Development of $PM_{2.5}$ short-term forecast model using Artificial Intelligence – Focused on Seoul.

Jeong-Beom Lee[1)], Geon-Woo Yun[1)], Youn-Seo Koo[1)], Hui-Young Yun[1)], Dae-Ryun Choi[1)], Ji-Seok Koo[2)]

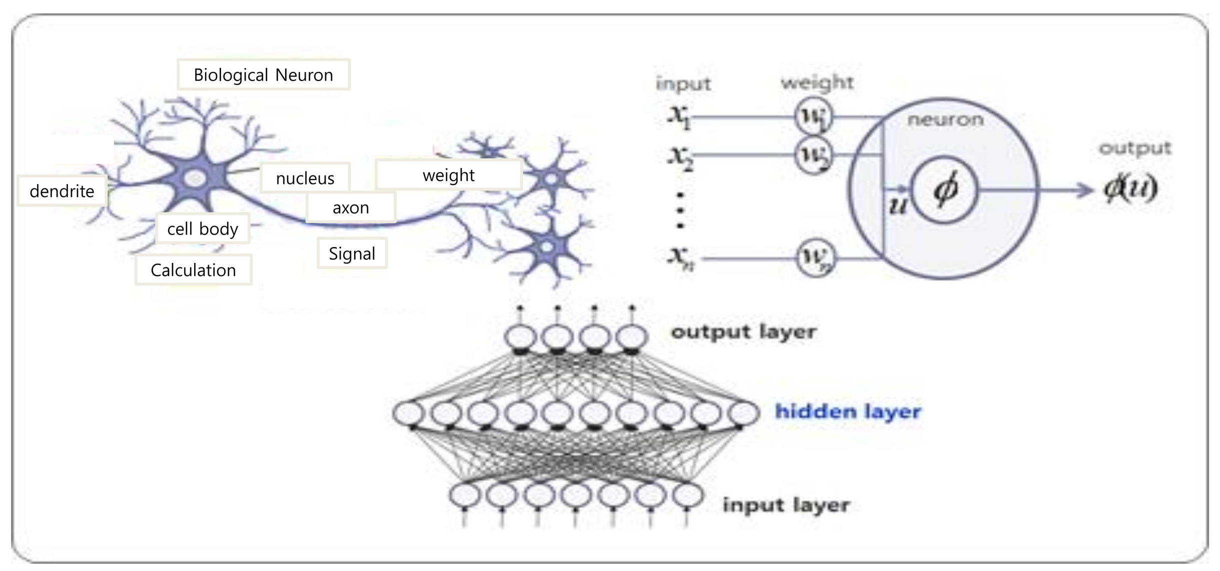
[1)]Dept. of Environmental & Energy Eng., Anyang University, Anyang, Korea, [2)]Enitech Co.,Ltd

Introduction

- Forecasting model performance for $PM_{2.5}$ using chemical transport model is often overestimated compared to the measurements(Koo et al, 2008;2012;2015, Choi et al, 2018;2019) in Korea.
- In odor to improve model performance for $PM_{2.5}$ forecasting, we developed PM forecasting system using artificial intelligence with big data such as air quality and weather observations as well as forecasting model data.
- It is important to number of high concentration of $PM_{2.5}$ data to accurately predict the episode. However, the number of high concentration of $PM_{2.5}$ data is insufficient. Therefore we created the data to improve model performance using AI for accuracy of high concentrations events of $PM_{2.5}$.
- We analyzed developed AI forecasting $PM_{2.5}$ model performance for 3-days in Seoul.

Methodology

- Deep Neural Network
  - The artificial intelligence technique used DNN(Deep Neural Network).
  - DNN is an extended model that includes multiple hidden layers between the input and output layers to enable deep learning in existing ANN.
  - The calculation of weights and biases between layers is key.

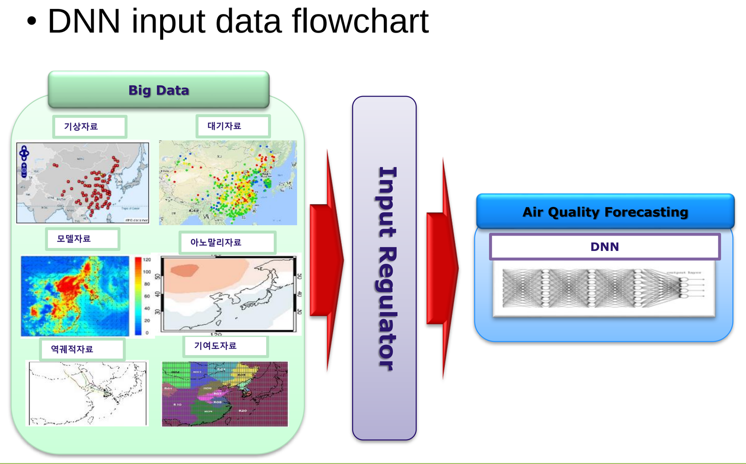


- DNN input data
  - Observed data is meteorological and six air pollutants(PM_{10}, $PM_{2.5}$, O_3, NO_2, SO_2, CO).
  - Numerical model value, WRF weather forecasts, Anomaly, Cosine similarity, Back trajectory, Contribution, Julian day were used.
  - The input data used for learning is normalized.

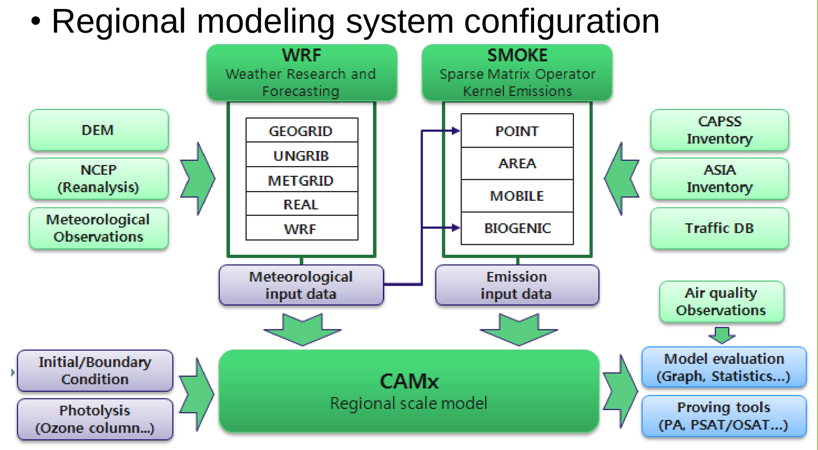
•Normalization equation

$$If \cdot \min(X_{raw}) < 0$$
$$X_{norm} = (X_{raw} - \min(X_{raw}))/(\max(X_{raw}) - \min(X_{raw})) - 0.5$$
$$If \cdot \min(X_{raw})\ is \geq 0$$
$$X_{norm} = (X_{raw} - \min(X_{raw}))/(\max(X_{raw}) - \min(X_{raw}))$$

• DNN input data flowchart

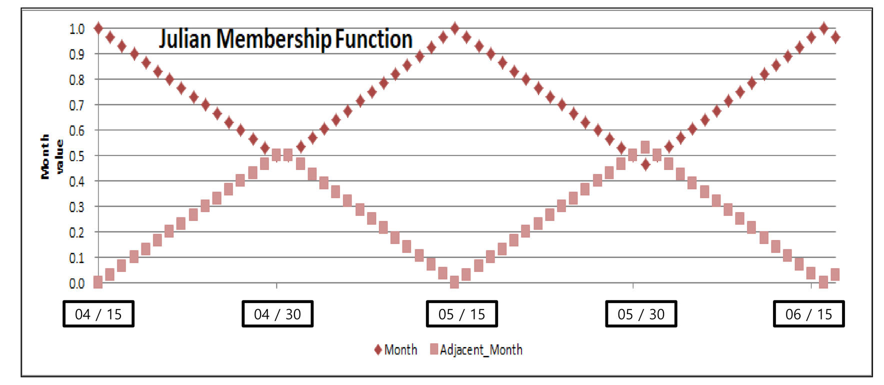


• Regional modeling system configuration



- Fuzzy theory to Julian day data (Jeon et al., 2018)

1. Adjacent Month definition
   - Ⓐ if(Day = 15) then Adjacent_Month is not exsite
   - Ⓑ else-if(Month = 1 and Day < 15) then Adjacent_Month = 12
   - Ⓒ else-if(Month = 12 and Day > 15) then Adjacent_Month = 1
   - Ⓓ else-if(Day < 15) then Adjacent Month = Month − 1
   - Ⓔ else-if(Day > 15) then Adjacent Month = Month + 1
2. Month value definition
   - Ⓐ if(Day < 15) then Month_Value= $\frac{1}{28} \times Day + \frac{13}{28}$
   - Ⓑ else-if(Day > 15) then Month_Value= $-\frac{1}{30} \times Day + \frac{3}{2}$
   - Ⓒ Day = 15, Month_value=1
3. Adjacent_Month_value definition
   - Adjacnet_Month_value = 1 − Month_value

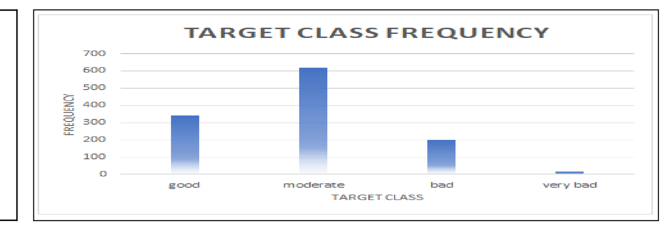


- Data Creation method
  - Most observed $PM_{2.5}$ data are Good and Moderate.
  - Learning is not done properly because the data is not balance.
  - Therefore, data creation method is implemented to create sufficient high concentration of $PM_{2.5}$ data.

$X_{new} = X_{old} + (X_{old} * 0.1 * uniform(x,y))$

X_{new} : Created Data. X_{old} : Original Data.

uniform(x,y) : random number from x to y



- Model performance evaluation method
  - Change the concentration value of $PM_{2.5}$ to the index value.
  - Reference : http://www.airkorea.or.kr/

• Index assessment

Column		Forecatsed			
		Good	Moderate	Bad	Very Bad
OBS	Good	a1	b1	c1	d1
	Moderate	a2	b2	c2	d2
	Bad	a3	b3	c3	d3
	Very Bad	a4	b4	c4	d4

Ⅰ: , Ⅱ: , Ⅲ: , Ⅳ:

Method	Equation
Accuracy (A)	$= \frac{(a1+b2+c3+d4)}{N} \times 100(\%)$
HIT rate (HIT)	$= \frac{(c3+d4)}{(III+IV)} \times 100(\%)$
Probability of Detection (POD)	$= \frac{IV}{(III+IV)} \times 100(\%)$
False Alarm Rate (FAR)	$= \frac{(II)}{(II+IV)} \times 100(\%)$

• Statistic assessment

$$MBIAS = \frac{1}{N}\sum_{1}^{N}(Model - Obs)$$
$$NMB = \frac{\sum_{1}^{N}(Model - Obs)}{\sum_{1}^{N}Obs} \times 100$$
$$IOA = 1 - \frac{\sum_{1}^{N}(Model - Obs)^2}{\sum_{1}^{N}(|Model - \overline{Obs}| + |Obs - \overline{Obs}|)^2}$$
$$R = \frac{\sum(Model - \overline{Model}) \times (Obs - \overline{Obs})}{\sqrt{\sum(Model - \overline{Model})^2 \times \sum(Obs - \overline{Obs})^2}}$$

Results and Discussion

- Results of model performance evaluation (not create data (Standard model))

		Observed $PM_{2.5}$ vs AI $PM_{2.5}$ Index Assessment				Observed $PM_{2.5}$ vs AI $PM_{2.5}$ Statistic Assessment			
Model name	Day	ACC	HIT	POD	FAR	MB	NMB	IOA	R
Standard	D+0	72.22	63.47	73.17	21.05	-0.09	-2.71	0.93	0.91
	D+1	71.43	52.27	70.45	16.22	-2.96	-8.57	0.86	0.90
	D+2	65.87	51.16	69.77	28.57	-2.66	-7.73	0.82	0.85

- Results of model performance evaluation (Create data(X_{old} = Julian day))

		Observed $PM_{2.5}$ vs AI $PM_{2.5}$ Index Assessment				Observed $PM_{2.5}$ vs AI $PM_{2.5}$ Statistic Assessment			
Model name	Day	ACC	HIT	POD	FAR	MB	NMB	IOA	R
Create Data (Standard value Julian day)	D+0	70.63	65.85	78.05	20.00	0.48	1.43	0.94	0.90
	D+1	64.29	63.64	81.82	33.33	0.81	2.37	0.83	0.85
	D+2	57.14	60.47	79.07	43.33	1.21	3.52	0.79	0.77

- Results of model performance evaluation (Create data(X_{old}=numerical_$PM_{2.5}$))

		Observed $PM_{2.5}$ vs AI $PM_{2.5}$ Index Assessment				Observed $PM_{2.5}$ vs AI $PM_{2.5}$ Statistic Assessment			
Model name	Day	ACC	HIT	POD	FAR	MB	NMB	IOA	R
Create Data (Standard value 'Numerical $PM_{2.5}$')	D+0	73.02	68.29	82.93	22.73	0.29	0.89	0.91	0.89
	D+1	74.60	50.00	65.91	6.45	-4.07	-11.82	0.87	0.90
	D+2	66.67	46.51	65.12	24.32	-3.09	-8.99	0.82	0.83

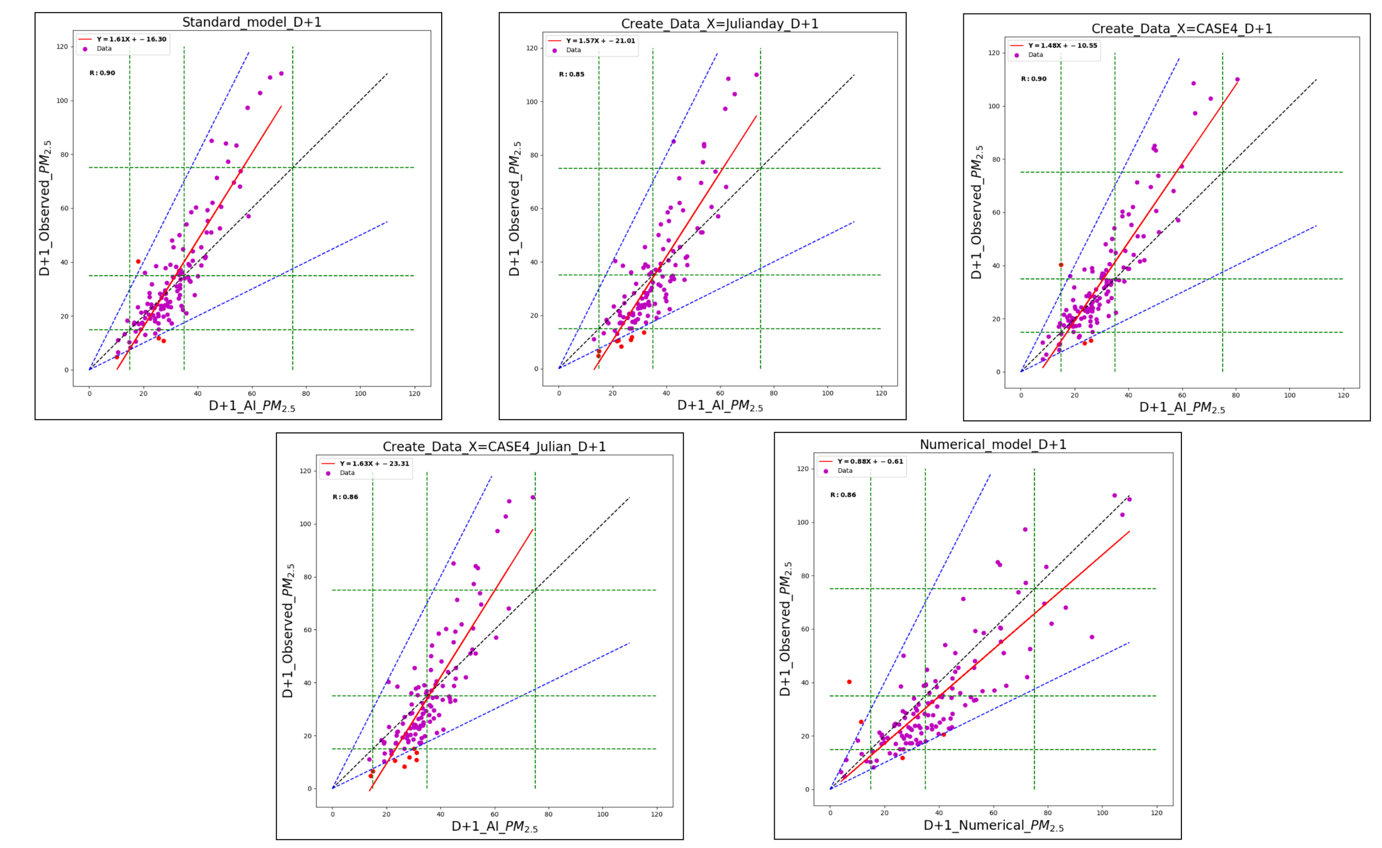
- Results of model performance evaluation (Create data)
  ($X1_{old}$ = Julian day // $X2_{old}$ = Numerical $PM_{2.5}$)

		Observed $PM_{2.5}$ vs AI $PM_{2.5}$ Index Assessment				Observed $PM_{2.5}$ vs AI $PM_{2.5}$ Statistic Assessment			
Model name	Day	ACC	HIT	POD	FAR	MB	NMB	IOA	R
Create Data (Standard value 'Numerical $PM_{2.5}$'&Julian day)	D+0	70.63	75.61	82.93	24.44	2.60	7.84	0.93	0.88
	D+1	65.08	61.36	79.55	31.37	0.89	2.58	0.83	0.86
	D+2	59.52	60.47	79.07	39.29	0.68	1.98	0.81	0.80

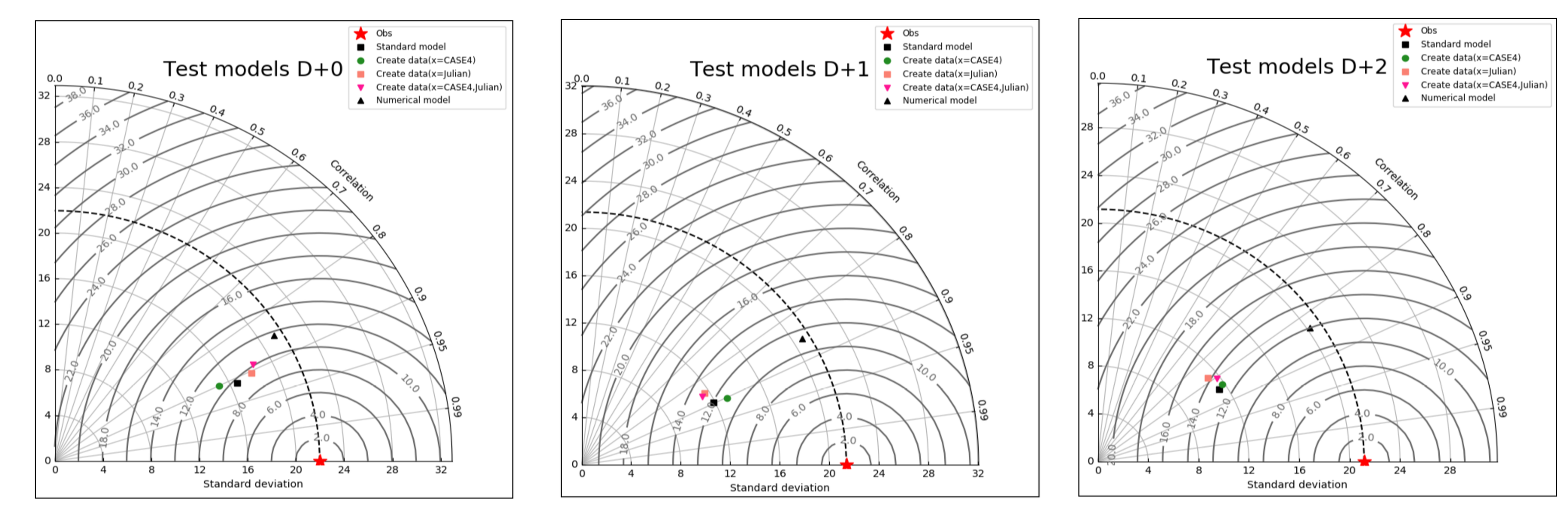
- Results of model performance evaluation compared to numerical model

		Observed $PM_{2.5}$ vs Numerical $PM_{2.5}$ Index Assessment				Observed $PM_{2.5}$ vs Numerical $PM_{2.5}$ Statistic Assessment			
Model name	Day	ACC	HIT	POD	FAR	MB	NMB	IOA	R
Numerical Model	D+0	64.29	68.29	85.37	33.96	3.84	11.58	0.92	0.86
	D+1	61.11	68.18	86.36	39.68	5.23	15.18	0.91	0.86
	D+2	60.32	65.12	79.07	43.33	4.60	13.37	0.90	0.83

- Comparison of correlation graphs in test models – Focused D+1.



- Comparison of Taylor-diagram



Conclusions

- AI results showed ACC increased and FAR decreased compared with numerical mode because AI tend to reduce overestimation of $PM_{2.5}$ of a numerical model.
- In D+0, the POD index of AI models with created high concentration of $PM_{2.5}$ events data is increased and ACC and FAR are similar compared with standard model.
- In D+1, the POD and FAR index of AI models with created high concentration of $PM_{2.5}$ events data using Julian-day or Julian-day&Numerical-$PM_{2.5}$ are increased, but ACC is decreased compared with standard model.
- The results of comparisons in various aspects in this study suggest that developed AI forecast model is able to replace numerical model for air quality $PM_{2.5}$ forecasting in Seoul.
- We believe further studies with development of data created method are necessary to improve performance of AI model.

Acknowledgements

This subject is supported by the National Institute of Environmental Research.