#RC 1

Major Comments:

The introduction lacks background on why and how $PM_{2.5}$ is created, removed from the atmosphere, or where it lingers in the atmosphere. This needs to be expanded upon along with its connection to AOD. Further, it needs to be discussed why AOD may not be an appropriate proxy for $PM_{2.5}$ at times and how this has been a limiting factor in the past to using AOD to effectively monitor $PM_{2.5}$ from space.

Reply:

- In the revised manuscript, we added the following background of $PM_{2.5}$, including how it is created, removed from the atmosphere and where it lingers in the atmosphere.
- "$PM_{2.5}$ in the atmosphere is either emitted directly or formed from gaseous precursors through complex gas phase, aqueous phase or heterogeneous chemical reactions (Cheng et al., 2016; Gao et al., 2016). $PM_{2.5}$ consists of complex composition, including sulfate, nitrate, organic carbon, elemental carbon, soil dust, and sea salt (Gao et al., 2016). It can stay in the boundary layer for a few days and in the free troposphere for a few weeks (similar to ozone). It can be efficiently removed out of atmosphere by precipitation, which is the major atmospheric sink (Jacob and Winner, 2009)."
- "Aerosol optical depth (AOD) is the vertical integration of aerosol extinction from the surface to the top of the atmosphere, while ground-based instrument only measures $PM_{2.5}$ concentrations near the ground. When long-range transport of particles occurs above the ground, high AOD does not always coincide with high $PM_{2.5}$ concentrations (Hu et al., 2022). Besides, an opposite seasonality of AOD and surface $PM_{2.5}$ over China was identified due to aerosol hygroscopic growth (Xu et al., 2019)."
- Cheng, Y., Zheng, G., Wei, C., Mu, Q., Zheng, B., Wang, Z., Gao, M., Zhang, Q., He, K., Carmichael, G. and Pöschl, U., 2016. Reactive nitrogen chemistry in aerosol water as a source of sulfate during haze events in China. *Science advances*, *2*(12), p.e1601530.
- Gao, M., Carmichael, G.R., Wang, Y., Saide, P.E., Yu, M., Xin, J., Liu, Z. and Wang, Z., 2016. Modeling study of the 2010 regional haze event in the North China Plain. *Atmospheric Chemistry and Physics*, *16*(3), pp.1673-1691.
- Jacob, D.J. and Winner, D.A., 2009. Effect of climate change on air quality. *Atmospheric environment*, *43*(1), pp.51-63.
- Hu, Q., Liu, C., Li, Q., Liu, T., Ji, X., Zhu, Y., Xing, C., Liu, H., Tan, W. and Gao, M., 2022. Vertical profiles of the transport fluxes of aerosol and its precursors between Beijing and its southwest cities. *Environmental Pollution*, *312*, p.119988.
- Xu, J. Han, F., Li, M., Zhang, Z., Du, X., Wei, P.: On the opposite seasonality of MODIS AOD and surface $PM_{2.5}$ over the Northern China plain, Atmospheric Environment, Volume 215, 2019, 116909, ISSN 1352-2310, https://doi.org/10.1016/j.atmosenv.2019.116909.

The setup of the model, data, and testing of the model is difficult to understand. The methodology portion needs major restructuring to understand what data was used, at

what scale, what meteorological inputs were used, if they were all at the same grid sized or kept somehow at their native resolutions, what time steps were included of each, how MODIS AOD at a single time step is integrated into the series of AOD from Himiwari-8, etc.

Reply:

- In the revised manuscript, we have revised the materials and methods following your valuable suggestions.
- In section 2.1, we introduced the input data of the model, which includes the data type, sample, resolution and data selection standard; in section 2.2 and 2.3, we introduced the model structure and training verification methods; section 2.2 introduces the feature extraction process of data and the fusion process between multi-source heterogeneous data, including the data fusion process between different spatial and temporal resolutions and dimensions; section 2.3 describes the training and test process of the model, and section 2.4 explains sensitivity analysis method used to quantitatively analyze the impact of the input data on the results.
- We added the time and space range of the model data.
  "Our research area is concentrated in the Middle East of China (Figure S3), with a time span of four years from 2017 to 2020."
- We added detailed introduction tables to the model input data. Table 6 shows the content and resolution of the dataset, and Table S12 describes the size of the dataset. The input data have different temporal and spatial resolutions. All input data have the same center position. Table S3 shows the spatial and temporal sizes of the data that input to the model. For different spatial resolution data, we use the upper sampling layer in the feature extraction process to make them have the same spatial size before data fusion.
- We added a description of how to perform data fusion for data with different characteristics.
- "First, feature extraction (nonlinear transformation) is performed on each data layer, and then data fusion is performed. The data fusion process can be expressed as:

$$Z_{fusion} = \sum_{i=1}^{c} K_i * X_{AOD_{himawari-8_{current}}} + \sum_{i=1}^{c} K_i * X_{AOD_{himawari-8_{closeness}}} + \sum_{i=1}^{c} K_i * X_{AOD_{himawari-8\,period}} + \sum_{i=1}^{c} K_i * X_{AOD_{MODIS}}$$

where * is convolution and Ki is learnable parameter."

Table 6. Descriptions of considered variables.

Product	Unit	Variable Definition	Spatial Resolution	Temporal Resolution
AOD(Himawari-8)		Aerosol optical depth	0.05°×0.05	1hour
AOD(MODIS)		Aerosol optical depth	0.05°×0.05	daily
Tempc	°C	Temperature	0.05°×0.05°×12 L	1hour
RH	%	Relative Humidity	0.05°×0.05°×12 L	1hour
HPBL	m	Planetary Boundary Layer Height	0.05°×0.05°	1hour

P	Hpa	Pressure	0.05°×0.05°×12 L	1hour
U	m/s	Wind Speed (U)	0.05°×0.05°×12 L	1hour
V	m/s	Wind Speed (V)	0.05°×0.05°×12 L	1hour
DEM	m	Digital Elevation Model	0.01°×0.01°	Annual
POI		Point of Interest	0.01°×0.01°	Annual
Traffic Network		Traffic Network	0.01°×0.01°	Annual
GDP	￥/km 2	Gross Domestic Product	0.01°×0.01°	Annual
TPOP	people /km2	population density	0.01°×0.01°	Annual
Land Cover Type		Land Cover Type	0.05°×0.05°	Annual
EVI		Enhanced Vegetation Index	0.05°×0.05°	Monthly
NDVI		Normalized Differnce Vegetation Index	0.05°×0.05°	Monthly

Table S3. The input data shape

Category	Name	shape type	shape
AOD data	Himawari-8 Current	width,length,time	32,32,4
	Himawari-8 Closeness	width,length,time	32,32,10
	Himawari-8 Period	width,length,time	32,32,7
	MODIS	width,length,band×time	32,32,3×7
Meteorology	rh	width,length,time	32,32,9
	temperature	width,length,time	32,32,9
	pressure	width,length,time	32,32,9
	hpbl	width,length,time	32,32,9
	u	width,length,time	32,32,9
	v	width,length,time	32,32,9
	rh	width,length,height	32,32,12
	temperature	width,length,height	32,32,12
	pressure	width,length,height	32,32,12
	hpbl	width,length,height	32,32,1
	u	width,length,height	32,32,12
	v	width,length,height	32,32,12
Geographic information data	POI	width,length,type	64,64,7
	Traffic Network	width,length,type	64,64,9
	DEM	width,length,type	64,64,1
	GDP	width,length,type	64,64,1
	Tpop	width,length,type	64,64,1
	Land Cover Type	width,length,type	32,32,17

	EVI	width,length,type	32,32,1
	NDVI	width,length,type	32,32,1

Table S12. Statistics of number of stations used for training and testing

	All Number(N)	Training Number(N)	Testing Number(N)	All Sample Number(N)			
				2017	2018	2019	2020
North China	176	150	26	1340252	1241169	1257882	1228487
East China	343	308	35	2663647	2525520	2478030	2503516
South China	237	213	24	1957469	1875939	1865894	1908516
Sichuan Basin	145	130	15	1132051	1043238	1044349	1062847
Shaanxi Province	177	159	18	1369405	1280065	1277902	1301866

Further, the section on the model configuration is extremely muddled. I do not understand how k-means was used, what a contingency table is, how the sensitivity analysis fits into the data/model configuration, and why only 10% of data is used as a test case instead of the standard 20% test, 20% validate, and 60% train. It seems as if they only use 10% to test their model, and given their numbers it wouldn't be surprising if that meant the model was overfit and not enough variability in the test samples existed to find that. They state that they did cross validation but the accuracy is never shown. The number of samples is never stated nor is the resolution or the exact inputs of the model clearly stated. The authors claim that they are predicting PM2.5 on an hourly timescale, but it is never clearly stated if that is what they actually trained their model to do. They use sensitivity analysis to test what inputs to use in their model then somehow also use that analysis to verify their model.

Reply:

- In the revised manuscript, we introduced the data selection process, the methods used, the training verification process and sensitivity analysis of the model in detail.
- Due to the multi-source heterogeneous data we used, different data have different space-time dimension characteristics. We used different statistical methods to pre-select data to determine whether the data needs to be included in the model input. For the two categorical variables, we use contingency table and Chi-square test to analyze the correlation between categorical variables.
- “Pearson correlation coefficient can test whether two continuous variables have potential correlation in statistics. For meteorological variables and satellite data, we used Pearson significance level to determine whether target variables are significant and to include as model inputs. Chi-square test is a commonly used method to test the statistical correlation between discrete variables. Geographic information variables were considered unchanged during our research period. First, we used

k-mean method to cluster them to reduce dimensions. Then we classified the annual average data of CNEMC $PM_{2.5}$ into groups (<10μg m^{-3}, 10~15μg m^{-3}, 15~25μg m^{-3}, 25~35μg m^{-3}, 35~50μg m^{-3}, 50~75μg m^{-3}, 75~100μg m^{-3}, >100μg m^{-3}) and used Chi-square test to test whether it has statistical correlation."

- We adopted a general sensitivity analysis method for black box model. The response relationship was obtained by adjusting the input data to analyze the output results. During the sensitivity analysis test, we only changed the data value range but not the data size to ensure the operation of the model.
- In our study, the 10-CV is enough. In the updated manuscript, we added the reason for using 10 -CV. Our huge dataset size ensures that our model will not be over fitted. And we used the Beijing site for independent verification. Figure S5 and Figure R2 show the verification results. The black points are the data of CNEMC sites participating in the training, and the red points are the testing data of Beijing control sites. The R^2 is above 0.86 and the RMSE is less than 24 μg·m^{-3}.

  "Compared with other studies, our hourly full coverage ground $PM_{2.5}$ concentration prediction has greatly increased the amount of data. The sample size exceeds 1 million in each study area and year (Table S11 and S12)."

  "10-CV has been proved to be a reasonable method to test the robustness of our model (Table S9). We also compared the results under different training and test proportions, and Table S16 shows that 10-CV is reasonable in this study (Table S16)."
- Cross validation refers to the random selection of training samples and validation samples according to the validation needs (samples, time, space) in the dataset (Tables S11,S12). The specific input information of the model is shown in Table 6 and Table S3.

  And we have carried out a variety of different tests, based on time-space samples and extrapolation in time and space. Figure 1, Table 1 and Table 2 show the detailed verification results. Figures 3 and 4 show the verification results under different conditions. Figures S5~S10 show the verification results under different scenarios. All our verifications are performed on our test dataset. We verified the R^2, RMSE, MAE and Slop of the model results.
- In our study, we didn't use the sensitivity analysis to test what inputs to use in their model. We conduct data pre-selection based on statistical tests, and then build the ST-NN model to obtain the ground $PM_{2.5}$ concentration with full space-time coverage and validate it. Then, we open the black box model through sensitivity analysis and visualization to analyze the quantitative impact of each input variable on the ground $PM_{2.5}$ concentration.

Table.S16. Validation results under different training and test proportions

test:train	1:9	2:8	3:7	4:6	5:5	6:4	7:3
R-squre	0.81	0.81	0.78	0.77	0.76	0.76	0.74
RMSE(μg/m³)	16.15	16.6	17.57	17.87	18.67	18.55	19.32

Table S11. Data quality control status.

	2017(N)	2018(N)	2019(N)	2020(N)	Rat

	outlier	not null	outlier	not null	outlier	not null	outlier	not null	e
North China	188547	1340252	186265	1241169	180120	1257882	175359	1228487	0.1
East China	420792	2663647	364102	2525520	395344	2478030	381681	2503516	0.1
South China	325190	1957469	280304	1875939	298656	1865894	342721	1908516	0.1
Sichuan Basin	170788	1132051	168210	1043238	167905	1044349	170817	1062847	0.1
Shaanxi Province	189610	1369405	184997	1280065	184688	1277902	199430	1301866	0.1

Table S12. Statistics of number of stations used for training and testing

	All	Training	Testing	All Sample Number(N)			
	Number(N)	Number(N)	Number(N)	2017	2018	2019	2020
North China	176	150	26	1340252	1241169	1257882	1228487
East China	343	308	35	2663647	2525520	2478030	2503516
South China	237	213	24	1957469	1875939	1865894	1908516
Sichuan Basin	145	130	15	1132051	1043238	1044349	1062847
Shaanxi Province	177	159	18	1369405	1280065	1277902	1301866

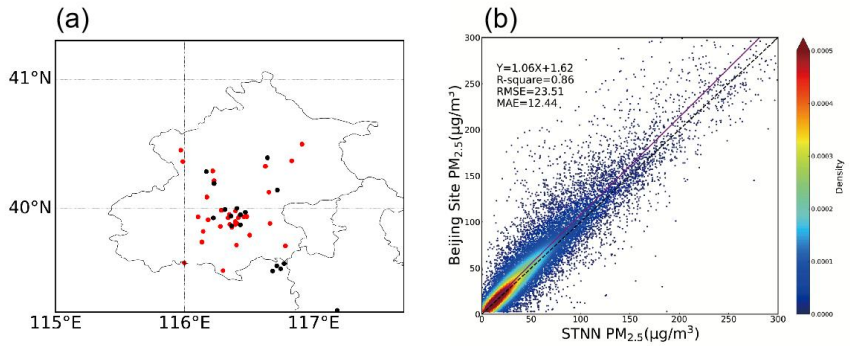

Figure R2 (a) Distribution of verified Beijing control sites (red) and CNEMC sites (black). (b) shows the verification results of the ST-NN model and Beijing sites.

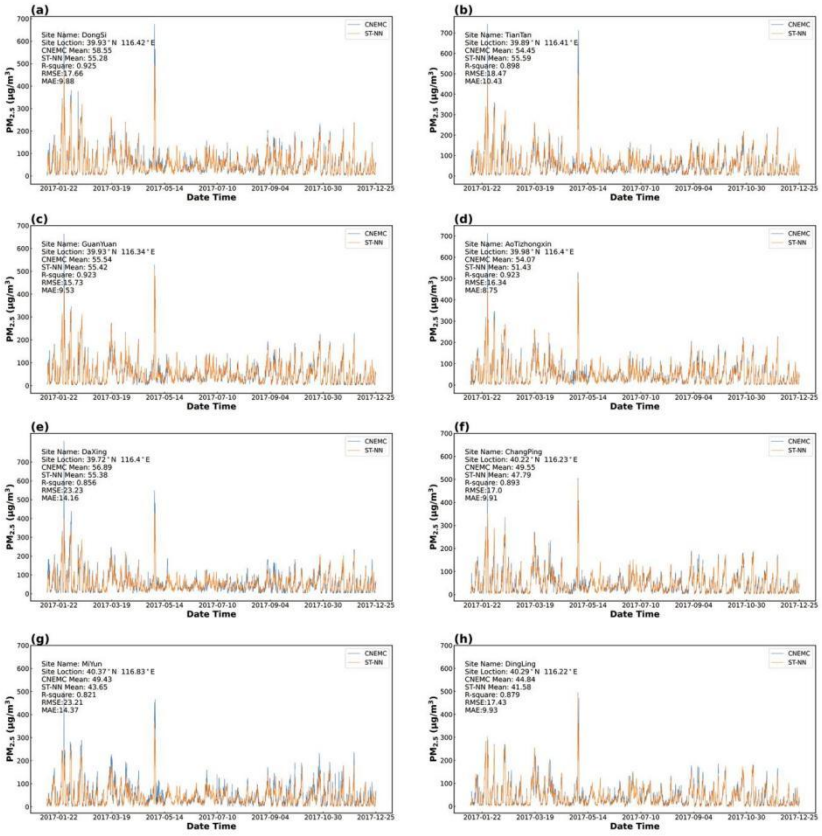

Figure S5. ST-NN model predicted and ground-level observed (not used in training) time series of PM2.5 in Beijing stations. (a) Dongsi station in Beijing. (a) Tiantan station in Beijing. (c) Guanyuan station in Beijing. (d) Aotizhognxin station in Beijing. (e) Daxing station in Beijing. (f) Changping station in Beijing. (g) Miyun station in Beijing. (h) Dingling station in Beijing. (a-d) stations are in city, and (e-h) are rural stations.

Minor comments:

Line 55: A map would be useful to understand where the locations are since I assume these were used as the true labels in training/testing.

Reply:

- Figure S13 shows the locations we used. We have conducted many experiments, and each experiment is randomly sampled for training and verification. The results are the average of many experiments.
- Figure R3 (a)~(e) shows the training and testing site in our experiment. The black point is the training site, and the red point is the testing site.
- Figure R4 shows the verification results of Beijing control sites. The black points are the data of CNEMC sites participating in the training, and the red points are the testing data of Beijing control sites. The R^2 is above 0.86 and the RMSE is less than 24 $\mu g \cdot m^{-3}$.

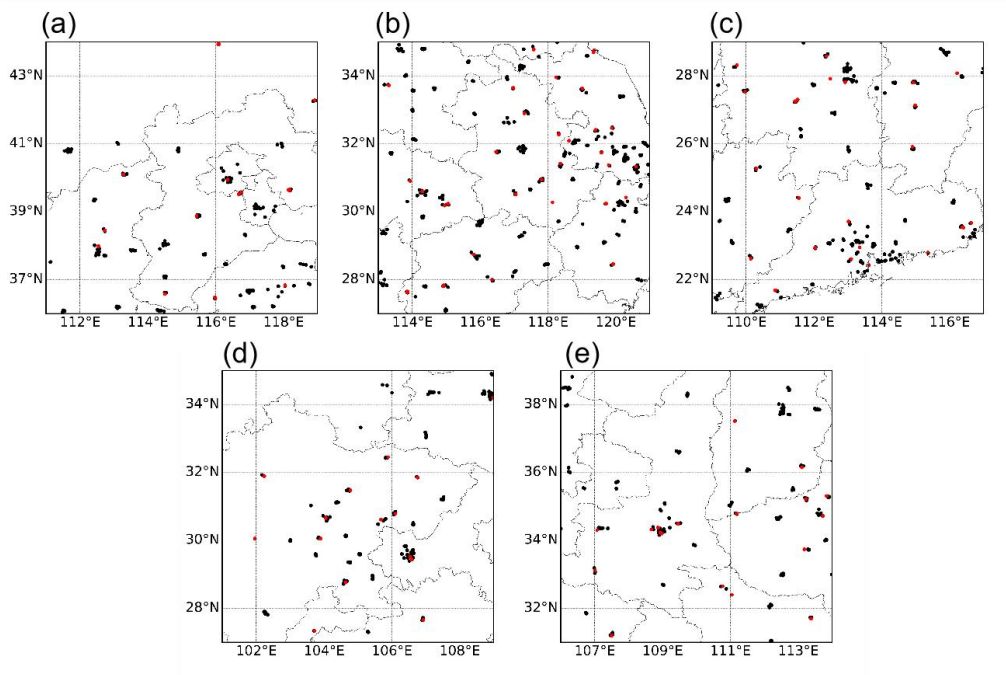

Figure R3. Distribution of training sites and testing sites in the experiment. Black points are training sites. Red points are testing sites.

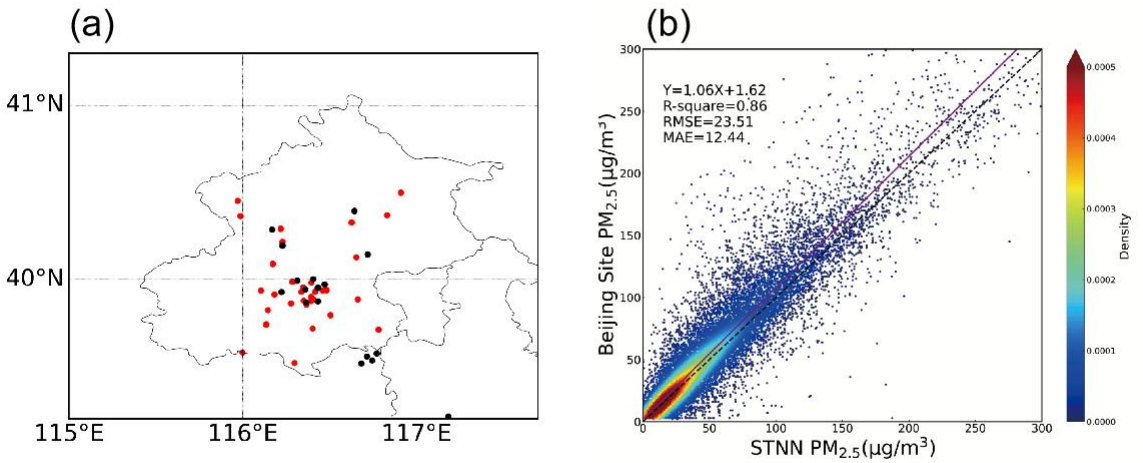

Figure R4 (a) Distribution of verified Beijing control sites (red) and CNEMC sites (black). (b) shows the verification results of the ST-NN model and Beijing sites.

Line 73: What limits these past studies from being able to fill gaps in time?

Reply:

- Due to observation limitations such as clouds, the AOD observed by satellite is lacking under cloud and night. Previous studies mostly focused on mining the relationship between the satellite AOD and ground $PM_{2.5}$ at the same spatial grid point, without fully mining the spatio-temporal correlation characteristics between the data, so there is a space gap, and most of them are filled in space through the accumulation of time.
- Also, the polar orbiting satellites are mostly used (MODIS AOD is the most commonly used data). Compared with the hourly resolution of geostationary satellites, their monitoring results are only once a day.

Line 88: Why is AOD only available for 33% of China?

Reply:

- The AOD observed by satellite is affected by cloud and surface conditions, which makes the satellite AOD have a high lack rate (Bi et al., 2019). Severe pollution weather occurs frequently in central and eastern China, and the retrieval of AOD by satellite sometimes fails under severe pollution weather (Wang et al., 2017).
- We mentioned it in the introduction.
- Bi, J., Belle, J. H., Wang, Y., Lyapustin, A. I., Wildani, A., and Liu, Y.: Impacts of snow and cloud covers on satellite-derived PM2. 5 levels, Remote sensing of environment, 221, 665-674, 2019.
- Wang, Y., Chen, L., Li, S., Wang, X., Yu, C., Si, Y., and Zhang, Z.: Interference of Heavy Aerosol Loading on the VIIRS Aerosol Optical Depth (AOD) Retrieval Algorithm, Remote Sensing, 9, 397, 2017.

Line 91: What causes haze in rural areas?

Reply:

- Emissions from transportation sources, coal and firewood burning process for rural heating and catering, adverse meteorological conditions, and primary and secondary aerosols in agricultural production (application of pesticides and fertilizers, dust and straw burning in harvest season) (Aunan et al., 2019; Sun et al., 2021; Ludwig et al., 2003).
- Aunan, K., Hansen, M. H., Liu, Z., and Wang, S.: The Hidden Hazard of Household Air Pollution in Rural China, Environmental Science & Policy, 93, 27-33, https://doi.org/10.1016/j.envsci.2018.12.004, 2019.
- Sun, J., Xie, C., Xu, W., Chen, C., Ma, N., Xu, W., Lei, L., Li, Z., He, Y., Qiu, Y., Wang, Q., Pan, X., Su, H., Cheng, Y., Wu, C., Fu, P., Wang, Z., and Sun, Y.: Light absorption of black carbon and brown carbon in winter in North China Plain: comparisons between urban and rural sites, Sci Total Environ, 770, 144821, 10.1016/j.scitotenv.2020.144821, 2021.
- Ludwig, J., Marufu, L. T., Huber, B., Andreae, M. O., and Helas, G.: Domestic Combustion of Biomass Fuels in Developing Countries: A Major Source of Atmospheric Pollutants, Journal of Atmospheric Chemistry, 44, 23-37, 10.1023/A:1022159910667, 2003.

Line 97: Spell out what ST-NN stands for.

Reply:

- ST-NN means Spatial Temporal Neural Network. It's the abbreviation of our model name.

Line 115: Define WRF as an acryonym as it is used later.

Reply:

- We have revise WRF to Weather Research and Forecasting (WRF). And we use its abbreviation later.

Line 118: A table would be useful of all the inputs since the meteorological inputs are never clearly stated.

Reply:

- Table.S3 shows the details of the input data, and Table. S17 shows the specific dimensions of the input data. And in the revised manuscript, we move the Table.S3 from supplement to the manuscript.

Table S3. Descriptions of considered variables.

Product	Unit	Variable Definition	Spatial Resolution	Temporal Resolution
AOD		Aerosol optical depth	0.05°×0.05	1hour
Tempc	°C	Temperature	0.05°×0.05°×12L	1hour
RH	%	Relative Humidity	0.05°×0.05°×12L	1hour

HPBL	m	Planetary Boundary Layer Height	0.05°×0.05°	1hour
P	Hpa	Pressure	0.05°×0.05°×12L	1hour
U	m/s	Wind Speed (U)	0.05°×0.05°×12L	1hour
V	m/s	Wind Speed (V)	0.05°×0.05°×12L	1hour
DEM	m	Digital Elevation Model	0.01°×0.01°	Annual
POI		Point of Interest	0.01°×0.01°	Annual
Traffic Network		Traffic Network	0.01°×0.01°	Annual
GDP	￥/km2	Gross Domestic Product	0.01°×0.01°	Annual
TPOP	people/km2	population density	0.01°×0.01°	Annual
Land Cover Type		Land Cover Type	0.05°×0.05°	Annual
EVI		Enhanced Vegetation Index	0.05°×0.05°	Monthly
NDVI		Normalized Differnce Vegetation Index	0.05°×0.05°	Monthly

Tabale.S17. The input data shape

Category	Name	shape type	shape
AOD data	Himawari-8 Current	width,length,time	32,32,4
	Himawari-8 Closeness	width,length,time	32,32,10
	Himawari-8 Period	width,length,time	32,32,7
	MODIS	width,length,band×time	32,32,3×7
Meteorology	rh	width,length,time	32,32,9
	temperature	width,length,time	32,32,9
	pressure	width,length,time	32,32,9
	hpbl	width,length,time	32,32,9
	u	width,length,time	32,32,9
	v	width,length,time	32,32,9
	rh	width,length,height	32,32,12
	temperature	width,length,height	32,32,12
	pressure	width,length,height	32,32,12
	hpbl	width,length,height	32,32,1
	u	width,length,height	32,32,12
	v	width,length,height	32,32,12
Geographic information data	POI	width,length,type	64,64,7
	Traffic Network	width,length,type	64,64,9
	DEM	width,length,type	64,64,1
	GDP	width,length,type	64,64,1
	Tpop	width,length,type	64,64,1
	Land Cover Type	width,length,type	32,32,17
	EVI	width,length,type	32,32,1

	NDVI	width,length,type	32,32,1

Line 135: I do not understand why the Pearson correlation is used to find "contain dimension of time"?

Reply:

- Pearson correlation coefficient can test whether two continuous variables have potential correlation in statistics. Meteorology, satellite AOD and CNEMC $PM_{2.5}$ data are time series data, we use Pearson's significance level to judge whether they have statistical correlation to determine if they are used as model input.

Line 136: What is CNEMC and why/how is a Chi-suqared test used?

Reply:

- In section 2, we define CNEMC.
- "We used hourly ground-level observations of PM2.5 from the Chinese National Environmental Monitoring Center (CNMEC)"
- We revise it to CNEMC $PM_{2.5}$ concentrations.
- Chi-suqared can be used to test whether there is statistical correlation between two discrete variables. For road types, buildings and other classification variables, we classify $PM_{2.5}$ stations, and then perform Chi-square test to determine the model input.

Line 143: Why is a k-means used? What does the discreteness of variables mean in this context? What and why is a contingency table used?

Reply:

- For geographic information variables (such as point of interest, road type, etc.), which are discrete, we use k-mean to reduce their dimensions. We use the annual average CNEMC $PM_{2.5}$ concentration data and classify it (10,15,25,35,50,75,100) according to the WHO Global Air Quality Guidelines (WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Executive summary). Then we use Chi-square test (contingency table) to test whether they have statistical correlation to determine it is used as model input or not.

Line 147: Where are these layers used? Why? How do they affect model performance?

Reply:

- These layers are the basic components of the model. Figure.S2 shows the model structure. The model approximates the complex nonlinear relationship through the neural network layer, uses massive data to drive the model, and continuously optimizes the parameters to obtain the desired results(Szegedy et al., 2017).
- Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A.: Inception-v4, inception-resnet and the impact of residual connections on learning, Thirty-first AAAI conference on artificial intelligence,

Line 161: How many samples? Are they all put into a common grid? What is the time resolution?

Reply:

- The sample size exceeds 1 million in each study area and year (Table. S12). The input data is multidimensional data, and they have the same central location. Different data have different time resolutions, as shown in Table. S3.

Table S3. Descriptions of considered variables.

Product	Unit	Variable Definition	Spatial Resolution	Temporal Resolution
AOD		Aerosol optical depth	0.05°×0.05	1hour
Tempc	°C	Temperature	0.05°×0.05°×12L	1hour
RH	%	Relative Humidity	0.05°×0.05°×12L	1hour
HPBL	m	Planetary Boundary Layer Height	0.05°×0.05°	1hour
P	Hpa	Pressure	0.05°×0.05°×12L	1hour
U	m/s	Wind Speed (U)	0.05°×0.05°×12L	1hour
V	m/s	Wind Speed (V)	0.05°×0.05°×12L	1hour
DEM	m	Digital Elevation Model	0.01°×0.01°	Annual
POI		Point of Interest	0.01°×0.01°	Annual
Traffic Network		Traffic Network	0.01°×0.01°	Annual
GDP	￥/km2	Gross Domestic Product	0.01°×0.01°	Annual
TPOP	people/km2	population density	0.01°×0.01°	Annual
Land Cover Type		Land Cover Type	0.05°×0.05°	Annual
EVI		Enhanced Vegetation Index	0.05°×0.05°	Monthly
NDVI		Normalized Differnce Vegetation Index	0.05°×0.05°	Monthly

Table S12. Statistics of number of stations used for training and testing

	All Number(N)	Training Number(N)	Testing Number(N)	All Sample Number(N)			
				2017	2018	2019	2020
North China	176	150	26	1340252	1241169	1257882	1228487
East China	343	308	35	2663647	2525520	2478030	2503516
South China	237	213	24	1957469	1875939	1865894	1908516
Sichuan Basin	145	130	15	1132051	1043238	1044349	1062847

Shaanxi Province	177	159	18	1369405	1280065	1277902	1301866

Section 2.4 Sensitivity Analysis: What do the levels mean? What does this section mean?

Reply:

- We open the neural network black box model by sensitivity analysis, and quantitatively analyze the impact of each input variable on the results. These analysis indexes are used to measure the impact of each input variable on the results. This section explains the sensitivity analysis method we adopted.
- Different indexes have certain differences in the distribution characteristics of inputs. Rg measures the impact of data range on results, Rg measures the impact of data gradient, Rv measures the impact of data variance, and Raad measures the impact of data dispersion. Compared with Rv, it is less sensitive to outliers (Cortez and Embrechts, 2013).
- Cortez, P. and Embrechts, M. J.: Using sensitivity analysis and visualization techniques to open black box data mining models, Information Sciences, 225, 1-17, 10.1016/j.ins.2012.10.039, 2013.

Line 231: What is 26 micrograms m^3 in terms of percentage? Or how does it compared to usual values? Is that significant?

Reply:

- From Table.2-1, we can see the results of the RMSE. According to the definition of pollution classification, the 24-hour average PM2.5 is classified according to: 0 to 35 μgm^{-3} is excellent, 35 to 75 μgm^{-3} is good, 75 to 115 μgm^{-3} is light pollution, 115 to 150 μgm^{-3} is medium pollution, 150 to 250 μgm^{-3} is heavy pollution, and more than 250 μgm^{-3} is serious pollution.
- For example, among the 12 CNEMC sites in Beijing in 2017, more than 23% of the data were more than 75 μgm^{-3}, and more than 6% of the data were more than 150 μgm^{-3}. For Shijiazhuang, more than 36% of the data were more than 75 μgm^{-3},more than 11% of the data were more than 150 μgm^{-3}.
- From Figure. S16, we can see that the relative error of the result validation is within 20% (except the high value and low value). But for a clean environment, our error has a significant impact.

Table 2-1. RMSE of cross validation with respect to spatial distribution.

	2017		2018		2019		2020	
	day	night	day	night	day	night	day	night
North China	19.77	22.59	19.92	19.86	16.53	18.44	16.46	13.99
East China	16.15	16.51	13.09	14.04	13.19	12.13	9.88	9.47
South China	11.11	12.81	10.38	11.38	9.52	11.41	6.00	8.96
Sichuan Basin	14.80	17.52	13.90	18.51	10.28	11.86	8.03	10.74
Shaanxi Province	20.15	22.79	15.47	18.88	15.14	17.13	12.01	12.33

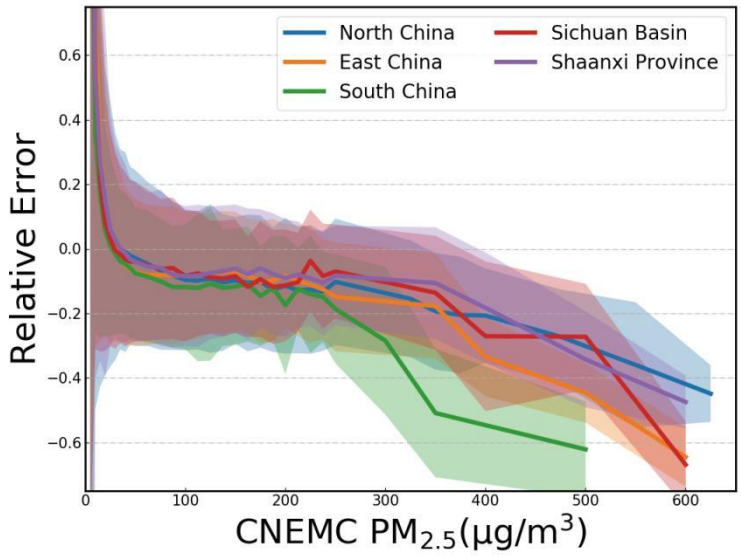

Figure. S16. Relative error varies with $PM_{2.5}$ concentrations in different regions. Five lines with different colors represent errors for North China, East China, South China, Sichuan Basin, and Shaanxi Province.

Line 234: Why is the model good at nighttime prediction?

Reply:

- The atmospheric lifetime of aerosols under different conditions is different. Under static and stable conditions, aerosols have a longer atmospheric lifetime. During monsoon and precipitation, atmospheric aerosols will settle faster. We found the potential temporal and spatial relationship between variables and atmospheric aerosols by mining the temporal and spatial characteristics of multi-source data, and then obtained the PM2.5 concentration at night.

Line 240: What does it mean the data are influenced by meteorological and aerosol data at 0.05?

Reply:

- We obtained the meteorological data with 0.05 ° resolution through WRF model, and the aerosol data of satellite remote sensing also has regrid as 0.05 °. These are the most important weighting factors. The temporal and spatial distribution characteristics of $PM_{2.5}$ are mainly affected by them. And we use ST-NN model to mine more refined geographic information data to obtain the temporal and spatial distribution of 0.01 ° surface $PM_{2.5}$.

Line 264: What are the AOD conditions in cloudy scenes? How does the model predict without AOD if it is one of the main predictands?

Reply:

- AOD cannot be retrieved by satellite remote sensing under cloudy conditions. Therefore, the satellite AOD is treated as NAN value in the pixel with cloud.
- Aerosols are spatiotemporal correlation. Compared with previous studies, we have made full use of the aerosol data in the spatiotemporal neighborhood. The

atmospheric lifetime and evolution process of aerosols are affected by their components and meteorological conditions (monsoon, precipitation, etc.). We obtain the ground $PM_{2.5}$ concentration under cloud by mining the potential relationship between meteorological data under different characteristic conditions, spatiotemporal neighborhood AOD data and various geographic information data (POI, road, etc.).

Line 306: Have any other studies ever used a NN or RF to predict PM2.5?

Reply:

- This is a hot research topic, and there are many related studies. Table.S9 shows typical studies in this field using different methods. The total number of citations of these articles was 2537, of which 9 studies cited more than 100. These studies are a good review of this field.
- In addition to Table. S9, Table.Review.1 also uses the random forest/neural network method to obtain the ground $PM_{2.5}$ concentration. In the revise manuscript, we add them to the Table.S9.
- Table.S9 shows typical studies in this field using different methods. The total number of citations of these articles was 2537, of which 9 studies cited more than 100.

Table.Review.1

Study	Model	Resolution	Study Area	Sample Validation			Space Validation			Time Validation		
				R^2	RMSE	Slop	R^2	RMSE	Slop	R^2	RMSE	Slop
Xiao et.al (2018)	Model	0.1°daily	China (2013~2017)	0.79	21	1	0.76	22	1	0.73	24	1
Mhawish et. al (2020)	LME+RF	1km daily	IGP (2018.7~2019.6)	0.87	28	0.84	0.87	28.1	0.84			
Guo et. Al (2021)	RF	1km daily	China (2017)	0.74	16.29	0.61	0.72	16.73	0.6	0.36	20.41	0.34
Schneider et. al (2020)	RF	1km daily	Great Britain (2008~2018)	0.77	4.042	1.05	0.66	2.24	0.99	0.8	3.38	1.06
Dong et. al (2020)	RF-BPNN	0.05°hourly (daytime)	China (2017)	0.8	18.54		0.76	20.27				

Xiao, Q., Chang, H. H., Geng, G., and Liu, Y.: An Ensemble Machine-Learning Model To Predict Historical PM2.5 Concentrations in China from Satellite Data, Environmental Science & Technology, 52, 13260-13269, 10.1021/acs.est.8b02917, 2018.

Mhawish, A., Banerjee, T., Sorek-Hamer, M., Bilal, M., Lyapustin, A. I., Chatfield, R., and Broday, D. M.: Estimation of High-Resolution PM2.5 over the Indo-Gangetic Plain by

Fusion of Satellite Data, Meteorology, and Land Use Variables, Environmental Science & Technology, 54, 7891-7900, 10.1021/acs.est.0c01769, 2020.
Guo, B., Zhang, D., Pei, L., Su, Y., Wang, X., Bian, Y., Zhang, D., Yao, W., Zhou, Z., and Guo, L.: Estimating PM2.5 concentrations via random forest method using satellite, auxiliary, and ground-level station dataset at multiple temporal scales across China in 2017, Science of The Total Environment, 778, 146288, https://doi.org/10.1016/j.scitotenv.2021.146288, 2021.
Schneider, R., Vicedo-Cabrera, A. M., Sera, F., Masselot, P., Stafoggia, M., de Hoogh, K., Kloog, I., Reis, S., Vieno, M., and Gasparrini, A.: A Satellite-Based Spatio-Temporal Machine Learning Model to Reconstruct Daily PM2.5 Concentrations across Great Britain, Remote Sensing, 12, 3803, 2020.
Dong, L., Li, S., Yang, J., Shi, W., and Zhang, L.: Investigating the performance of satellite-based models in estimating the surface PM2.5 over China, Chemosphere, 256, 127051, https://doi.org/10.1016/j.chemosphere.2020.127051, 2020.

Line 316: How do your inputs compare to past studies?
Reply:

- Compared with previous studies, our input data has higher spatial and temporal resolution characteristics. In the past, NCEP or ECMWF meteorological field data were used directly, with a spatial resolution of 0.25 or 0.125 ° and a temporal resolution of 3 or 6 hours. We used them as the initial boundary field, and used WRF to conduct three-layer nested simulation to obtain a meteorological field with a resolution of 0.05 ° and an hourly resolution. For geographic information data, we classified roads and buildings. We conduct more pre-processing based on prior knowledge.

Line 323: How good of a proxy is AOD for PM2.5 or PM10?
Reply:

- Figure. S14 shows the weight of each input data. As the largest influencing variable of PM2.5, AOD accounts for more than 30%.

Figure. S14. Relative importance indicators (R_r: Relative range; R_g: Relative gradient; R_v: Relative variance; R_{AAD}: Relative average absolute deviation) of input variables for different regions. (a-d) North China. (e-h) East China. (i-l) South China. (m-p) Sichuan Basin. (q-t) Shaanxi Province.