Master's Thesis

The estimation of ground-level nitrogen dioxide (NO_2) and ozone (O_3) concentrations using Real-Time Learning (RTL)-based machine learning approach

Minso Shin

Department of Urban and Environmental Engineering

(Environmental Science and Engineering)

Ulsan National Institute of Science and Technology

2022

The estimation of ground-level nitrogen dioxide (NO_2) and ozone (O_3) concentrations using Real-Time Learning (RTL)-based machine learning approach

Minso Shin

Department of Urban and Environmental Engineering

(Environmental Science and Engineering)

Ulsan National Institute of Science and Technology

The estimation of ground-level nitrogen dioxide (NO_2) and ozone (O_3) concentrations using Real-Time Learning (RTL)-based machine learning approach

A thesis/dissertation submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Master of Science

Minso Shin

07.15.2022
Approved by

_____

Advisor
Jungho Im

The estimation of ground-level nitrogen dioxide (NO_2) and ozone (O_3) concentrations using Real-Time Learning (RTL)-based machine learning approach

Minso Shin

This certifies that the thesis/dissertation of Minso Shin is approved.

07.15.2022

Signature

_____

Advisor: Jungho Im

Signature

_____

Thesis Committee member: Chang-Keun Song

Signature

_____

Thesis Committee member: Myong-In Lee

Abstract

Nitrogen dioxide (NO_2) and ozone (O_3) are the significant components of gaseous air pollutants that have harmful effects on human health. The monitoring and analysis of air pollutant exposure and persistence, and short-term forecasts are necessary for efficient public health management. In this study, the estimation model for the ground-level O_3 and NO_2 concentrations was developed which are spatially continuous over the land and ocean. The ground-level estimation was developed using the RTL-based machine learning technique with various satellite data and numerical model data as input variables. Three models were tested to build an accurate model using the most available data. 1) the ocean model using only ocean variables that have values for all regions; 2) the land model using all available data with assigning constant values to ocean variables; 3) the combined model that combines the results of the ocean model for sea area and the results of the land model for land area. Since NO_2 and O_3 have a relatively short lifespan, the real-time learning model is effective in estimating accurate ground-level concentrations.

Keywords: O_3, NO_2, ground-level NO_2 concentration, ground-level O_3 concentration, Real-Time Learning, Machine Learning

Contents

List of figures

List of tables

1. Introduction

Nitrogen dioxide (NO_2) and ozone (O_3) are the significant components of gaseous air pollutants that have harmful effects on human health. World Health Organization (WHO) (2006) offers air quality guidelines for policy development and risk reduction about common air pollutants including NO_2 and O_3. The short-term and long-term NO_2 and O_3 exposure have a causal relationship with respiratory diseases such as increased airway hypersensitivity and allergic inflammation. There is also a suggestive causal relationship between cardiovascular effects and total mortality (U.S. EPA., 2016, 2020). NO_2 also acts as a precursor to other air pollutants. It combines with volatile organic compounds (VOCs) through photochemical reactions to produce ozone or fine particulate matters in the form of NO_2 (National Research Council, 1991). Ozone naturally increases in concentration due to bio-VOC and stratospheric ozone, but it also increases anthropogenically due to NO_2 increased by human activities such as NO_2 comes from fossil fuel combustion. O_3 is also known to produce secondary aerosol particles, which affects the increase in ultra-fine PM concentrations. Therefore, monitoring and analysis of air pollutant exposure and persistence, and short-term forecasts are necessary for efficient public health management.

The East Asian region has been growing rapidly in recent decades, resulting in increased emissions and damage to air pollutants. In the case of air pollution, not only local effects but also the effects of air circulation on the neighboring regions and countries. Therefore, it is necessary to continuously monitor the ground-level NO_2 and O_3 concentrations. South Korea and China are conducting station-based monitoring. Many stations are concentrated around urban areas where air pollutions are more severe than a rural area. Because it is point-based station information not only for rural areas with few stations but also for urban areas, it is difficult to obtain spatially continuous concentration information. Station-based studies that do not use satellite data are those that interpolate the station and carry some uncertainty. In addition, spatially continuous concentration estimation results are required over the sea to analyze the characteristics of the movement tendency of air quality between countries. Satellite infrastructure can perform over large areas. Monitoring using satellite data is possible even in areas where no stations exist, such as in the ocean. However, little research has been done on the ocean.

Satellite outputs generally used in previous studies are data on the concentration of vertical columns of NO_2 and O_3, and in particular, the Ozone Monitoring Instrument (OMI) data have been widely used since it has been providing information continuously in the past. As a successor to OMI, the TROPOspheric Monitoring Instrument (TROPOMI) data has been used recently, which has a high spatial resolution (Cooper et al., 2020). Although many satellite-based monitoring studies have been conducted using OMI data, the focus was on trend analysis performed based on column concentration,

and many estimation studies on direct ground concentration were not conducted (Oner & Kaynak, 2016). In the case of O_3, since there is a limit to estimating the ground concentration from the satellite-based total column concentration (Zoogman et al., 2014), not many studies have been conducted focusing on satellite data. On the other hand, NO_2 has been used for monitoring the long-term trend of column density itself because the vertical column density of NO_2 has a relatively high correlation with NO_2 concentration (Xu et al., 2019). Previous studies suggest that tropospheric NO_2 column data are sufficient to track spatial patterns in ground-level NO_2 (Bechle et al., 2015; Knibbs et al., 2014).

Satellite-based studies on estimating ground-level NO_2 concentrations frequently used relatively simple statistical techniques such as land-use regression (LUR) and mixed-effect model (MEM) based on multi-linear regression analysis and chemical transport model (CTM) based model (de Hoogh et al., 2016; Freddy Grajales & Baquero-Bernal, 2014; Hoek et al., 2015; Kharol et al., 2015; Liu et al., 2018; Meng et al., 2018). The LUR models are regression models that include variables for land cover or such land information and are widely used when geographic information such as NO_2 and O_3 are directly related to air pollution concentrations. Although LUR is a simple technique, it has been continuously used by increasing the relevant land use variables in various ways through GIS programs (e.g., taking buffers) and extracting appropriate land use variables with high correlation. The CTM-based proportional method was first applied in the particulate matter estimation study and was calculated by applying it to NO_2 in the same way (Y. Zhang et al., 2018). In addition, ground concentration estimation through statistical-based empirical models such as GTWR (Qin et al., 2017) and Bayesian maximum entropy (Jiang & Christakos, 2018) was used for ground NO_2 estimation.

Recently, studies using machine learning have been conducted to estimate ground concentration by applying machine learning. Zhan et al. (2018) estimated daily NO_2 exposure in China using Random Forest and spatiotemporal kriging model with OMI NO_2 vertical column density data. Z. Zhang et al. (2018) suggested a national-scale air pollutant concentration estimation model using Generalized additive mixed models (GAMM) and land use regression for $PM_{2.5}$, PM_{10}, and NO_2. Yeganeh et al. (2018) estimated spatiotemporal variation of NO_2 concentration using adaptive neuro-fuzzy inference system (ANFIS) and LUR variables. In many cases, land use variables were used with machine learning techniques.

Research on the estimation of ground air quality has relatively low accuracy because it is a past data-based study or model-based study. Since NO_2 and O_3 have a relatively short lifespan, it is expected to show more improved results when reflecting real-time learning-based machine learning that provides information about current information.

In this study, spatially continuous ground NO_2 and O_3 concentrations were estimated for land and oceans in East Asia using real-time learning-based machine learning techniques, based on various input

variables such as satellite data, numerical model data, and land use data. As the two pollutants are related to each other by photochemical reactions, the temporal and spatial characteristics of each pollutant were analyzed through the spatial distribution map calculated from the ground concentration estimation model.

2. Study area and data

2.1 Study area

This study was conducted in East Asia, which can be observed by GOCI satellites, from May 2018 to March 2021. The East Asian region where the study was conducted corresponds to 20-50°N and 110-150°E, including the Korean Peninsula, East China, and Japan (Figure 1). The study area is in the midlatitude region that is affected by the air quality of neighboring countries due to the influence of the westerly wind. China is rapidly industrialized, resulting in relatively severe air pollution. From the eastern China where many industrial complexes are located, Korea Peninsular and Japan are located to the east.

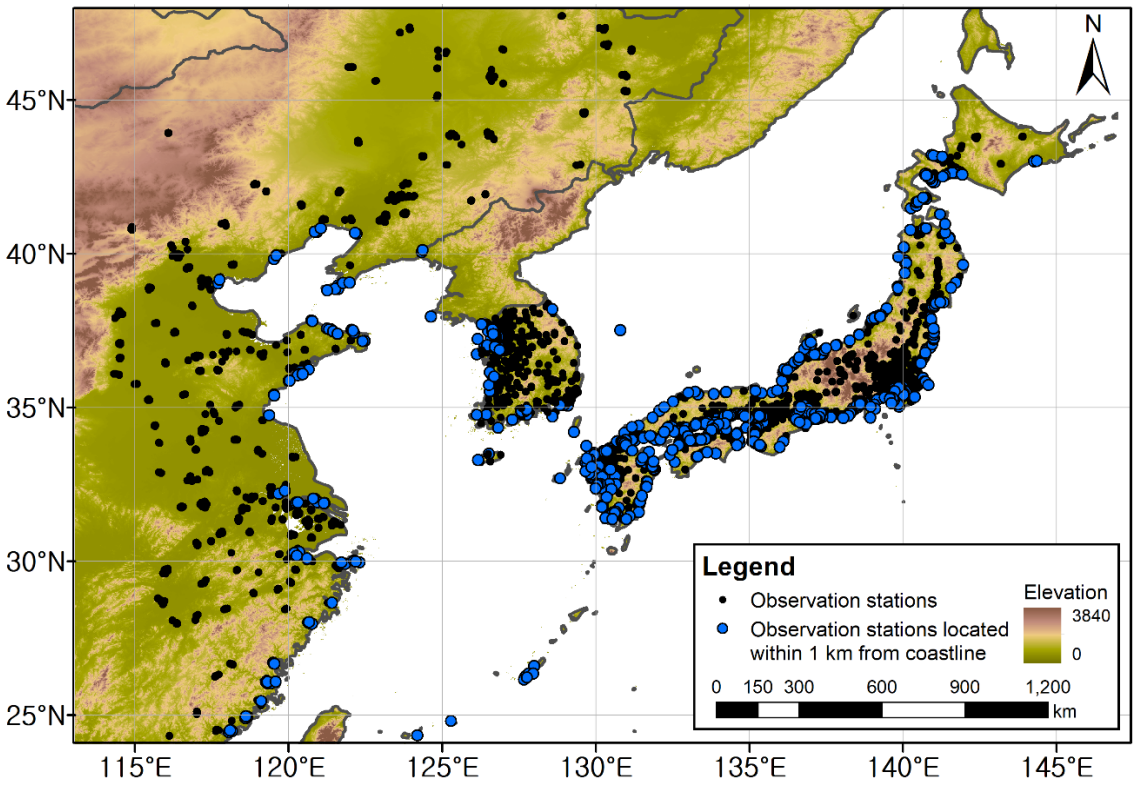

Figure 1. Study area and the location of stations.

Blue points are the stations located within 1km of the coastal line.

2.2 Data

2.2.1 Station data

The NO_2 and O_3 concentration values measured at the station were used as target variables. The ground-level NO_2 and O_3 concentration values were measured hourly at 443 stations in South Korea, 733 stations in eastern China, and 2073 stations in Japan (3249 stations in total) (Table 1). Since ground stations are distributed only in the land areas, it is impossible to verify sea areas. In this study, for the result evaluation of the estimated concentrations over the sea, the stations existing within 1 km of the coastline were considered as representative values at sea (**오류! 참조 원본을 찾을 수 없습니다.**). The hourly station data were extracted and used from 00 UTC to 07 UTC, which corresponds to the Geostationary Ocean Color Imagery (GOCI) providing time.

Table 1. The number of stations in the study area

	The number of stations	**The number of stations located within 1 km of coastline**
South Korea	443	48
Eastern China	733	30
Japan	2073	244

The ground observation data were provided in the form of real-time data and confirmed data with quality checks. The confirmed data were used preferentially when provided by country and period, otherwise, real-time data are used. The periods using the confirmed data are the entire period in the South Korea and the period before 2019 in Japan. For the period after 2020 in Japan and the entire period in China, real-time data were used.

The NO_2 and O_3 observation data with a concentration of 400 or more were considered outliers and removed. Since real-time data did not go through a quality check by the data provider, abnormal values were removed through additional steps. When the number of hourly data obtained at each station from 00 to 07 UTC was less than four, all data from that station were removed. In the case of stations that obtained more than four observation data, the outliers were removed in such a way as to remove values that exceed 99.9% of the normal distribution of the measured values of all observations within

8 hours.

2.2.2 Satellite data

Table 2. The satellite data used in this study

Data		Satellite	Spatial Resolution	Temporal Resolution
L2__NO2___	NO_2 vertical column density	TROPOMI	3.5 x 7.0 km 3.5 x 5.5 km (since 6 Aug 2019)	Daily
L2__O3____	O_3 vertical column density	TROPOMI	3.5 x 7.0 km 3.5 x 5.5 km (since 6 Aug 2019)	Daily
AOD	Aerosol Optical Depth	GOCI	6 km	8/day
FMF	Fine-Mode Fraction	GOCI	6 km	8/day
SSA	Single Scattering Albedo	GOCI	6 km	8/day
AE	Ångström Exponent	GOCI	6 km	8/day
MYD13A2	Normalized Difference Vegetation Index (NDVI)	MODIS	1 km	16 days
MDC12Q1	Land cover ratio (LCurban, LCbarren, LCcrop, LCveg)	MODIS	250 m	Yearly
DEM	Digital Elevation Model	SRTM	90 m	-

Table 2 shows satellite-based data among the input variables used in this study. The NO_2 and O_3 vertical column concentrations of TROPOMI were used as major variables. With TROPOMI data, the GOCI aerosol products including aerosol optical depth (AOD), the fine-mode fraction (FMF), Ångström exponent (AE), and single scattering albedo (SSA) were used. Since the TROPOMI which is the most relevant data for estimating NO_2 and O_3 concentration provides daily data, it is difficult to estimate the air quality concentration every hour. By using the GOCI hourly data, the changes over time

can be monitored. In addition, two types of the Modulate Resolution Imaging Spectroradiometer (MODIS) products (i.e., 16-day NDVI (Normalized Difference Vegetation Index) and landcover product) were used. Using MODIS land cover data provided in 17 classes, the four land cover ratio variables (land cover ratio of barren, crop, vegetation, and urban) were calculated. After reclassification as binary for each class, the ratio within a radius of 3 km was calculated and used as an input variable. In the case of vegetation classes, 10 classes, including forest, shrub lands, savanna, and grasslands, were grouped and used as one class.

2.2.3 Model-based meteorological data

Table 3 shows meteorological data based on the United Model-Regional Data Assessment and Precision System (UM-RDAPS) numerical model required for estimating ground-level NO_2 and O_3 concentrations. The UM-RDAPS model has a spatial resolution of 12 km and provides analysis field data four times a day (00, 06, 12, and 18:00 UTC). The hourly values of UM-RDAPS were generated by temporal interpolation using the analysis field. Wind speed (WS) and wind direction (Wsin, Wcos) were calculated using U-wind and V-wind variables. A total of 22 model-based meteorological variables were used as input variables, including 18 RDAPS variables highly associated with air pollutants and the additionally calculated cumulative maximum wind speed for 1, 3, 5, and 7 days.

Table 3. Meteorological variables from UM-RDAPS numerical model data

Variable	Description
Temp	Temperature
Dew	Dew-point temperature
RH	Relative Humidity
MaxWS	Maximum wind speed (3 Hour Maximum)
Stacked_MaxWS (1,3,5,7-days)	Accumulated maximum wind speed
PBLH	Planetary Boundary Layer Height
Visibility	Visibility height above ground

P_srf	Pressure surface
Tmax	Maximum temperature
Tmin	Minimum temperature
Tsrf	Temperature surface
FrictionalVelocity	Frictional velocity
PotentialEnvergy	Convective available potential energy
SurfaceRoughness	Surface roughness
LatentHeatFlux	Latent heat net flux
SpecificHumidity	Specific humidity
WS	Wind speed
Wcos	Cosine value of wind direction
Wsin	Sine value of wind direction

2.2.4 Ancillary data

In addition to satellite-based and numerical model-based data, the population density, road density, day of the year (DOY), and hour of the day (HOD) were used as other ancillary variables. The population density for each year calculated by linearly interpolating 2015 and 2020 data among Gridded Population of the World (GPW) v4 data was used. Road density data used Global Roads Inventory Project (GRIP4) data provided with a spatial resolution of 8 km (Meijer et al., 2018). DOY and HOD were used by applying a sine function to convert it to a value between -1 and 1 to reflect the continuity of time.

3. Methodology

3.1 Machine learning (random forest)

In this study, the ground-level NO_2 and O_3 concentration estimation were performed using random forest, one of machine learning. Random forest is a proposed method to overcome overfitting problems arising from a tree model and is based on the classification and regression trees (CART) model (Breiman, 2001). Based on different datasets generated through bagging, numerous independent trees are generated and then ensemble to produce final results (Figure 2). Random forest uses out-of-bag (OOB) data, which is not used for model learning, to provide variable importance, which is information about whether variables contribute to model development.

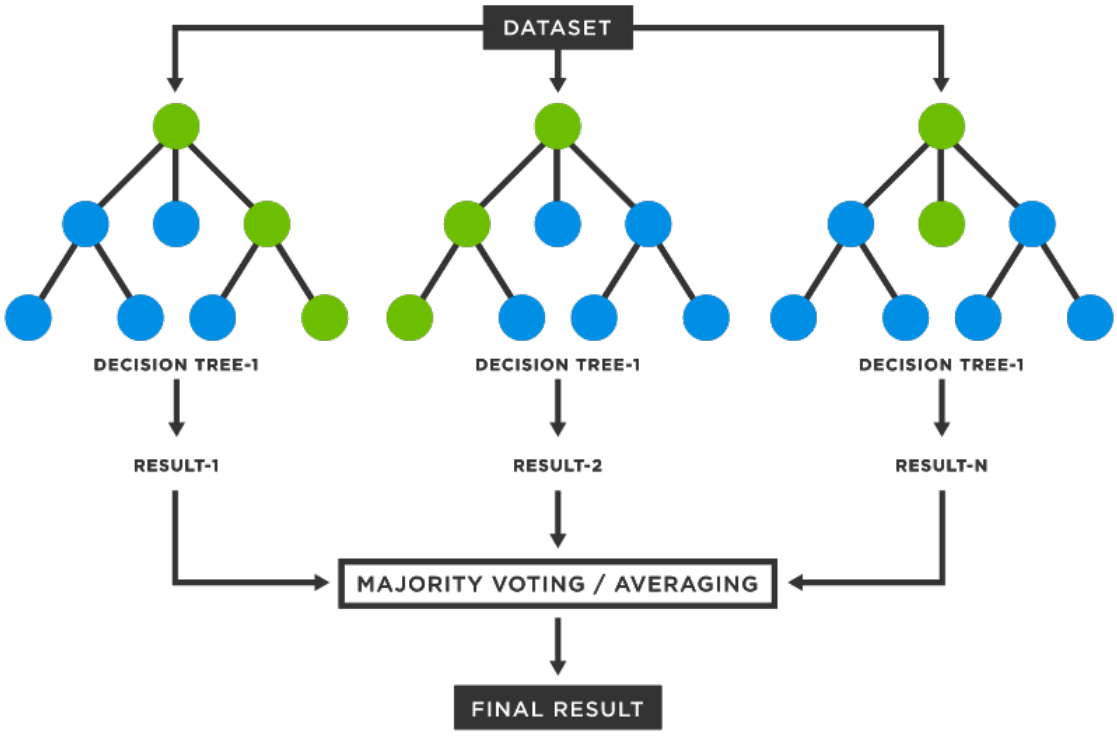

Figure 2. Concept of random forest algorithm (TIBCO, 2022)

3.2 Oversampling and subsampling

As shown in Figure 3, the distribution of O_3 concentration is biased toward medium and low concentrations, and the distribution of NO_2 concentration is biased toward low concentrations. In addition, the concentration distribution varies greatly depending on the season (Figure 4). To balance the concentration distribution in the unbalanced dataset, over-/sub-sampling and sample adjustments were performed to adjust the concentration distribution of NO_2 and O_3. After the over-/sub-sampling ratio was divided into 30 ppb intervals for each pollutant's dataset, the distribution of the number of samples in each interval and the entire number of samples were considered. The subsampling was performed in the intervals with many samples, and oversampling was performed in the intervals with few samples.

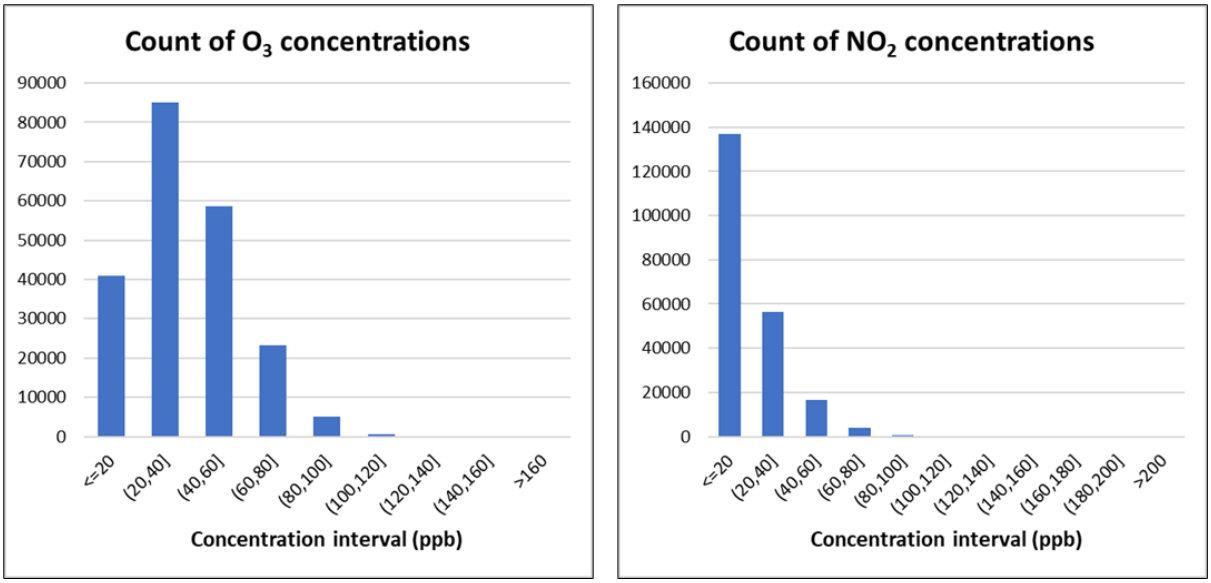

Figure 3. Sample distribution of O_3 and NO_2 concentration

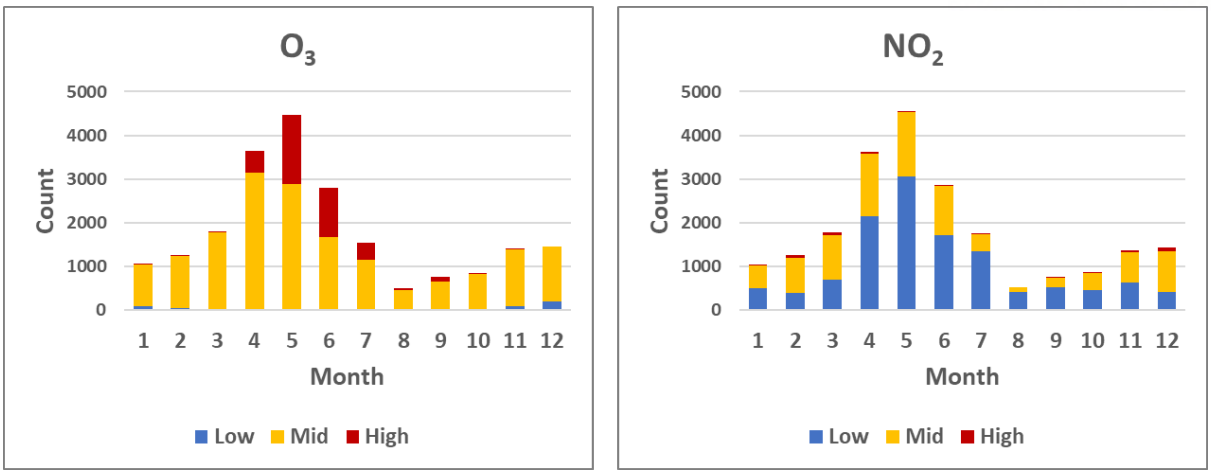

Figure 4. Distribution of low, mid-high concentration samples of O_3 and NO_2 by month

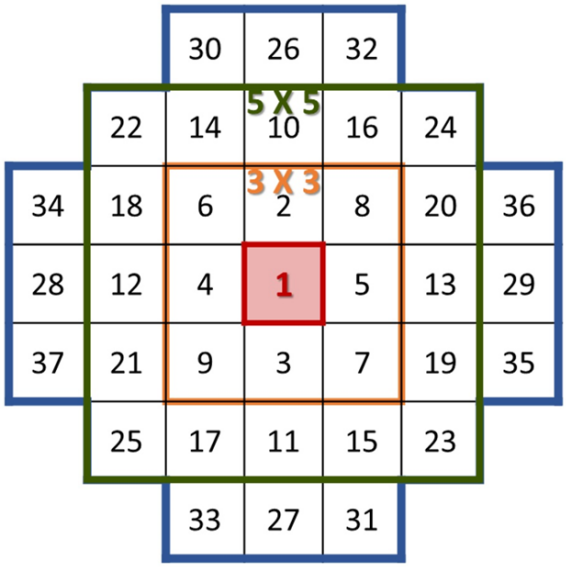

Figure 5. Oversampling patch

The oversampling technique was used under the assumption that the concentration of air pollutants in nearby areas is similar and can be used to increase the number of samples for relatively insufficient intervals. After extracting a sample of surrounding pixels within a certain distance from the station where the sample exists, a value equivalent to around 5% of the observation value can be arbitrarily allocated to generate up to 37 times the training data (Figure 5). The subsampling technique can be applied to the intervals with an excessive number of samples compared to other intervals. The

distribution of samples is controlled by randomly removing samples by an appropriate ratio. In the case of NO_2, subsampling was applied in the low concentration section and oversampling in the high concentration section. And in the case of O_3, only oversampling was applied in the low concentration and high concentration sections.

3.3 Offline model

The offline model is a model made using samples in the entire study period, from May 2018 to March 2021. Unlike the RTL-based model, it is possible to build a model when not only data from the past few days but also data from a long period are accumulated. All samples were divided into two groups by predefined dates. The predefined dates were randomly divided into 80% of model development dates and 20% of prediction dates, considering the date with many high concentration samples. It was intended to prevent high-concentration samples from being concentrated on one side of the sample for training data or prediction. Data for model development were divided into training datasets (80%) for model construction and test datasets (20%) for model validation. Considering the unbalanced concentration distribution, oversampling and subsampling techniques were equally applied to offline models. The oversampled sample is used only for model training and not for verification.

3.4 Real-Time Learning (RTL)-based dataset construction

Real-time learning is a technique that increases model accuracy by using real-time data as training data for machine learning models. RTL-based modeling uses data for a certain period of recent time, unlike the existing model that uses samples for the entire research period. Figure 6 shows the structure of building the dataset for the real-time learning method. To weighting to the latest samples (i.e., temporal weighting), a 1 % reduction rate was applied for past times on the same day and 10 % for other days. In this study, real-time training datasets were built by accumulating the samples of recent N days including all GOCI data available time (9 – 16 KST). Recent 30, 15, 7, and 3 days were tested for accumulating periods.

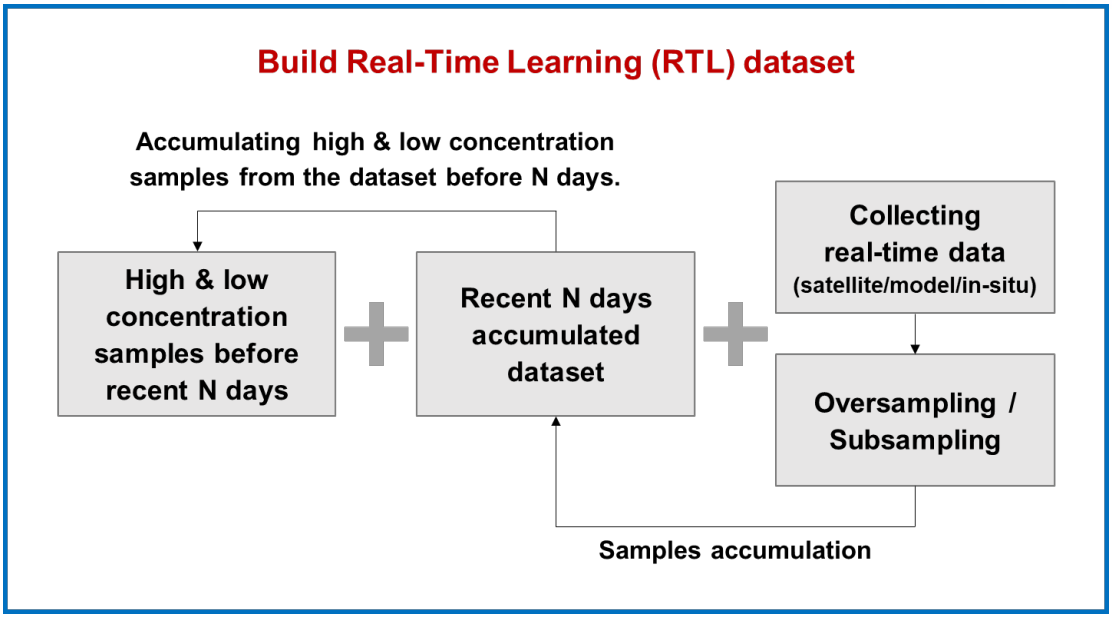

Figure 6. Structure of building real-time learning dataset

When training using only recent data, it is advantageous to immediately reflect real-time information to estimate ground-level air pollutant concentrations. However, there is a disadvantage that the model may vary greatly in real-time incoming samples. In particular, the ratio of high and low concentrations of samples was additionally considered to compensate for the lack of estimation ability in the case of the sudden inflow of samples with high and low concentrations.

The sample adjustments for high and low concentrations were carried out in three stages. The first sample adjustment was applied when the low and high concentration ratios were under 30 % and the total number of samples was more than 10,000. The oldest oversampled samples were removed, which were lower than the high concentration threshold and higher than the low concentration threshold. The high and low concentration thresholds were 80 and 20 ppb for O_3, and 40 and 10 ppb for NO_2. Second sample adjustment was applied when a low or high concentration ratio is greater than 30%. For high concentration, oversampled samples among all oldest samples were removed and for low concentration, the low concentration samples from the oldest samples were removed. The third sample adjustment was applied when the low or high concentration ratio is greater than 30% after the above two sample adjustments. The oldest accumulated high concentration samples during the cumulative period and the low concentration samples from the samples immediately before leaving were removed. When the low and high concentration ratio is under 30%, the high and low concentration samples were accumulated and discarded at the end of the cumulative period.

3.5 Land and ocean modeling

In the case of existing ground-level air pollutant concentration estimation studies, only information on land areas is presented. The air pollutants move according to the flow of the atmosphere and are emerging as a problem in various countries. Estimating the concentration of ground-level air pollutants at sea can help identify their trend of movement of them. Therefore, in this study, continuous ground-level NO_2 and O_3 concentrations were calculated not only on land but also at sea.

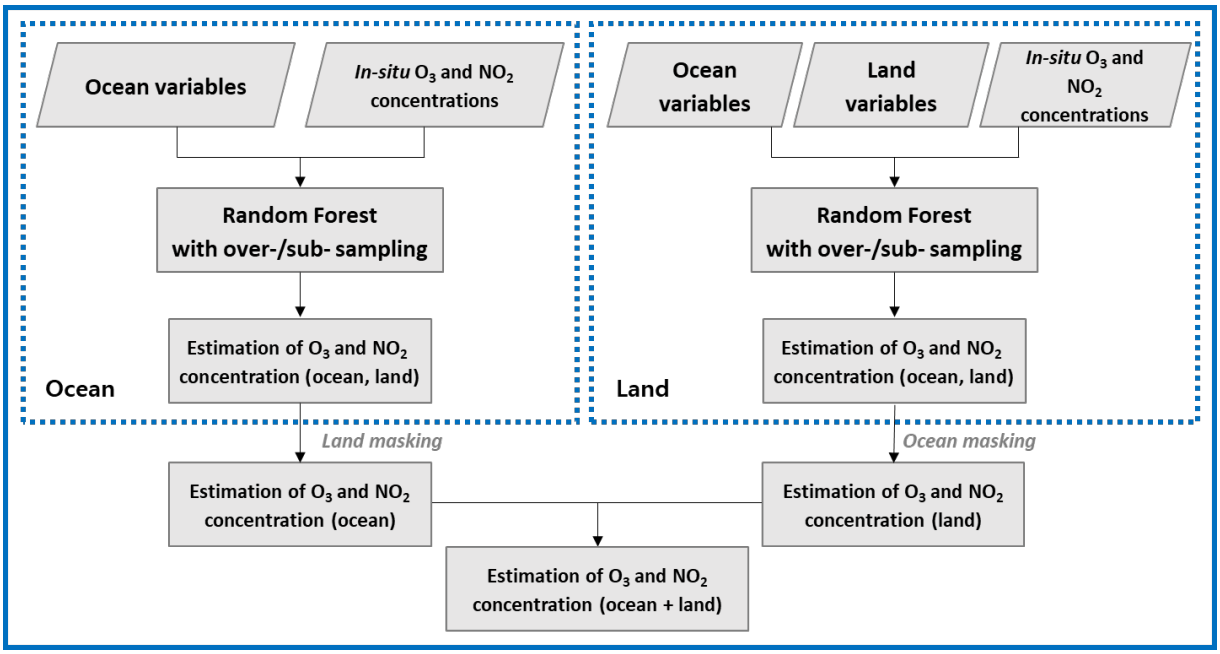

Figure 7. Flowchart for estimating ground-level NO_2 and O_3 concentrations over both land and ocean.

Input variables used to estimate ground NO_2 and O_3 concentrations can be divided into variables that provide values only for land (land variables) and variables that provide values for both land and ocean (ocean variables). The land variables used in this study are NDVI, land cover ratio variables, population density, road density, and DEM data, and the other variables are marine variables. Since the land variable has a NaN value in the ocean, the concentration value cannot be obtained even in the estimated ground concentration result.

Figure 7 shows the flow chart of the algorithm for calculating concentrations of air pollutants for the land and the ocean. In this study, ground-level air pollutant concentration calculation models (Scheme 3) including the ocean were developed by fusing a model (Scheme 1) constructed with ocean variables and a model (Scheme 2) constructed with both ocean and land variables. In Schemes 1 and 2,

over-/sub-sampling was applied based on the variables corresponding to each scheme, and then the ground-level concentrations were estimated using the random forest. Scheme 3 did not build a separate model but extracted the ocean area of Scheme 1 result and the land area of Scheme 2 result to calculate the ground-level NO_2 and O_3 concentration distribution.

3.6 Model validation

For model validation of the offline model, the previously divided validation dataset (16% of all samples) and prediction dataset (20% of all samples) were used. The coefficient of determination (R^2), root mean squared error (RMSE), and relative RMSE (rRMSE) were used as indicators for validation. Indicators are calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2}$$

$$\mathrm{rRMSE} = \frac{\mathrm{RMSE}}{\bar{y}} \times 100\ \%$$

where y_i is the observed data, y is the mean of the observed data, f_i is an estimated value, and N is the number of observations. The rRMSE is the RMSE normalized by the mean of the observed data, which is useful for comparing results on different scales.

In the case of real-time learning-based models, it is difficult to apply the model validation method (model construction 80% and model validation 20%) used in the existing offline models since all sample data are used for model training. Instead of constructing models using separated calibration and validation datasets, the estimated models were trained with all samples and the model validations were performed with the newly entered sample at that time.

In addition, a station-based 10-fold cross-validation (CV) was performed to validate the RTL-based model. The stations to be used for CV were divided into 10 groups in advance. It was intended to give unity to the 10-fold CV dataset generated every hour.

In the case of ground air pollution concentration, it is difficult to validate marine areas because in-

situ observations are provided only for land areas. For the evaluation of the results of the estimation of air pollutant concentration at sea, the station existing within 1 km of the coastline was considered as a representative value of the observation at sea, and validation for the ocean model was performed using it.

4. Results and discussion

4.1 Comparison of the offline models and the RTL-based models

In this study, the ground-level O_3 and NO_2 concentrations were estimated based on an RTL-based machine learning approach using various satellite outputs and meteorological models. The model was constructed by dividing it into the land model and ocean model according to the use of land variables. For comparison with RTL-based models, estimation using the offline model was performed together. Table 4 shows the model validation results of offline and RTL-based models for O_3 and NO_2 concentrations.

Table 4. Accuracy assessment results of offline and RTL-based model for estimating O_3 and NO_2 concentration

Target	Model	R^2	RMSE (ppb)	rRMSE (%)	Slope	Intercept
O_3	Offline (validation)	0.58	12.10	28.5	0.57	16.27
	RTL (30 days)	0.95	4.78	11.7	0.89	3.31
	RTL (15 days)	0.95	4.76	11.5	0.89	3.22
	RTL (7 days)	0.94	5.23	12.8	0.87	4.15
	RTL (3 days)	0.98	3.22	7.3	0.95	1.28
NO_2	Offline (validation)	0.26	10.20	105.8	0.18	4.98
	RTL (30 days)	0.92	4.09	36.0	0.74	1.84
	RTL (15 days)	0.92	4.03	35.4	9.74	1.79
	RTL (7 days)	0.91	4.34	38.5	0.71	2.03
	RTL (3 days)	0.96	2.37	20.7	0.89	0.92

Figure 8 and Figure 10 show the results of model validation and prediction using the offline model for O_3 and NO_2, respectively. Figure 9 and Figure 11 show the model validation results for O_3 and NO_2 using the RTL-based model with four cumulative periods. The scatterplots were expressed by gathering samples of the coastal station from the ocean model and inland station samples from the land model.

As shown in Figure 8, the prediction result tested for the untrained day was underestimated compared to the model validation result. Although the concentration distribution of the dataset for the offline model was adjusted by over- and sub-sampling, the accuracy of the estimation for the untrained day was poor (R^2 = 0.50, RMSE = 14.08 ppb, rRMSE = 36.0%). In the case of a prediction result of the ground O_3 concentration estimation model, the overestimation and underestimation results were shown in the low concentration and high concentration sections to which oversampling was applied, respectively. The increase in the number of samples in the low and high concentration sections through oversampling greatly contributed to the improvement of accuracy in the model validation, but the degree of contribution to the improvement of accuracy seems to be low in prediction for untrained dates.

The RTL-based models show higher accuracy than the offline model regardless of cumulative periods (Figure 9). For all cumulative periods, R^2 was 0.94 or higher, and rRMSE was 12.8% or less, showing relatively high accuracy. Similar results were obtained for the four cumulative periods. Overall, the shorter the cumulative period, the higher the accuracy tends to increase. The results with a 3-days accumulated dataset showed the best performances with the shortest running time. However, it cannot be said that the accuracy simply improves depending on the cumulative period since the results with 7 days show a poor accuracy. More tests may be needed to optimize the accumulation period.

In the case of NO_2 estimation, the regression slopes are low overall. As for the sample distribution, many NO_2 samples are in the low concentration section. Even considering that the performance of the NO_2 offline model is very poor. The NO_2 estimation using an offline model could hardly be estimated. As a prediction result of the ground NO_2 concentration estimation model, underestimation occurred in the high concentration section. The model validation result shows 0.26 of R^2, 105.8% of rRMSE, and 0.18 of slope (Figure 10). It is necessary to test more oversampling and subsampling schemes when looking at offline models. The oversampling and subsampling schemes have also been determined through several tests, but they seem to need to be further optimized.

The RTL-based models for NO_2 show higher performance than the offline model regardless of cumulative periods (Figure 11). For all cumulative periods, R^2 was 0.91 or higher, and rRMSE was 38.5% or less. It is difficult to say that it is a good performance compared to the results of ozone, but it is a significant result compared to the offline model of NO_2. Similar results were obtained for the four cumulative periods. The shorter the cumulative period, the higher the accuracy tends to increase, except for 7 days. The results with 3 days accumulated dataset showed the best performances with the shortest

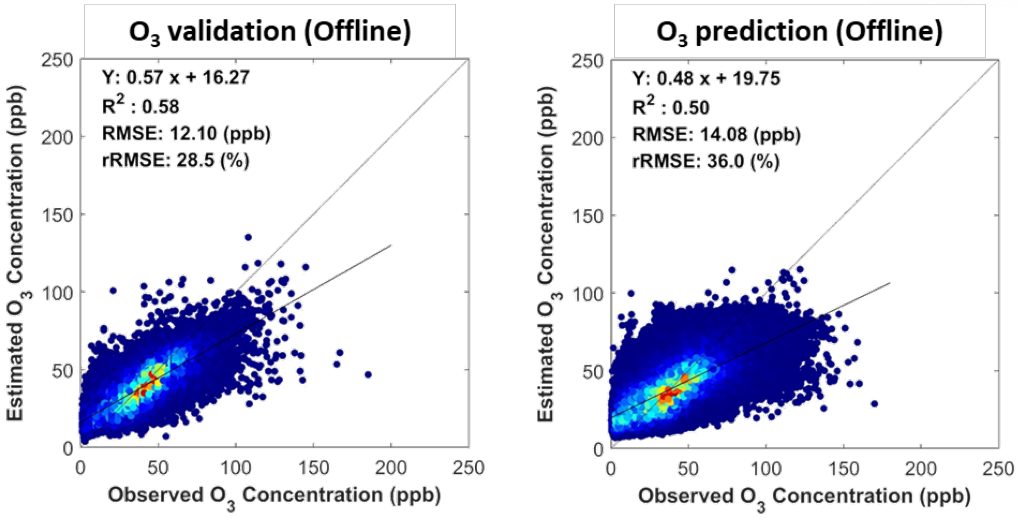

Figure 8. The model validation and prediction results of the O_3 estimation using the offline model.

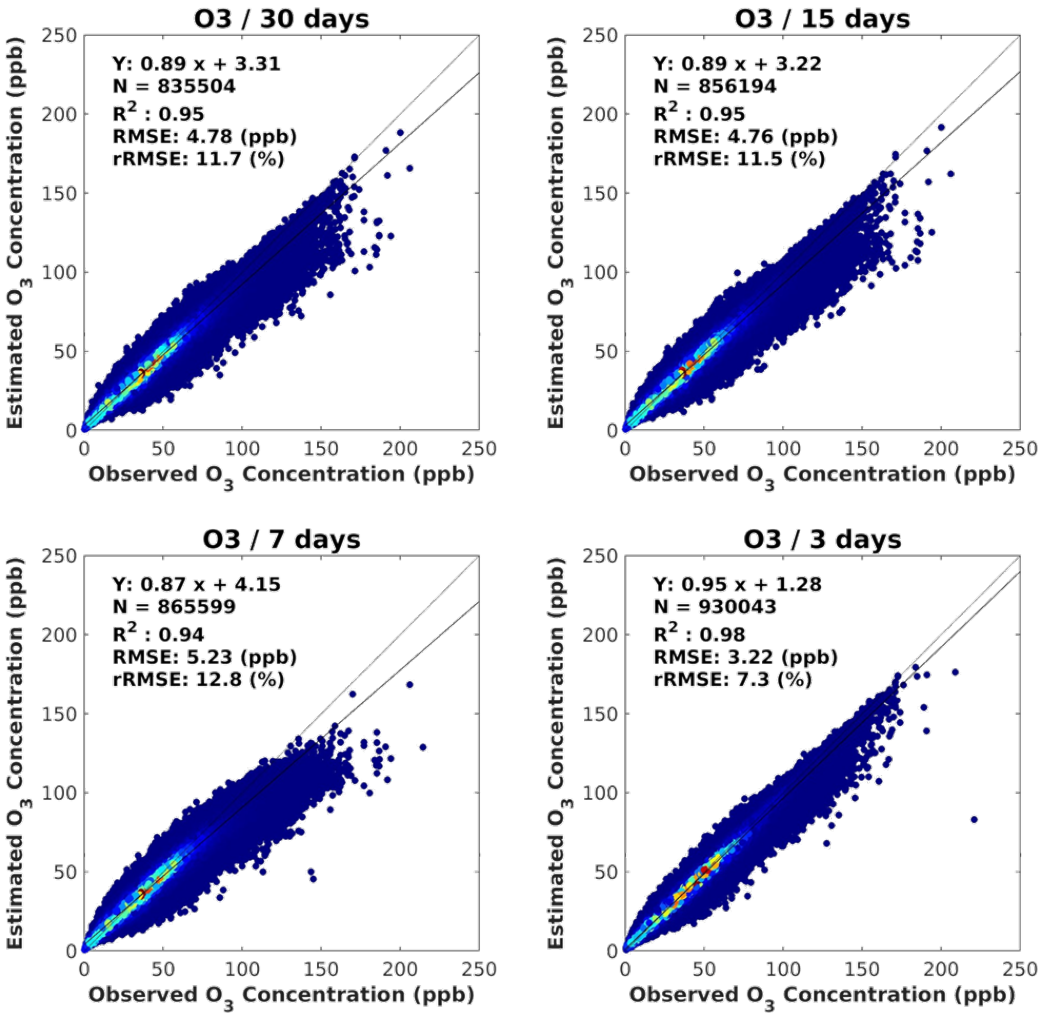

Figure 9. The model validation results of the O_3 estimation using the RTL-based model.

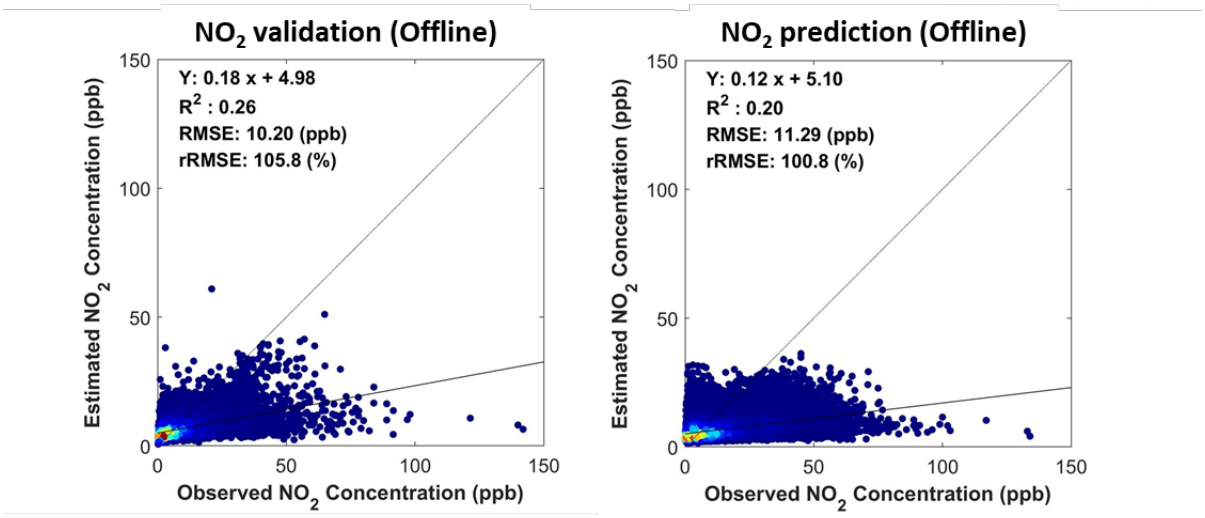

Figure 10. The model validation and prediction results of NO_2 estimation using the offline model.

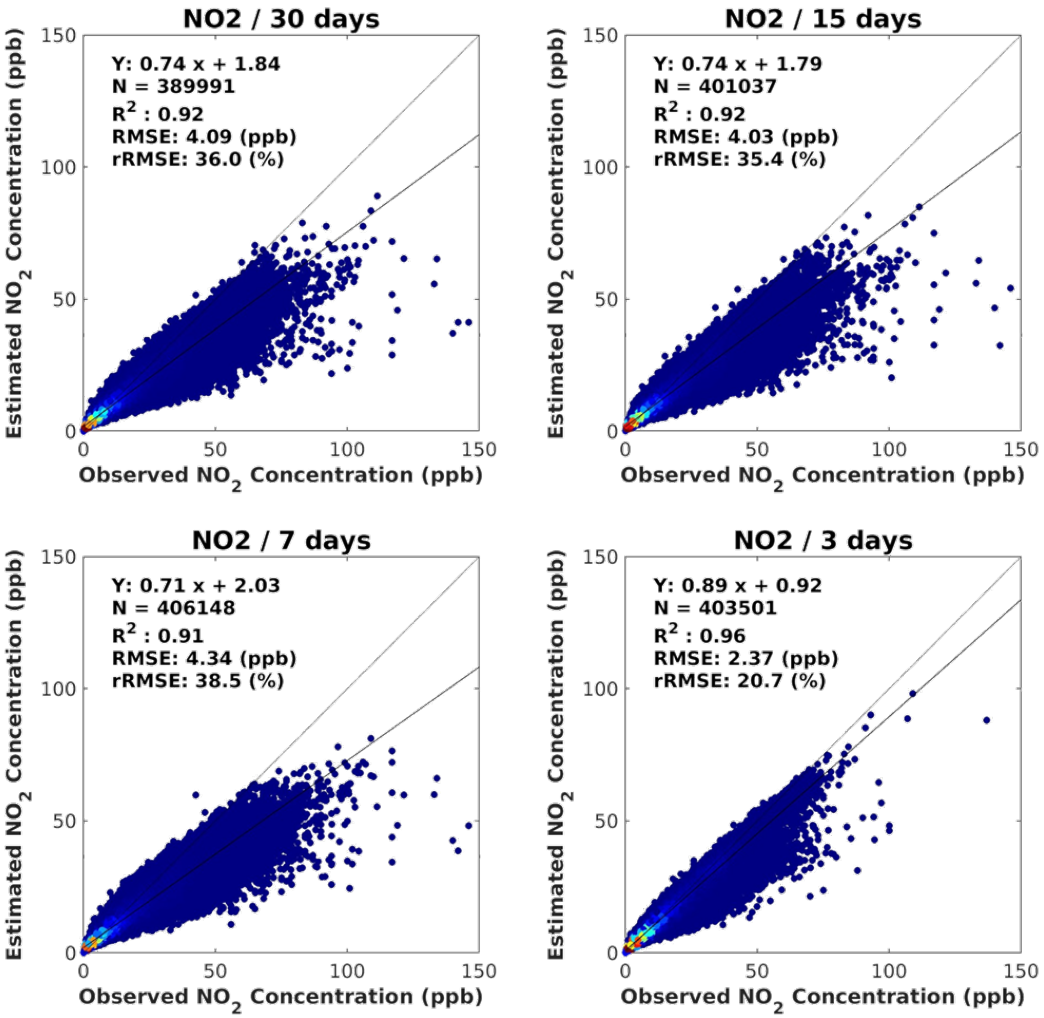

Figure 11. The model validation results of the NO_2 estimation using an RTL-based model

running time.

Both O_3 and NO_3 showed a significant increase in accuracy in RTL-based model results compared to the offline model verification results. Both bias and variance have decreased, which means that the ground concentration estimation model is well simulated for the date when it is not trained.

4.2 Station-based 10-fold cross-validation

4.2.1 Comparison of offline models and the RTL-based models

In the case of real-time learning-based models, it is difficult to apply the model validation method since all sample data are used for model training. The station-based 10-fold cross-validation was performed to validate the RTL-based model. The stations to be used for cross-validation were divided into 10 groups in advance. It was intended to give unity to the 10-fold cross-validation dataset generated every hour. The 10-fold cross-validation results of the offline and RTL-based model for O_3 and NO_2 concentrations are presented in Table 5.

Figure 12 and Figure 13 show the 10-fold cross-validation results for O_3 using the offline model and RTL-based model with four cumulative periods, respectively. The result of the offline model shows a similar distribution of scatterplot to the RTL-based model result with a 30-days cumulative period. The offline model resulted in R^2 values of 0.62, 13.12 ppb of RMSE, and 31.3% of rRMSE. The best results with the 3-day RTL-based model show R^2 values of 0.83, 9.20 ppb of RMSE, and 20.6% of rRMSE. There was a significant improvement in the RTL-based model compared to the offline model. The results of large variance due to over-estimated and under-estimated at medium and high concentrations showed the highest accuracy when the 3-day cumulative period was applied.

In the 30-, 15-, and 3-days models, there are the samples estimated to be similar to the 1:1 line in the high concentration part. But, in the scatterplot of the RTL-based 7-day model, the part was estimated to be low resulting in slightly different plot shapes. Unlike the prediction result of the offline model, the station-based 10-fold cross-validation results of the offline model were not significantly different from the model validation result. It refers that the offline model is robustly constant to the station.

Table 5. The 10-fold cross-validation results of offline and RTL-based model for estimating O_3 and NO_2 concentration

Target	Model	R^2	RMSE (ppb)	rRMSE (%)	Slope	Intercept
O_3	Offline	0.62	13.12	31.3	0.66	12.67
	RTL (30 days)	0.64	12.84	31.3	0.70	10.60
	RTL (15 days)	0.64	12.71	30.7	0.71	10.27
	RTL (7 days)	0.57	13.62	33.5	0.65	12.38
	RTL (3 days)	0.83	9.20	20.6	0.86	4.77
NO_2	Offline	0.37	9.32	82.2	0.32	5.33
	RTL (30 days)	0.44	8.91	77.1	0.42	5.00
	RTL (15 days)	0.43	8.79	77.0	0.42	5.05
	RTL (7 days)	0.35	9.33	82.8	0.35	5.65
	RTL (3 days)	0.71	6.02	51.0	0.70	3.04

Figure 14 and **오류! 참조 원본을 찾을 수 없습니다.** show the 10-fold cross-validation results for NO_2 using the offline model and RTL-based model with four cumulative periods, respectively. The result of the offline model shows a similar distribution of scatterplot to RTL-based model result with 30- and 15-day cumulative period. Unlike the model validation results, the variances were large in the 10-fold cross-validation results of the RTL-based model for NO_2. The offline model resulted in R^2 values of 0.37, 9.32 ppb of RMSE, and 82.2% of rRMSE. The 30-day RTL-based model shows R^2 values of 0.44, 8.91 ppb of RMSE, and 77.1% of rRMSE. Although the RTL models were improved from the offline model, they were still less accurate on the 30-, 15-, and 7-day models. The samples were largely underestimated in the high concentration section and were overestimated in the low concentration section. However, the 3-day model shows the improved accuracy resulted in R^2 values of 0.71, 6.02 ppb of RMSE, and 51.0% of rRMSE. Its slope was increased, and the variance was decreased.

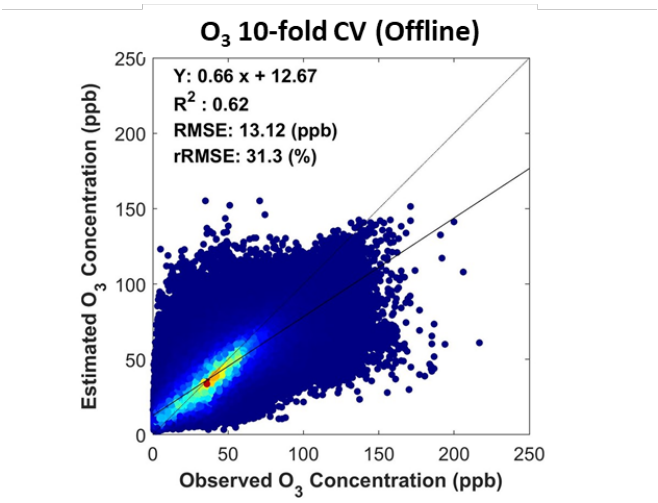

Figure 12. The 10-fold cross-validation results of the O_3 estimation using the offline model.

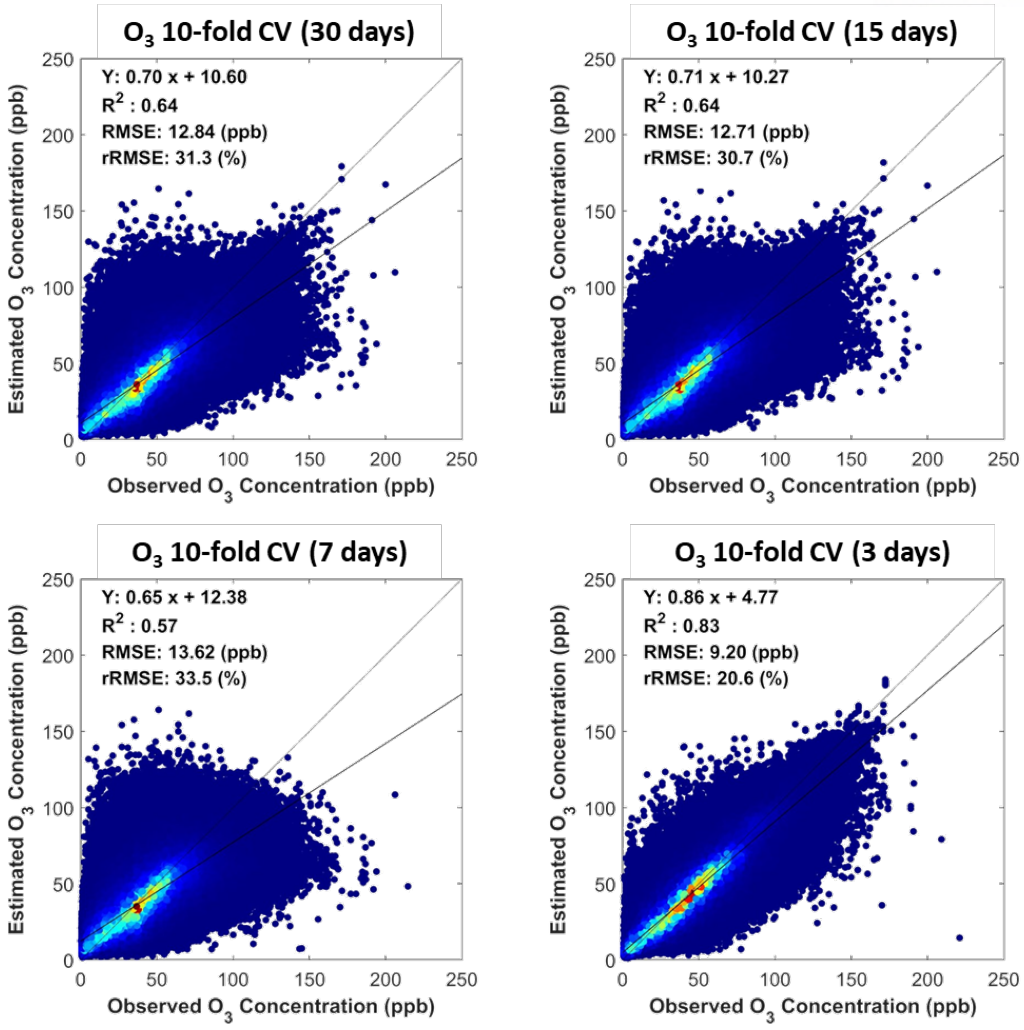

Figure 13. The 10-fold cross-validation results of the O_3 estimation using the RTL-based model.

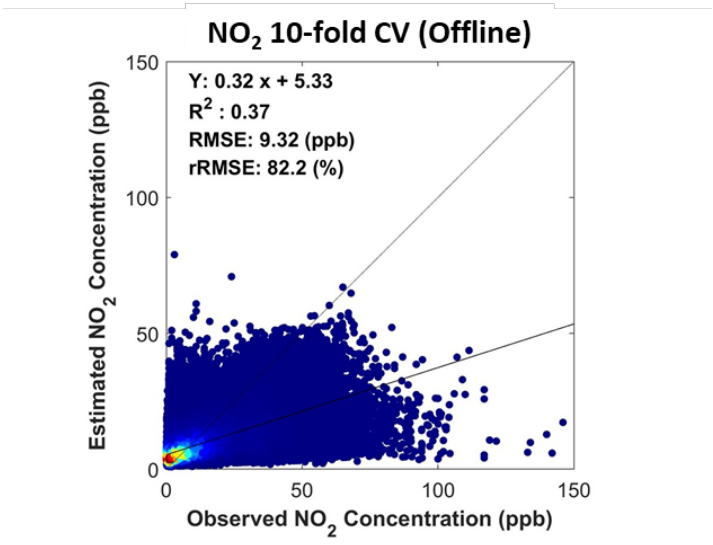

Figure 14. The 10-fold cross-validation results of the NO_2 estimation using the offline model.

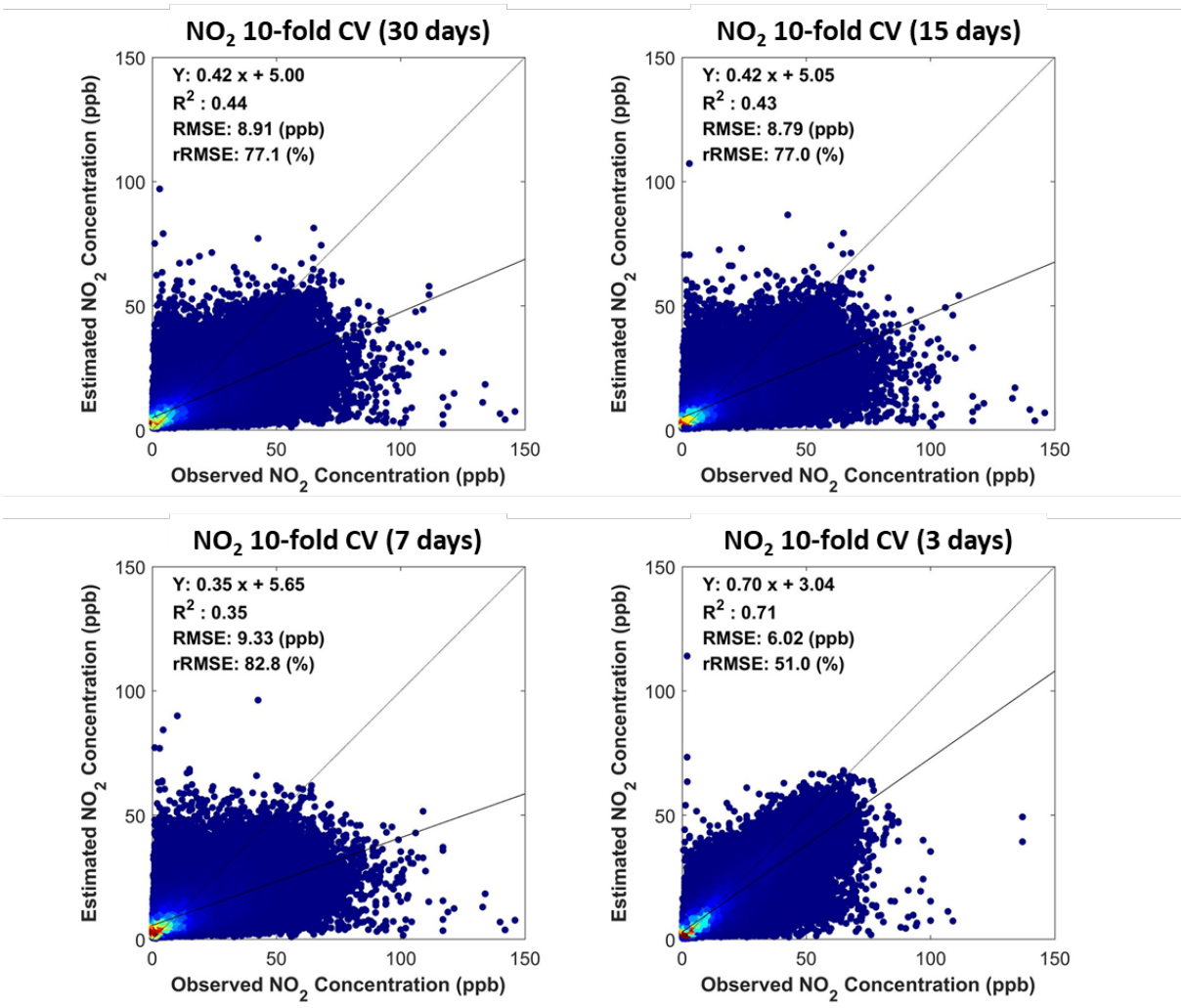

Figure 15. The 10-fold cross-validation results of the NO_2 estimation using the RTL-based model.

4.2.2 Comparison of ocean model and land model

The RTL-based model results for the coastline and the 10-fold cross-validation results were checked together. The results for ground-level O_3 concentration are shown in Figure 16 and for NO_2 concentration are shown in Figure 17. In the figure, the station data located within 1 km of the coast are shown in red triangles and other stations in blue circles. The black indicators in the lower right corner of each figure are the accuracy of the combined models including the coastal samples from the ocean model and inland samples from the land model.

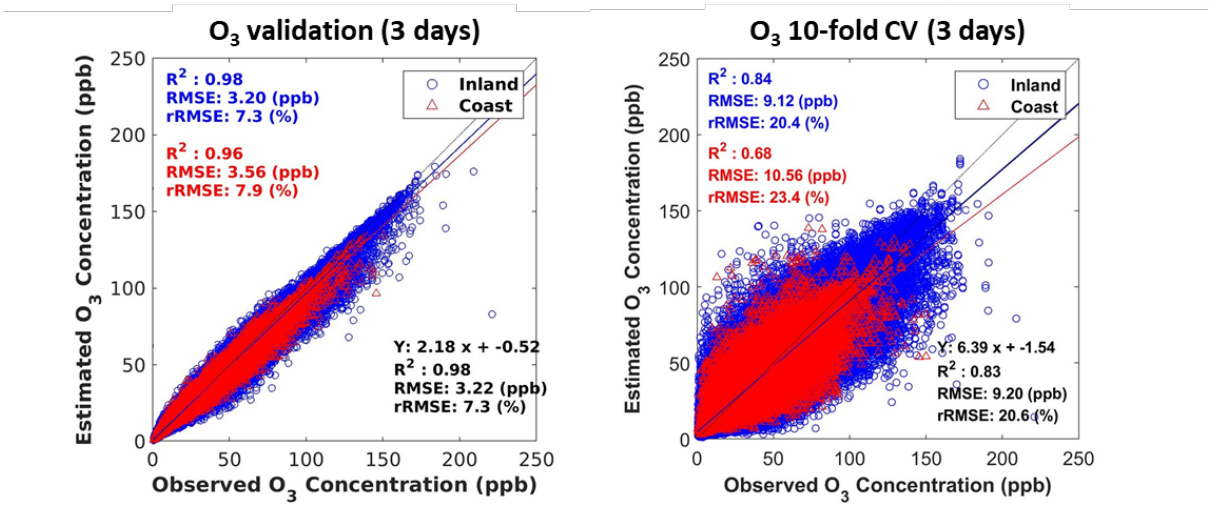

Figure 16. Model validation and 10-fold cross-validation results of RTL-based O_3 estimation model with inland and coast samples

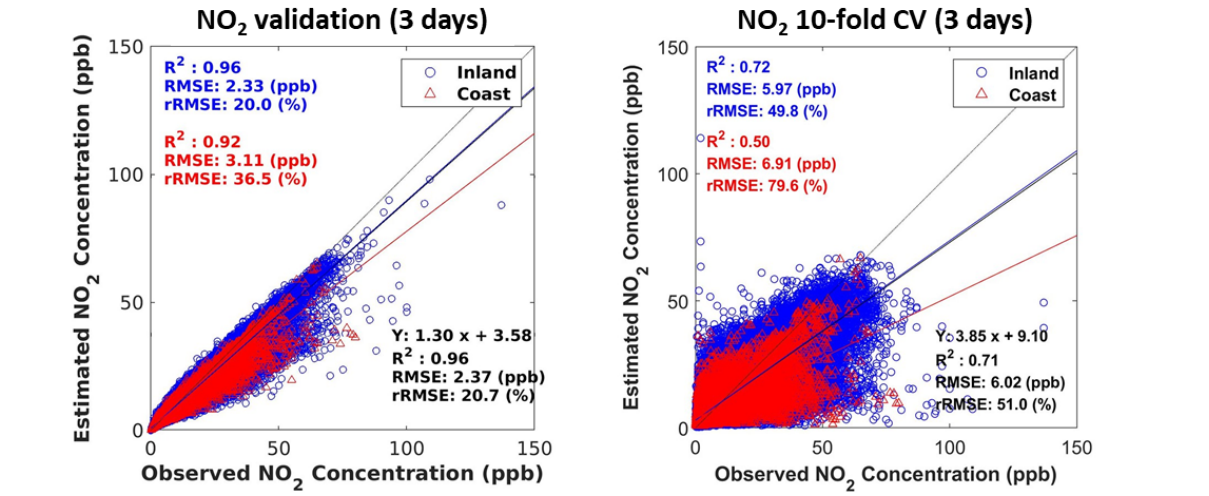

Figure 17. Model validation and 10-fold cross-validation results of RTL-based NO_2 estimation model with inland and coast samples

The validation result using only the station located on the coast and only the station located in the rest of the region show a similar distribution overall, and the accuracy is also similar. The coastal samples in the 10-fold cross-validation result of the ocean model have a relatively high rRMSE value because the concentration distribution range is narrower than that of the inland samples from the land model. In the case of O_3, compared to the coastal samples from the ocean model and inland samples from the land model, the distribution of samples concentrated on the 1:1 line.

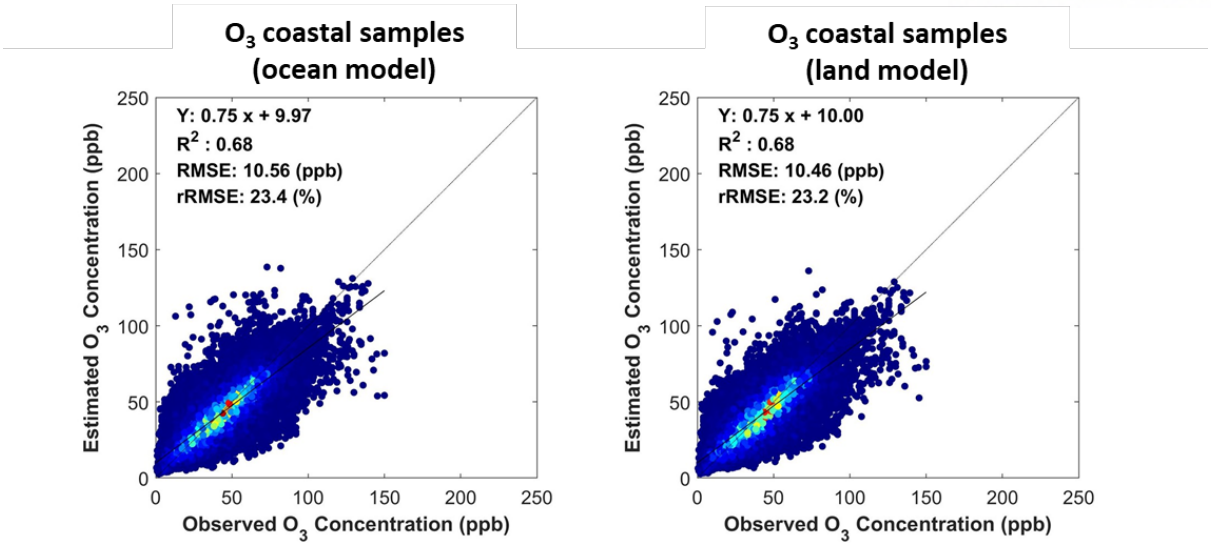

Figure 18. The 10-fold cross-validation results of coastal samples from RTL-based ocean model and land model for O_3

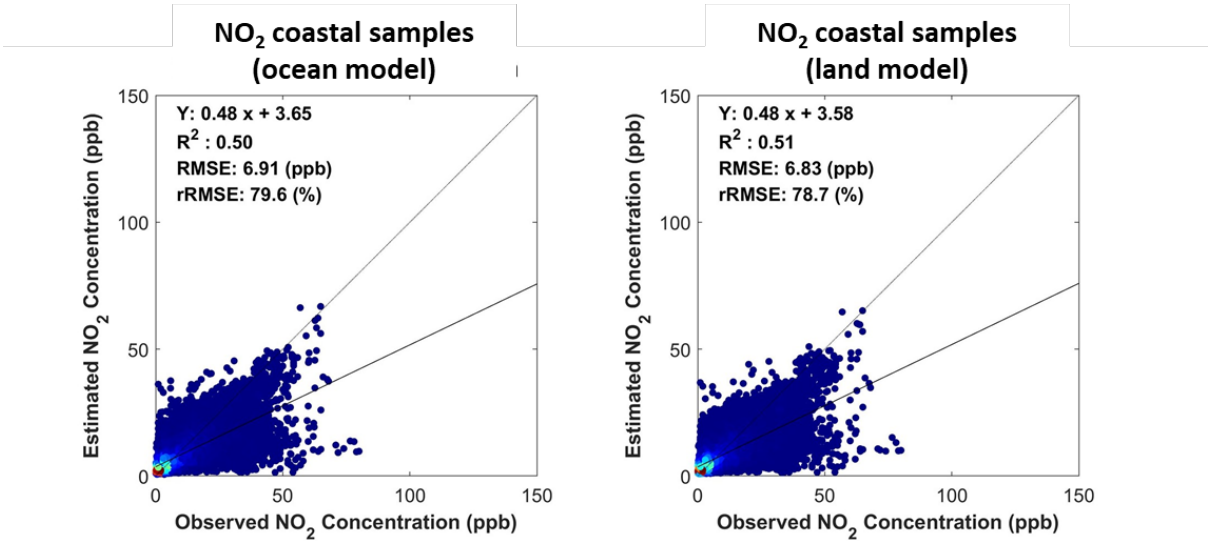

Figure 19. The 10-fold cross-validation results of coastal samples from RTL-based ocean model and land model for NO_2

In the RTL-based model results, the accuracy difference of model validation results between the coastal samples and inland samples were similar. On the other hand, in the 10-fold cross-validation results, the accuracy in coast station samples was slightly lower than that of inland station samples. It is regarded as an error caused by a relatively small number of samples.

In the case of ocean models, it was expected to be simulated well at coastal stations because it was a model excluding land variables. But both inland and coastal stations showed lower performance than the land model. The model results of O_3 showed that the performance of ocean model and land model

were almost similar (ocean model: R^2 = 0.68, RMSE = 10.56 ppb, rRMSE = 23.4% / land model: R^2 = 0.68, RMSE = 10.46 ppb, rRMSE = 23.2%). In the case of NO_2, the effect of land variables is greater than that of O_3, so the difference in accuracy between ocean and land model seems to be larger (ocean model: R^2 = 0.50, RMSE = 6.91 ppb, rRMSE = 79.6% / land model: R^2 = 0.51, RMSE = 6.83 ppb, rRMSE = 78.7%) (Figure 18, Figure 19).

Although the validation was performed using a 1km coastal station, it is difficult to say that these station data are directly representative of the observation of the ocean. This is because the influence of the cities near the coast is not small, and they are inland stations actually. It may be necessary to extract and analyze only stations located on the island away from the inland.

4.3 Spatial and temporal distribution

4.3.1 Annual map

Figure 20 and Figure 21 are annual mean concentration maps calculated using the RTL-based ground-level O_3 and NO_2 concentration estimation models. The study period is not included from January to December every year. So the period used on annual mean map was adjusted. For 2018, the hourly estimated results from June 2018 to February 2019 were used. The results from March 2019 to February 2020 were used for the annual average of 2019, and the results from March 2020 to February 2021 were used for 2020.

Over the years, the O_3 concentration tends to increase overall. In the case of the map of 2018, data from March to May, the spring period, were not included. The 2018 O_3 map showed a lower average concentration value than in other years since the O_3 tends to have a high concentration in spring and summer. In the case of NO_2, the overall value was low in the 2019 annual mean map, including the winter of 2019 (December 2019 to February 2020), when the pandemic occurred due to the outbreak of COVID-19.

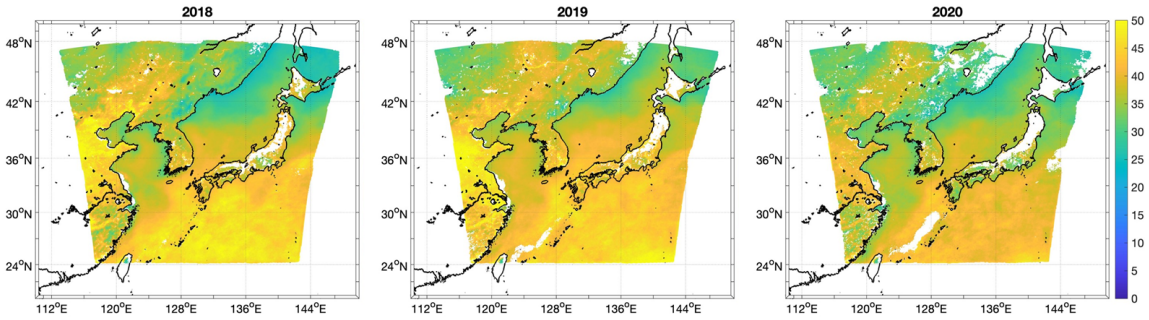

Figure 20. Annual map of O_3 estimation using RTL-based model (3 days)

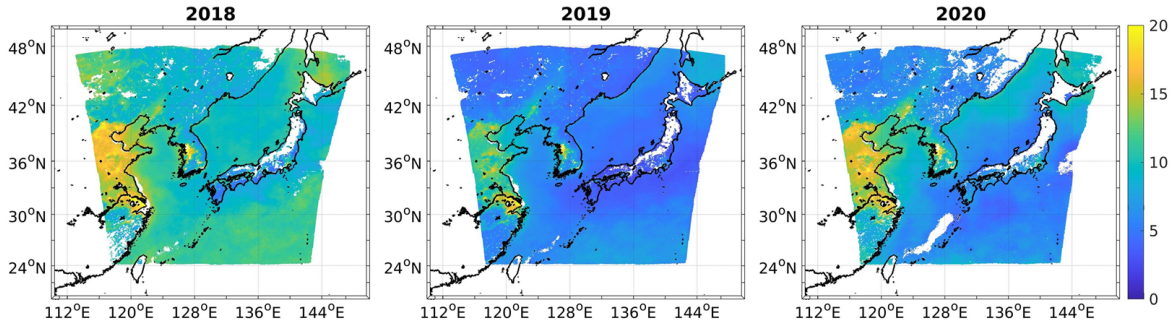

Figure 21. Annual map of NO_2 estimation using RTL-based model (3 days)

4.3.2 Seasonal map

Figure 22 and Figure 23 are the maps showing the spatial distribution of seasonal average about estimated concentrations of ground O_3 and NO_2 from 2018 summer to 2020 winter (including January and February 2021).

In the O_3 estimation model, the difference between the ocean and land model was not as large as the model validation result. The overall result of combining the two models showed similar results. The distribution of O_3 concentration at sea was confirmed on the map with seasonal fluctuations. In particular, the spatial distribution of high concentration values on the east coast of Japan in spring was shown. This is a spatial distribution characteristic that was difficult to identify if ground-level O_3 concentration estimation was not performed at sea. The cause of this phenomenon needs to be investigated further. In the seasonal average concentration distribution map of O_3, the spatial pattern was not prominent, but the seasonal characteristics of the high concentration value in spring and summer were clearly shown.

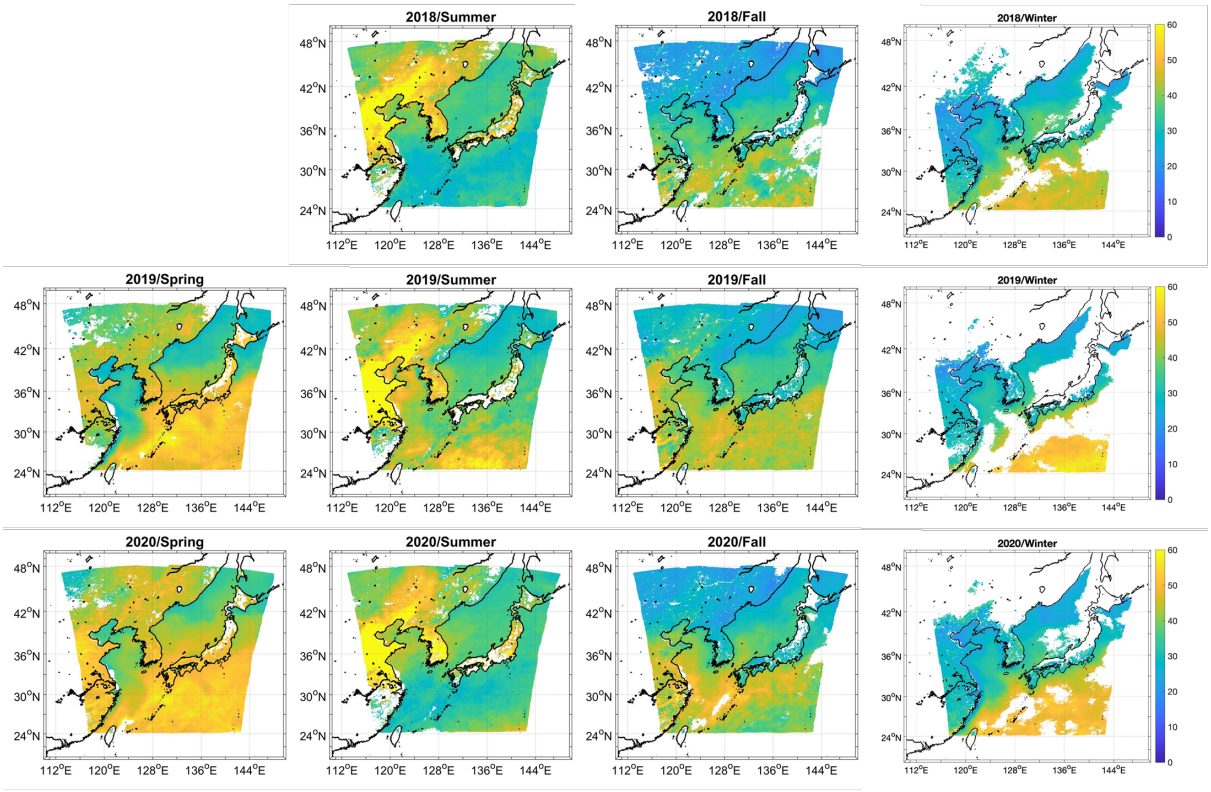

Figure 22. Seasonal map of O_3 estimation using RTL-based model (3 days)

In the case of NO_2, seasonal and spatial patterns were evident on land, and there was no significant change in concentration value at sea. In all models, seasonal patterns showing high concentrations in autumn and winter and spatial patterns showing high concentration values in urban areas were well simulated. In the results of the ocean model, land variables were excluded. The estimation of ground-level NO_2 concentration on land was underestimated.

Using the estimating model of ground-level air pollutant concentration, it is expressed spatially continuous ground-level concentrations including land as well as sea. Through this, it is possible to monitor the pattern of movement of the concentration of air pollutants. The hourly concentration map can be produced, but it is difficult to see the spatial pattern of every hour due to the existence of clouds. In the case of GOCI data, it contributed to improving the accuracy of the model. But in the spatial distribution analysis, the number of empty samples was very large due to GOCI, making it difficult to analyze the space. For spatial distribution, the analysis may be required using results excluding GOCI data. Using Gap-filling satellite data would produce more spatially continuous results. If the ground concentration of all-sky is estimated through a more developed model, its utilization will be even higher.

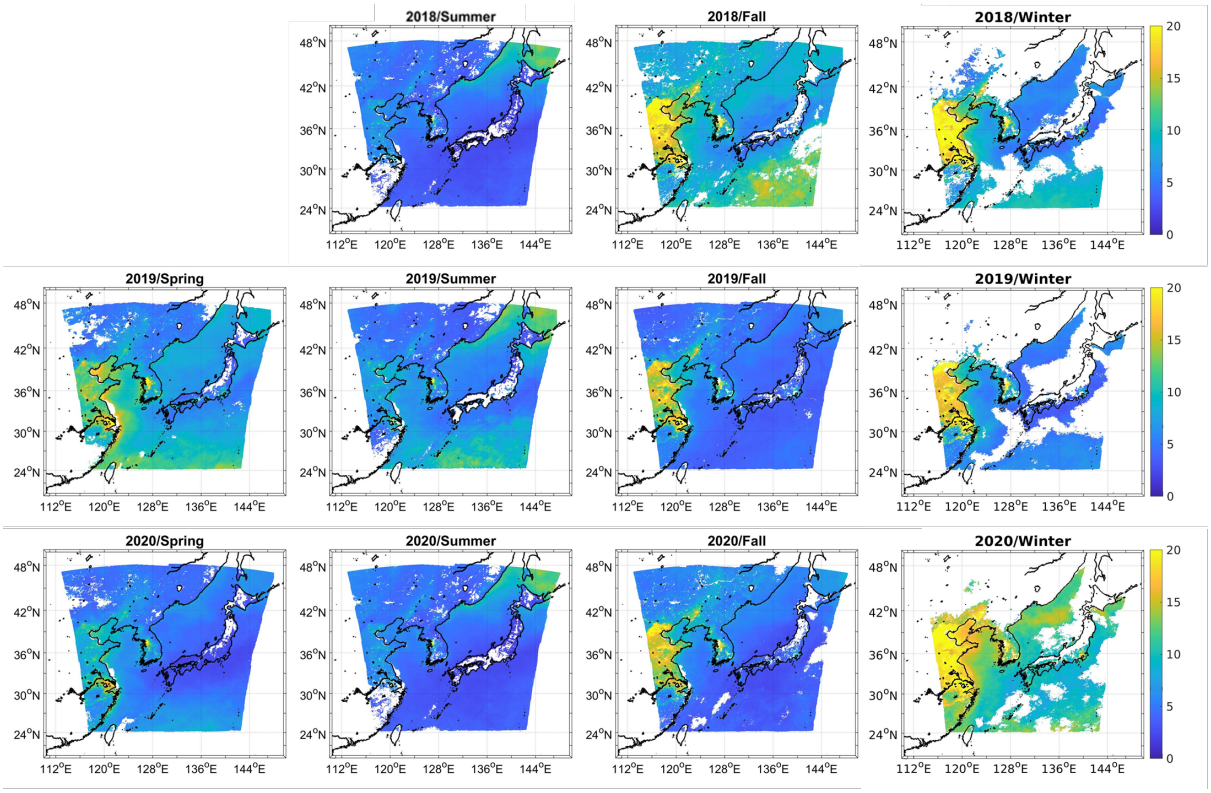

Figure 23. Seasonal map of NO_2 estimation using RTL-based model (3 days)

5. Conclusions

In this study, the ground-level O_3 and NO_2 concentration estimation model was developed using the RTL-based machine learning technique with various satellite data and numerical model data as input variables. Among the input variables, some variables did not provide pixel values at sea. For the spatially continuous distribution of O_3 and NO_2 concentration, the estimation models were constructed over both land and sea. Three models were tested to build an accurate model using the most available data.

The study was conducted on the ocean model using only ocean variables which have values for all regions; land model using all available data which could estimate only over the land; and combined model that combines the results of the ocean model for sea area and the results of the land model for land area. The results of the land model showed higher accuracy than the ocean model. But the uncertainty is greater since the estimates at sea calculated from land models are generated by assigning constant values to some variables. Considering this problem, the combined model was developed in estimating the ground-level concentration over land and ocean.

The studies for estimating the trend of satellite vertical column density data and estimating ground-level concentration using a CTM-based ratio were conducted previously. To evaluate the harmfulness of O_3 and NO_2, a more accurate ground-level concentration is necessary. The model developed through this study can be valuable in evaluating the harmfulness of O_3 and NO_2 on the ground. An error analysis is needed to make good use of the results of the developed model in the field. In addition, when applying these RTL-based ground-level O_3 and NO_2 concentration estimation models to the field, it will helpful for the public to prepare for the damage from air pollutants by continuously producing a ground-level concentration of air pollutants in semi-real time.

In this study, the ground-level air pollutant concentrations were estimated for 8 times a day based on the TROPOMI and GOCI satellite data. Although the estimated spatial distribution can be produced every hour, it may be difficult to monitor changes over time. Because TROPOMI is a polar orbit satellite data, is provided once a day. In addition, the GOCI data is sometimes provided with spatially omitted pixels due to cloud issues. It is expected that more accurate and immediate real-time monitoring will be possible when utilizing geostationary environmental satellite data such as GEMS which provides hourly data 10 times a day.

References

Bechle, M. J., Millet, D. B., & Marshall, J. D. (2015). National Spatiotemporal Exposure Surface for NO2: Monthly Scaling of a Satellite-Derived Land-Use Regression, 2000–2010. *Environmental Science & Technology*, *49*(20), 12297-12305. https://doi.org/10.1021/acs.est.5b02882

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Cooper, M. J., Martin, R. V., McLinden, C. A., & Brook, J. R. (2020). Inferring ground-level nitrogen dioxide concentrations at fine spatial resolution applied to the TROPOMI satellite instrument. *Environmental Research Letters*, *15*(10), 104013. https://doi.org/10.1088/1748-9326/aba3a5

de Hoogh, K., Gulliver, J., van Donkelaar, A., Martin, R. V., Marshall, J. D., Bechle, M. J., Cesaroni, G., Pradas, M. C., Dedele, A., & Eeftens, M. (2016). Development of West-European PM2.5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental research*, *151*, 1-10.

Freddy Grajales, J., & Baquero-Bernal, A. (2014). Inference of surface concentrations of nitrogen dioxide (NO2) in Colombia from tropospheric columns of the ozone measurement instrument (OMI). *Atmósfera*, *27*(2), 193-214. https://doi.org/https://doi.org/10.1016/S0187-6236(14)71110-5

Hoek, G., Eeftens, M., Beelen, R., Fischer, P., Brunekreef, B., Boersma, K. F., & Veefkind, P. (2015). Satellite NO2 data improve national land use regression models for ambient NO2 in a small densely populated country. *Atmospheric environment*, *105*, 173-180. https://doi.org/https://doi.org/10.1016/j.atmosenv.2015.01.053

Jiang, Q., & Christakos, G. (2018). Space-time mapping of ground-level PM2.5 and NO2 concentrations in heavily polluted northern China during winter using the Bayesian maximum entropy technique with satellite data. *Air Quality, Atmosphere & Health*, *11*(1), 23-33.

Kharol, S., Martin, R., Philip, S., Boys, B., Lamsal, L., Jerrett, M., Brauer, M., Crouse, D., McLinden, C., & Burnett, R. (2015). Assessment of the magnitude and recent trends in satellite-derived ground-level nitrogen dioxide over North America. *Atmospheric environment*, *118*, 236-245.

Knibbs, L. D., Hewson, M. G., Bechle, M. J., Marshall, J. D., & Barnett, A. G. (2014). A national satellite-based land-use regression model for air pollution exposure assessment in Australia. *Environmental research*, *135*, 204-211. https://doi.org/https://doi.org/10.1016/j.envres.2014.09.011

Liu, F., Eskes, H., Ding, J., & Mijling, B. (2018). Evaluation of modeling NO2 concentrations driven by satellite-derived and bottom-up emission inventories using in situ measurements over China. *Atmospheric Chemistry and Physics*, *18*(6), 4171-4186.

Meijer, J. R., Huijbregts, M. A., Schotten, K. C., & Schipper, A. M. (2018). Global patterns of current and future road infrastructure. *Environmental Research Letters*, *13*(6), 064006.

Meng, K., Xu, X., Cheng, X., Xu, X., Qu, X., Zhu, W., Ma, C., Yang, Y., & Zhao, Y. (2018). Spatio-temporal variations in SO2 and NO2 emissions caused by heating over the Beijing-Tianjin-Hebei Region constrained by an adaptive nudging method with OMI data. *Science of the Total Environment*, *642*, 543-552. https://doi.org/https://doi.org/10.1016/j.scitotenv.2018.06.021

National Research Council. (1991). *Rethinking the Ozone Problem in Urban and Regional Air Pollution*. The National Academies Press. https://doi.org/doi:10.17226/1889

Oner, E., & Kaynak, B. (2016). Evaluation of NOx emissions for Turkey using satellite and ground-based observations. *Atmospheric Pollution Research*, *7*(3), 419-430. https://doi.org/https://doi.org/10.1016/j.apr.2015.10.017

Qin, K., Rao, L., Xu, J., Bai, Y., Zou, J., Hao, N., Li, S., & Yu, C. (2017). Estimating Ground Level NO2 Concentrations over Central-Eastern China Using a Satellite-Based Geographically and Temporally Weighted Regression Model. *Remote Sensing*, *9*(9), 950.

TIBCO. (2022). *what is a random forest*. Retrieved May 15 from https://www.tibco.com/reference-center/what-is-a-random-forest

U.S. EPA. (2016). *Integrated Science Assessment (ISA) For Oxides of Nitrogen – Health Criteria (Final Report)*. (EPA/600/R-15/068). Washington, DC: U.S. Environmental Protection Agency

U.S. EPA. (2020). *Integrated Science Assessment (ISA) for Ozone and Related Photochemical Oxidants (Final Report)*. (EPA/600/R-20/012). Washington, DC: U.S. Environmental Protection Agency

World Health Organization (WHO). (2006). Air Quality Guidelines—Global Update 2005. In: WHO Copenhagen.

Xu, H., Bechle, M. J., Wang, M., Szpiro, A. A., Vedal, S., Bai, Y., & Marshall, J. D. (2019). National PM2.5 and NO2 exposure models for China based on land use regression, satellite measurements, and universal kriging. *Science of the Total Environment*, *655*, 423-433. https://doi.org/https://doi.org/10.1016/j.scitotenv.2018.11.125

Yeganeh, B., Hewson, M. G., Clifford, S., Tavassoli, A., Knibbs, L. D., & Morawska, L. (2018). Estimating the spatiotemporal variation of NO2 concentration using an adaptive neuro-fuzzy inference system. *Environmental Modelling & Software*, *100*, 222-235.

Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M. L., & Di, B. (2018). Satellite-Based Estimates of Daily NO2 Exposure in China Using Hybrid Random Forest and Spatiotemporal Kriging Model. *Environmental Science & Technology*, *52*(7), 4180-4189.

Zhang, Y., Wang, Y., Crawford, J., Cheng, Y., & Li, J. (2018). Improve observation-based ground-level ozone spatial distribution by compositing satellite and surface observations: A

simulation experiment. *Atmospheric environment*, *180*, 226-233. https://doi.org/https://doi.org/10.1016/j.atmosenv.2018.02.044

Zhang, Z., Wang, J., Hart, J. E., Laden, F., Zhao, C., Li, T., Zheng, P., Li, D., Ye, Z., & Chen, K. (2018). National scale spatiotemporal land-use regression model for PM2.5, PM10 and NO2 concentration in China. *Atmospheric environment*, *192*, 48-54. https://doi.org/https://doi.org/10.1016/j.atmosenv.2018.08.046

Zoogman, P., Jacob, D. J., Chance, K., Worden, H. M., Edwards, D. P., & Zhang, L. (2014). Improved monitoring of surface ozone by joint assimilation of geostationary satellite observations of ozone and CO. *Atmospheric environment*, *84*, 254-261. https://doi.org/https://doi.org/10.1016/j.atmosenv.2013.11.048