We thank the editor and the reviewers for the time and effort put in towards the review of this manuscript. The insightful comments and suggestions have helped improve the manuscript significantly. We have incorporated several changes based on the suggestions of the reviewers. The detailed responses to the reviewers' comments are given below.

Section 1. Major comments.

Question 1. It's not clear what the contributions of the paper are versus what already exists. Did the authors run the WRF and CMAQ models to generate the training data or were these data obtained from some other source?

Answer 1. We understand the concern raised by the reviewer. We directly generated the training data using the WRF and CMAQ. We will make revisions to the manuscript to clarify this point.

Question 2. Similarly there is no justification for the choice of model. I would like to see the approach benchmarked against simpler models (ARIMA, Random Forest, … basically anything from the statistical or machine learning family to compare against the deep learning approach). The volume of data that the model is trained on are not huge so it is not apparent that a DNN is the best choice of algorithm.

Question 2-1. Similarly, there is no justification for the choice of model. I would like to see the approach benchmarked against simpler models (ARIMA, Random Forest, … basically anything from the statistical or machine learning family to compare against the deep learning approach).

Answer 2-1. Among the statistical models mentioned by the reviewer, the prediction performances of the Random Forest (RF) have been evaluated and compared with that of the DNN-ALL. Table 1 shows the results of the statistical evaluations, and Table 2 lists the results of the Air Quality Index (AQI) evaluation. Compared to the results of the RF model, the Root Mean Square Error (RMSE) value of the DNN-ALL model decreased by 0.6 to 1.9 μgm^{-3}, and the Correlation Coefficient (R) and Index of Agreement (IOA) values increased slightly. The Accuracy (ACC) of the DNN-ALL model increased by approximately 2–13 %p compared to the RF model, and the F1-score decreased by 1 %p at D+1 but increased by 1 %p and 9 %p at D+0 and D+2, respectively. A comparison of the performance results showed that the DNN-ALL model outperformed the RF model. We will this information in the revised manuscript to include these results.

Table 1. Statistical performance of the DNN-ALL and Random Forest models.

Model	Day	MSE ((μgm^{-3})^2)	RMSE (μgm^{-3})	R	IOA
DNN-ALL	D+0	53.3	7.3	0.91	0.95
	D+1	81.0	9.0	0.85	0.90
	D+2	112.4	10.6	0.79	0.86
Random Forest	D+0	62.4	7.9	0.90	0.93
	D+1	106.1	10.3	0.83	0.85
	D+2	156.3	12.5	0.73	0.76

Table 2. Categorical performance of the DNN-ALL and Random Forest model.

Model	Day	ACC (%)		POD (%)		FAR (%)		F1-score (%)
DNN-ALL	D+0	77.8	70/90	72.7	16/22	11.1	2/18	80
	D+1	64.4	58/90	71.4	15/21	31.8	7/22	70
	D+2	61.1	55/90	76.2	16/21	40.7	11/27	67
Random Forest	D+0	75.6	68/90	77.3	17/22	19.0	4/21	79
	D+1	61.1	55/90	76.2	16/21	33.3	8/24	71
	D+2	48.9	44/90	71.4	15/21	50.0	15/30	58

Question 2-2. The volume of data that the model is trained on are not huge so it is not apparent that a DNN is the best choice of algorithm.

Answer 2-2. We agree that the volume of data in this paper is not sufficiently huge to be applied to artificial intelligence (AI). Nevertheless, the reason for choosing DNN algorithm is to take into account the scalability of the model, which can reflect training data expansion to forecast the segmentation with a 1-h interval and the future data growth over time. Therefore, the performance of the AI is expected to improve as the training data increases.

Question3. Many details on the DNN model setup are presented with no real justification for their choice e.g. using the membership function for temporal features, the choice of DNN architecture such as number of layers is not explained. On line 50 – 55 the authors 1) note the advantages of RNN for time series forecasting and 2) that Kim et al. (2019) developed an RNN model to predict PM2.5 concentrations at two locations in Seoul. Why was RNN not considered for this study rather than DNN and how does the performance of this model compare to that reported by Kim et. al. Similarly it is not clear if the autoregressive features of the data were expressed in any form? Of course RNN expresses these implicitly but in other models it can be advantageous to feature engineer the autoregressive dependencies. Were any feature combinations other than those reported explored in the paper. The authors need to justify the choice of algorithm and how the DNN was designed detailing such information as feature selection, number of layers/nodes.

Question 3-1. Many details on the DNN model setup are presented with no real justification for their choice e.g. using the membership function for temporal features.

Answer 3-1. . In this paper, the membership function was used to reflect these monthly change characteristics. As shown in Figure 1 (Figure 5 in the paper), $PM_{2.5}$ concentration in Seoul is high in January, February, March, and December, and low from August to October. $PM_{2.5}$ concentration has a characteristic that changes gradually from month to month. The examples of how membership function is applied are described in lines **151–153** of the paper. The membership function was applied based on the results presented by Yu et al. (2019). Yu et al. (2019) performed training that reflected monthly change characteristics to improve the high-concentration PM_{10} forecast performance. As indicated by the experiment results presented in Table 3, the POD performance of the training model reflecting the characteristics of the monthly change was improved by 25 %p. The information related to this will be added in the paper.

Table 3. Results of artificial intelligence model performance evaluation when using and without the membership function presented in Yu et al. (2019).

Model	Day	ACC (%)	POD (%)	FAR (%)
Using Membership function	D+1	70	75	48
Without Membership function	D+1	76	50	33

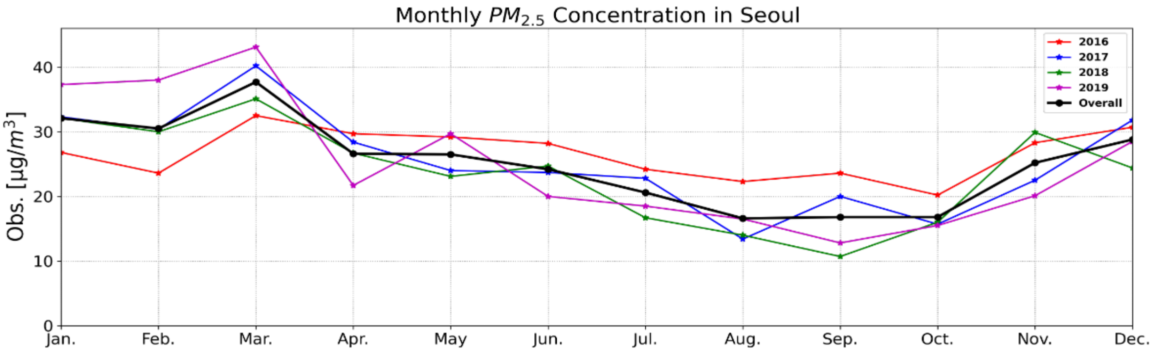

Figure 1. Time series of the average monthly $PM_{2.5}$ concentrations from 2016 to 2019.

Question 3-2. The choice of DNN architecture such as number of layers is not explained. The authors need to justify the choice of algorithm and how the DNN was designed detailing such information as feature selection, number of layers/nodes. Were any feature combinations other than those reported explored in the paper.

Answer 3-2. In order to provide the justification for the layer selection mentioned by the reviewer, we presented the evaluation results according to the number of layers. The statistical and AQI evaluation results of the DNN-ALL model based on the layer are presented in Tables 4 and 5, respectively. The results of the 4-layer and 5-layer models show that the performance is similar. However, compared with the 4-layer model, the RMSE of the 5-layer decreases by approximately 0.1 μgm^{-3} to 1 μgm^{-3} at D+0 to D+2, and the ACC of the 5-layer model increases by approximately 1 %p to 6 %p at D+0 to D+2. Therefore, the 5-layer model shows the best performance. The 6-layer and 8-layer models contain errors that converge without decreasing during the training process of the model (vanishing gradient problem). The authors believe that the cause of this problem is the activate function. Therefore, as the layer becomes deeper, the value of the last output cannot be significantly changed due to the sigmoid function. We will include this information in the revised manuscript.

Table 4. Statistical evaluation results according to the number of layers.

Model	Day	MSE $((\mu gm^{-3})^2)$	RMSE (μgm^{-3})	R	IOA
2-layer	D+0	59.3	7.7	0.91	0.94
	D+1	92.1	9.6	0.86	0.89
	D+2	156.3	12.5	0.75	0.80

4-layer	D+0	54.7	7.4	0.91	0.95
	D+1	88.3	9.4	0.86	0.90
	D+2	134.5	11.6	0.77	0.84
5-layer (DNN-ALL)	D+0	53.3	7.3	0.91	0.95
	D+1	81.0	9.0	0.85	0.90
	D+2	112.4	10.6	0.79	0.86
6-layer	D+0	174.2	13.2	0.81	0.66
	D+1	292.4	17.1	0	0.17
	D+2	292.4	17.1	0	0.17
8-layer	D+0	302.7	17.4	0	0.15
	D+1	292.4	17.1	0	0.17
	D+2	292.4	17.1	0	0.17

Table 5. AQI evaluation results according to the number of layers.

Model	Day	ACC (%)		POD (%)		FAR (%)		F1-score (%)
2-layer	D+0	70.0	63/90	81.8	18/22	28.0	7/25	77
	D+1	55.6	50/90	81.0	17/21	39.3	11/28	69
	D+2	51.1	46/90	81.0	17/21	50.0	17/34	61
4-layer	D+0	71.1	64/90	81.8	18/22	28.0	7/25	76
	D+1	60.0	54/90	85.7	18/21	35.7	10/28	73
	D+2	60.0	54/90	81.0	17/21	45.2	14/31	65
5-layer (DNN-ALL)	D+0	77.8	70/90	72.7	16/22	11.1	2/18	80
	D+1	64.4	58/90	71.4	15/21	31.8	7/22	70
	D+2	61.1	55/90	76.2	16/21	40.7	11/27	67
6-layer	D+0	55.6	50/90	50	11/22	8.3	1/12	64
	D+1	47.8	43/90	0	0/21	0	0/0	0
	D+2	47.8	43/90	0	0/21	0	0/0	0
8-layer	D+0	45.6	41/90	0	0/22	0	0/0	0
	D+1	47.8	43/90	0	0/21	0	0/0	0
	D+2	47.8	43/90	0	0/21	0	0/0	0

Question 3-3. On line 50 – 55 the authors 1) note the advantages of RNN for time series forecasting. and 2) that Kim et al. (2019) developed an RNN model to predict PM2.5 concentrations at two locations in Seoul. Why was RNN not considered for this study rather than DNN and how does the performance of this model compare to that reported by Kim et. al.

Answer 3-3-1. (Reason why RNN was not considered) There are very few studies and relatively less research to predict air quality using AI such as DNN, RNN and CNN, although it has increased recently. Therefore, the purpose of this study is to evaluate the performance of fine dust prediction when using the DNN among various AI algorithms. The RNN is known to have the advantage of time series prediction, and the DNN is known to have the advantage of extracting characteristics of training data well. There is no convergent result confirming which of the two algorithms is better when applied to fine dust prediction. Therefore, we first performed the simulation using the DNN rather than the RNN in order to maximize the advantages of the DNN for predicting fine dust. In the future, we plan to perform comparative evaluation with the DNN results presented in this paper through the development of RNN models.

Answer 3-3-2. (Comparison with Kim et al.) We compared the results obtained by Kim et al. (2019) with those obtained in our study. Kim et al. (2019) performed a $PM_{2.5}$ concentration prediction for two out of 41 measuring stations that are located in the Seoul area. However, in this paper, the average $PM_{2.5}$ concentration prediction for 41 measuring stations in Seoul was performed. In other words, there is a spatial difference for the area to be predicted. In addition, the periods of prediction for the two papers are different. The forecast period considered by Kim et al. (2019) was four months, from January 2016 to April 2016, and the forecast period in this study was three months, from January 2021 to March 2021. Although it is difficult to directly compare the two studies because of the existence of temporal and spatial differences, the results of the prediction performance are presented in Table 6. Because Kim et al. (2019) performed only the D+1 prediction, the comparison of the prediction performance with this paper was conducted for D+1. The values indicate that the RMSE is decreased and the IOA is increased compared to other models.

Table 6. Statistical performance of the DNN-ALL and Random Forest models.

Model	Day	RMSE (μgm^{-3})	IOA
DNN-ALL	D+1	9.0	0.90
Seoul-1 (Kim et. al (2019))	D+1	12.5	0.71
Seoul-2 (Kim et. al (2019))	D+1	15.1	0.77

Question 3-4. Similarly, it is not clear if the autoregressive features of the data were expressed in any form? Of course, RNN expresses these implicitly but in other models it can be advantageous to feature engineer the autoregressive dependencies.

Answer 3-4. The RNN algorithm implicitly reflects autoregressive features, but the DNN algorithm does not reflect autoregressive features. This study did not consider any autoregressive features.

Question4. The manuscript could be improved to enhance readability and replicability of the study. I appreciate the authors making code and data available on Zenodo. I would however encourage them to create a GitHub repository with some documentation to allow people easily replicate the results. As mentioned in 1) authors could be more descriptive when detailing data sources. Some parts of the paper could be explained better, e.g. line 161 “average weather and air quality prediction data”. What is meant by average here? Spatial or temporal. Are WRF and CMAQ data extracted from the entire Seoul area domain or subset corresponding to location of the observation point. Line 166 – 167 “ensure that the training data were not biased” – feature scaling does not ensure unbiased datasets, it simply helps the model learn better. The data could still be biased. Line 173: “undergoes feature scaling through the backpropagation algorithm” – not clear what is meant by feature scaling in this context. Line 128 “16 meteorological forecast variables were created by the WRF model” – I believe what is meant here is that 16 variables were extracted as features, but many more variables were generated by the WRF model.

Question 4-1. Line 161 “average weather and asir quality prediction data”. What is meant by average here? Spatial or temporal. Are WRF and CMAQ data extracted from the entire Seoul area domain or subset corresponding to location of the observation point.

Answer 4-1. In “average weather and air quality prediction data” - "average" refers to conversion of 1-h interval data into 6-h interval data. In addition, spatially, it means the average of 9 km grids corresponding to Seoul. We have clarified the meaning and revised it in the paper.

Question 4-2. Line 166 – 167 “ensure that the training data were not biased” – feature scaling does not ensure unbiased datasets, it simply helps the model learn better. The data could still be biased.

Answer 4-2. We thank the reviewer for highlighting this issue. We will incorporate changes based on the suggestion of the reviewer.

Original: Feature scaling, involving standardization and normalization, was used to convert the data into a uniform format, ensure that the training data were not biased and that equal learning took place for the DNN model in each T-step.

Revise: The feature scaling, including standardization and normalization, was implemented to transform data into uniform formats, reduce data bias of training data, and ensure equal learning for the DNN model at each T-step.

Question 4-3. Line 173: “undergoes feature scaling through the backpropagation algorithm” – not clear what is meant by feature scaling in this context.

Answer 4-3. The phrase “undergoes feature scaling through the backpropagation algorithm” means that feature scaling data is used as training data for the DNN model. We will clarify the meaning in the revised manuscript.

Question 4-4. Line 128 “16 meteorological forecast variables were created by the WRF model” – I believe what is meant here is that 16 variables were extracted as features but many more variables were generated by the WRF model.

Answer 4-4. "16 meteorological forecast variables were created by the WRF model" - In the paper, the reason for using the weather forecast data was explained through several reference papers in section 2.1. Additionally, $PM_{2.5}$ is discharged from the ground, and it moves at an altitude of 1.5 km or less. Therefore, in this paper, lower altitude data were used. We will add this content in the revised manuscript.

Question 4-5. The manuscript could be improved to enhance readability and replicability of the study. I appreciate the authors making code and data available on Zenodo. I would however encourage them to create a GitHub repository with some documentation to allow people easily replicate the results. As mentioned in 1) authors could be more descriptive when detailing data sources.

Answer 4-5. As suggested by the reviewer, we upload the code to GitHub. (https://github.com/GercLJB/GMD)

Question 5. I really don’t see the relevance of section 4.2. The models have already been compared and evaluated in terms of predictive skill in regression. Then you take the same models and evaluate in terms of a classification model but only whether they predicted within those bounds (i.e. the model and results are the same the only thing that changes are the interpretation)

Answer 5. In Korea, the PM2.5 forecast results are categorized and provided to the public as good (PM$_{2.5}$ ≤ 15 μgm^{-3}), moderate (16 μgm^{-3} ≤ PM$_{2.5}$ ≤ 35 μgm^{-3}), bad (36 μgm^{-3} ≤ PM$_{2.5}$ ≤ 75 μgm^{-3}), and very bad (76 μgm^{-3} ≤ PM$_{2.5}$) Therefore, both the statistical and category evaluations are necessary to determine whether the DNN model developed in this paper is suitable for forecasting. Section 4.2 presents the comparison of the category performance of the DNN-ALL model and that of the CMAQ model to identify the superior model for actual prediction.

Section 2. Minor comments.

Question 1. What is the membership function defined in line 144? Is this the generation of temporal features described in subsequent lines? Why was time data encoded in this manner? It seems more standard to represent as integer values or to convert those integer values to cyclic features (i.e. so that month 12 and month 1 are close to each other rather than far away). I haven't seen this approach used previously and would like to understand the motivation and/or justification.

Answer 1. As explained in A3-1 among the answers to Q3 (Section 1), the data was expressed stochastically through the membership function to reflect the characteristics of the monthly change.

Question 2. The test period is quite short – 3 months out of 51 months. Was there a reason for this?

Answer 2. The data from 2016 to 2018 were used as training data, and those from 2019 were used as evaluation data. The data from January to March 2021 were used as test data to find out the performance when the actual DNN model was predicted.

Question 3. Line 159 – 163: This is a quite confusing way to present forecast horizons. I'd suggest to just use hours and present forecast horizons as T06, T12, T18, T24, ... Mixing days + hours and having different chunks within each day is confusing to the reader.

Answer 3. The T-step presented in Table 3 of the paper was revised and is shown in Table 7.

Table 7. Statistical performance of the DNN-ALL and Random Forest models.

Day	T-Step	Time	Composition of learning data
D+0	T12	07:00 to 12:00	01~06‘o clock Observation data of D+0 + Forecast data of Tx(x : 06, 12, 19, 24) from CMAQ and WRF
	T18	13:00 to 18:00	
	T24	19:00 to 00:00	
D+1	T06	01:00 to 06:00	
	T12	07:00 to 12:00	
	T18	13:00 to 18:00	
	T24	19:00 to 00:00	
D+2	T06	01:00 to 06:00	
	T12	07:00 to 12:00	
	T18	13:00 to 18:00	
	T24	19:00 to 00:00	

Question 4. Line 167 – 168: I don't quite understand why data was both standardized and normalized? Did this improve performance versus just using normalization (if you wished to have bounded between 0 and 1) or indeed versus the unscaled data? Generally people chose either standardization or normalization so I'm curious why you did both.

Answer 4. The normal distribution of input variables was standardized through standardization. The normalization was applied thereafter to ensure that the scale of each variable is equal. The reason why both normalization and standardization were applied was to train the characteristics of input variables equally to the DNN model.

Question 5. Figure 4: What does Epoch_n = Epoch_n-1 + 1 mean? What does Epoch_n-1 of validation cost > Epoch_n of validation cost mean? Should it be Validation cost of Epoch_n?

Answer 5. While addressing the concern raised by the reviewer, we found out that the formula was incorrect. The modified picture is shown in Fig 1. First, "Epoch_n=Epoch_{n-1}+1" expresses that the epoch increases by one as the algorithm is repeated. We modified this part to "Epoch_{n+1} = Epoch_n + 1". In addition, "Epoch_{n-1} of Validation cost > Epoch_n of Validation cost" is an incorrect expression, and we will revise it as "Validation cost of Epoch_{n-1}> Validation cost of Epoch_n". In addition, in this part, we found that the inequality sign was incorrectly marked, and it was corrected to "Validation cost of Epoch_{n-1} < Validation cost of Epoch_n".

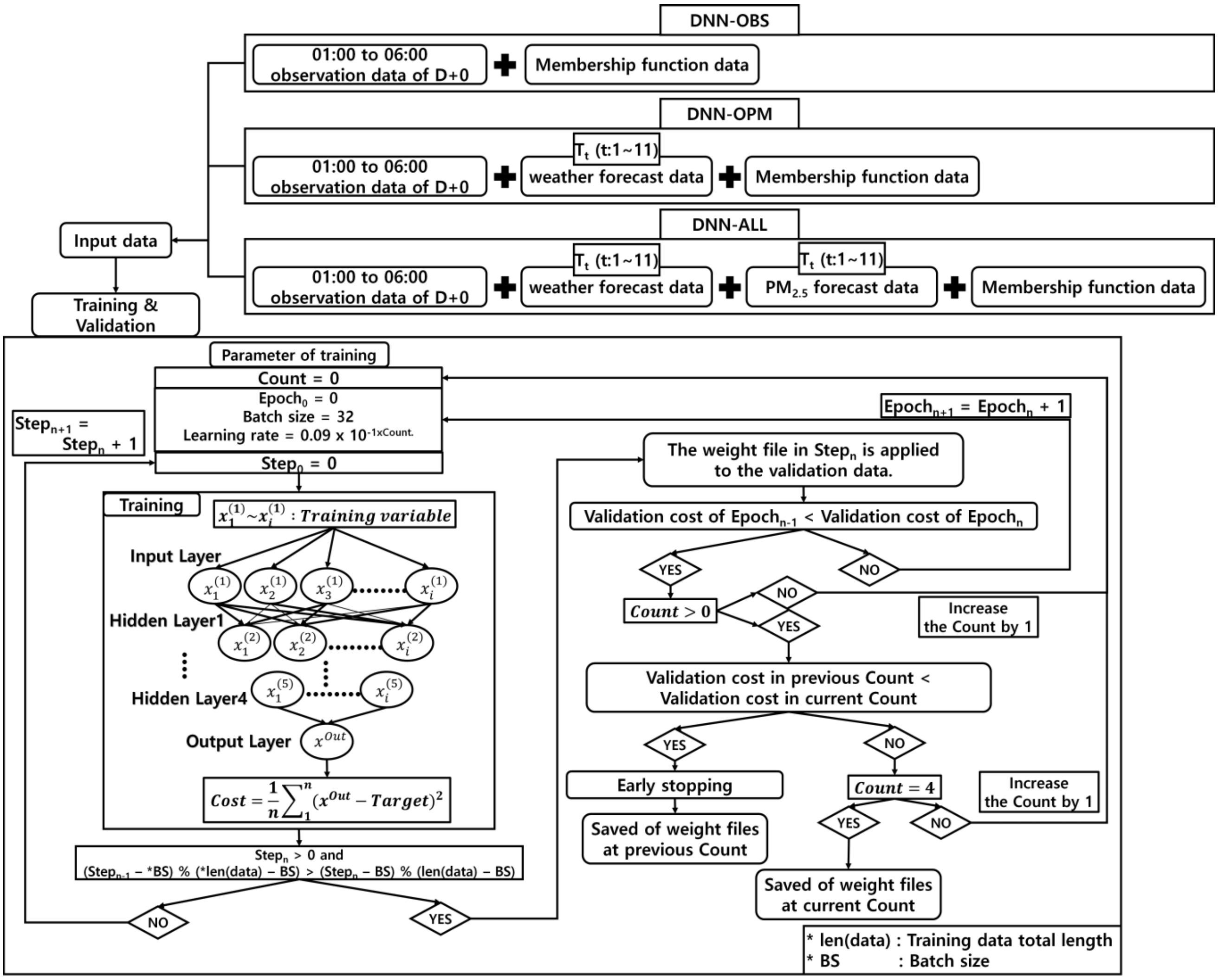
DNN-OBS
01:00 to 06:00 observation data of D+0
Membership function data
DNN-OPM
01:00 to 06:00 observation data of D+0
T_t (t:1~11)
weather forecast data
Membership function data
DNN-ALL
01:00 to 06:00 observation data of D+0
T_t (t:1~11)
weather forecast data
T_t (t:1~11)
PM2.5 forecast data
Membership function data
Input data
Training & Validation
Parameter of training
Count = 0
Epoch0 = 0
Batch size = 32
Learning rate = 0.09 x 10^-1xCount
Step0 = 0
Stepn+1 = Stepn + 1
Training
x1(1)~xi(1) : Training variable
Input Layer
Hidden Layer1
Hidden Layer4
Output Layer
x^Out
Cost = 1/n Σ (x^Out − Target)^2
Stepn > 0 and
(Stepn-1 − *BS) % (*len(data) − BS) > (Stepn − BS) % (len(data) − BS)
NO
YES
Epochn+1 = Epochn + 1
The weight file in Stepn is applied to the validation data.
Validation cost of Epochn-1 < Validation cost of Epochn
YES
NO
Count > 0
NO
YES
Increase the Count by 1
Validation cost in previous Count < Validation cost in current Count
YES
NO
Early stopping
Saved of weight files at previous Count
Count = 4
Increase the Count by 1
YES
NO
Saved of weight files at current Count
* len(data) : Training data total length
* BS : Batch size

Figure 2. Structure of DNN model training process.

Question 6. Equation 11 and 12, I’m not sure the use of both MSE and RMSE is necessary and could probably drop one.

Answer 6. In this paper, the MSE was used as the cost function of the DNN model. Therefore, it was intended to indicate the degree of difference in MSE for each model. In addition, the RMSE was presented to compare the model errors in units of the $PM_{2.5}$ concentration.

Question 7. Line 225 – 230: In classification problems, accuracy, precision and recall are the standard metrics presented. I would suggest in this paper you also include (you already do accuracy so I suggest adding precision and recall)

Answer 7. Based on the suggestion of the reviewer, we have added the precision and recall in Table 8. The precision and recall of all categories for the DNN-ALL model in D+0 is presented to be 0.12 equally higher than that of the CMAQ model. However, in D+1 and D+2, the precision of the DNN-ALL and the CMAQ models is found to be similar, and the recall of the DNN-ALL model shows a slight decrease compared to the CMAQ model. The reason for these results is that the two indexes of the good and moderate categories of DNN-ALL are reduced compared to the CMAQ model.

On the other hand, the precision and recall of the DNN-ALL model in the bad category are presented to be higher than those of the CMAQ model. In the bad category of D+0, the precision and recall of DNN-ALL are greater than those of the CMAQ model by 0.24 and 0.04, respectively. In addition, in the very bad category, the precision and recall of DNN-ALL are to be 1.0 equally higher than those of the CMAQ model. In D+1, the precision of DNN-ALL in the bad category is greater than that of the CMAQ model by 0.1, but the recall is similar to the CMAQ model. In D+2, the precision and recall for the bad category of DNN-ALL increased by 0.14 and 0.20 compared to the CMAQ model, respectively. These results show that the performance of the DNN-ALL model is superior to that of the CMAQ model for predicting high concentrations that affect the health of the people.

Table 8. Precision and recall of DNN-ALL and CMAQ models by four categories : "good" (PM2.5 ≤ 15 µgm-3), "moderate" (16 µgm-3 ≤PM2.5 ≤35 µgm-3), "bad" (36 µgm-3 ≤ PM2.5 ≤ 75 µgm-3), and "very bad" (76 µgm-3 ≤ PM2.5).

Model	Day	Precision					Recall				
		Good	Moderate	Bad	Very bad	Total	Good	Moderate	Bad	Very bad	Total
DNN-ALL	D+0	0.83	0.71	0.88	1.0	0.86	0.70	0.85	0.71	1.0	0.82
	D+1	0.83	0.61	0.64	0.0	0.52	0.38	0.79	0.70	0.0	0.47
	D+2	0.79	0.59	0.56	0.0	0.49	0.42	0.67	0.75	0.0	0.46
CMAQ	D+0	0.74	0.67	0.64	0.0	0.51	0.64	0.68	0.67	0.0	0.50
	D+1	0.85	0.69	0.54	0.0	0.52	0.65	0.67	0.70	0.0	0.51
	D+2	0.82	0.69	0.42	0.0	0.48	0.69	0.63	0.55	0.0	0.47

Question 8. Line 280 – 281: This is a difficult narrative to support. You are using the CMAQ model output as the training data and then saying the usage of that training data allows the model to better represent the long-term-transport-induced phenomenon.

Answer 8. In the DNN model (DNN-OBS) that uses only observation data, the RMSE for three months is 11.4 μgm^{-3} in D+0. The predictive performance is improved by decreasing the RMSE to 9.4 μgm^{-3} for three months, excluding high concentration cases (February 11-14, 2021) by long-distance transport. These results imply that the prediction of high concentration occurrence due to long-distance transportation is insufficient in the case of the DNN model when only the observation data are used. On the other hand, when both the observation data and prediction data are used in the DNN model (DNN-ALL), the RMSE for three months is 7.3 μgm^{-3}, and the RMSE for excluding high concentration cases by long-distance transport is 7.0 μgm^{-3}, showing no significant difference. In addition, the RMSE of DNN-ALL is reduced compared to the CMAQ model, showing a superior performance. Therefore, it is inferred that the use of prediction data produced by CMAQ improved the predictive performance of high concentration phenomena by long-distance transport.

Question 9. What do the dashed lines in the residual figures represent in Figure 7.

Answer 9. In Fig. 7 (a1), (b1), and (c1), each of the dashed lines represents values of 15.5 μgm^{-3}, 35.5 μgm^{-3}, 75.5 μgm^{-3}, and average value of observation (27 μgm^{-3}). Moreover, 15.5 μgm^{-3} is the boundary between “Good” and “Moderate,” 35.5 μgm^{-3} is the boundary between “Moderate” and “Bad,” and 75.5 μgm^{-3} is the boundary between “Bad” and “Very bad.”

In Fig. 7 (a2), (b2) and (c2), the dashed lines represent the standard deviation of observed $PM_{2.5}$ as negative and positive.