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. In Section 1, the answers to major comments are provided, and in Section 2, answers to minor comments are given.

Section 1. Major comments.

Q1. Line 75~80: "In addition, the membership function was used to reflect temporal information." More information is needed about "membership function". How does this function reflect temporal information?

A1. The concept of the membership function is derived from the fuzzy theory, and it defines the probability that a single element belongs to a set. In this study, the probability that the date (element) belongs to 12 months (set) was calculated using the membership function. The date change probability was trained as a factor that reflected the characteristics of the monthly change. As shown in Figure 1 (Figure 5 in the paper), the $PM_{2.5}$ 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. In this paper, the membership function was used to reflect these monthly change characteristics. The examples of how membership function is applied are described in lines **151–153** of the paper.

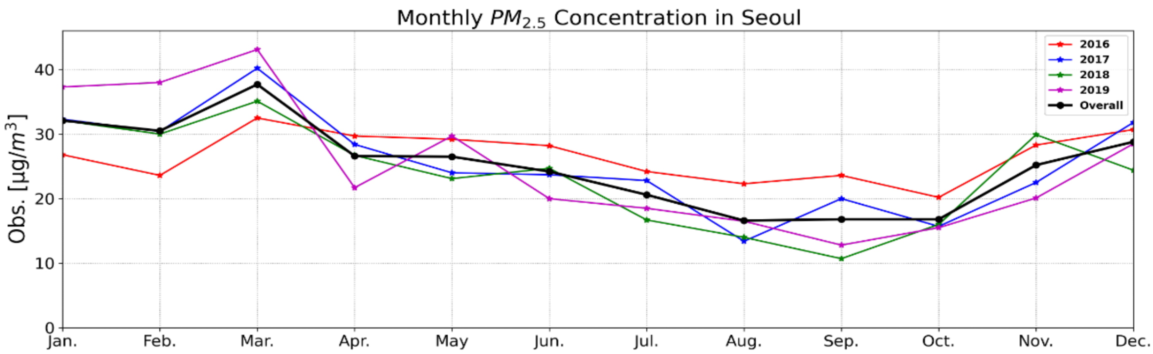

Figure 1. Time series of the monthly average $PM_{2.5}$ concentrations from 2016 to 2019.

Q2. Line 145-155: The authors want to predict PM2.5 within 3 days. Why do you need to add the time information ("adjacent month") of the next month that hasn't happened yet in Eq. (2)? If you know the information of next month, you can predict PM2.5 of next month. Why only forecast PM2.5 within 3 days. This is very difficult to understand.

A2. As described in A1, the probability of reflecting the characteristics of the monthly change was calculated using the membership function. The calculated probability was referred to as "adjacent month" and "month." Therefore, "adjacent month" is not a factor that provides information for the next month, but the one that represents the characteristics of the monthly change according to the corresponding date.

Q3. In Section 2.2, what are the super parameters of DNN model? Why only use five stacked-layer DNN model? Generally, a neural network model with more than 8 hidden layers is considered as a deep neural network (Hinton et al., 2012).

A3. 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 indicate similar performance. However, compared to the 4-layer model, the RMSE of the 5-layer model 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 a superior performance.

The 6-layer and 8-layer models have a problem of errors that converge without any decrease in the training process of the model (vanishing gradient problem). The authors believe that the reason for 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.

Table 1. 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 2. 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

Q4. Line 210~215: The input data of the three experiments (DNN-OBS, DNN-OPM and DNN-ALL) are not very clear. Why should the predicted PM2 5 into the model (DNN-ALL)? Reason?

A4. The measurement variables presented in Table 1 in Section 2.1 of the paper were used as common variables in the three experiments (DNN-OBS, DNN-OPM, and DNN-ALL). The DNN-OBS used the observation data as the sole training data. Among the predictors shown in Table 2 in Section 2.1 of the paper, the variables produced in the WRF model were used in the DNN-OPM and DNN-ALL experiments, whereas the variables produced in the CMAQ model were used only in the DNN-ALL experiments.

The predicted PM2.5 from CMAQ tends to be over-simulated than the observed PM2.5, but the correlation appears to be good. Therefore, it was judged as training data that can reflect the features of observed PM2.5. The predicted PM2.5, the predicted weather data from WRF, and observation data were studied together to improve PM2.5 prediction performance using DNN-ALL.

Q5. Line 230~240: There's something wrong with Eq. (18). The commonly used expression for F1-score is (2*ACC*Recall)/(ACC+Recall). F1-score is for one category. My understanding is that there are four categories (Good, Moderate, Bad and Very bad). Has anyone else used it like this? More explanation is needed.

A5. The authors agree that F1-score is generally referred to as (2*Precision*Recall)/(Precision + Recall). The F1-score used in this paper is not an evaluation of four categories, but an index that simultaneously considers (1-FAR) and POD to evaluate the prediction performance for exceeding 35 μgm^{-3} as a bad criterion. Tables 3 and 4 show the intervals corresponding to the four and two categories for POD and FAR calculation, respectively. The I in Table 4 is a corresponding category for the conditions of a1, a2, b1, and b2 of Table 3. Similarly, II in Table 4 correspond to c1, c2, d1, and d2, III in Table 4 correspond to a3, a4, b3, b4, and IV in Table 4 correspond to c3, c4, d3, and d4. Eq. (7) represents a ratio when prediction concentration in the model corresponds to the observation category in the case that the observation concentration appears in the bad or very bad category. Eq. (8) is the ratio when observation concentration is in the good or moderate category in the case that the prediction concentration appears in bad or very bad category. The POD means Recall, and FAR means (1-precision). Therefore, we use F1-score to reflect the harmonious mean of POD and (1-FAR).

$$POD\ (\%) = \frac{(c3+c4+d3+d4)}{(a3+a4+b3+b4+c3+c4+d3+d4)} \times 100\ ,$$ (1~~5~~)

$$FAR\ (\%) = \frac{(c1+c2+d1+d2)}{(c1+c2+c3+c4+d1+d2+d3+d4)} \times 100\ ,$$ (2~~6~~)

Table 3. Intervals corresponding to the four categories for calculating POD and FAR: "good" ($PM_{2.5} \leq 15$ μgm^{-3}), "moderate" (16 μgm^{-3} $\leq PM_{2.5} \leq 35$ μgm^{-3}), "bad" (36 $\mu gm^{-3} \leq PM_{2.5} \leq 75$ μgm^{-3}), and "very bad" (76 $\mu gm^{-3} \leq PM_{2.5}$).

Level		Model forecast			
		Good	Moderate	Bad	Very bad
Observation	Good	a1	b1	c1	d1
	Moderate	a2	b2	c2	d2
	Bad	a3	b3	c3	d3
	Very bad	a4	b4	c4	d4

$$POD\ = \frac{\text{Ⅳ}}{\text{Ⅲ} + \text{Ⅳ}}\ ,$$ (3~~7~~)

$$FAR\ = 1 - \frac{\text{Ⅳ}}{\text{Ⅱ} + \text{Ⅳ}}\ ,$$ (4~~8~~)

Table 4. Intervals corresponding to the two categories for calculating POD and FAR : "good and moderate" ($PM_{2.5} \leq 35$ μgm^{-3}), "bad and very bad" ($PM_{2.5} \geq 36$ μgm^{-3}).

Level		Model forecast	
		Good and moderate	Bad and very bad
Observation	Good and Moderate	Ⅰ	Ⅱ
	Bad and Very bad	Ⅲ	Ⅳ

Q6. In Table 2, why are 925hPa and 850hPa variables selected? Why not consider 700hPa and 500hp variables? Is there any reason?

A6. Various forecast data for each altitude are produced in the WRF model. However, the reason why the upper layer altitude (700 and 500 hPa) was not used in this study is that the emission of PM2.5 mainly occurs on the ground and moves up to an altitude of 1.5 km. Therefore, we only used the lower altitude forecast data.

Q7. Table 5 only provides the performance of the model in the test set (January–March 2021) and it is recommended to add the performance of the model in the training set (2016 to 2018) and validation set (2019).

A7. Table 3 shows the statistical evaluation results of three experiments (DNN-OBS, DNN-OPM, and DNN-ALL) and CMAQ models from 2016 to 2018. In D+0 to D+2, the DNN-ALL model performs the best in terms of all statistical indicators. In addition, the values of all three experiments indicate a decrease in the RMSE compared to the CMAQ model.

Table 4 presents the statistical evaluation results of the three experiments (DNN-OBS, DNN-OPM, and DNN-ALL) and CMAQ models for 2019. The DNN-OBS model shows similar performance for D+0 compared to the CMAQ model but decreased performance owing to an increased RMSE of D+1 and D+2 by 2 μgm^{-3} and 2.2 $\mu gm^{-3,}$ respectively. The DNN-OPM model shows an increase in performance owing to a decrease in the RMSE of D+0 by 3 μgm^{-3} compared to the CMAQ model. Moreover, the RMSE of D+1 and D+2 decrease by 0.4 μgm^{-3} and 0.4 μgm^{-3} compared to the CMAQ model, respectively, indicating that the performance is similar. For the DNN-ALL model, the RMSE from D+0 to D+2 decreased by 4.6 μgm^{-3}, 2.7 μgm^{-3}, and 2.1 μgm^{-3}, compared to the CMAQ model, which shows an improved performance.

Table 5~~3~~. Statistical evaluation results of CMAQ, DNN-OBS, DNN-OPM, and DNN-ALL models from 2016 to 2018.

Model	Day	MSE (($\mu gm^{-3})^2$)	RMSE (μgm^{-3})	R	IOA
CMAQ	D+0	136.9	11.7	0.76	0.86
	D+1	146.4	12.1	0.74	0.84
	D+2	185.0	13.6	0.67	0.80
DNN-OBS	D+0	79.2	8.9	0.78	0.87
	D+1	139.2	11.8	0.54	0.65
	D+2	158.8	12.6	0.43	0.54
DNN-OPM	D+0	53.3	7.3	0.86	0.92
	D+1	88.4	9.4	0.75	0.83
	D+2	108.2	10.4	0.68	0.77
DNN-ALL	D+0	39.7	6.3	0.90	0.94
	D+1	57.8	7.6	0.84	0.90
	D+2	72.3	8.5	0.80	0.87

Table 6~~4~~. Statistical evaluation results of CMAQ, DNN-OBS, DNN-OPM, and DNN-ALL models for 2019.

Model	Day	MSE (($\mu gm^{-3})^2$)	RMSE (μgm^{-3})	R	IOA
CMAQ	D+0	123.2	11.1	0.82	0.90
	D+1	132.3	11.5	0.80	0.89
	D+2	156.3	12.5	0.75	0.86
DNN-OBS	D+0	92.2	9.6	0.84	0.88
	D+1	182.3	13.5	0.63	0.65
	D+2	216.1	14.7	0.52	0.52
DNN-OPM	D+0	65.6	8.1	0.89	0.92
	D+1	123.2	11.1	0.78	0.81
	D+2	166.4	12.9	0.66	0.72

DNN-ALL	D+0	42.3	6.5	0.93	0.95
	D+1	77.4	8.8	0.88	0.90
	D+2	108.2	10.4	0.81	0.84

Q8. In Table 5: The DNN-ALL model uses the forecast variable (F_PM2.5 predicted by CMAQ). However, IOA of F_PM2.5 in CMAQ is 0.9, 0.9 and 0.85 respectively, and IOA of PM2.5 in DNN-ALL is 0.95, 0.9 and 0.86 respectively. Could it be understood that compared with CMAQ, IOA in DNN-ALL model is improved by 0.05, 0.0 and 0.01 respectively? More explanation is needed.

A8. The denominator of IOA indicates the trends of the model and observation based on the average of observation, and the numerator of IOA represents the deviation of the model and observation. In other words, the IOA can be interpreted as an indicator that considers trends and quantitative differences. Therefore, the quantitative difference (error) of the DNN-ALL model decreases compared to the CMAQ model, but the trend toward the mean of observation is similar between the two models, showing no significant difference in IOA.

Q9. In Table 6: From T04 to T11, why does the indicators (RMSE and IOA) not decrease monotonically? The IOA of T09 is larger than T04. Meanwhile, the mean IOA of D+2 is 0.79 ((0.77+0.85+0.74+0.80)/4.0) and IOA of D+2 in table 5 is 0.86, What are the reasons for the unequal values?

A9. There could be a difference in the performance of the model according to the conditions of target time such as daytime, nighttime, high concentration, and low concentration. As shown in the CMAQ results, the prediction performance of the DNN-ALL model degrades or improves monotonically over time.

The IOA of D+2 is not equal to (0.77+0.85+0.74+0.80)/4.0. The IOA of D+2 refers to the value calculated using the IOA method after calculating the daily average concentration using the predicted concentration of each T-step such as T08, T09, T10, and T11. (0.77+0.85+0.74+0.80)/4.0 is simply averaged after calculating IOA using the predicted concentration by T-step.

Section 2. Minor comments.

Q1. Line 19: “IOA” should be "index of agreement (IOA)". The first abbreviation needs to give the complete name. Please check other parts of the paper.

A1. We thank the reviewer for highlighting this issue. We have included the complete name at the first mention of the abbreviation.

Q2. Line 100 Figure 3: Nested-grid is often used in models. It is recommended to combine figure 2 and figure 3 into one figure.

A2. Figure 2 shows the location information of the measuring station where the measurement data are obtained, and Figure 3 depicts the domain of the model. Therefore, the information conveyed by the two images is different.

Q3. Add the temporal and spatial resolution of the variables in Tables 1 and 2.

A3. Based on the suggestion of the reviewer, the descriptions are added to Table 5 and Table 6.

Table 7~~5~~. Training variables in the $PM_{2.5}$ prediction system using a DNN based on surface-weather observations. Air quality variables are obtained from 41 air quality measurement equipment in Seoul. Surface weather variables are obtained from ASOS in Seoul. Observation data are collected every hour.

Observation Variable	Description	Unit
O_SO_2	Sulfur dioxide	ppm
O_NO_2	Nitrogen dioxide	ppm
O_O_3	Ozone	ppm
O_CO	Carbon monoxide	ppm
O_PM_{10}	Particulate matter (aerodynamic diameters ≤ 10 μm)	μgm^{-3}
O_$PM_{2.5}$	Particulate matter (aerodynamic diameters ≤ 2.5 μm)	μgm^{-3}
O_V	Vertical wind velocity	m/s
O_U	Horizontal wind velocity	m/s
O_RN_ACC	Accumulative precipitation	Mm
O_RH	Relative humidity	%
O_Td	Dew point temperature	℃
O_Pa	Pressure	hPa
O_Radiation	Solar radiation	0.01 MJ per hr-m^3
O_Ta	Air temperature	℃

Table 8~~6~~. Training variables in the $PM_{2.5}$ prediction system using a DNN based on the WRF and CMAQ models. WRF and

CMAQ model results are obtained from 9 km horizontal grid resolution. These values are collected on an hourly interval.

Model	Forecast Variable	Description	Unit
CMAQ	$F\_PM_{2.5}$	Particulate matter (aerodynamic diameter ≤ 2.5 μm)	μgm^{-3}
WRF	F_V	Vertical wind velocity at surface	m/s
	F_U	Horizontal wind velocity at surface	m/s
	F_RN_ACC	Accumulative precipitation	mm
	F_RH	Relative humidity at surface	%
	F_Pa	Pressure at surface	pa
	F_Ta	Air temperature at surface	K
	F_MH	Mixing height	m
	F_925hpa_gpm	Position altitude at 925 hPa	m
	F_925hpa_V	Vertical wind velocity at 925 hPa	m/s
	F_925hpa_U	Horizontal wind velocity at 925 hPa	m/s
	F_850hpa_gpm	Position altitude at 850 hPa	m
	F_850hpa_V	Vertical wind velocity at 850 hPa	m/s
	F_850hpa_U	Horizontal wind velocity at 850 hPa	m/s
	F_850hpa_RH	Relative humidity at 850 hPa	%
	F_850hpa_Ta	Potential temperature at 850 hPa	Θ
	F_Temp_ 850hpa-925hpa	Potential temperature difference between 850 hPa and 925 hPa	Θ

Q4. The font in Figure 4 is too small to see clearly.

A4. We have increased the font size to improve the clarity of the figure.

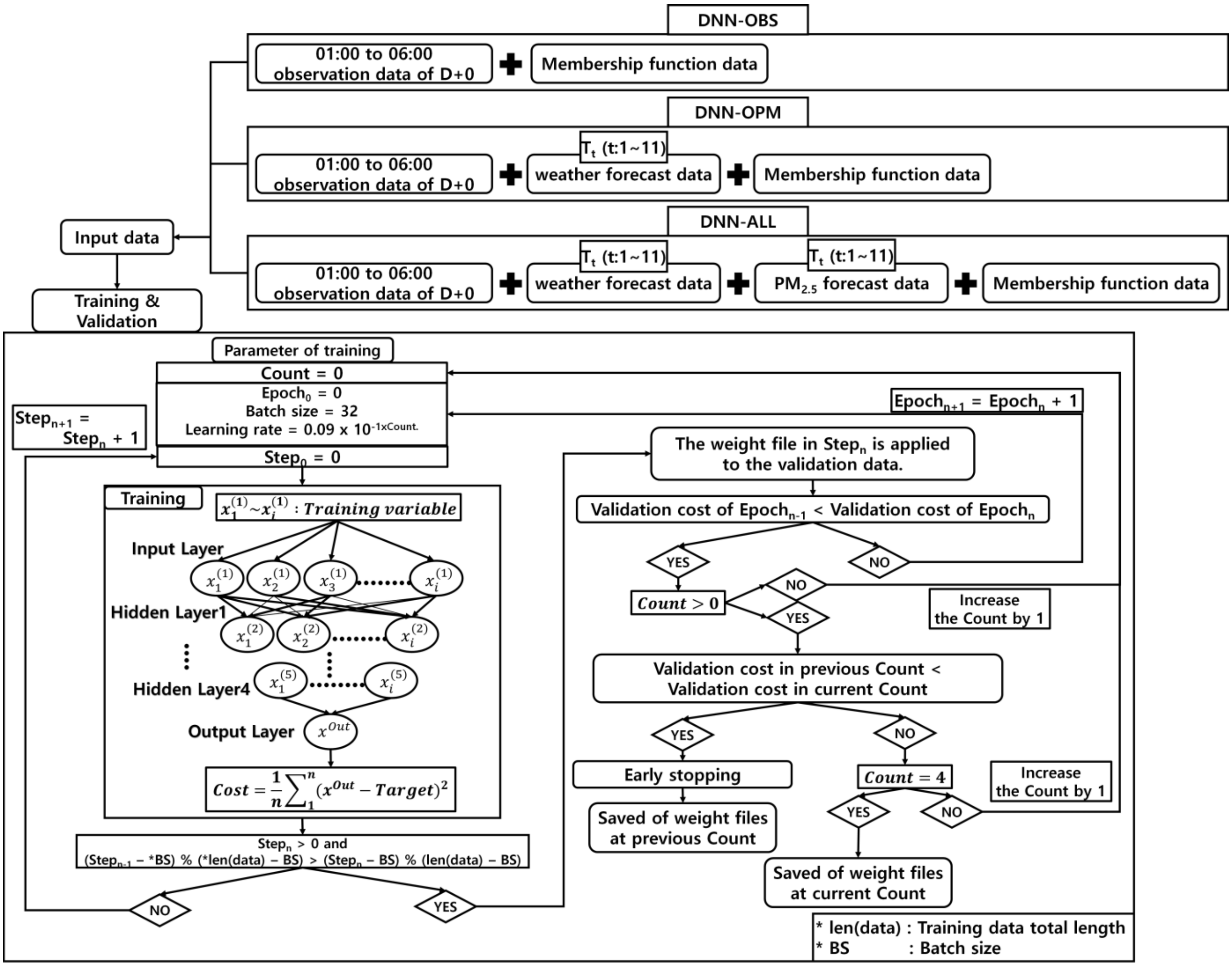

Figure 2. Structure of DNN model training process.

Q5. It is suggested to add the content between DNN model and other machine learning models (https://doi.org/10.5194/amt-14-5333-2021, https://doi.org/10.1016/j.scitotenv.2021.150338).

A5. Based on the paper mentioned by the reviewer, a comparative evaluation is performed between the DNN-ALL model and the Random Forest (RF) model, which is a machine learning model. Tables 7 and 8 show the results of the statistical evaluations and that of the AQI evaluation, respectively. The RMSE value of the DNN-ALL model decreased from 0.6 to 1.9 μgm^{-3} compared to the RF model and the R and IOA values increased slightly. The ACC of the DNN-ALL model increased by about 2 to 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. From the results, it is observed that the DNN-ALL model showed a superior performance compared to the RF model. The machine learning method was selected according to the scalability of the model for future data growth and 1-h forecast time segmentation.

Table 9~~7~~. 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 10~~8~~. 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