Supporting Information

Identifying surface-level O_3 photochemical regimes using meteorology-adjusted satellite observations and statistical and machine learning models in South Korea

Chaoreum Kang[1], Hyung Joo Lee[1]

[1]Division of Environmental Science and Engineering, Pohang University of Science and Technology (POSTECH), Pohang, Gyeongbuk 37673, Republic of Korea

Correspondence to: Hyung Joo Lee (hyungjoolee@postech.ac.kr)

Contents of this file:

Text S1

Tables S1 to S2

Figures S1 to S8.

Text S1. Hyperparameter Optimization for XGBoost.

We employed Tree-structured Parzen Estimator (TPE) algorithm, a Bayesian optimization method implemented in Python framework '*optuna*', to optimize hyperparameters of XGBoost models (Watanabe, 2023; Akiba et al., 2019). The hyperparameter optimization proceeded as follows: (1) the TPE algorithm suggests a hyperparameter set in the predefined ranges by fitting two Gaussian mixture models (GMMs) to parameter values associated with better and worse objective values from previous trials, and selecting new candidates by maximizing their density ratio; (2) 10-fold cross validation (CV) is conducted with the suggested hyperparameters; (3) steps (1) and (2) are repeated 50 times; (4) the hyperparameter set yielding the lowest CV root mean squared error (RMSE) is selected as the optimized hyperparameters.

Table S1. Hyperparameters of province-specific XGBoost models optimized by TPE algorithm.

	XGBoost Hyperparameters					
Provinces	**n_estimators**	**max_depth**	**learning_rate**	**subsample**	**colsample_bytree**	**min_child_weight**
South Korea	383	10	0.131	0.793	0.898	6
Busan	292	9	0.057	0.932	0.713	4
Chungcheongbuk-do	337	10	0.043	0.914	0.947	4
Chungcheongnam-do	239	10	0.077	0.797	0.931	3
Daegu	234	10	0.047	0.842	0.784	2
Daejeon	328	10	0.034	0.692	0.903	2
Gangwon-do	201	10	0.039	0.892	0.953	5
Gwangju	321	9	0.052	0.889	0.929	4
Gyeonggi-do	374	10	0.067	0.796	0.704	6
Gyeongsangbuk-do	280	10	0.079	0.797	0.999	2
Gyeongsangnam-do	351	10	0.022	0.837	0.735	1
Incheon	396	10	0.021	0.787	0.993	6
Jeollabuk-do	343	10	0.050	0.886	0.726	1
Jeollanam-do	332	10	0.027	0.789	0.674	2
Sejong	359	9	0.034	0.818	0.749	3
Seoul	237	9	0.068	0.806	0.837	1
Ulsan	290	10	0.052	0.840	0.806	7

Table S2. Performance of province-specific cubic polynomial models for daily surface-level O_3 (with out-of-sample 10-fold CV) and 100 binned surface-level O_3.

Province	R^2 daily	R^2 binned	Province	R^2 daily	R^2 binned
Busan	0.043	0.75	Gyeongsangbuk-do	0.046	0.90
Chungcheongbuk-do	0.127	0.94	Gyeongsangnam-do	0.035	0.87
Chungcheongnam-do	0.062	0.92	Incheon	0.075	0.72
Daegu	0.094	0.70	Jeollabuk-do	0.034	0.84
Daejeon	0.114	0.82	Jeollanam-do	0.009	0.52
Gangwon-do	0.018	0.68	Sejong	0.152	0.72
Gwangju	0.081	0.71	Seoul	0.132	0.88
Gyeonggi-do	0.137	0.98	Ulsan	0.019	0.62

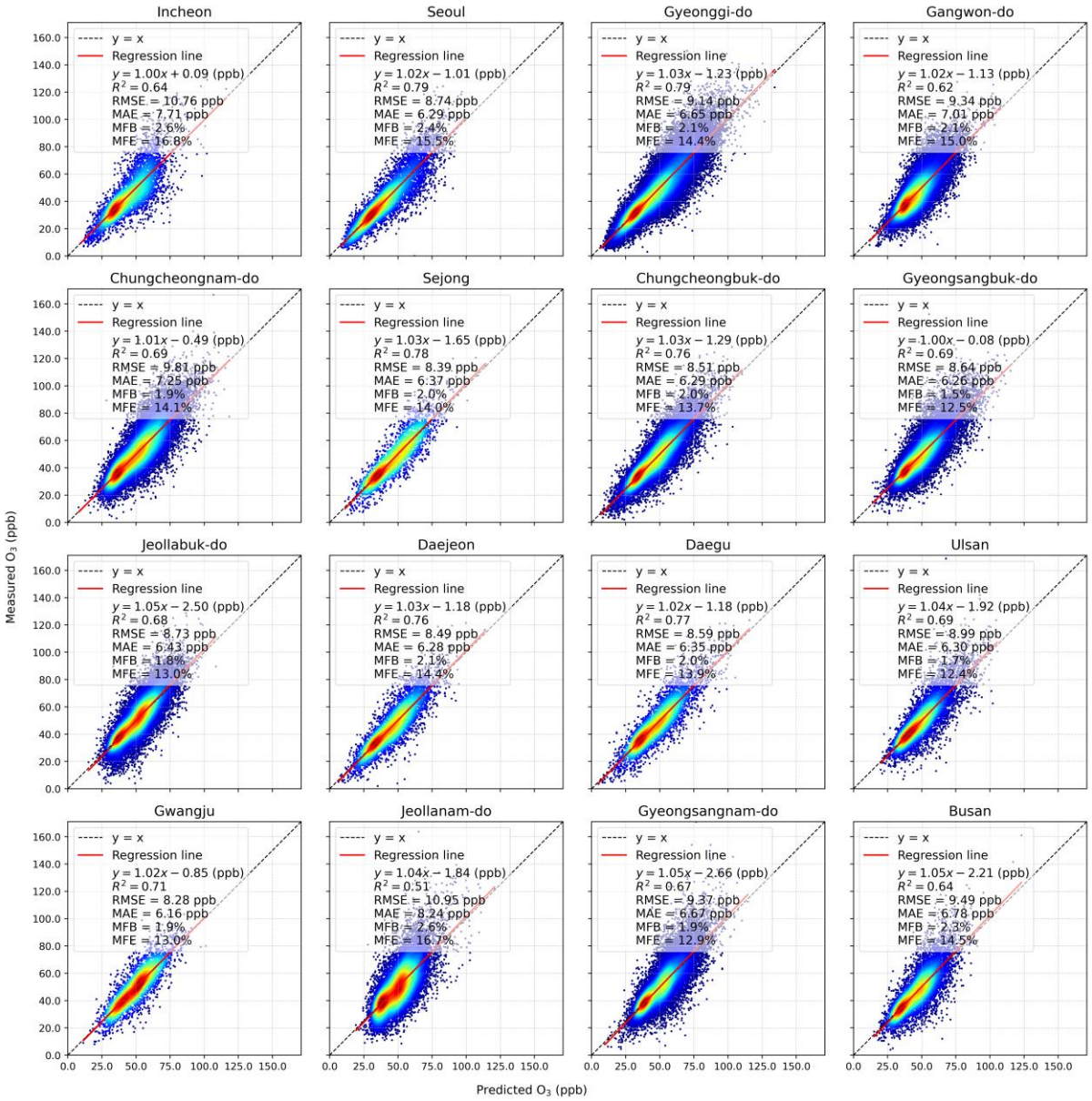


Figure S1. Performance of province-specific XGBoost models evaluated by nested out-of-sample 10-fold CV.

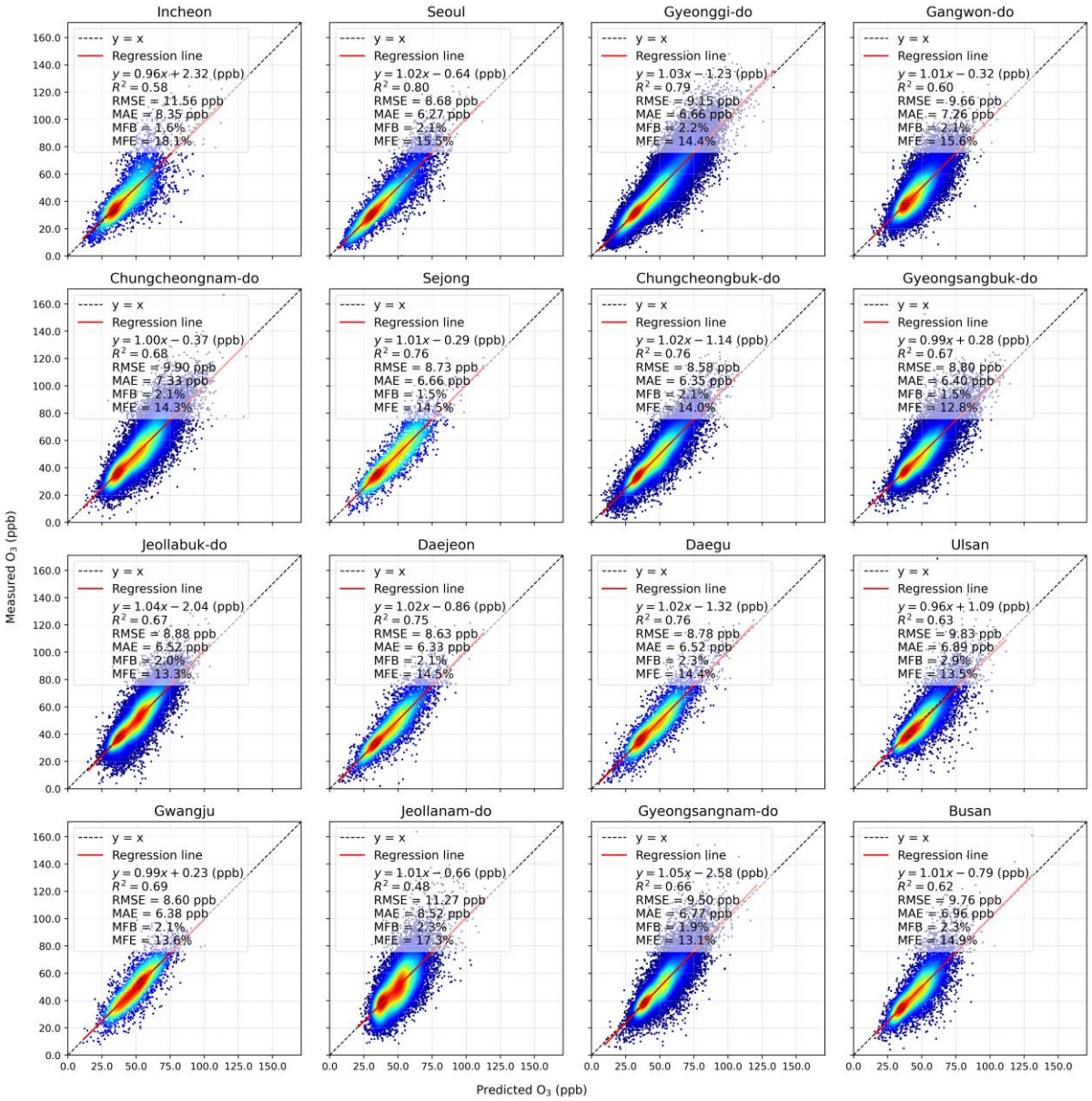


Figure S2. Performance of province-specific XGBoost models evaluated by nested out-of-grid 10-fold CV.

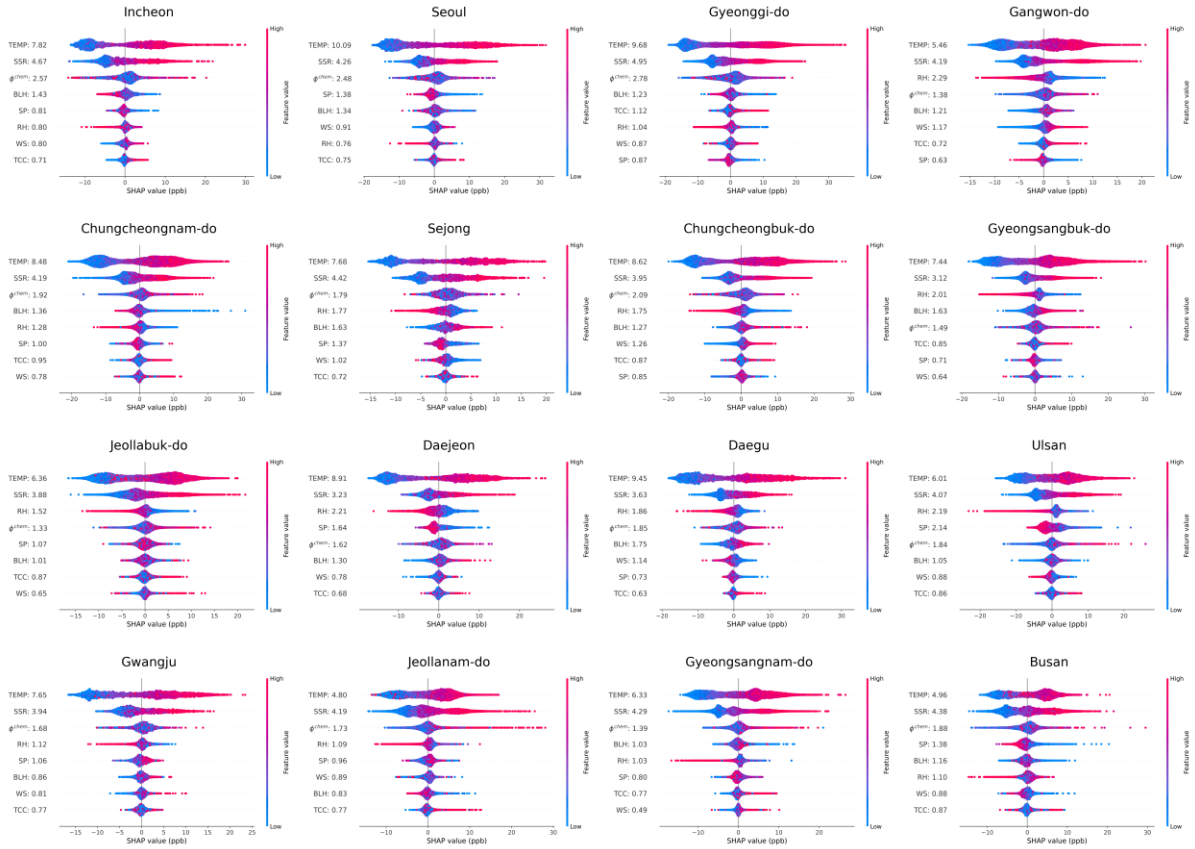


Figure S3. SHAP summary plots for the XGBoost models. Abbreviations denote temperature (TEMP), surface net solar radiation (SSR), SHAP-chemistry (ϕ^{chem}), relative humidity (RH), boundary layer height (BLH), total cloud cover (TCC), surface pressure (SP), and wind speed (WS). In each subplot, variables are ranked from top to bottom by decreasing mean absolute SHAP value, with higher values indicating greater feature importance. These values are displayed next to the variable abbreviations.

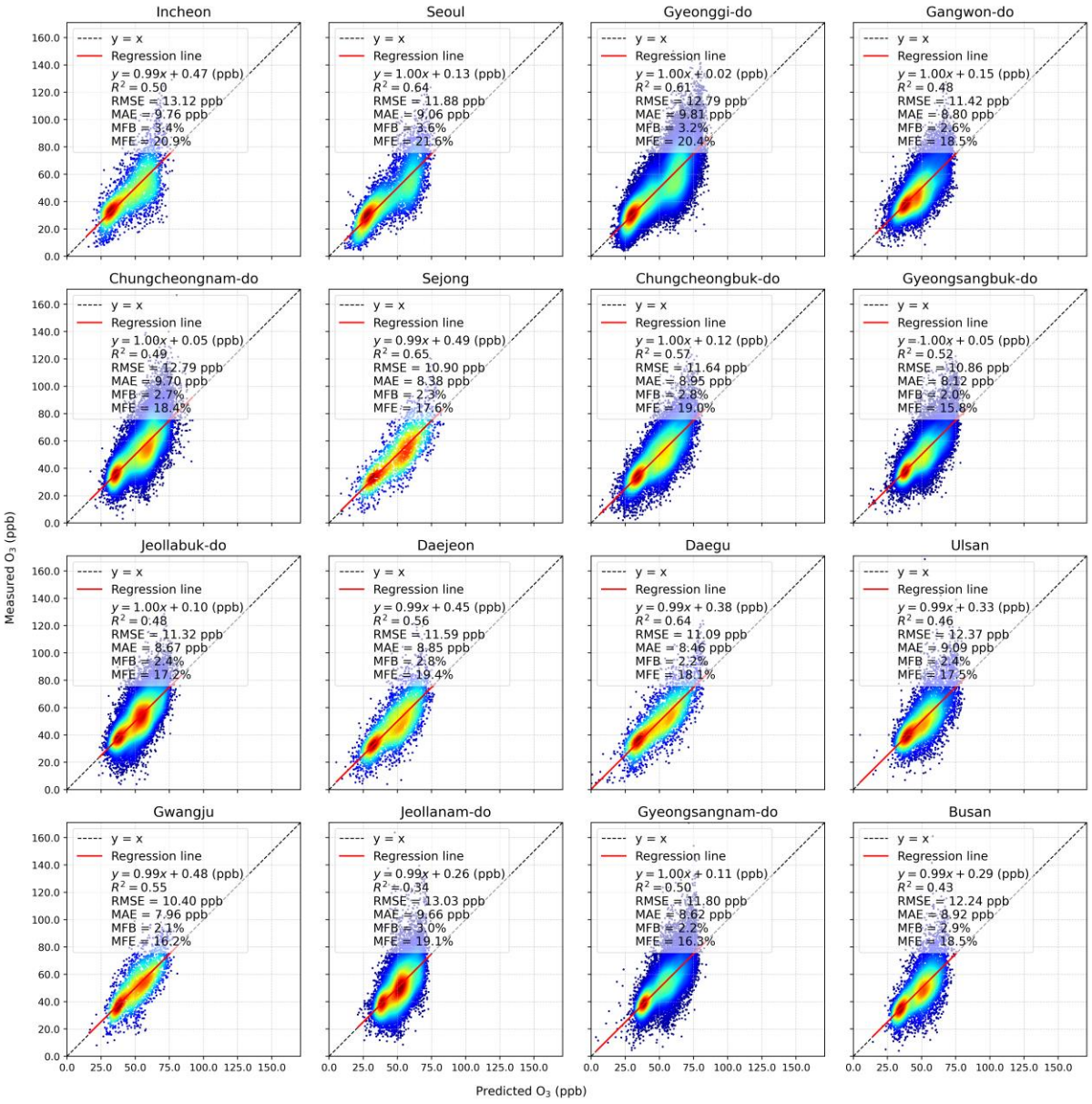


Figure S4. Performance of province-specific GAMs evaluated by out-of-sample 10-fold CV.

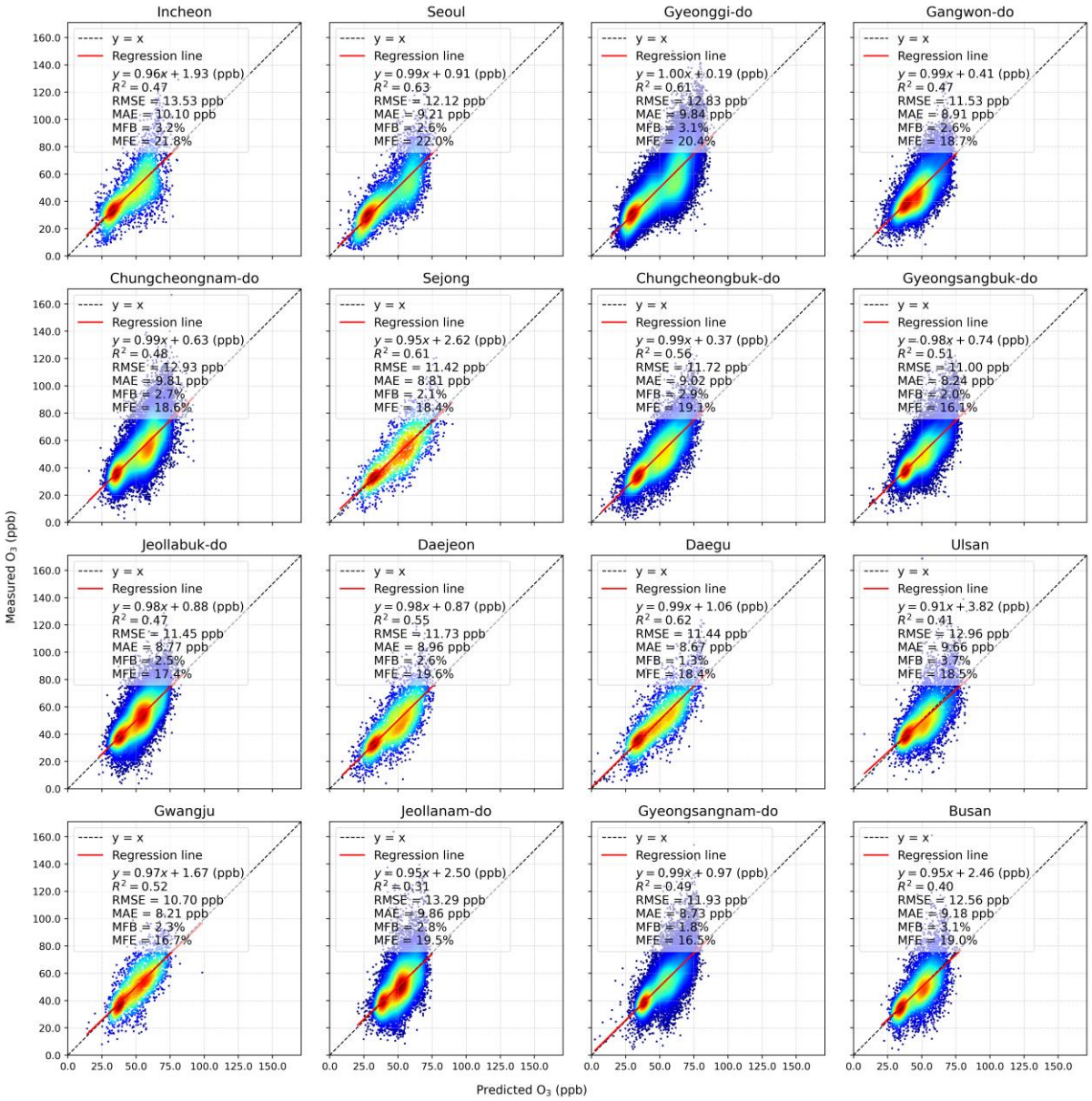


Figure S5. Performance of province-specific GAMs evaluated by out-of-grid 10-fold CV.

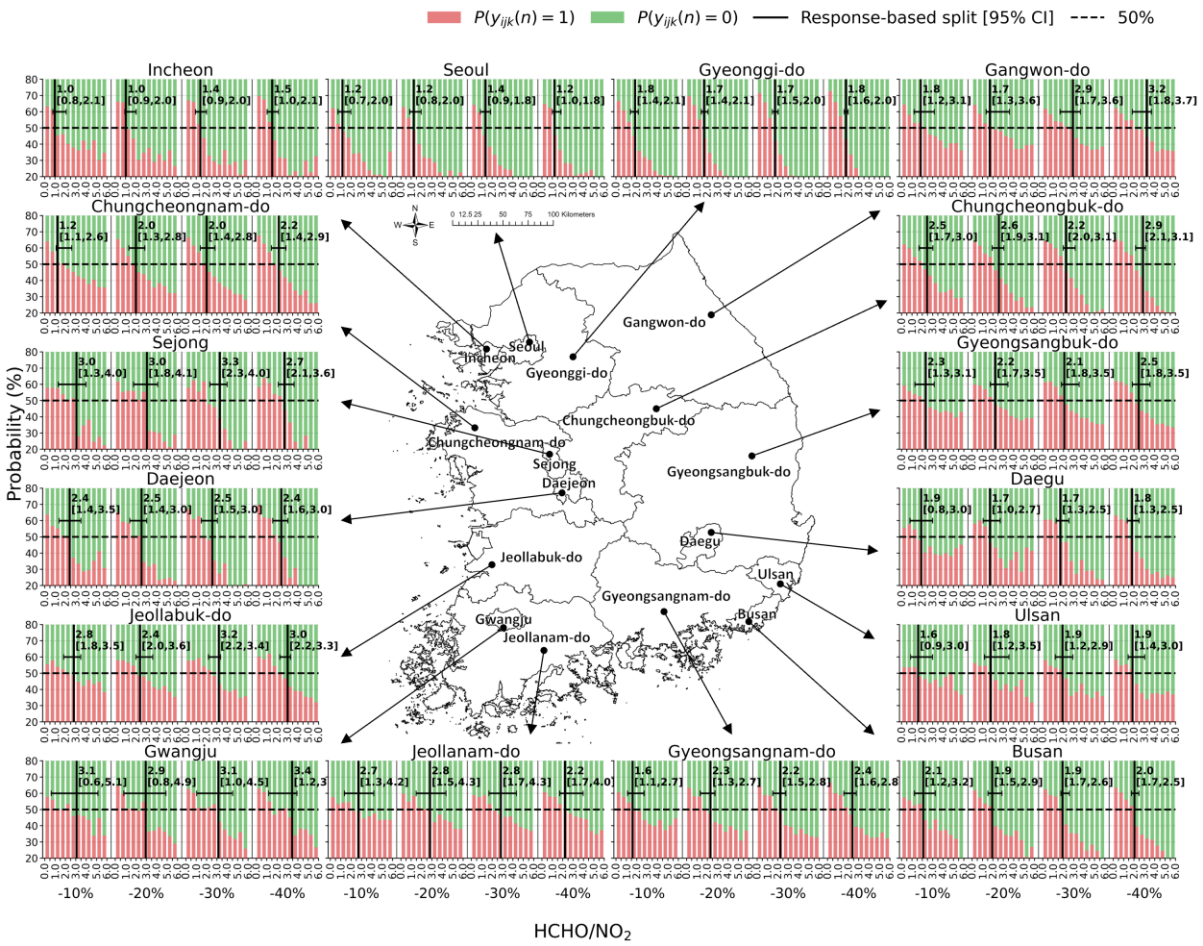


Figure S6. O_3 response probabilities within HCHO/NO_2 intervals of width 0.5 and O_3 response-based splits with 95% CIs by province and NO_2 reduction scenario. The probabilities are calculated as the weighted proportions of samples showing O_3 increases ($y_{ijk}(n) = 1$) and decreases ($y_{ijk}(n) = 0$) within each interval. The 95% CIs for the response-based splits are derived from 500 bootstrap replicates.

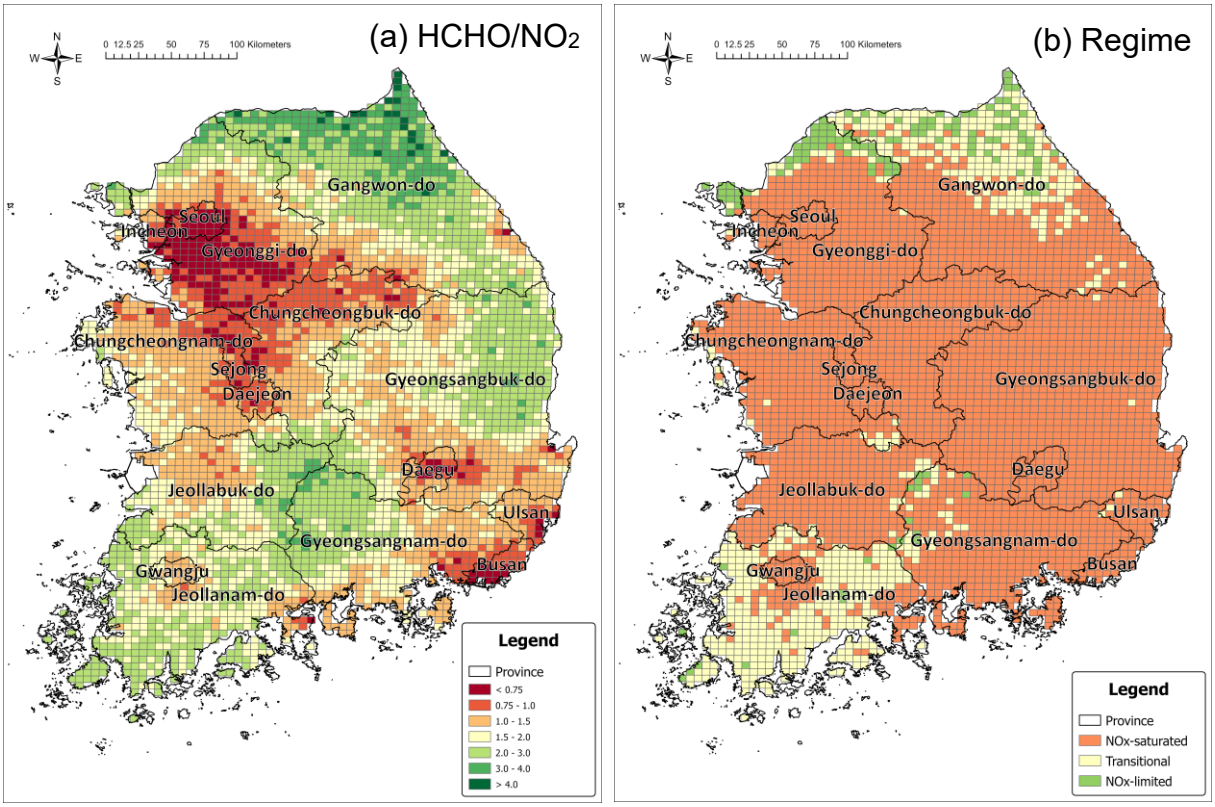


Figure S7. Spatial distributions (5 km × 5 km grid cell) of (a) TROPOMI HCHO/NO_2 ratios and (b) O_3 photochemical regimes based on province-specific regime thresholds derived using the XGB+SHAP approach during the O_3 non-peak season (from October 16 to April 14) for 2019–2023.

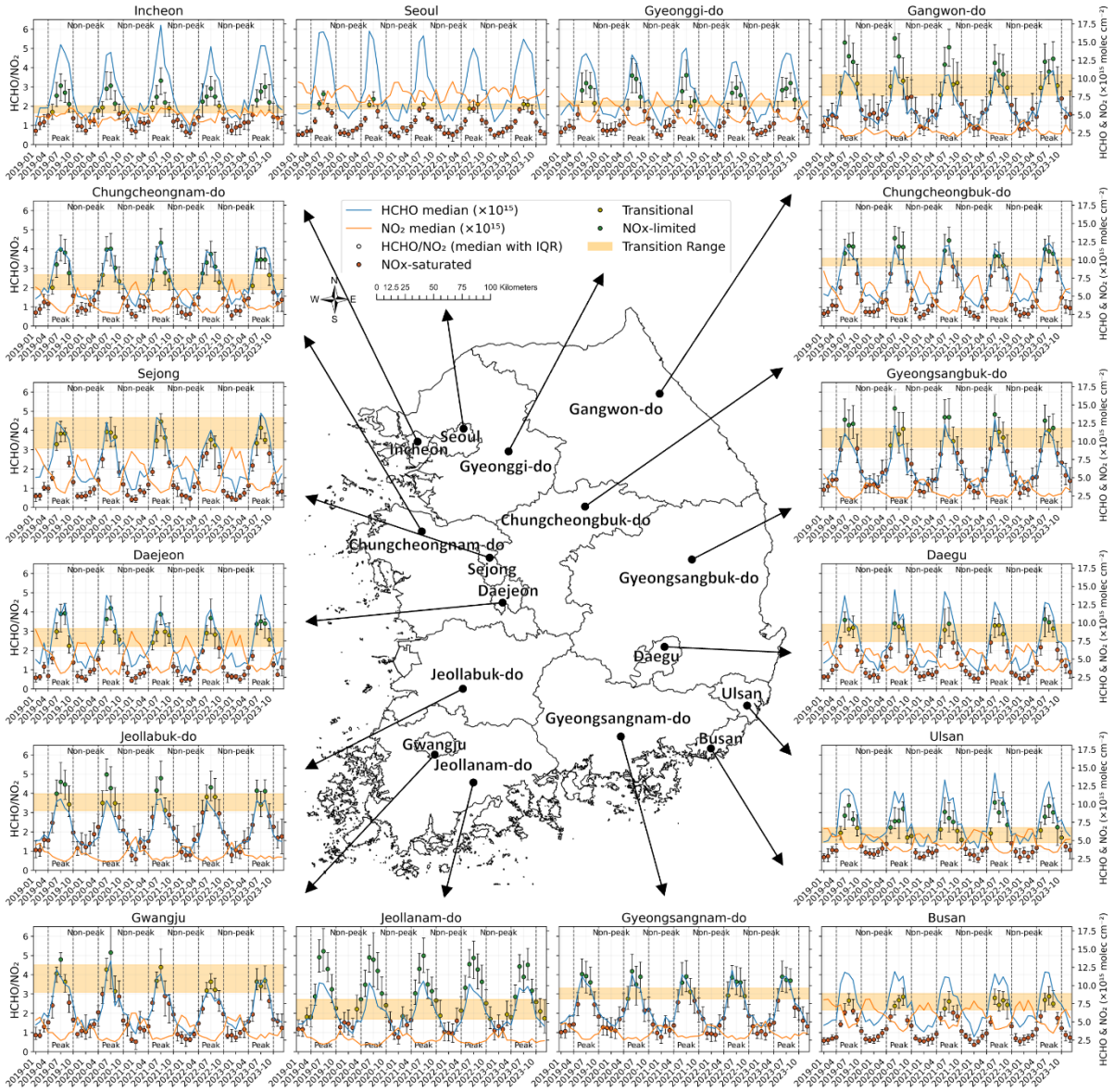


Figure S8. Month-to-month variations in TROPOMI HCHO, NO_2, $HCHO/NO_2$ ratios (with interquartile ranges), and O_3 photochemical regimes by province in South Korea (2019–2023).

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A next-generation hyperparameter optimization framework, Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2623-2631,

Watanabe, S.: Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance, arXiv preprint arXiv:2304.11127, 2023.