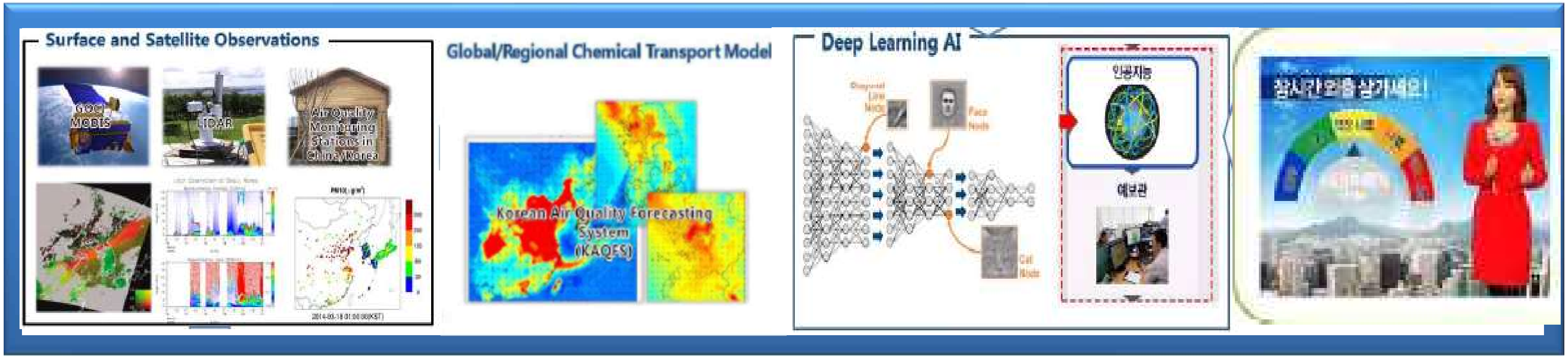
A development of PM2.5 forecasting system

using physicochemical models and ensemble machine learning

2021. 11.

Youn-Seo Koo
Department of Environmental and Energy Engineering
Anyang University



Research Outline

Research Team

- Anyang Uni., Environmental Eng., Prof. Youn-Seo Koo
  : Chemical Transport Model (CTM), Big Data(BD), Test-bed running
- Anyang Uni., Computer Eng., Prof. Hee-Yong Kwon
  : Development of Machine Learning algorithm and optimization (DNN, CNN)
- SNU, Earth and Environmental Sciences, Prof. Chang-Hoi Ho
  : Cluster analysis of BD, RNN development and optimization
- NextSoft Co., Dr. Hee-Cheol Kwon
  : Development of AI forecasting system

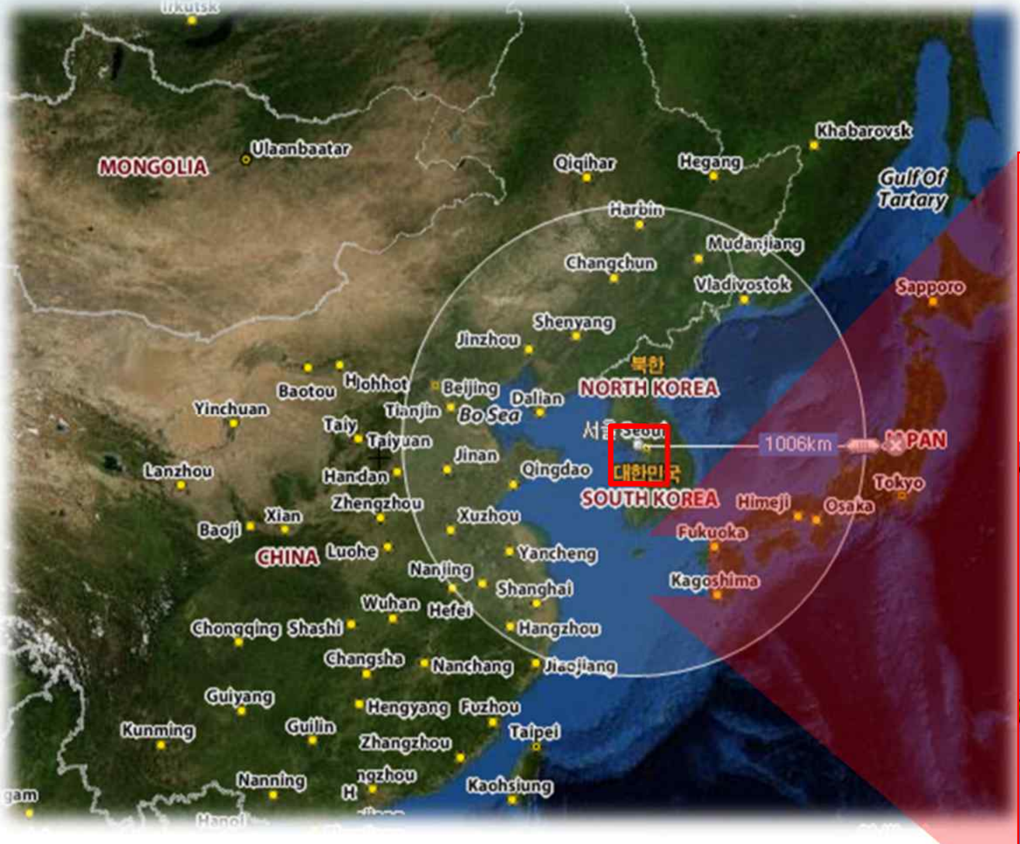
Funded by NIER

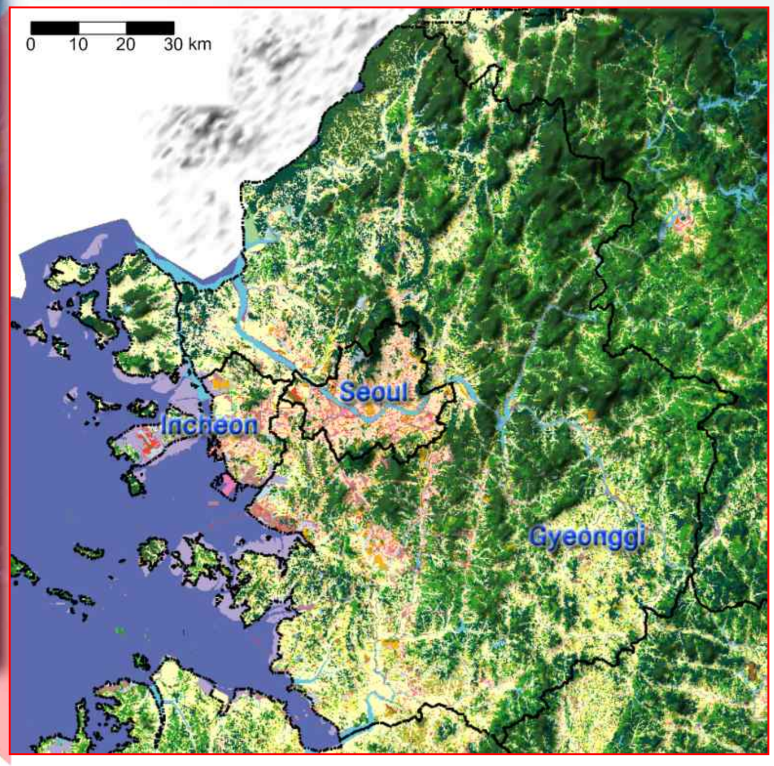
: A Development of Short-term Prediction Tool for PM_{10} and $PM_{2.5}$ Concentrations using Artificial Intelligence(AI)

Contents

1. Introduction

- The Northeast Asian region may be in the same air basin
- The air quality in Korea : long-range transport + local emissions
- Big data and ML algorithms reflecting the long-range transport as well as locally characteristics





1. Introduction

Visions

- Accuracy to meet the public needs to protect health in advance

Target

- Performance for PM2.5 forecast
  - Final : Accuracy : 90%, POD : 80% , FAR : 30% or less
- AI forecast that communicates with the public
  : explanation of the cause of high concentration by case classification

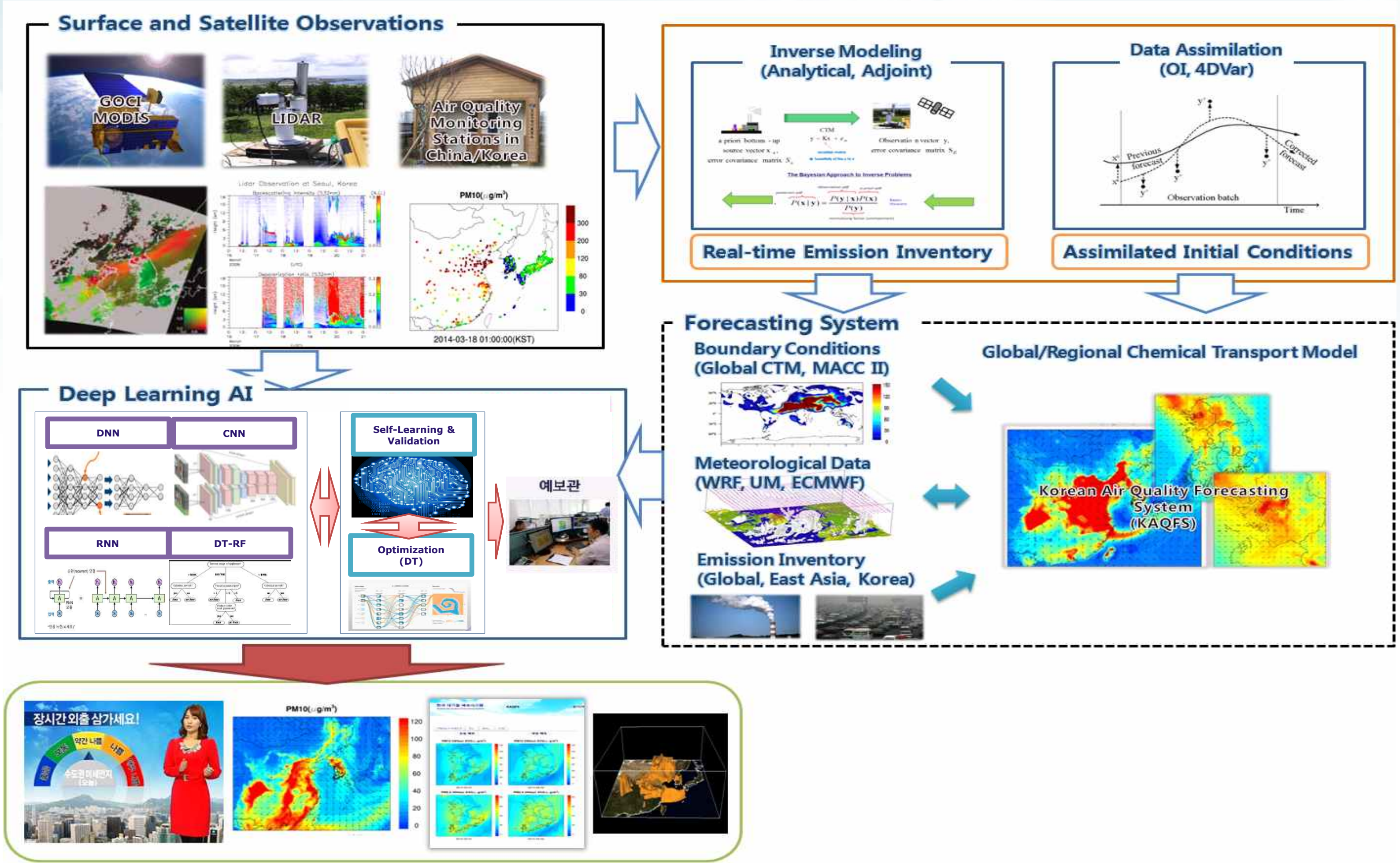
Why ML ?

- CTM forecast limitation : Uncertainties in input data & CTM itself
- Forecaster's conceptual forecast: subjective bias, cognitive limitations
- Recent measurements and forecast data become Big Data: Systematic processing and analysis are required

1. Introduction

Diagram of ML forecasting system in Korea

- **Development of ML forecasting system reflecting the long-range transport in Northeast Asia and local characteristics of PM.**

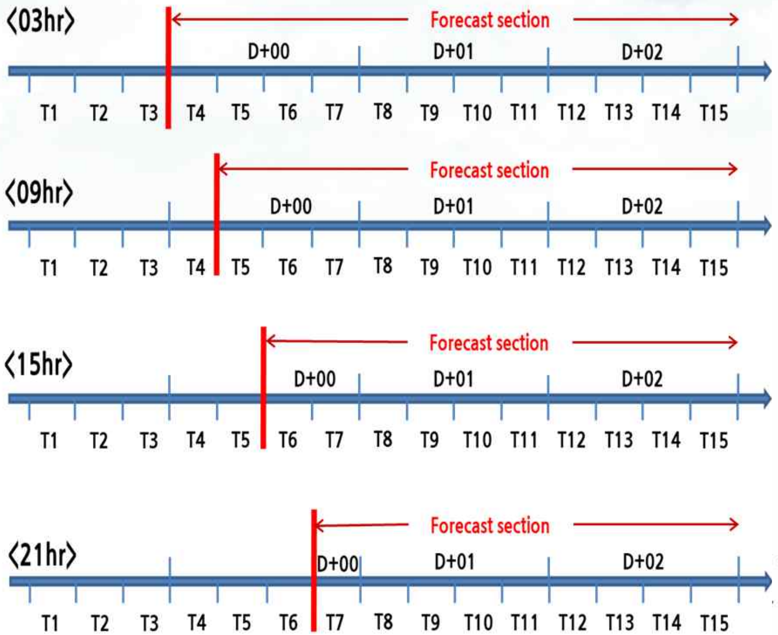


1. Introduction

Forecast time schedule and regions

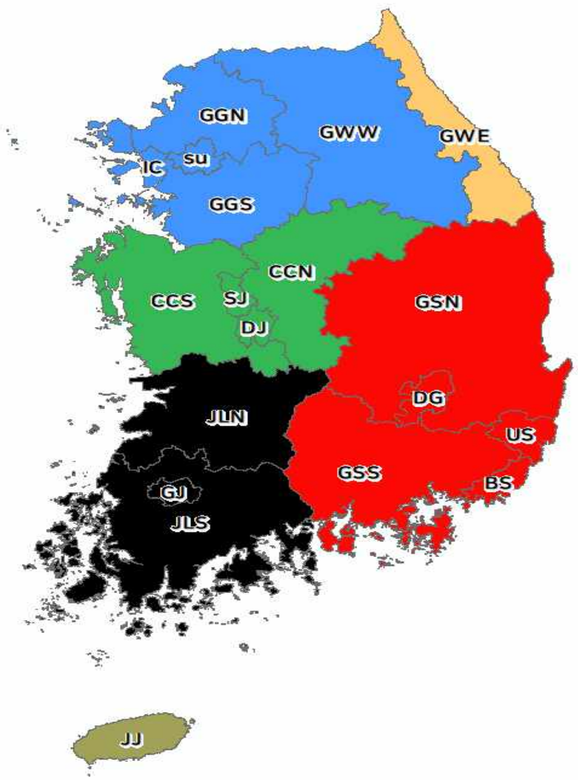
- 19 forecast regions with 6 cluster groups for cluster analysis
- Ti(i=1,15) : basic time interval for ML forecasting is 6 hours rather than 1 hour
- Forecast is 4 times a day for 3 days

<Forecast time schedule>



<19 forecast regions with 6 cluster groups>

(Seoul Metropolitan Area + Yeongseo,Chungcheong Area, Gyeongsang Area, Jeolla Area, Yeongdong Area, Jeju Area).



① South Gyeonggi, North Gyeonggi, Seoul, Yeongseo, Incheon
② Daejeon, Sejong, Chungnam, Chungbuk
③ Gyeongnam, Gyeongbuk, Daegu, Busan, Ulsan
④ Gwangju, Jeonnam, Jeonbuk
⑤ Yeongdong
⑥ Jeju

1. Introduction

Evaluation method

<A verification statistics used to evaluate category forecasts>

Observed \ Forecasted	No	Yes
No	a	b
Yes	c	d

PM2.5 (μg/m³): 35

Observed \ Forecasted	Good	Moderate	Bad	Very bad
Good	k	l	m	n
Moderate	o	p	q	r
Bad	s	t	u	v
Very bad	w	x	y	z

PM2.5 (μg/m³): 15, 35, 75

- **Accuracy(ACC, %))**
  = (k+p+u+z)/N*100
- **Probability of Detection(POD, %)**
  = d/(c+d) *100
- **False Alarm Rate(FAR, %)**
  = b/(b+d) *100

<Statistical Analysis Method>

Index Of Agreement(IOA)

$$= 1 - \frac{\sum_1^n (Model - Obs)^2}{\sum_1^n \left(\left| Model - \overline{Obs} \right| - \left| Obs - \overline{Obs} \right| \right)^2}$$

Normalized Mean Bias(NMB)

$$= \frac{\sum_1^n (Model - Obs)}{\sum_1^n Obs} \times 100$$

- **Root Mean Square Error(RMSE)**

$$= \sqrt{\frac{1}{N} \sum_1^n (Model - Obs)^2}$$

2. Primary and secondary big data

Primary BD – Air quality and weather monitoring data

- Used for CMAQ Data Assimilation
- 2 D. concentration map by kriging the point data for CNN input

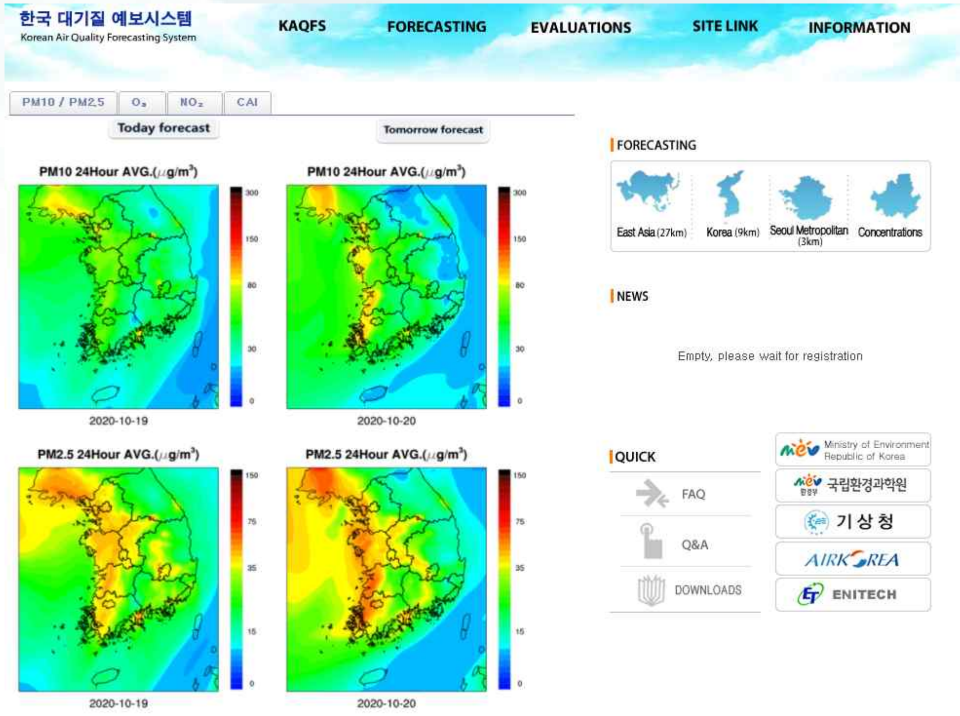
<Air quality monitoring stations>

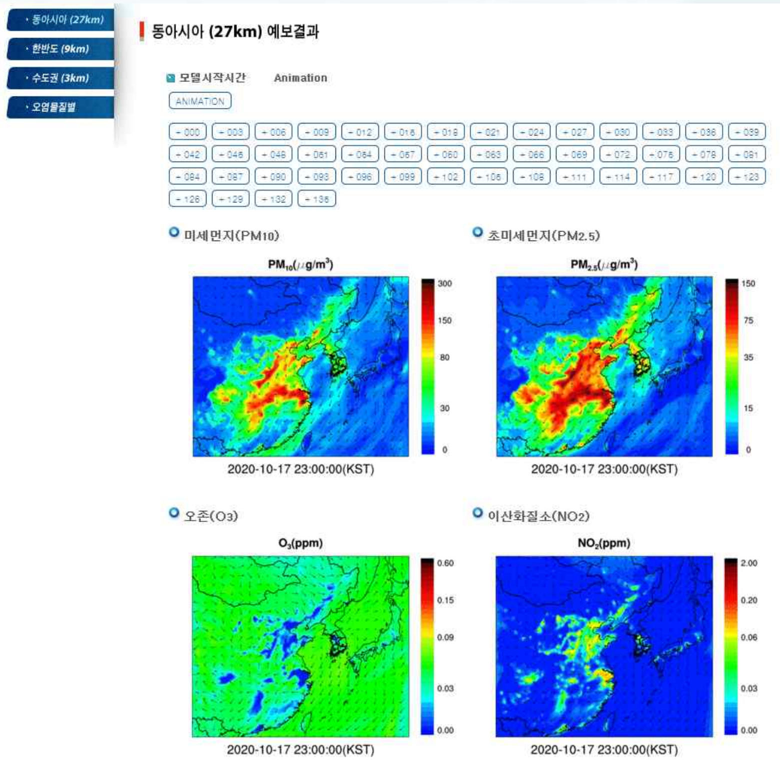
<Weather monitoring stations>

2. Primary and secondary big data

Primary BD– Korean Air Quality Forecasting System(KAQFS)

- CASE4 (CMAQ) : WRF V3.6.1 with GFS, CMAQ v4.7.1
- Asia emission inventory : MEIC (2010), REAS (2008), Korea : CAPSS (2011)
- Data Assimilations with surface air quality measurements in China and Korea
- It detects PM episode in advance and open to the public (www.kaq.or.kr) since 2007

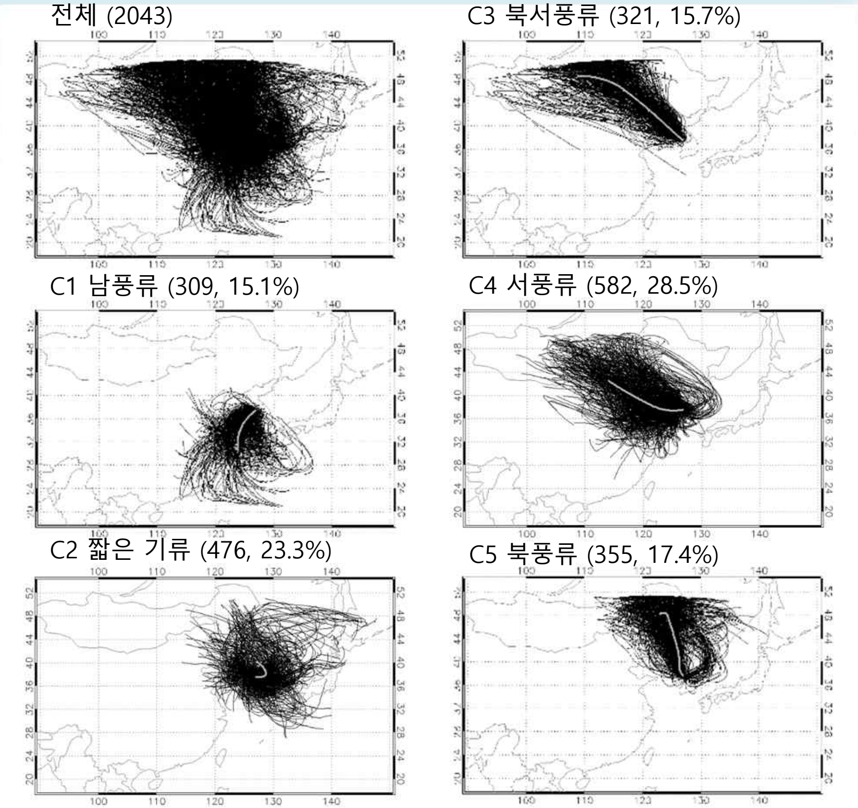




2. Primary and secondary big data

Secondary BD – Back trajectory and Cluster analysis

- clustering analysis by back trajectories
- C4 and C3 are representing the long-range transport from China to Korea



No. of days for PM episode

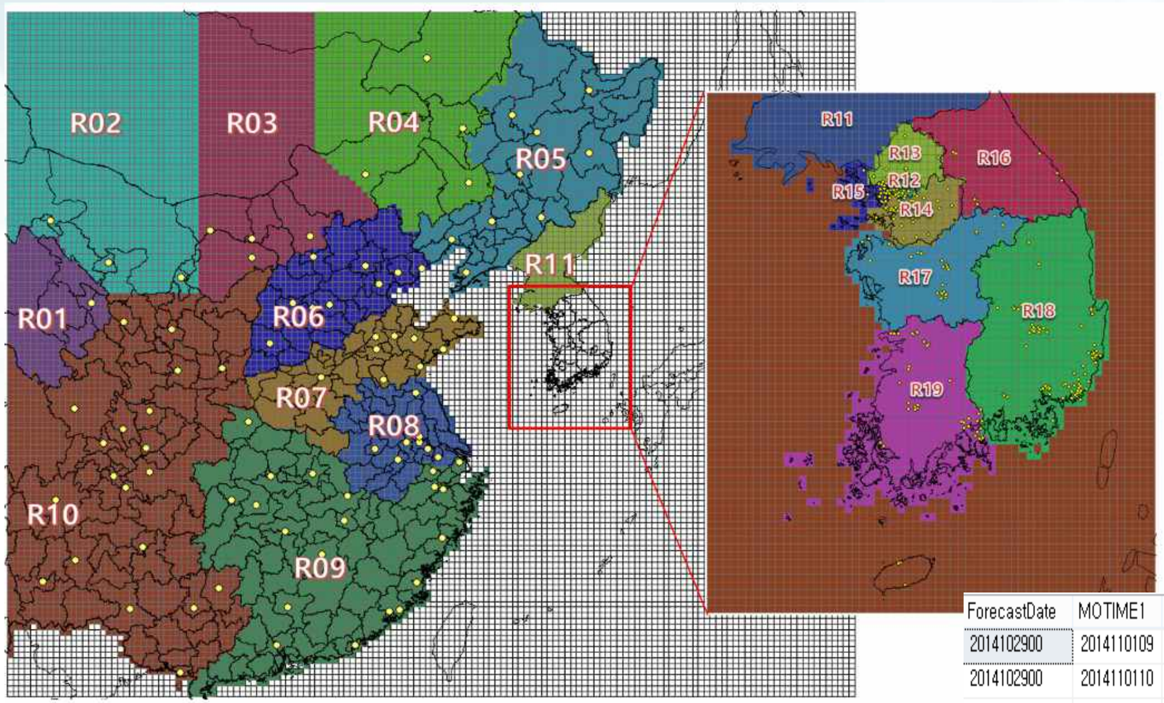
	C1	C2	C3	C4	C5
PM episode days	28	16	37	86	29
Yellow sand days	8	1	23	6	9

<Air quality monitoring stations>

2. Primary and secondary big data

Secondary BD – Regional Contributions

- **Model : CAMx PSAT, 20 source regions including China and Korea**
- **Regional emission source (R1-R20) contribution on 19 forecasting regions**



ForecastDate	MOTIME1	Pollutant	Region	BC	IC	R01	R02	R03	R04	R05	R06	R0
2014102900	2014110109	PM10	12	27,10771	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,
2014102900	2014110110	PM10	12	22,20646	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,
2014102900	2014110111	PM10	12	20,47342	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,
2014102900	2014110112	PM10	12	17,97756	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,
2014102900	2014110109	PM2,5	12	21,15480	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,
2014102900	2014110110	PM2,5	12	17,75178	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,
2014102900	2014110111	PM2,5	12	16,34840	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,
2014102900	2014110112	PM2,5	12	14,29166	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,
2014102912	2014110109	PM10	12	27,10771	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,
2014102912	2014110110	PM10	12	22,20646	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,

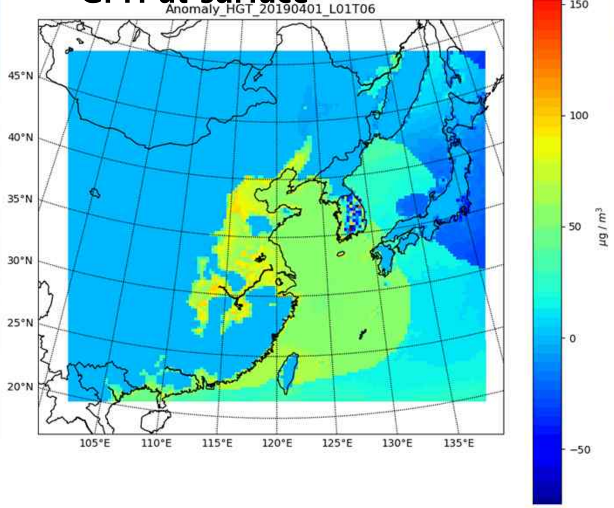
<Source regions and contribution data>

2. Primary and secondary big data

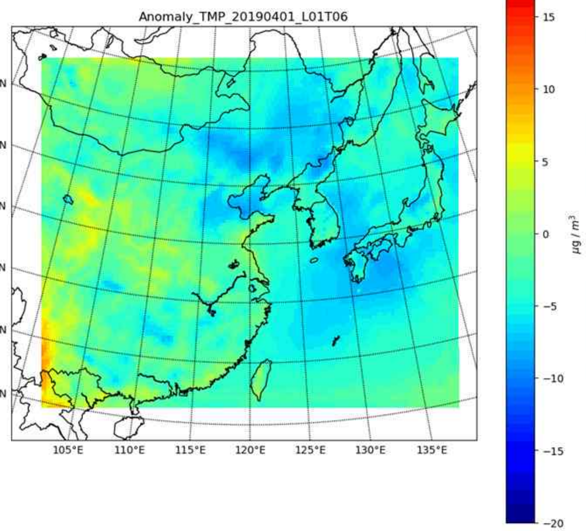
Secondary BD – Anomaly

- **WRF model predictions**

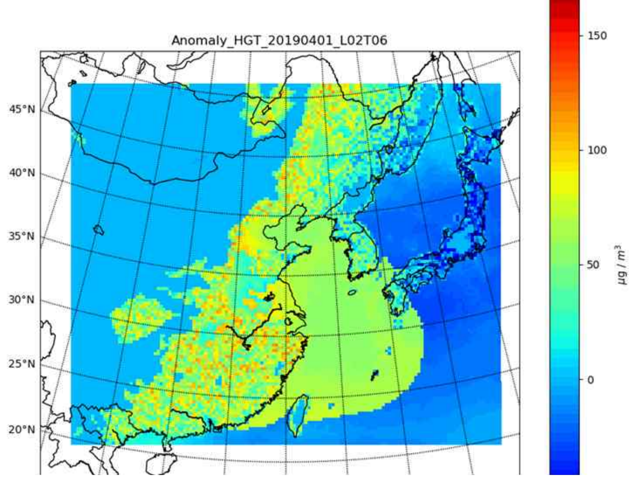
GPH at surface



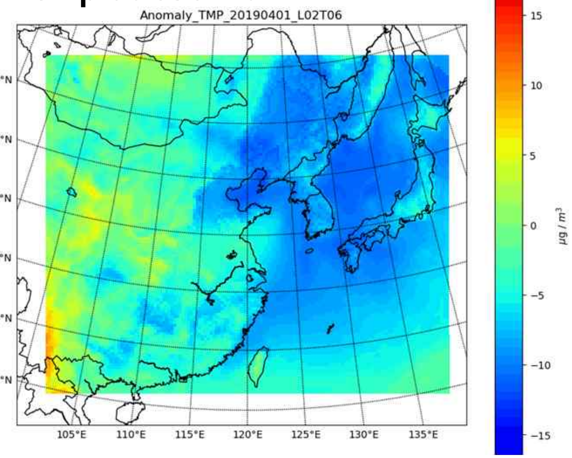
Temp at surface



GPH at 850hPa



Temp at 850hPa



2. Primary and secondary big data

Secondary BD – Cosine similarity

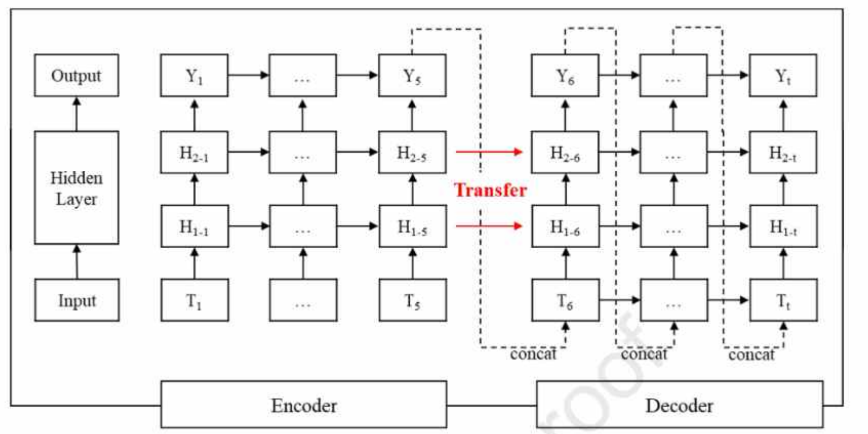
- $\cos^t$ using three-month climatology and anomaly patterns.

$$\cos^t = \frac{A_i^t \cdot A_i^{clm-high}}{\| A_i^t \| \cdot \| A_i^{clm-high} \|}$$

인자		원시값					정규화 변수		비고
		최소값	최대값	평균값	중앙값	표준편차	최소값	최대값	
유사도	CS_TA_1000	-0.89	0.88	0.00	0.00	0.42	-0.80	0.80	선택인자
	CS_GEOH_1000	-0.64	0.75	0.00	0.00	0.23	-0.60	0.60	
	CS_RH_1000	-0.88	0.90	0.03	0.04	0.39	-0.80	0.80	선택인자
	CS_U_1000	-0.64	0.72	-0.01	-0.01	0.21	-0.60	0.60	
	CS_V_1000	-0.72	0.79	-0.01	0.00	0.23	-0.60	0.60	
	CS_W_1000	-0.70	0.65	0.00	0.00	0.18	-0.60	0.60	
	CS_TA_925	-0.77	0.77	-0.01	-0.01	0.28	-0.80	0.80	선택인자
	CS_GEOH_925	-0.60	0.63	0.00	0.00	0.20	-0.60	0.60	
	CS_RH_925	-0.89	0.89	0.02	0.04	0.36	-0.80	0.80	선택인자
	CS_U_925	-0.66	0.78	-0.01	0.00	0.21	-0.60	0.60	
	CS_V_925	-0.76	0.77	0.00	0.00	0.24	-0.60	0.60	
	CS_W_925	-0.52	0.58	0.00	0.00	0.13	-0.40	0.40	
	CS_TA_850	-0.63	0.68	0.00	0.00	0.20	-0.60	0.60	선택인자
	CS_GEOH_850	-0.52	0.57	0.00	0.00	0.17	-0.60	0.60	
	CS_RH_850	-0.87	0.86	0.02	0.04	0.33	-0.80	0.80	선택인자
	CS_U_850	-0.70	0.73	0.00	0.00	0.22	-0.60	0.60	
	CS_V_850	-0.82	0.78	0.00	0.01	0.27	-0.80	0.80	
	CS_W_850	-0.42	0.53	0.00	0.00	0.10	-0.40	0.40	
	CS_TA_700	-0.63	0.60	0.02	0.02	0.17	-0.60	0.60	
	CS_GEOH_700	-0.61	0.59	0.00	0.00	0.18	-0.60	0.60	
	CS_RH_700	-0.85	0.85	0.01	0.02	0.31	-0.80	0.80	
	CS_U_700	-0.65	0.73	0.00	0.00	0.24	-0.60	0.60	
	CS_V_700	-0.84	0.81	0.01	0.01	0.30	-0.80	0.80	
	CS_W_700	-0.32	0.46	0.00	0.00	0.08	-0.30	0.30	
	CS_TA_500	-0.52	0.60	0.00	0.00	0.17	-0.60	0.60	
	CS_GEOH_500	-0.61	0.59	0.00	0.00	0.19	-0.60	0.60	선택인자
	CS_RH_500	-0.87	0.86	0.01	0.01	0.30	-0.80	0.80	
	CS_U_500	-0.67	0.67	0.00	0.01	0.24	-0.70	0.70	
	CS_V_500	-0.86	0.84	0.01	0.01	0.30	-0.80	0.80	
	CS_W_500	-0.34	0.48	0.00	0.00	0.09	-0.40	0.40	
	CS_TA_300	-0.74	0.73	0.01	0.01	0.23	-0.70	0.70	
	CS_GEOH_300	-0.57	0.54	0.00	-0.01	0.16	-0.60	0.60	선택인자
	CS_RH_300	-0.91	0.88	0.00	0.00	0.41	-0.80	0.80	
	CS_U_300	-0.80	0.85	0.00	-0.01	0.30	-0.80	0.80	
	CS_V_300	-0.89	0.90	0.01	0.01	0.32	-0.80	0.80	
	CS_W_300	-0.39	0.49	0.00	0.00	0.11	-0.40	0.40	

3. Results & Discussions – RNN

Long Short-Term Memory (LSTM) model – SNU, Prof. Chang-Hoi Ho



- LSTM model has encoder and decoder parts which are equivalent to observation and model data.
- LSTM model has 2-stacked layers and outputs of that are **$PM_{2.5}$ concentrations for two-day forecast**

➢ **3-month separated models**

Jan–Dec	→	Jan–Feb–Mar	Feb–Mar–May	…	Nov–Dec–Jan	Dec–Jan–Feb

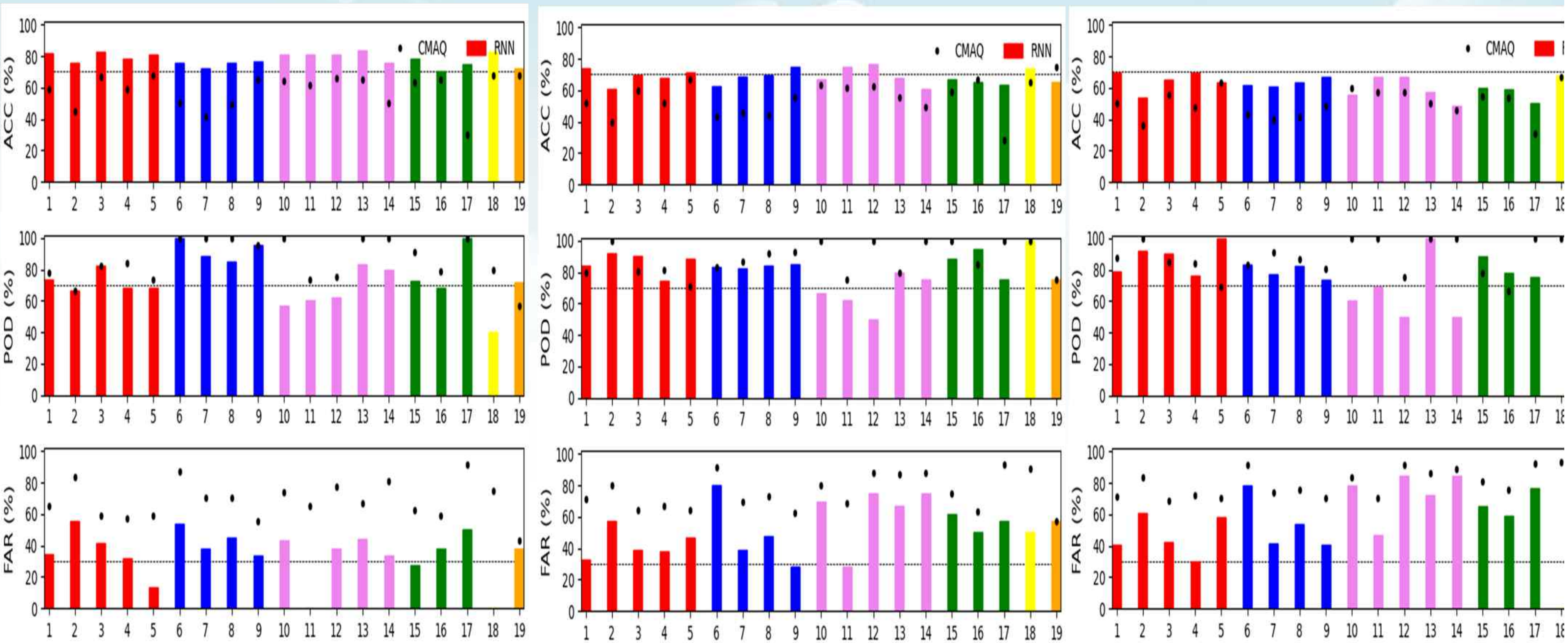
- We use separate one year to **12 sets of consequence 3-month periods** because atmospheric conditions that affect $PM_{2.5}$ concentrations vary with season.

➢ **Ho, Chang-Hou et al., "Development of a $PM_{2.5}$ prediction model using a recurrent neural network algorithm for the Seoul metropolitan area, Republic of Korea" , AE, 2021.**

3. Results & Discussions – RNN

Daily prediction for 19 forecasting regions (2020.1-2020.5)

Day+0 **Day+1** **Day+2**

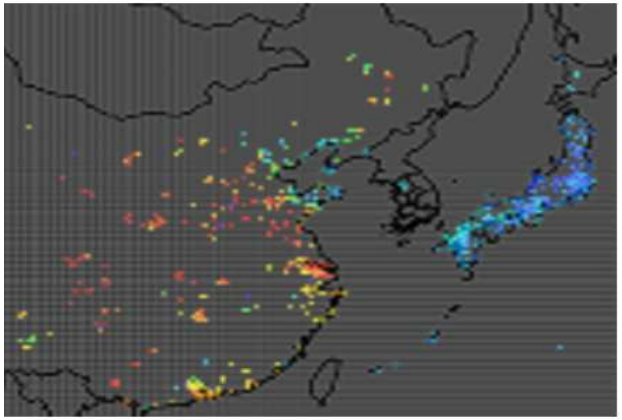


- 1: SU, 2: IC, 3: GGN, 4: GGS, 5: GWW, 6: DJ, 7: SJ, 8: CCN, 9: CCS, 10: BS, 11: DG, 12: US, 13: GSN, 14: GSS, 15: GJ, 16: JLN, 17: JLS, 18: GWE, 19: JJ
- Bar and dot indicates RNN and CMAQ(CASE4) results, respectively
- The accuracy of RNN decreased below 70 % at one- and two- day forecast.
- The POD of RNN has high performance, but the FAR have also high values.
- In validation and test periods, RNN got better results than CMAQ.

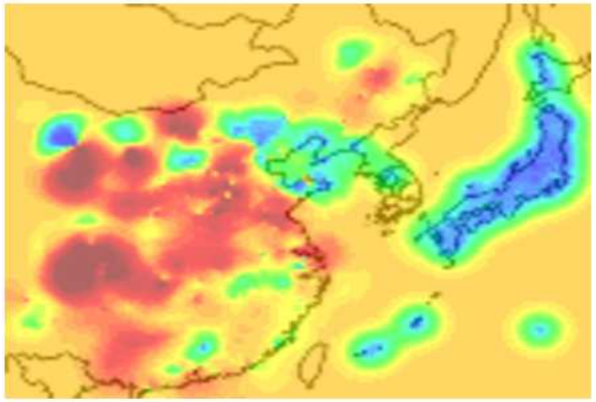
3. Results & Discussions – CNN

Input variables and Kriging for 2 D. maps

Classification	Input variable
Grid data	**Observed surface PM concentrations, kriging** : PM (T1 ~ T5)
	Meteorology, kriging : U, V, TA, RH (T1 ~ T5)
	Forecasted PM concentrations : PM2.5 (T6 ~ T15)
Regional Point data	**Observation at T5** : O_U, O_V, O_Pa, O_T, O_RH, O_RNN, O_rad, O_PM10, O_PM2.5, SO2, O_NO2, O_O3, O_CO, etc.
	Forecast data at T*i*(i : 6 ~ 15): f_PM2.5, f_U, f_V, f_MH, f_GPH_850hPa, f_U_850hPa, f_U_850hpa, f_V_850hpa, f_RH_850hpa, f_T_850hpa, f_GHP_925hpa, f_U_925hpa, f_V_925hap, f_DT_850-925



Observed PM2.5 data

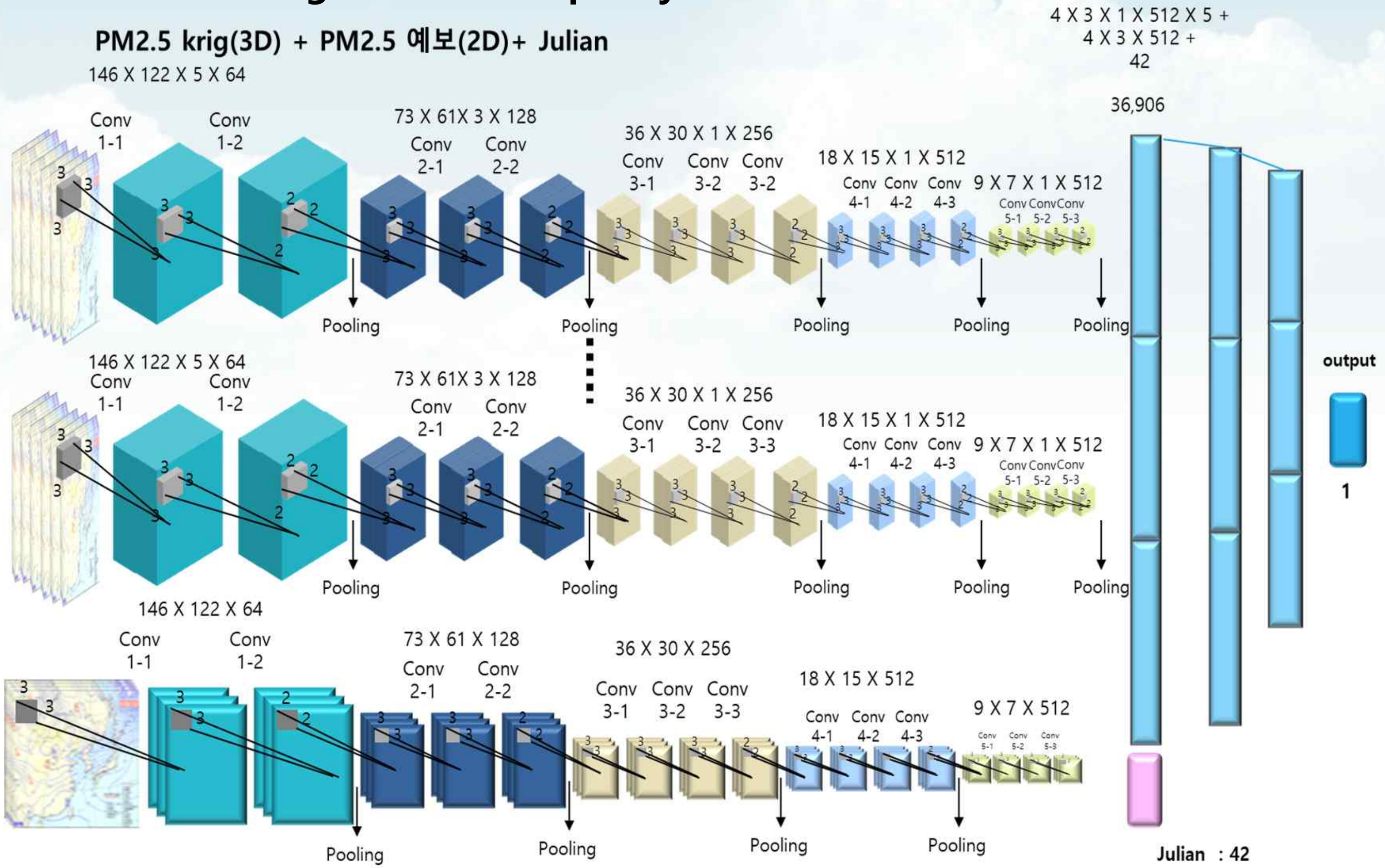


2 dimensional PM2.5 map by Kriging interpolation

3. Results & Discussions – CNN

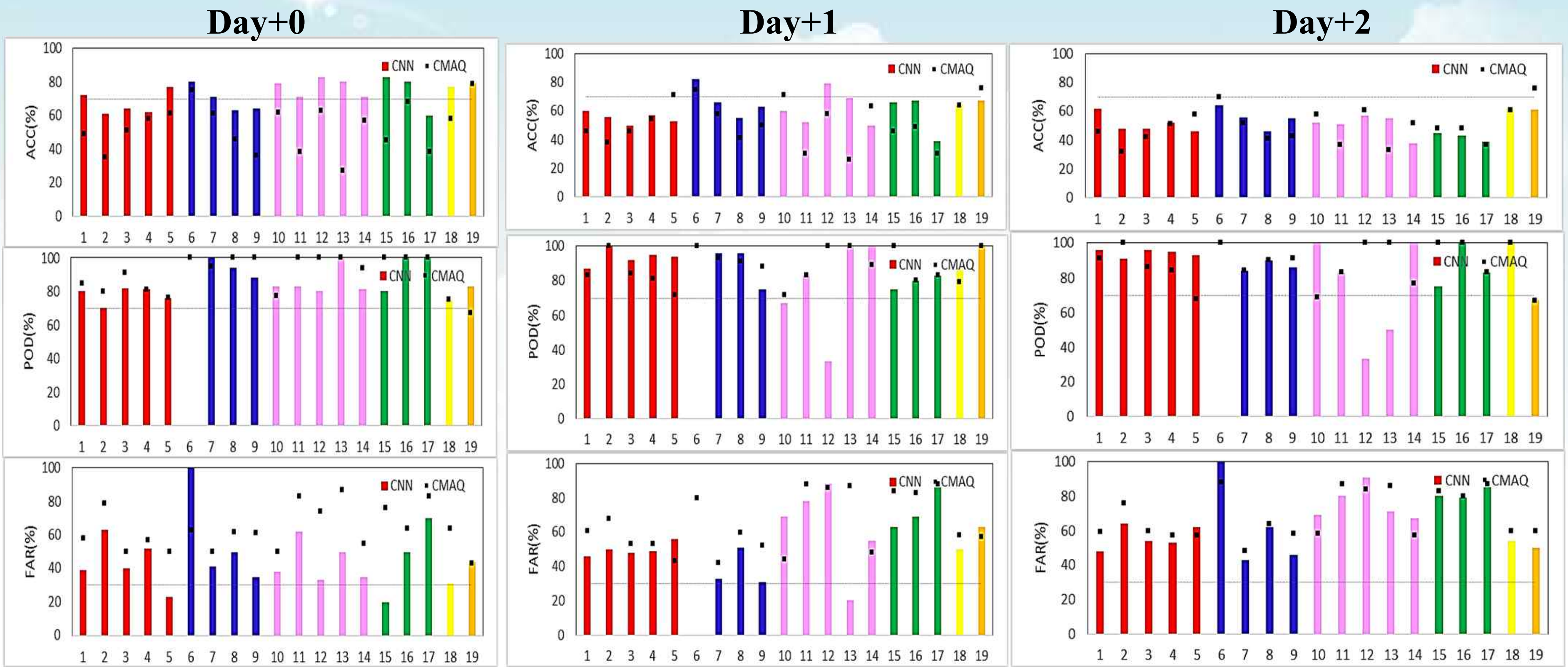
CNN with VGG 16 network

- **VGG 16 Net**
  : Convolution layer (3x3 Filter, padding size 1), Max pooling (2x2 filter, 2 stride)
- **Data balancing for bad air quality index**



3. Results & Discussions – CNN

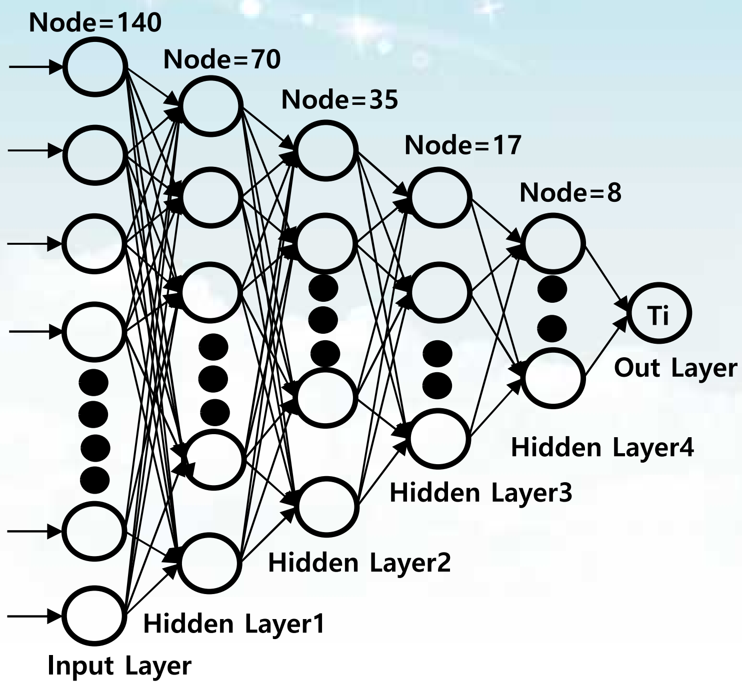
Daily prediction for 19 forecasting regions (2020.1-2020.3)



- ACC : 60 - 80%, POD : 60-90%, FAR : 20-60%
- CNN improves ACC and FAR comparing to CMAQ

3. Results & Discussions – DNN

Input Variables and Hyperparameters

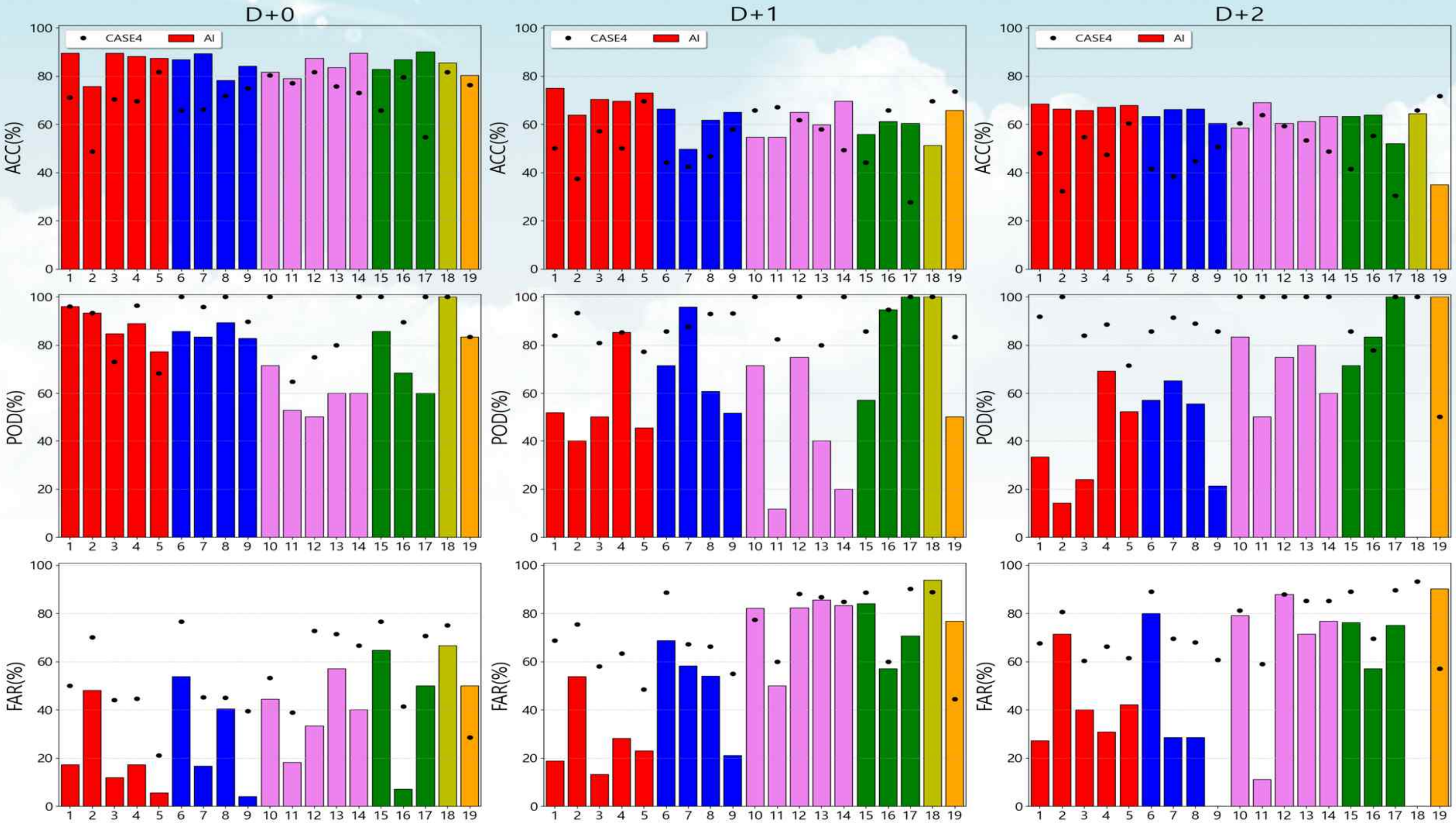


- Hyper parameters
  → EPOCH : 100,000 epoch
  → Learningrate : 0.009
  → Layer : 5 layer
  → Node : 140-70-35-17-8-1
  → Batch_size : 1
  → Seed_number : 1
  → Activation function : sigmoid
  → Cost function : MSE

$PM_{2.5}$ AI Input variables				
Primary data		Secondary data		
Observation	Prediction	Pattern/Regional contribution	Anomaly	Cosine similarity
T5	Ti (i:6~15)	Ti (i:6~15)	Ti (i:6~15)	Ti (i:6~15)
O_U	f_PM2.5	cluster_pattern_P01	a_1000hpa_gpm	cs_1000hpa_gpm
O_V	f_RH	cluster_pattern_P02	a_1000hpa_RH	cs_1000hpa_RH
O_Pa	f_MH	cluster_pattern_P03	a_1000hpa_Ta	cs_1000hpa_Ta
O_Ta	f_Pa	cluster_pattern_P04	a_1000hpa_U	cs_1000hpa_U
O_Td	f_Ta	cluster_pattern_P05	a_1000hpa_V	cs_1000hpa_V
O_RH	f_U	PM2.5_01	a_1000hpa_W	cs_1000hpa_W
O_RN_ACC	f_V	PM2.5_02	a_925hpa_gpm	cs_925hpa_gpm
O_Radiation	f_RN_ACC	PM2.5_03	a_925hpa_RH	cs_925hpa_RH
O_PM2.5	f_850hpa_gpm	PM2.5_04	a_925hpa_Ta	cs_925hpa_Ta
O_O3	f_850hpa_U	PM2.5_05	a_925hpa_U	cs_925hpa_U
O_NO2	f_850hpa_V	PM2.5_06	a_925hpa_V	cs_925hpa_V
O_CO	f_850hpa_RH	PM2.5_07	a_925hpa_W	cs_925hpa_W
O_SO2	f_850hpa_Ta	PM2.5_08	a_850hpa_gpm	cs_850hpa_gpm
O_PM10	f_925hpa_gpm	PM2.5_09	a_850hpa_RH	cs_850hpa_RH
	f_925hpa_U	PM2.5_10	a_850hpa_Ta	cs_850hpa_Ta
	f_925hpa_V	PM2.5_11	a_850hpa_U	cs_850hpa_U
	f_temp_850_925	PM2.5_12	a_850hpa_V	cs_850hpa_V
		PM2.5_13	a_850hpa_W	cs_850hpa_W
		PM2.5_14	a_700hpa_gpm	cs_700hpa_gpm
		PM2.5_15	a_700hpa_RH	cs_700hpa_RH
		PM2.5_16	a_700hpa_Ta	cs_700hpa_Ta
		PM2.5_17	a_700hpa_U	cs_700hpa_U
		PM2.5_18	a_700hpa_V	cs_700hpa_V
		PM2.5_19	a_700hpa_W	cs_700hpa_W
		PM2.5_20	a_500hpa_gpm	cs_500hpa_gpm
			a_500hpa_RH	cs_500hpa_RH
			a_500hpa_Ta	cs_500hpa_Ta
			a_500hpa_U	cs_500hpa_U
			a_500hpa_V	cs_500hpa_V
			a_500hpa_W	cs_500hpa_W
			a_300hpa_gpm	cs_300hpa_gpm
			a_300hpa_RH	cs_300hpa_RH
			a_300hpa_Ta	cs_300hpa_Ta
			a_300hpa_U	cs_300hpa_U
			a_300hpa_V	cs_300hpa_V
			a_300hpa_W	cs_300hpa_W

3. Results & Discussions – DNN

Daily prediction for 19 regions (2020.1-2020.3)

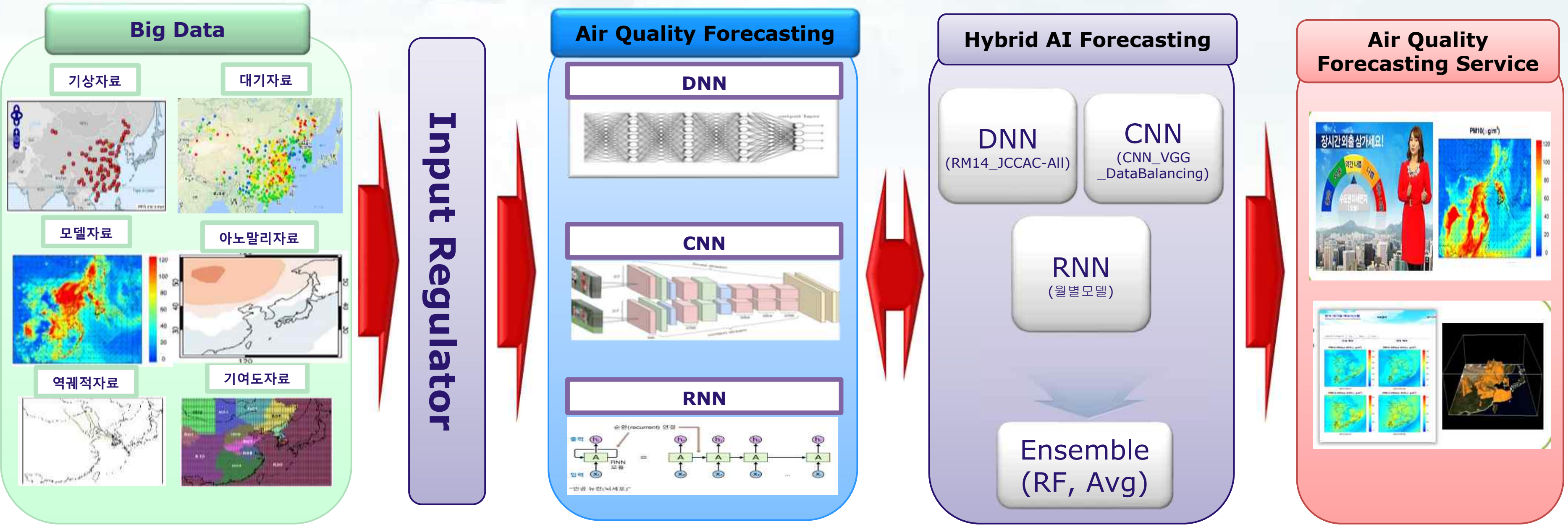


- ACC : 60 - 80%, POD : 20-80%, FAR : 10-60%
- DNN performance varies with forecasting regions
- DNN improves ACC in general

3. Results & Discussions – Ensemble

Ensemble forecasting

- Hybrid-type forecast module of Ensemble concept that comprehensively reflects the predictive characteristics of each of the developed DNN, CNN, and RNN
- Ensemble : Average



<Hybrid (Ensemble) Forecasting system>

3. Results & Discussions – AI Test bed

Real time AI forecasting Test bed

- **Real time forecasting performance for DNN, RNN, CNN and Ensemble**

AIFM AI Forecasting Master | 권역별예보분석 | 전국예보분석 | 예보정확도분석

예보지역 서울 인천 경기북부 경기남부 대전 세종 충북 충남 강원영동 강원영서 대구 부산 광주 전북 전남 경북 경남 울산 제주

예보시간 03h 09h 15h 21h

예보물질 PM2.5 PM10

분석종류 입력자료분석 예보결과분석 예보결과보고서

예보결과분석(서울) : 2020-03-19, 15h, PM2.5

예보일자 : 2020-03-19

■ 일별 예보결과

구분	OBS	CASE4	DNN	RNN	CNN	Ensemble
당일(D)	29.0	23.5	32.1	32.5	31.3	32.0
내일(D+1)		23.7	28.3	29.2	30.4	29.3
모레(D+2)		47.3	27.5	34.2	41.3	34.3

■ 시간별 예보결과

구분		OBS	CASE4	DNN	RNN	CNN	Ensemble
당일(D)	T4 (01:00~06:00)	36.0	34.6				
	T5 (07:00~12:00)	41.0	30.7				
	T6 (13:00~18:00)	10.0	12.6	30.7	28.0	22.9	27.2
	T7 (19:00~24:00)		16.0	20.8	24.9	25.4	23.7
내일(D+1)	T8 (01:00~06:00)		28.2	22.2	27.3	26.3	25.2
	T9 (07:00~12:00)		32.8	30.4	32.4	38.2	33.6
	T10 (13:00~18:00)		17.4	29.2	29.1	33.8	30.7
	T11 (19:00~24:00)		16.3	31.4	28.2	23.5	27.7
모레(D+2)	T12 (01:00~06:00)		20.9	12.6	24.4	21.8	19.6
	T13 (07:00~12:00)		51.2	25.6	34.2	52.0	37.3
	T14 (13:00~18:00)		48.9	28.7	33.6	40.3	34.2
	T15 (19:00~24:00)		68.3	43.1	44.6	50.9	46.2

구분	좋음	보통	나쁨	매우나쁨
PM2.5(μg/m³)	0-15	16-35	36-75	76이상

3. Results & Discussions – AI Test bed

Forecasting performance on Real time AI forecasting Test bed

- Period : 2020.1.1 – 2020.5.20
- DNN, RNN, CNN, Ensemble performance for D+1, D+2.

예보정확도분석(서울) : 2020-01-01~2020-05-20, 내일(D+1), 15h, PM2.5   분석기간 : 2020-01-01 ~ 2020-05-20

지수평가결과

구분	CASE4	DNN	RNN	CNN	Ensemble
적중율 (A)	48.9	73.4	67.6	69.8	75.5
	68 / 139	102 / 139	94 / 139	30 / 43	105 / 139
고농도 적중률 (HIT)	72.0	52.0	88.0	0.0	80.0
	18 / 25	13 / 25	22 / 25	0 / 0	20 / 25
감지확률 (POD)	84.0	56.0	88.0	0.0	80.0
	21 / 25	14 / 25	22 / 25	0 / 0	20 / 25
오경보율 (FAR)	67.7	30.0	42.1	100.0	23.1
	44 / 65	6 / 20	16 / 38	3 / 3	6 / 26

통계평가결과

구분	CASE4	DNN	RNN	CNN	Ensemble
평균	34.8	25.9	28.9	24.4	27.2
편차	±16.66	±11.25	±11.91	±7.9	±10.4
MBIAS	10.38	1.36	4.44	3.85	2.7
NMB	42.13	5.5	18.04	4.83	10.96
IOA	0.75	0.86	0.87	0.71	0.89
R	0.76	0.75	0.8	0.2	0.82

예보정확도분석(서울) : 2020-01-01~2020-05-20, 모레(D+2), 15h, PM2.5   분석기간 : 2020-01-01 ~ 2020-05-20

지수평가결과

구분	CASE4	DNN	RNN	CNN	Ensemble
적중율 (A)	47.8	71.0	65.2	61.9	73.2
	66 / 138	98 / 138	90 / 138	26 / 42	101 / 138
고농도 적중률 (HIT)	66.7	41.7	83.3	0.0	70.8
	16 / 24	10 / 24	20 / 24	0 / 0	17 / 24
감지확률 (POD)	91.7	45.8	83.3	0.0	70.8
	22 / 24	11 / 24	20 / 24	0 / 0	17 / 24
오경보율 (FAR)	66.2	35.3	47.4	100.0	34.6
	43 / 65	6 / 17	18 / 38	6 / 6	9 / 26

통계평가결과

구분	CASE4	DNN	RNN	CNN	Ensemble
평균	36.0	25.9	29.7	24.2	27.4
편차	±18.89	±12.9	±12.59	±7.36	±11.27
MBIAS	11.78	1.49	5.33	4.19	3.09
NMB	48.05	6.06	21.75	5.2	12.58
IOA	0.7	0.79	0.8	0.63	0.83
R	0.72	0.64	0.71	0.13	0.72

4. Conclusion

- Machine learning improves PM forecasting performance by minimizing the bias errors of CTM forecast.
- Continuous efforts are still necessary to develop additional big data and improve their accuracy.
- In addition, it is judged that the accuracy of the AI forecast can be improved through an integrated effort for artificial intelligence algorithms.
- Artificial intelligence forecasts under development need to be continuously verified and supplemented through real-time forecasts via Test Bed operation.