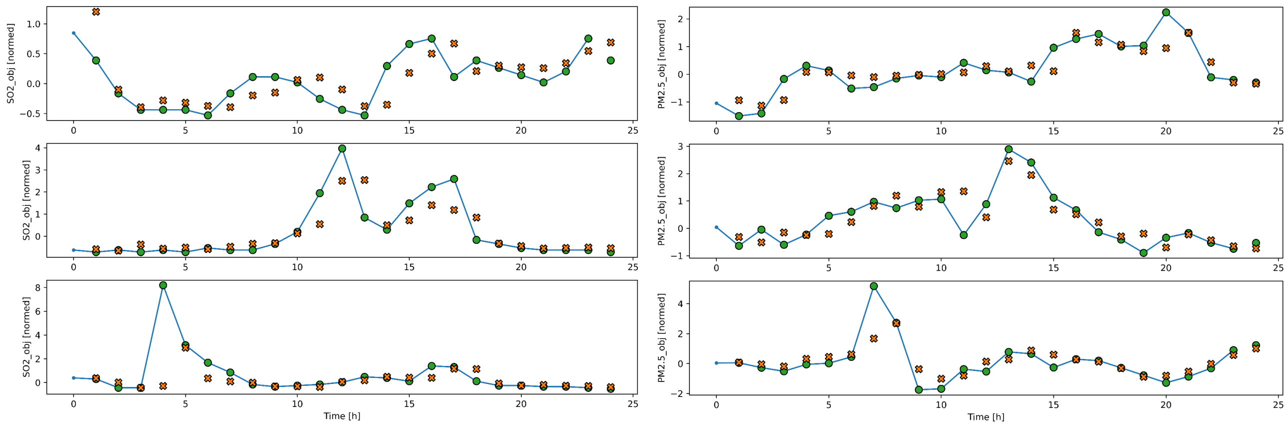
The input data is shown in Figure C in section 3.4, where all the data is input to this integrated model to find any possible residual error that could be explained by other mechanisms, such as transportation, emission, atmospheric chemical reaction, and diffusion.

Table 4. Summary of Integrated Air Quality Predictor by RNN with LSTM



```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_12 (Dense)            (None, 24, 16)            528

 dense_13 (Dense)            (None, 24, 16)            272

 dense_14 (Dense)            (None, 24, 16)            272

 lstm_3 (LSTM)               (None, 24, 32)            6272

 dense_15 (Dense)            (None, 24, 1)             33

=================================================================
Total params: 7,377
Trainable params: 7,377
Non-trainable params: 0
_________________________________________________________________
```

The structure of the Integrated Air Quality Predictor is shown in Table 4. Due to the even more complicated relationship between the comprehensive input data, two more hidden layers with 16 neurons and Relu activation function are added before the layer of internal state.



(a)

(b)

Figure 19. Prediction Result of Integrated Air Quality Model with GRU on

(a) SO2 concentration, (b) PM2.5 concentration

Figure 19 shows an example of the prediction result of the Integrated Air Quality Model with GRU on SO2 concentration and PM2.5 concentration. Compared to the previous prediction methods, the prediction of the Integrated Air Quality Model is even more accurate. Included with the data representing various mechanisms that could potentially affect the air quality, this model is able to accurately predict most of the times, except for some sudden and dramatic change, which could only be fully explained by some pollution events.

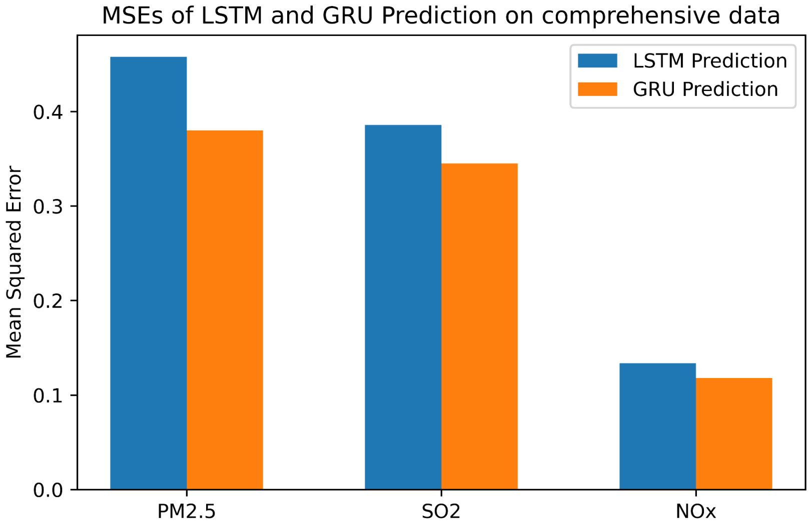


Figure 20. Mean Square Error of Integrated Air Quality Predictor on comprehensive data

Quantitative analysis on the MSEs of Integrated Air Quality Model with GRU and LSTM on test dataset is shown in Figure 20, where the model with GRU shows a better result of predicting the concentration of the pollutants. It could indicate that GRU is more advanced or is more suitable for this case, but maybe it is caused by some random factors. To derive the actual reason, further analysis on the process of two methods should be made.

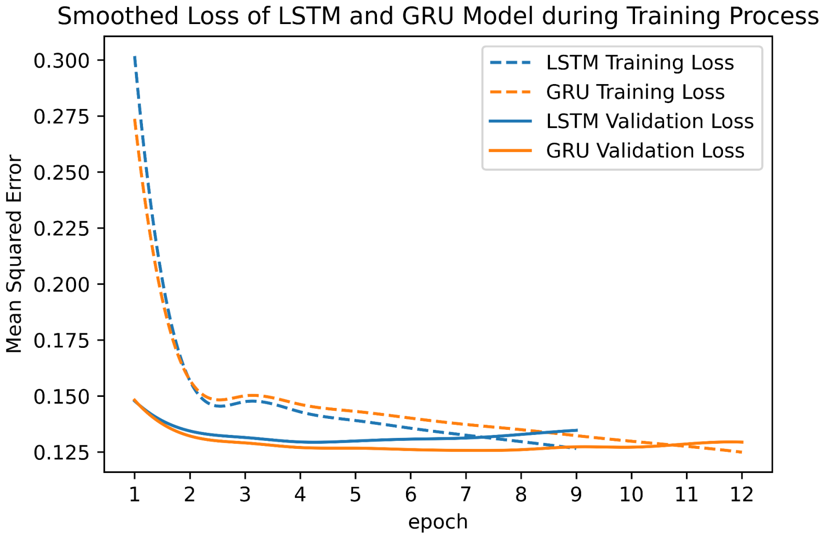


Figure 21. Smoothed Training and Validation MSEs during training process of GRU and LSTM

The changing training loss and validation loss of two approaches during the whole training process is shown in Figure 21. To make it more clear, scattered data points are processed into smooth curves through interpolation. Although the training loss of LSTM decreases more rapidly than GRU, the validation loss of LSTM could hardly decrease with the training epoch like GRU does. So the reason why GRU has a better performance compared to LSTM could be, due to the absence of output gate and the presence of fewer parameters, GRU has a less tendency for overfitting in this case, thus it could have a better prediction on test dataset and a slightly longer training process.

4.4. Summary

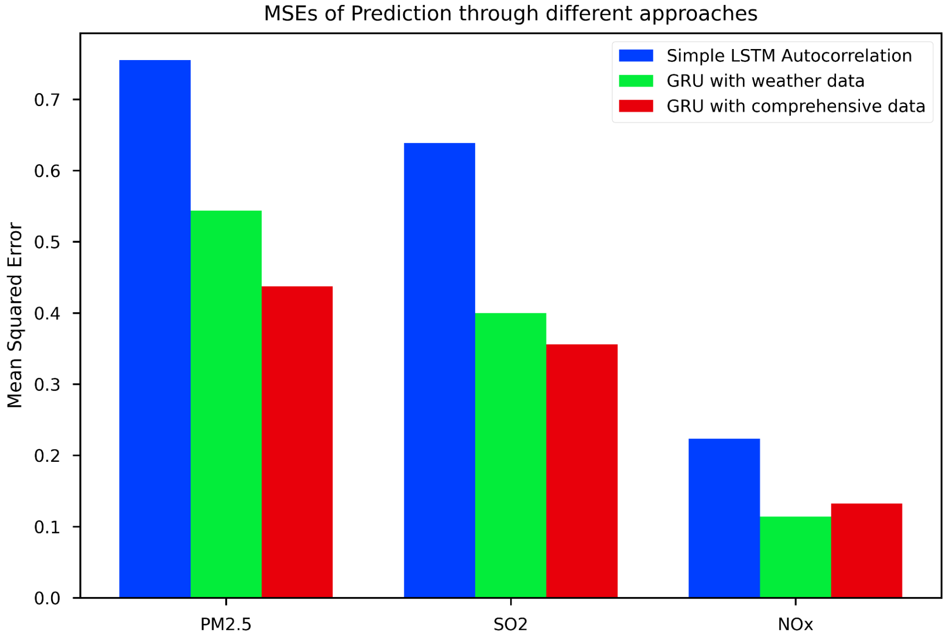


Figure 22. Mean Square Error on test dataset by different prediction approaches and models.

Figure 22 summarizes the performance of different prediction models in predicting the concentration of AQI pollutants in the atmosphere in Los Angeles from 2016 to 2017. Among the three approaches investigated in this project, the Integrated Air Quality Predictor with GRU shows the best performance in predicting PM2.5 and SO2, while RNN prediction model with only historical AQI and weather data best predicts the concentration of nitrogen oxides. It could be inferred that NOx is a pollutant with a relatively long lifespan in the atmosphere, so the historical concentration and weather data may suffice to make a good prediction.

In general, the Integrated Air Quality Predictor with GRU has shown its obvious advantage over other different approaches in this case, so it is the method that would be proposed by this project to make further predictions for other cases and locations.

5. Conclusions

In this project, several potential approaches to predict the air quality in the near future are presented and discussed. Then three different recurrent neural networks with different kinds of input data, neural structure, and gating mechanism are trained and tested with real hourly data in Los Angeles from 2016 to 2017. After the preliminary selection of the potential method and the reasonable preprocessing of the original data, a simple autoregression with LSTM on historical concentrations is conducted in the first part of this project, in order to determine the autoregression ability of the concentrations of AQI pollutants. And the huge difference between the prediction of validation data and test data indicates the fact that some important factors, which may change with time, have not yet been included in this model.

Furthermore, a RNN prediction model with LSTM and GRU on historical AQI and weather data is set and trained, in order to discover the effectiveness of applying weather data to predict the concentration of AQI pollutants, as well as compare the performance of LSTM and GRU. The prediction result shows an obvious improvement than simple autoregression, showing the necessity to integrate weather data into the model for an accurate prediction, while no apparent difference of using LSTM and GRU is discovered in this part.

Finally, an integrated air quality prediction model by RNN on various courses of data is established to take the effect of mechanism in the atmospheric process into the consideration of the prediction. It turns out that, with the further consideration of these mechanisms, this model is able to provide the most accurate predictions among the methods in this project. And GRU shows a better understanding of the changing trend of pollutants' concentrations across different seasons and years in this case.

After a comprehensive analysis and comparison of their prediction results, the Integrated Air Quality Predictor by RNN with GRU layer is selected as the optimal method in this case, and is proposed as a reliable and accurate method to predict local air quality based on integrated historical data.

Due to the lack of data availability, the most pitiful part of this project is that the data representing atmospheric stability is not included in the input data. Atmospheric stability is an indicator used to characterize the convective strength of the atmospheric flow, which is an essential factor for the diffusion mechanism of pollutants and is described by the height of the boundary layer or the temperature difference between different heights. Unfortunately, they are not the conventional meteorological indicator that will be measured during daily weather observation, and that's why they are not included in the project for even more reliable predictions.

And still, incorporating our prior knowledge of those atmospheric mechanisms to the model would be an efficient way to improve the quality and reduce the data requirement of the prediction process. Thus the development of a self-defined RNN structure, which is able to preserve a certain arithmetic relationship between some of the intermediate neurons, is very useful in this case, which shows a perfect combination of data-driven and theory-driven approach to predict the air quality.

References

Kumar, A., & Goyal, P. (2013). Forecasting of Air Quality Index in Delhi Using Neural Network Based on Principal Component Analysis. *Pure and Applied Geophysics, 170*(4), 711-722. doi:10.1007/s00024-012-0583-4

Lin, K. P., Liao, G. L., Huang, Y., Chen, J. B., Chen, B., & Destech Publicat, I. (2015, May 17-18). *A Genetic Neural Network Model for AQI Prediction.* Paper presented at the International Conference on Computer Science and Environmental Engineering (CSEE), Beijing, PEOPLES R CHINA.

Liu, Z. Y., Yang, Y. T., & Cai, Q. D. (2019). Neural network as a function approximator and its application in solving differential equations. *Applied Mathematics and Mechanics-English Edition, 40*(2), 237-248. doi:10.1007/s10483-019-2429-8

Rackauckas, C., Ma, Y. B., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., Edelman, A. (2020). Universal Differential Equations for Scientific Machine Learning. arXiv:2001.04385v3.

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics, 378*, 686-707. doi:10.1016/j.jcp.2018.10.045

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature, 566*(7743), 195-204. doi:10.1038/s41586-019-0912-1

Wang, H. Y., Wang, J. Y., & Wang, X. H. (2017, Jul 17-18). *An AQI Level Forecasting Model Using Chi-square Test and BP Neural Network.* Paper presented at the 2nd International Conference on Intelligent Information Processing (IIP), Bangkok, THAILAND.