First-time estimation of global near-surface methane mixing ratio distributions using machine learning with the TROPOMI data

Dagyo Lee(blessed_bunny@naver.com)[1], Wonei Choi[2], and Hanlim Lee[3*]

1 Pukyong National University, Division of Earth and Enviromental System Sciences Department of Spatial Information Engineering 2 NASA Goddard Space Flight Center 3 Pukyong National University Department of Spatial Information Engineering

Abstract

Although the total amount of methane (CH4) is much smaller than that of carbon dioxide (CO2), methane is known to accelerate global warming due to its high Global Warming Potential (GWP) and recent increasing trend. While satellites can observe total column density, surface methane mixing ratios are largely influenced by fresh emissions. Global distribution data of surface methane mixing ratios can be valuable for identifying methane source locations worldwide and for accurately estimating methane emissions. In this study, we used TROPOMI methane data from 2018 to 2024 as the main independent variable, and surface methane mixing ratio data from WDCGG sites for the same period as the dependent variable, to train a Random Forest model. The model was then used to estimate the global distribution of surface methane mixing ratios from 2018 to 2024, using data that was not included in the training. When comparing the global model estimates using all WDCGG sites with surface methane concentrations not used in the training, the models showed similar performance with a correlation coefficient (R) of 0.70 and a root mean square error (RMSE) of 0.034 ppm. The previously significant performance gap seen in local models using about three years of data samples was improved by extending the study period, resulting in consistent performance across all models. This underscores the importance of a large amount of global training data in enhancing the performance of surface methane estimation models, particularly when WDCGG data is insufficient.

Data

- **Research Scope** - TROPOMI, ERA5 : Global (180°, -90°, 180°, 90°)
  - WDCGG : all methane observation sites
- **Study Period** - 2018. 05. 01 – 2024. 05. 31

- TROPOMI instrument

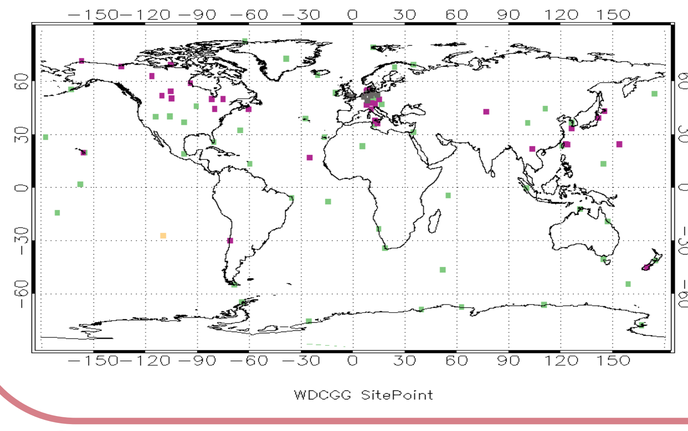
Spectral range / Spectral resolution(FWHM)	UV-VIS	270 - 500 nm / 0.5 nm
	NIR	675 - 775 nm / 0.5 nm
	SWIR	2305 - 2385 / 0.23 nm
Signal to noise ratio(SNR)		100(SWIR) – 1200(UV – VIS)
Swath		2600 km
Spatial resolution		7 × 7 km² / 5.5 × 7 km²
Temporal resolution		1 day

- ERA5

Spatial resolution	0.25° × 0.25°
Temporal resolution	1 hour

- WDCGG

locations of the WDCGG methane sites



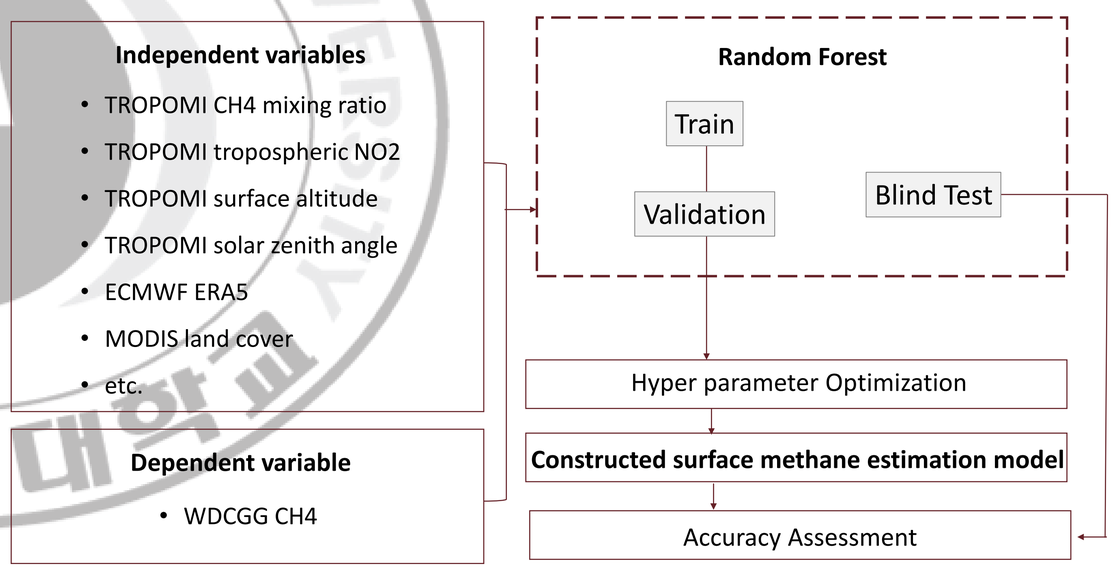
WDCGG Spatial and Temporal Resolution

CH_4 in-situ	
Temporal resolution	1 hour

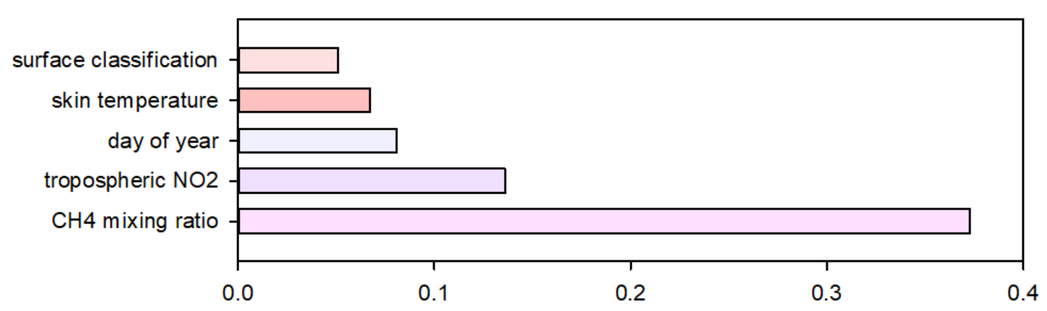
Platform
Surface:
Tower:

Method

- **Machine Learning Model Development Flowchart**
  - Outlier removal and resolution adjustment processes are performed on all input data prior to model development.
  - For each WDCGG in situ station, any day with missing satellite data or ERA5 meteorological data was excluded from the analysis.
  - The total number of data samples is 5,438.

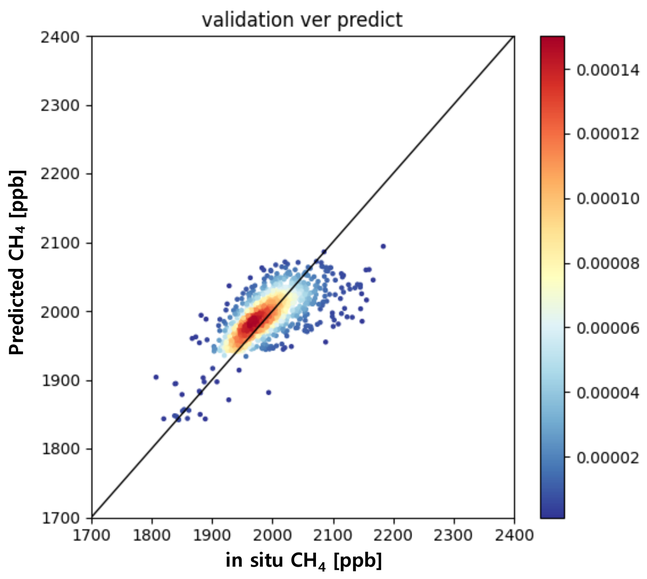


- **Feature importance**

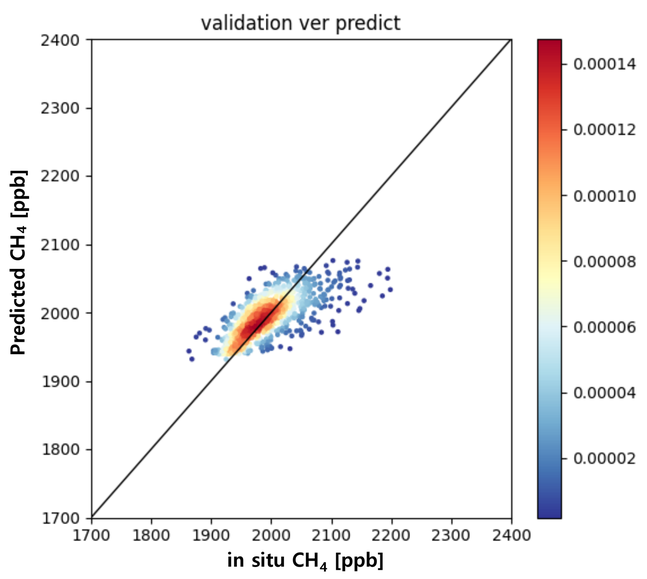


Result

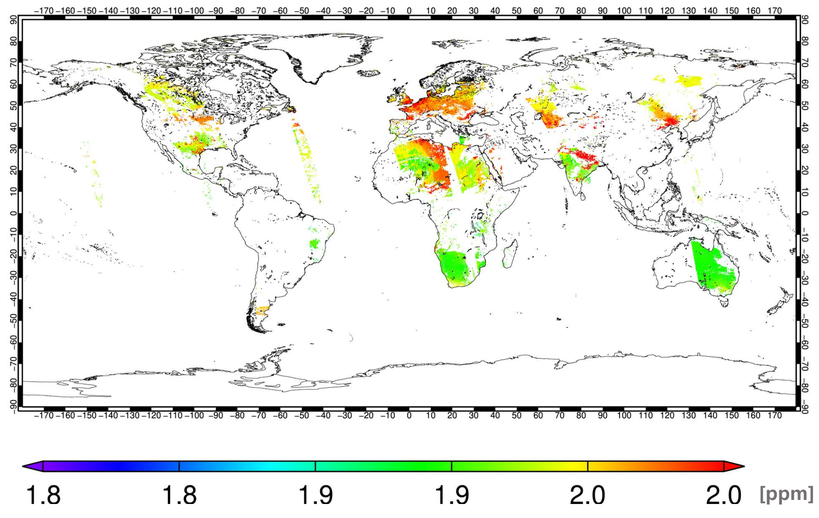
- **Global model**



- **Local model (Canada & WestEurope)**



- **Spatial distribution of the near-surface methane mixing ratio estimated by the global model** (2018.05.06)



Conclusion

By extending the study period from the original (January 1, 2019, to December 31, 2022), the total number of data samples increased from 2,573 to 5,529, leading to noticeable differences in the study results. The performance of the Global model, the Canada & West Europe model, and the West Europe model remained consistent. However, the Canada model, which initially showed relatively lower performance, saw a significant improvement in the coefficient of determination (R^2), increasing from 0.31 to 0.50. This enhancement enabled us to develop a model with consistent performance across all regions. The importance of the independent variables is ranked as follows: 'CH_4 mixing ratio', 'Tropospheric NO_2', 'Day of Year (DOY)', 'Skin Temperature', and 'Surface Classification'. Additionally, the availability of a large amount of global training data was found to play a crucial role in the performance of surface methane estimation models, particularly in scenarios where WDCGG data is insufficient.

The research was supported by the Remote Equipment-Based Satellite Verification Tool Development Research Project, funded by the National Institute of Environmental Research.