# Diurnal variation of surface Nitrogen Dioxide mixing ratios over South Korea estimated using the machine learning with the GEMS observations data

**Seonyeong Park**[1], Hanlim Lee[1*], Wonei Choi[2], and Yeonjin Jung[1]

[1] Division of Earth and Environmental System Science, Major of Spatial Information Engineering, Pukyong National University, [2] NASA Goddard Space Flight Center

## Abstract

Nitrogen dioxide ($NO_2$) is an important target for monitoring atmospheric quality. $NO_2$ is a precursor in the formation of aerosols and tropospheric ozone, and it directly impacts on human health and the environment. In particular, surface $NO_2$ mixing ratio serves as an indicator for air quality assessment in many countries. Surface $NO_2$ mixing ratio is typically measured by in-situ instruments. It is, however, difficult to identify its spatial characteristics with the in-situ monitors due to the lack of their spatially resolved measurements. Recently, there have been some efforts to estimate surface $NO_2$ mixing ratios over large areas using LEO satellite observations showing good agreement between the estimates and those of in-situ measurements. Our study aims, for the first time, to estimate diurnal variation of surface $NO_2$ distributions over South Korea with the Geostationary Environment Monitoring Spectrometer (GEMS) observation data. Satellite-based variables, numerical model-based meteorological variables, and auxiliary variables are used to train several machine learning models, including Random Forest, XGBoost, and LightGBM, for the estimation of surface $NO_2$ mixing ratios over South Korea for a one-year period from 2021 through 2022. The Random Forest model shows the best performance against the other models based on evaluations via comparisons with the in-situ measurement data at AirKorea sites which are excluded from the model training dataset. The R, RMSE, and MAE obtained between the in-situ data and those estimated by the Random Forest model are 0.87, 2.08 ppb and 1.24 ppb, respectively. Seasonal and area type dependency of estimation performance has also been investigated. The best agreement was found between in-situ measurement data and those of estimates in urban type areas, while discrepancies were found in urban-port mixed type areas. In terms of seasonal performance, the best agreement was found in winter when $NO_2$ levels and the magnitude of diurnal variation are larger than other seasons. Poor agreement was found in summer. This study demonstrates a high potential to provide diurnally-spatially resolved surface $NO_2$ mixing ratio information using the GEMS observation data.

## Data & Method

### Data

- **2021.12 ~ 2022.11, South Korea (33-39°N, 124-132°E)**
- **AirKorea observation sites in South Korea**
  - Type : Urban (503), National background (9), Road (54), Suburb (27), Port (21)

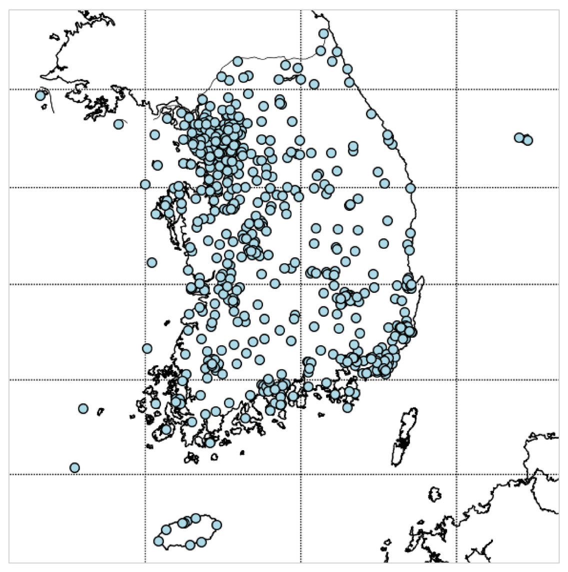


Fig. Location of AirKorea observation sites

| Dataset | Variables | Source |
|---|---|---|
| In-situ measurements | AirKorea NO2 mixing ratio | AirKorea |
| Satellite-based | GEMS NO2 vertical column density (Total, Tropospheric) | NIER |
| | GEMS O3 vertical column density (Total) | |
| | GEMS Terrain height | |
| | MODIS Land cover | NASA |
| Model-based | ERA5 Meteorological variables | ECMWF |
| Auxiliary | KST | - |

Table. List of input variable candidates

### Method

- **Input variables pre-processing**
  - Used data with quality flags indicating high quality and filtered data of low quality (e.g. geometry, cloud)
  - Resampled to new grid with GEMS spatial resolution (3.5 km X 8 km)
  - Meteorological variables : Resampled to GEMS resolution over South Korea using a linear interpolation
- **Machine learning approach and training process**
  - Random Forest (RF), XGBoost (XGB), LightGBM (LGBM) model construction (train 80%, test 20%)
  - Investigated the contribution of input variable candidate
  - Determined optimal hyperparameters based on a 5-fold cross validation
- **Assessment of surface $NO_2$ mixing ratio estimation model**
  - Statistical assessment : R, MAE(ppb), RMSE(ppb)

## Result

### Model validation and comparison

- **Machine Learning model (RF, XGB, LGBM) validation result (All)**

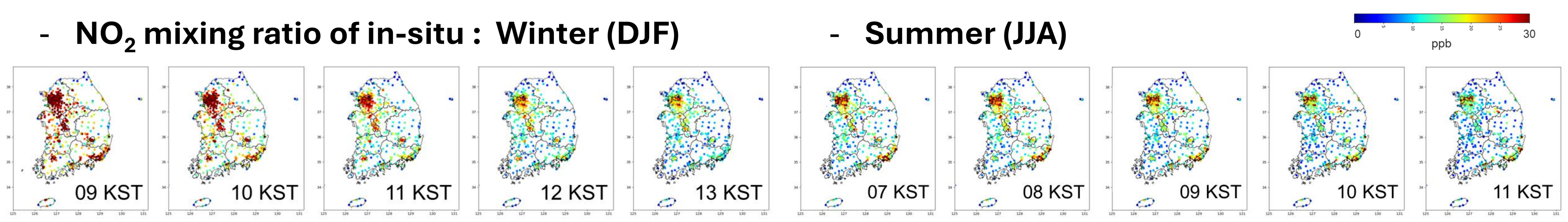
| | MLR | RF | XGB | LGBM |
|---|---|---|---|---|
| N | 170,324 | 170,324 | 170,324 | 170,324 |
| R | 0.60 | 0.86 | 0.86 | 0.83 |
| MAE(ppb) | 1.39 | 1.24 | 1.28 | 1.32 |
| RMSE(ppb) | 8.01 | 2.08 | 2.83 | 2.94 |

- **Random Forest (RF) model validation result (Area type)**

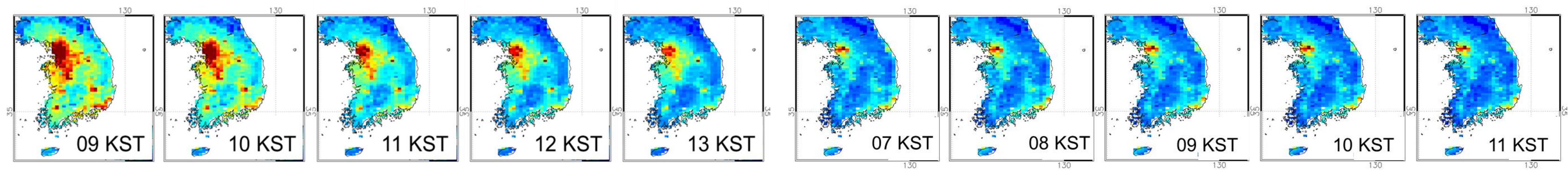
| | RF | | | | |
|---|---|---|---|---|---|
| | Urban | Urban-background | Urban-road | Urban-suburb | Urban-port |
| N | 141,443 | 143,818 | 156,310 | 148,603 | 145,922 |
| R | 0.87 | 0.86 | 0.85 | 0.87 | 0.83 |
| MAE(ppb) | 1.24 | 1.28 | 2.03 | 1.31 | 5.81 |
| RMSE(ppb) | 2.08 | 2.11 | 4.21 | 2.06 | 7.94 |

### Diurnal variation of estimated surface $NO_2$ (Seasonal)

- **$NO_2$ mixing ratio of in-situ : Winter (DJF)** - **Summer (JJA)**




- **The estimates from GEMS : Winter (DJF)** - **Summer (JJA)**

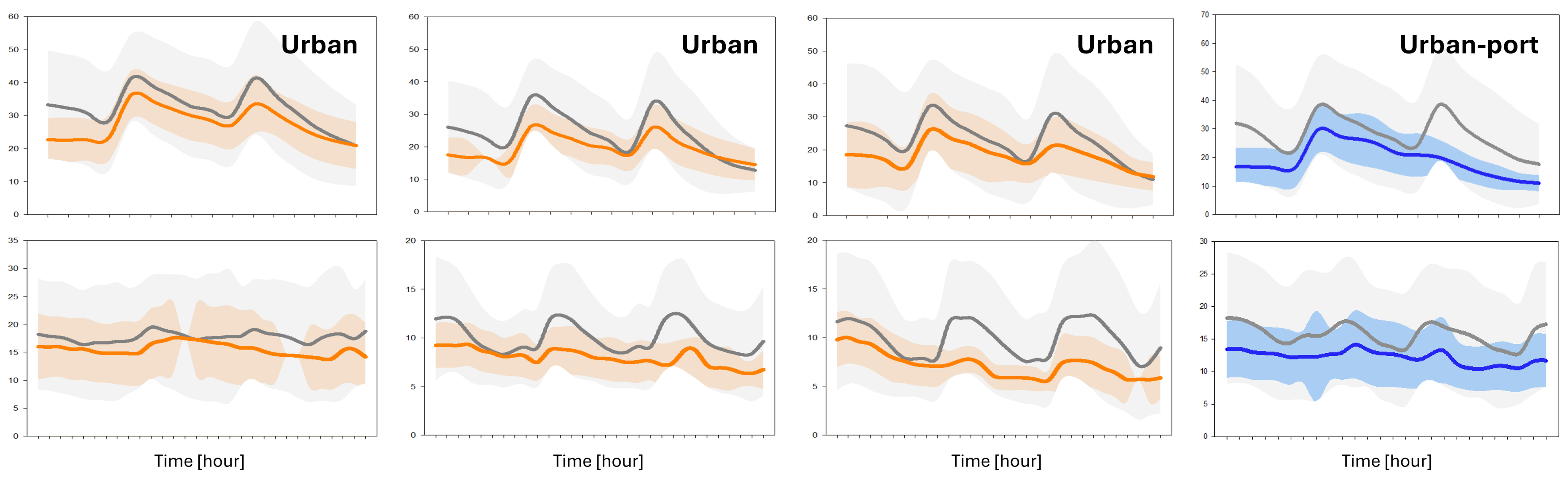



### Diurnal variation of estimated surface $NO_2$ (Area type)

- **Seoul** - **Daejeon** - **Daegu** - **Incheon**

Winter (DJF)

Summer (JJA)




## Conclusion

We estimated diurnal variation of surface $NO_2$ distributions over South Korea for a one-year period from 2021 through 2022 with the GEMS data. The RF-based model shows the best performance against the other models via comparisons with the in-situ measurement data at AirKorea sites The R, RMSE, and MAE obtained between the in-situ data and those estimated by the RF-based model are 0.87, 2.08 ppb and 1.24 ppb, respectively. Seasonal and area type dependency of estimation performance has also been investigated. The best agreement was found between in-situ measurement data and those of estimates in urban type areas, while discrepancies were found in urban-port mixed type areas. In terms of seasonal performance, the best agreement was found in winter when $NO_2$ levels and the magnitude of diurnal variation are larger than other seasons. This study demonstrates a high potential to provide diurnally-spatially resolved surface $NO_2$ mixing ratio information using the GEMS data.