Supplement of Atmos. Chem. Phys., 25, 13527–13545, 2025
https://doi.org/10.5194/acp-25-13527-2025-supplement

Atmospheric Chemistry and Physics
Open Access
EGU

Supplement of

Estimating surface sulfur dioxide concentrations from satellite data over eastern China: Using chemical transport models vs. machine learning

Zachary Watson et al.

Correspondence to: Zachary Watson (zw0033@uah.edu) and Shan-Hu Lee (shanhu.lee@uah.edu)

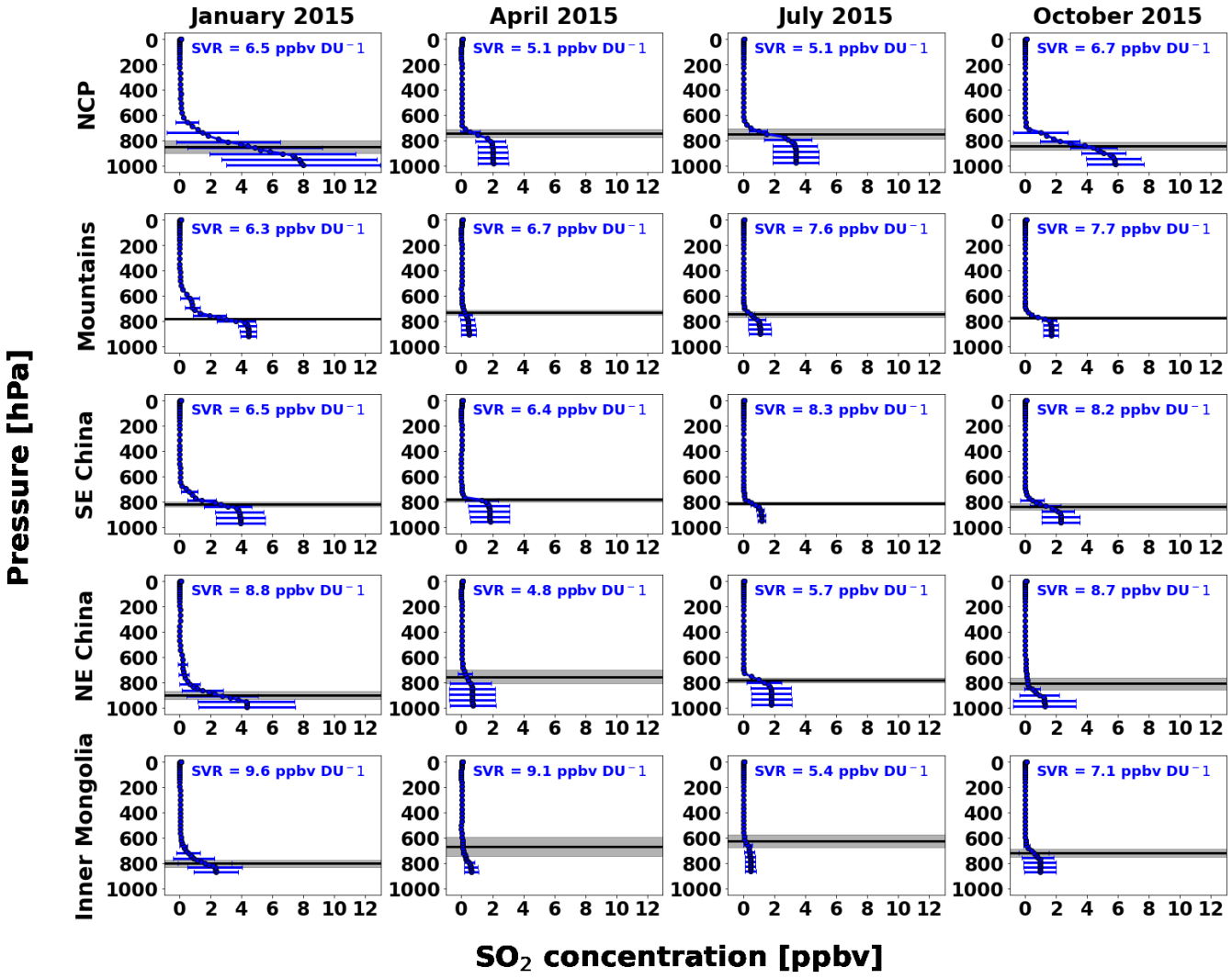


Figure S1: Monthly averaged vertical SO_2 profiles (blue lines) from each of the 2015 GEOS-Chem simulations (from left to right: January, April, July, and October) with 1 standard deviation error bars, and GEOS-FP boundary layer heights (black line with 1 standard deviation shading) at five locations in different parts of the study region including, from top to bottom, the North China Plain (NCP; 115 °E, 38 °N), the Qin Mountains (107.5 °E, 32 °N), southeastern China (115 °E, 26 °N), northeastern China (122.5 °E, 44 °N), and Inner Mongolia (107.5 °E, 40 °N).

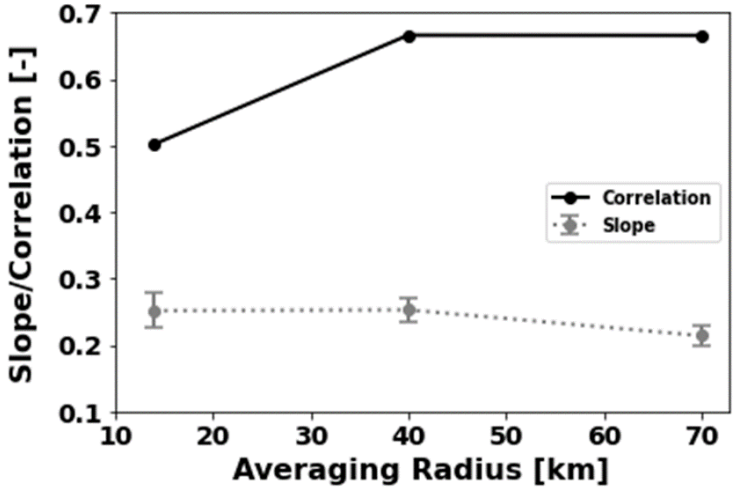


Figure S2: Correlations and slopes of 2015 annual mean CTM-derived surface SO_2 concentrations compared to CNEMC in-situ measurements as a function of the averaging radius around each CNEMC site. Error bars represent the 95% confidence interval in slope based on the standard error of the linear regression fit. Radii selected are 14 km (nearest OMI pixel), 40 km (nearest 3x3 grid of OMI pixels centered on the station) and 70 km (nearest 5x5 grid of OMI pixels centered on the station).

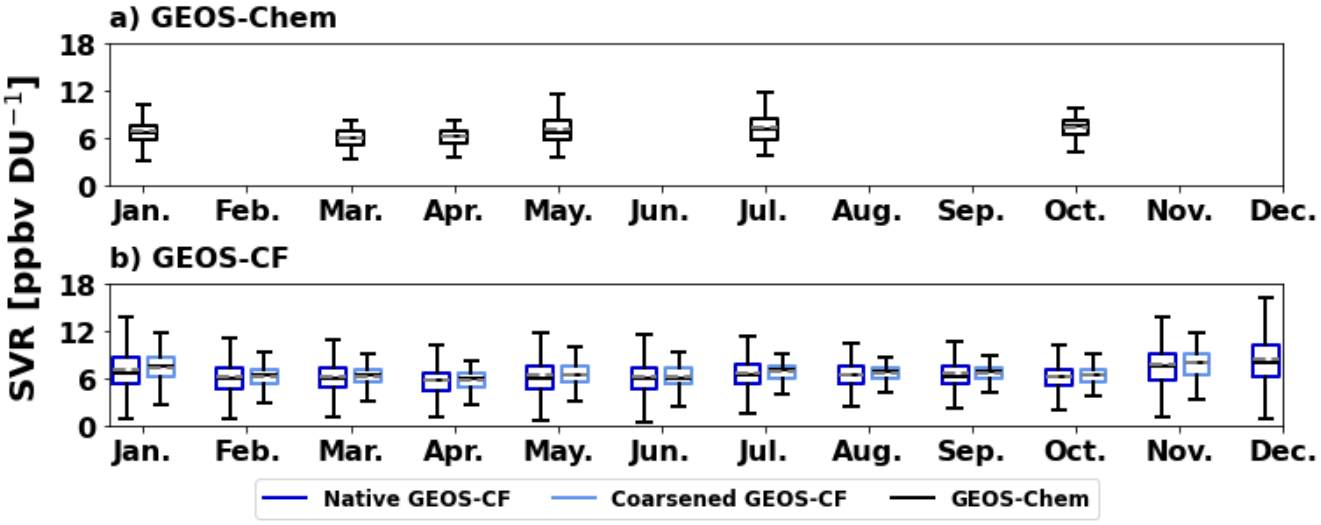


Figure S3: Boxplots showing monthly average (a) SVRs from the 201 GEOS-Chem simulations at 2.5° x 2° resolution (black), and (b) SVRs from the archived 2018 GEOS-CF simulations at native (0.25° x 0.25°; dark blue) and coarsened (2.5° x 2°; light blue) resolution. The solid black and dashed gray lines represent the median and mean, respectively. GEOS-Chem simulations were available for January, March, April, May, July, and October. Months without a boxplot were not simulated.

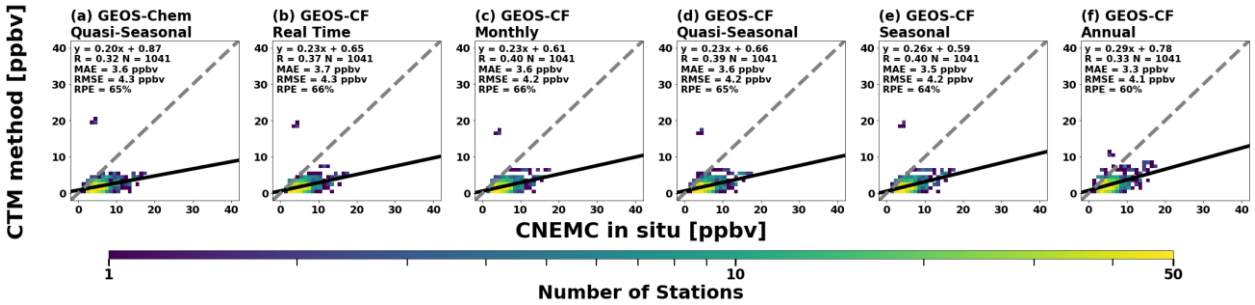


Figure S4: Scatterplots of the 2018 annual mean surface SO_2 concentrations from the CTM-based method against the CNEMC in situ measurements for different models including (a) GEOS-Chem (2.5° x 2° horizontal resolution) and (b-f) GEOS-CF (0.25° x 0.25° horizontal resolution) for different temporal sampling of the CTM SVRs when combined with daily OMI data to calculate daily surface SO_2 concentrations. (b) "Real Time" sampling indicates that daily SVRs were used to calculate daily surface concentrations. (c) "Monthly" sampling indicates that monthly averaged SVRs were used to calculate daily surface concentrations within that given month. (a,d) "Quasi-Seasonal" sampling indicates that January, April, July, and October average SVRs were used to calculate daily surface concentrations within the winter, spring, summer, and autumn months, respectively. (e) "Seasonal" sampling indicates that SVRs averaged over DJF, MAM, JJA, and SON were used to calculate daily surface concentrations within that given season. (f) "Annual" sampling represents that the annual average SVR was used to calculate daily surface concentrations for within that year. Scatterplots are binned every 1 ppbv and colored by the number of stations. Each panel contains a linear regression analysis with the best fit line (solid lines), best-fit equation, correlation coefficient, total number of stations, 1:1 line (black dashed line), MAE, RMSE, and RPE.

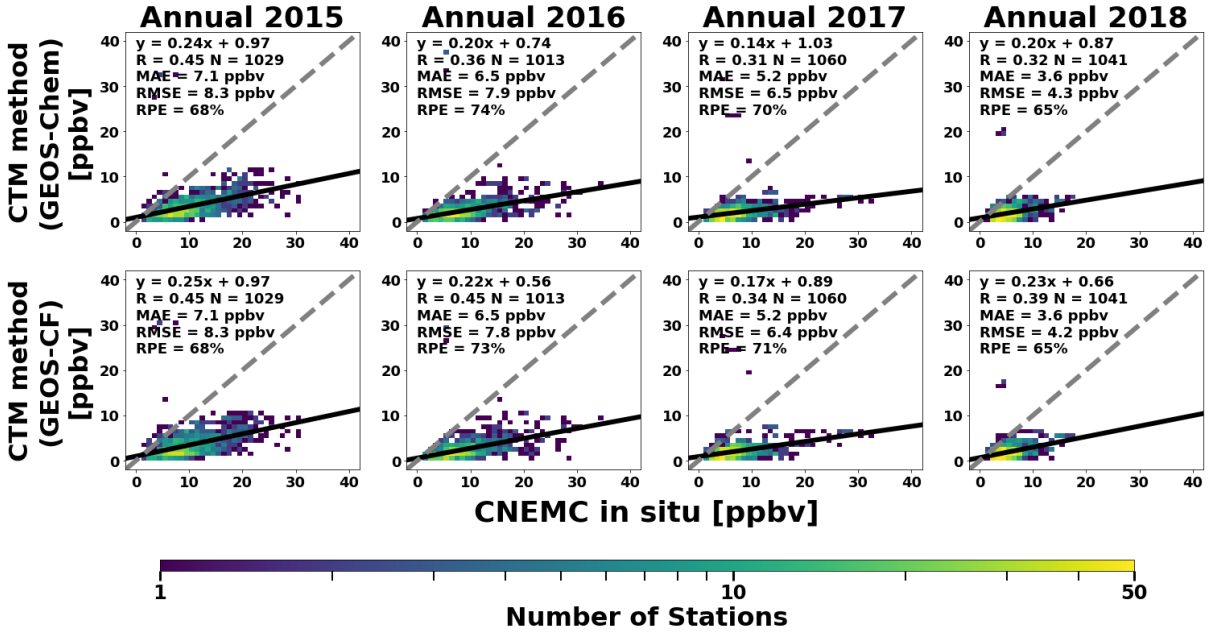


Figure S5: Scatterplots of the annual mean surface SO_2 concentrations from the CTM-based method against the CNEMC in situ measurements for 2018 GEOS-Chem (top row) and GEOS-CF model simulations. Both models had the same temporal sampling (quasi-seasonal) but different spatial resolutions (2.5° x 2.0° and 0.25° x 0.25°, respectively). Each column represents a different year of the study period (from left to right: 2015, 2016, 2017, and 2018). Scatterplots are binned every 1 ppbv and colored by the number of stations. Each panel contains a linear regression analysis with the best fit line (solid lines), best-fit equation, correlation coefficient, total number of stations, 1:1 line (black dashed line), MAE, RMSE, and RPE.

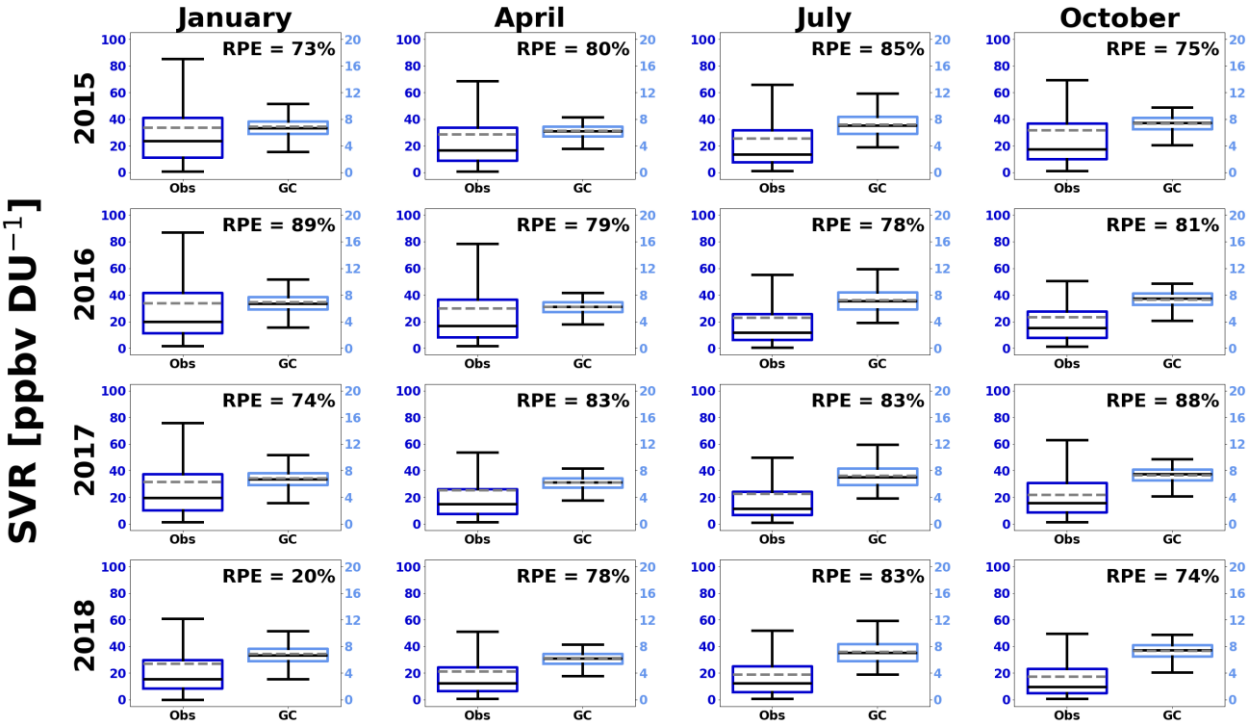


Figure S6: Boxplots showing the monthly averaged observed SVRs calculated from CNEMC in situ surface SO_2 concentrations and OMI SO_2 VCDs (Obs; dark blue), and SVRs from the GEOS-Chem model (GC; light blue). Each column represents a different monthly average (from left to right: January, April, July, and October), and each row represents a different year in the study period (from top to bottom: 2015, 2016, 2017, and 2018). The solid black and dashed gray lines represent the median and mean SVR, respectively. Each panel also contains the average RPE between the observed and GEOS-Chem SVRs. Note that the GEOS-Chem SVRs are only from the 2015 simulations. Note that the y-axes for the Obs SVRs are a factor of 5 larger than for the GC SVRs, suggesting that the GC SVRs are significantly less than those calculated from CNEMC surface SO_2 concentration and OMI SO_2 VCD observations.

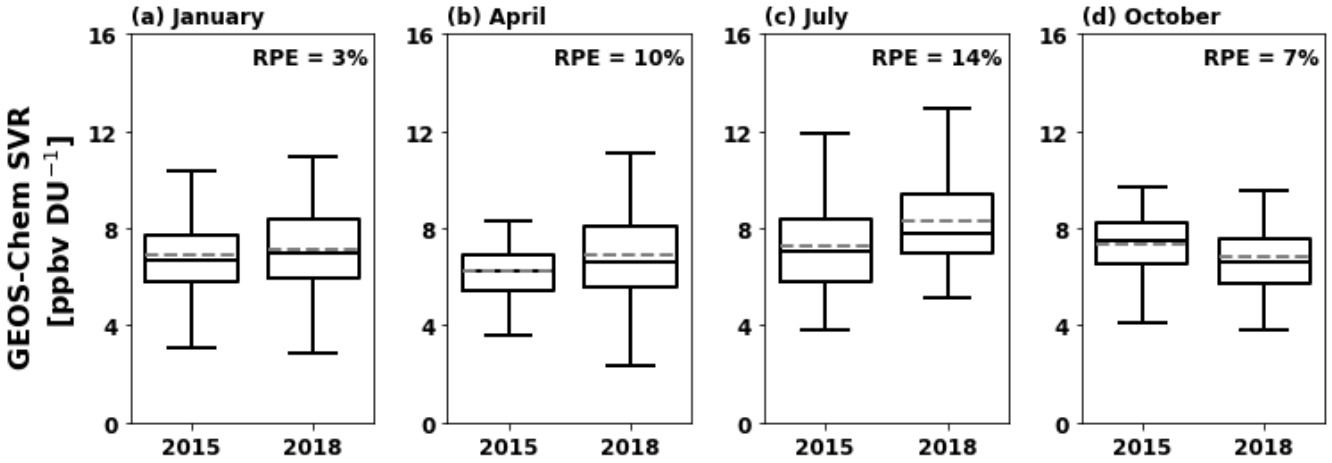


Figure S7: Boxplots showing the monthly averaged SVR from the 2015 and 2018 GEOS-Chem simulations for a) January, b) April, c) July, and d) October. Each panel contains the average RPE between the SVRs from the two simulations.

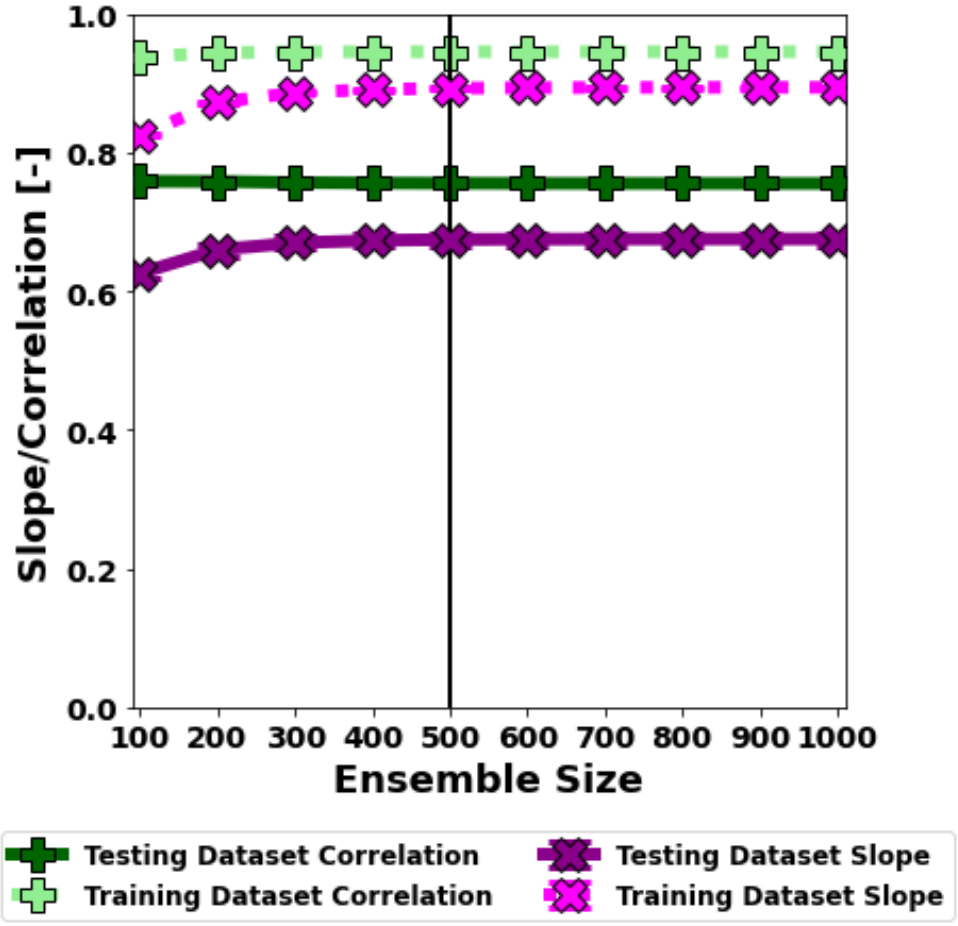


Figure S8: Performance of the XGBoost model measured by the slope and correlation of a linear regression analysis between the daily ML-predicted and CNEMC in situ surface SO_2 concentrations as a function of ensemble size. Increasing the ensemble size slightly improved the performance of the model until stabilizing around a value of 500 (black line), so this value was used for the model.

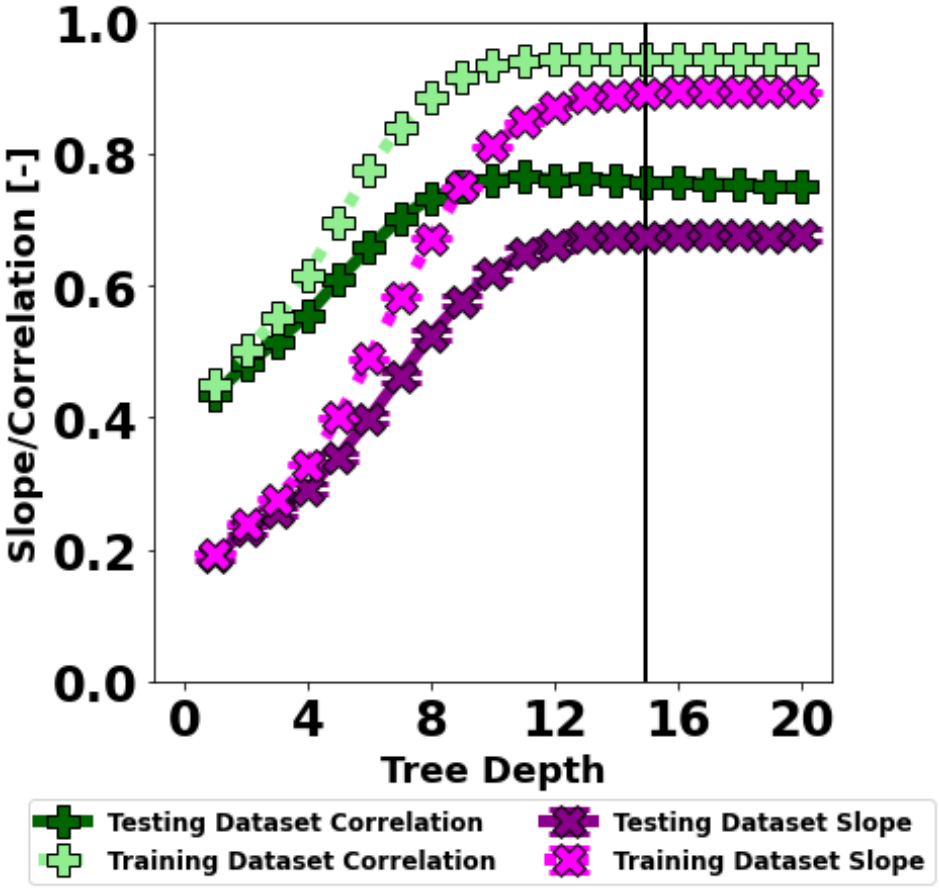


Figure S9: Performance of the XGBoost model measured by the slope and correlation of a linear regression analysis between the daily ML-predicted and CNEMC in situ surface SO_2 concentrations as a function of tree depth. Increasing the tree depth significantly improved the performance of the model until stabilizing around a value of 15 (black line), so this value was used for the model.

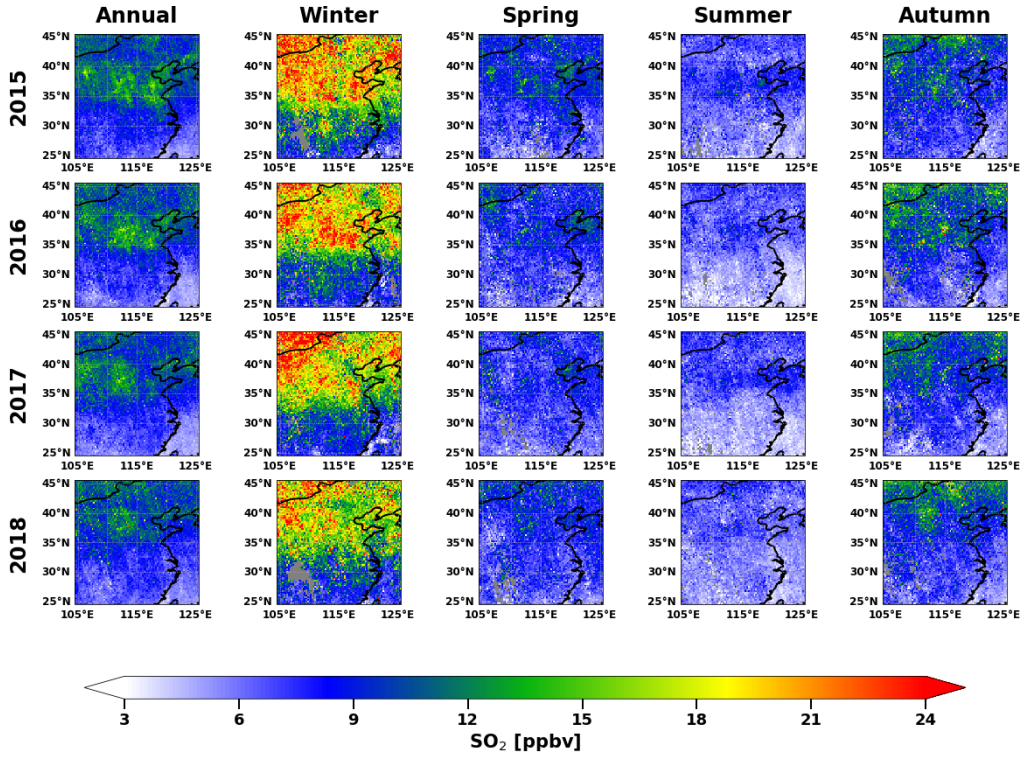
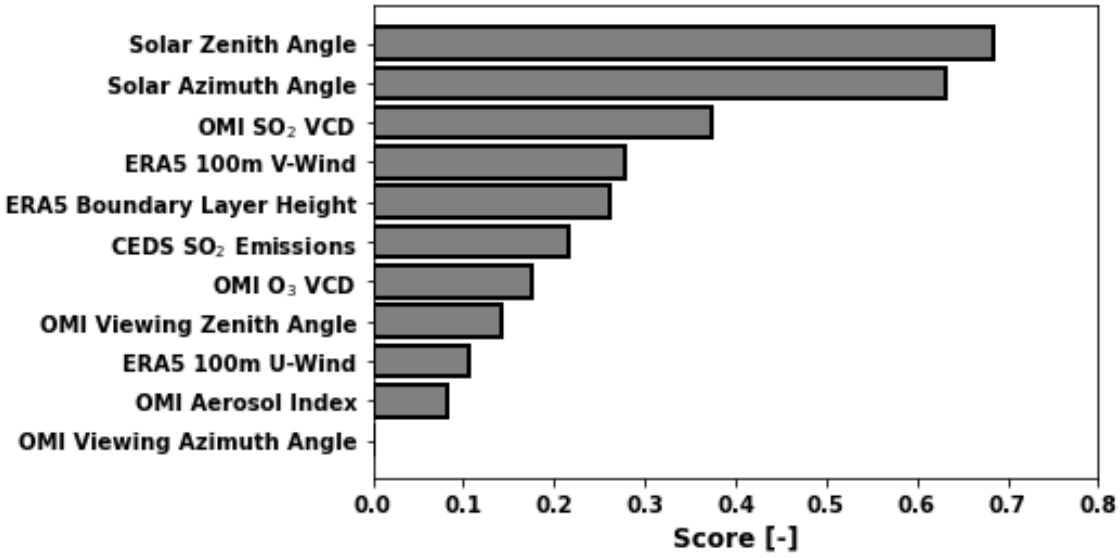


Figure S10: Maps of surface SO_2 concentrations at 0.25° x 0.25° horizontal resolution from a previous version of the XGBoost model trained on 11 predictors including SO_2 VCDs, ozone VCDs, aerosol index, solar zenith angle, solar azimuth angle, viewing zenith angle, and viewing azimuth angle from the OMI product, 100 m u-wind, 100 m v-wind, 2 m temperature, 2 m dew point temperature, and boundary layer height from ERA5, and SO_2 emissions from the CEDS emission inventory. Each column represents a different averaging period, and each row represents a different year of the study period. The spatial distribution here is significantly different to OMI SO_2 VCDs and CNEMC in situ measurements, suggesting that this version of the model did not produce accurate results.



120 **Figure S11: Permutation importance analysis performed on a previous version of the model trained on 11 predictors including SO_2 VCDs, ozone VCDs, aerosol index, solar zenith angle, solar azimuth angle, viewing zenith angle, and viewing azimuth angle from the OMI product, 100 m u-wind, 100 m v-wind, 2 m temperature, 2 m dew point temperature, and boundary layer height from ERA5, and SO_2 emissions from the CEDS emission inventory. The permutation importance shows the dominant influence solar geometry on the predicted surface SO_2 concentrations shown in Fig. S10.**

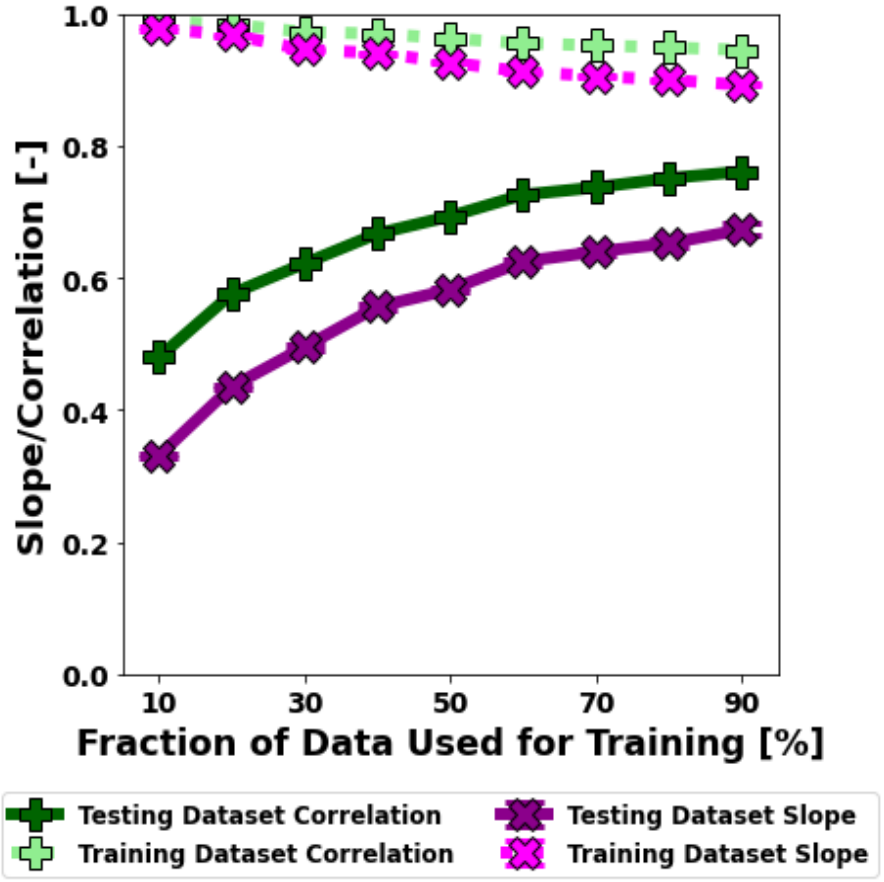


Figure S12: Performance of the XGBoost model measured by the slope and correlation of a linear regression analysis between the daily ML-predicted and CNEMC in situ surface SO_2 concentrations as a function of the percent of data used for model training. As the size of the training dataset increased, the performance of the independent testing dataset improved.

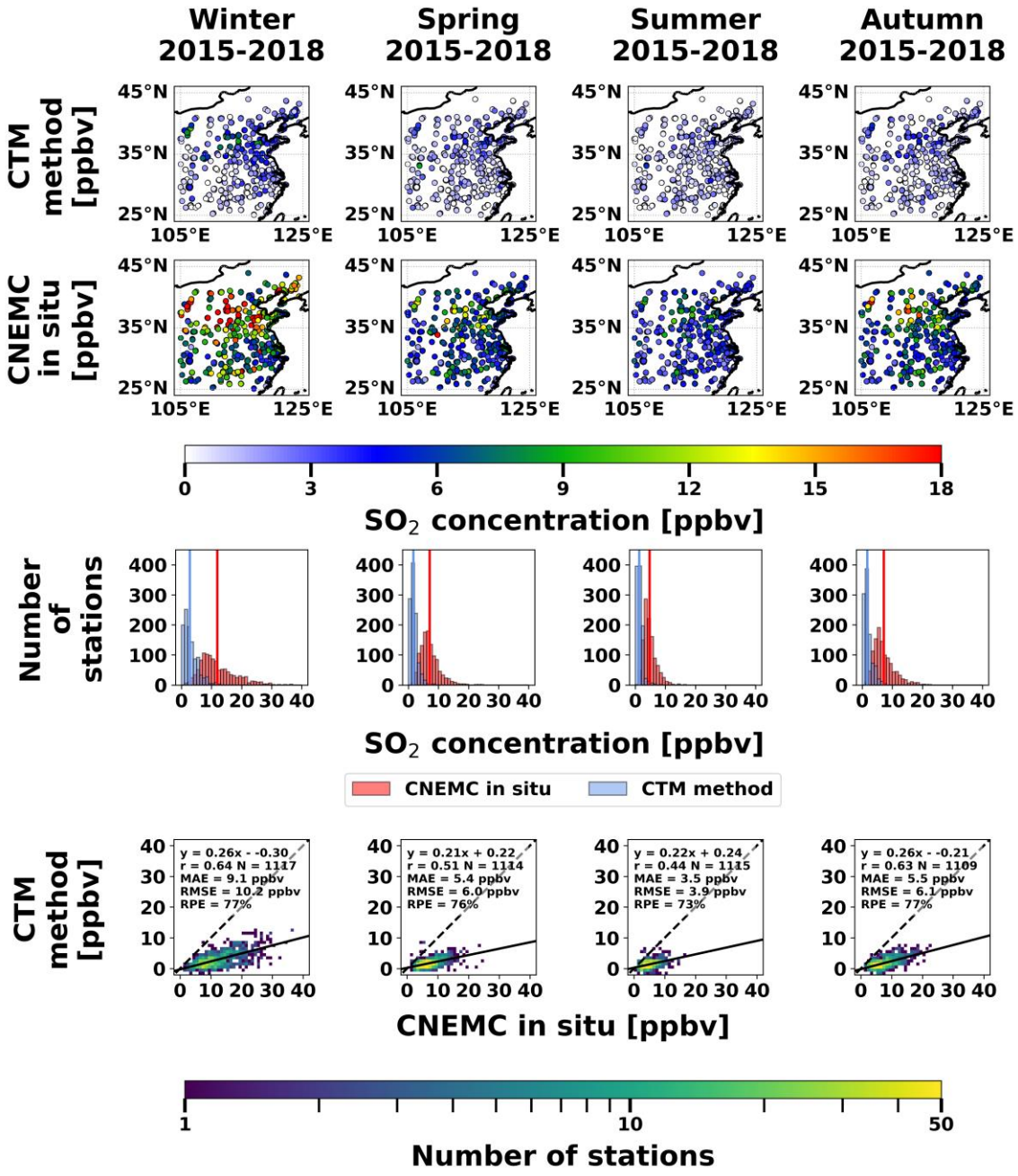


Figure S13: Spatial distributions of the seasonal surface SO_2 concentrations averaged from 2015-2018 for the CTM-based method (top row) and CNEMC in-situ measurements (second row), histograms of the surface concentrations from each dataset with vertical bars representing the means (third row), and scatterplots between the two datasets (bottom row). Each column represents a different year in the study period. Histograms and scatterplots are binned every 1 ppbv. Each scatterplot is colored by the number of stations in each bin and includes a linear regression analysis with the best fit line (solid line), best-fit equation, correlation coefficient, total number of stations, 1:1 line (black dashed line), MAE, RMSE, and RPE.

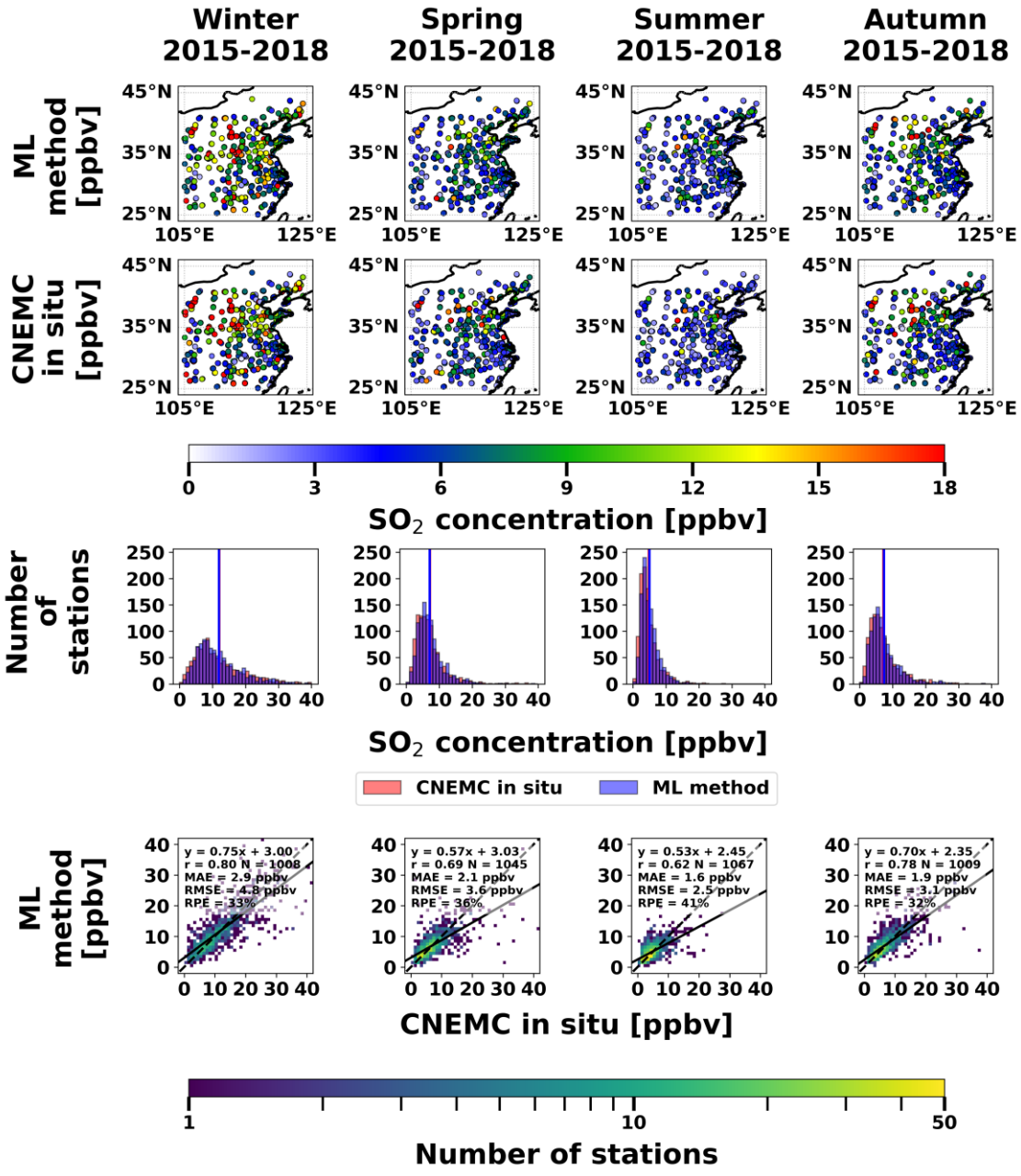
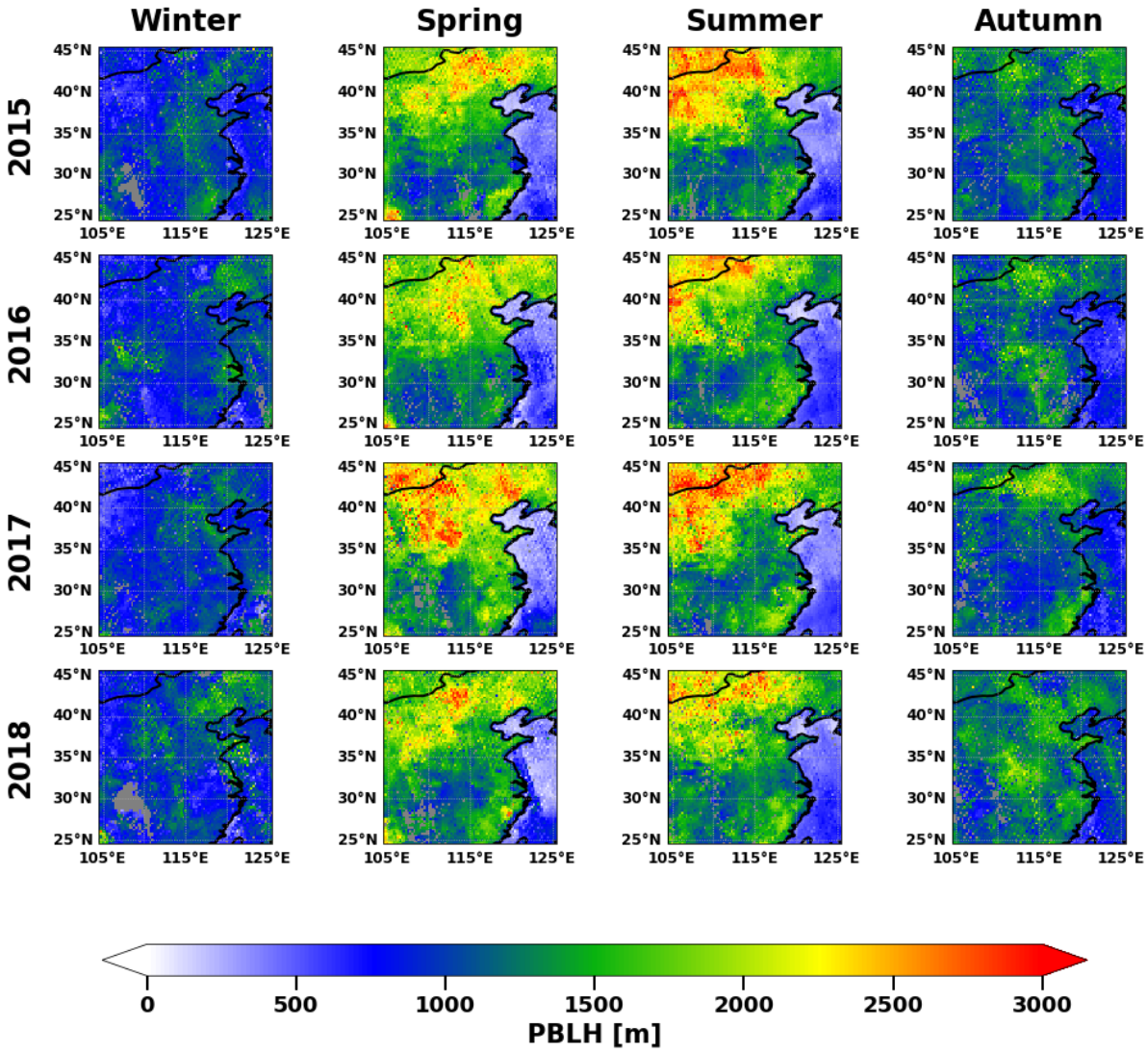


Figure S14: Spatial distributions of the seasonal surface SO_2 concentrations averaged from 2015-2018 for the ML-based method (top row) and CNEMC in-situ measurements (second row), histograms of the surface concentrations from each dataset with vertical bars representing the means (third row), and scatterplots between the two datasets (bottom row). Each column represents a different year in the study period. Histograms and scatterplots are binned every 1 ppbv. Each scatterplot is colored by the number of stations in each bin and includes a linear regression analysis with the best fit line (solid line), best-fit equation, correlation coefficient, total number of stations 1:1 line (black dashed line), MAE, RMSE, and RPE.



145

Figure S15: Maps of the seasonal mean ERA5 boundary layer heights at 0.25° x 0.25° horizontal resolution in m. Each column represents a different season, and each row represents a different year of the study period.

Table S1: Annual and seasonal average surface SO_2 concentrations from the CNEMC in situ measurements, CTM-based method, and ML-based method from the independent testing dataset from 2015-2018. Slopes and correlations are from the validatation of the CTM-based and ML-based surface SO_2 concentrations against the in situ measurements. Data supports Fig. 7 from main text.

	CNEMC In Situ	CTM-Based Method			ML-Based Method		
Timescale	**Surface SO_2 Concentration [ppb]**	**Surface SO_2 Concentration [ppb]**	**Slope**	**Correlation**	**Surface SO_2 Concentration [ppb]**	**Slope**	**Correlation**
Annual 2015	9.4	2.7	0.24	0.41	9.5	0.67	0.73
Annual 2016	8.5	1.9	0.16	0.29	8.3	0.64	0.76
Annual 2017	6.8	1.5	0.17	0.31	7.0	0.73	0.78
Annual 2018	5.4	1.5	0.14	0.18	5.6	0.73	0.79
Winter 2015	15.7	4.4	0.19	0.30	15.7	0.65	0.68
Spring 2015	9.3	2.7	0.19	0.32	9.2	0.51	0.63
Summer 2015	6.1	1.8	0.19	0.26	6.2	0.42	0.53
Autumn 2015	8.9	2.5	0.21	0.32	9.3	0.67	0.71
Winter 2016	13.9	2.6	0.16	0.30	13.3	0.68	0.75
Spring 2016	7.7	1.8	0.13	0.20	7.6	0.47	0.60
Summer 2016	5.2	1.2	0.08	0.13	5.3	0.40	0.52
Autumn 2016	9.0	2.4	0.22	0.34	9.2	0.74	0.76
Winter 2017	11.7	2.6	0.17	0.30	11.5	0.75	0.81
Spring 2017	6.0	1.3	0.11	0.13	6.1	0.59	0.65
Summer 2017	4.5	1.2	0.15	0.17	4.8	0.45	0.55
Autumn 2017	6.3	1.4	0.10	0.18	6.7	0.74	0.77
Winter 2018	8.3	2.6	0.15	0.18	8.5	0.64	0.72
Spring 2018	5.8	1.3	0.08	0.12	5.7	0.49	0.65
Summer 2018	3.4	1.2	0.08	0.06	3.8	0.72	0.65
Autumn 2018	5.4	1.4	0.12	0.14	5.6	0.56	0.68

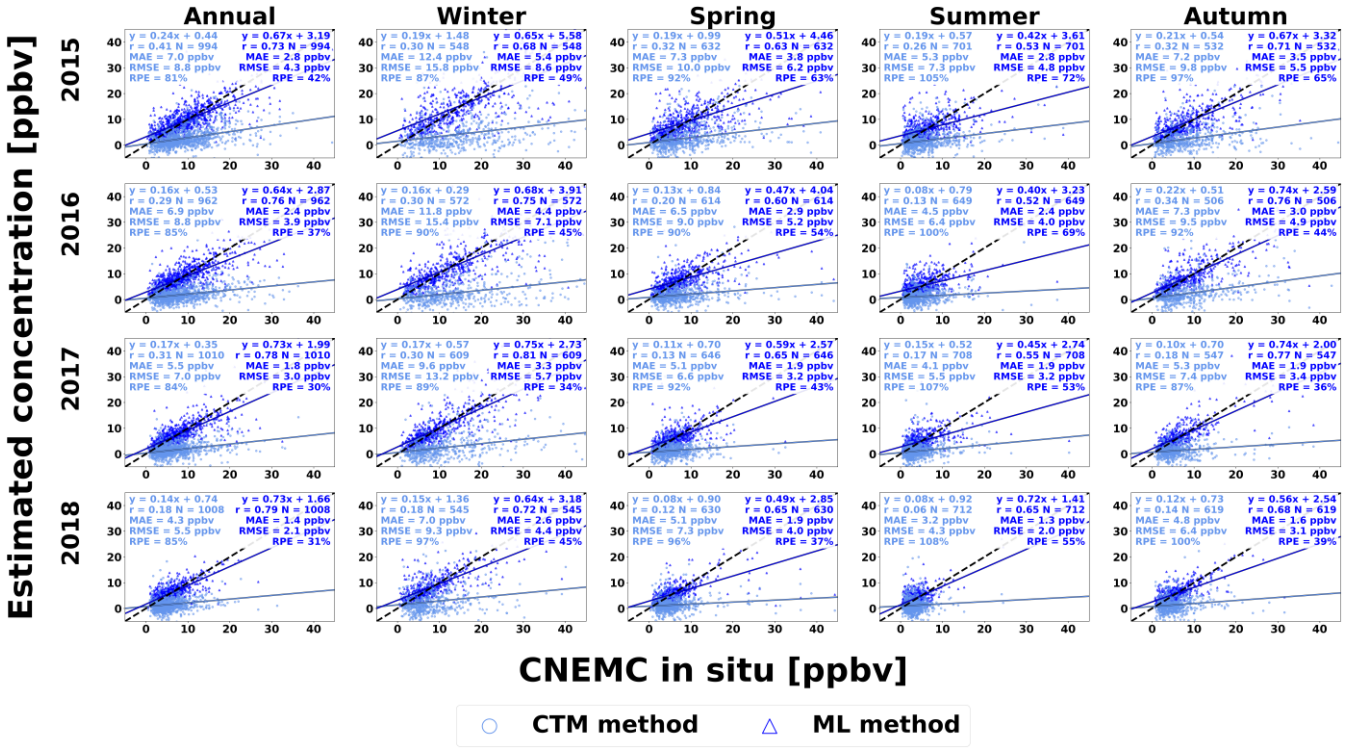


Figure S16: Scatterplots showing the estimated surface SO_2 concentrations from the CTM method (light blue circles) and ML method (dark blue triangles) against the in-situ measurements for individual years and seasons during the study period. Each column represents a different averaging period, and each row represents a different year of the study period. Each scatterplot includes a linear regression analysis with the best fit line (solid line), best-fit equation, correlation coefficient, total number of stations, 1:1 line (black dashed line), MAE, RMSE, and RPE.

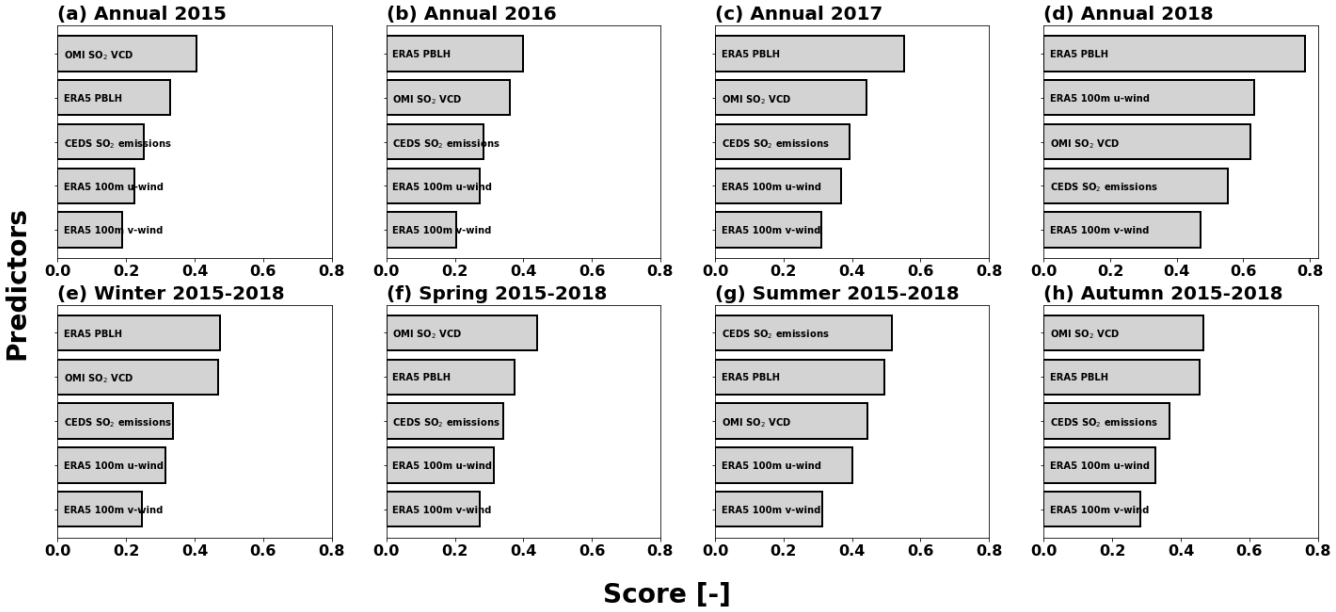


Figure S17: Permutation importance for different time periods including (a-d) individual years of the study period and (e-f) each season combined from all four years of the study period. Higher scores indicate a larger impact on the results.

165

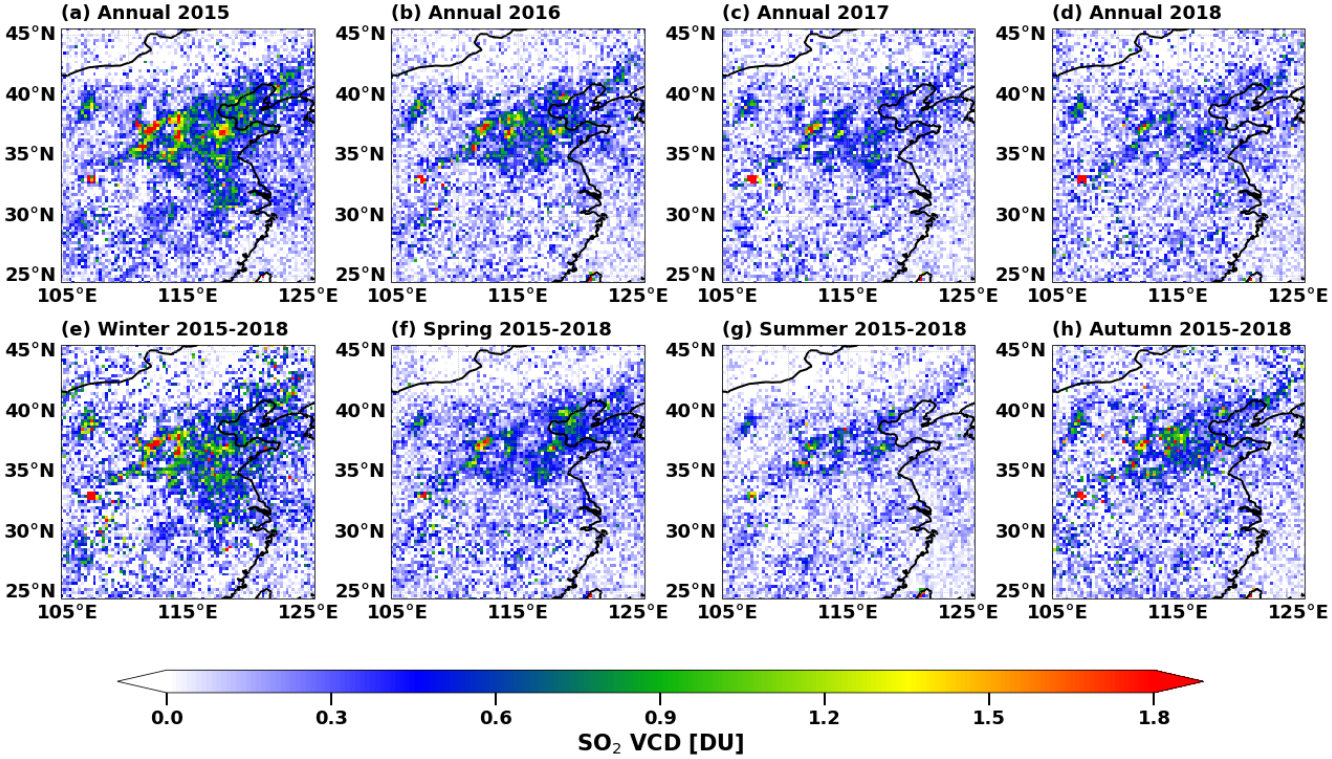


Figure S18: Spatial distribution of the annual mean and seasonal mean OMI PBL SO_2 VCDs in Dobson Units (DU) located in our study region at 0.25° x 0.25° horizontal resolution. Gray pixels represent missing data. Panels (a-d) represent the annual mean SO_2 VCDs for each year in the study period, and panels (e-f) represents the seasonal means averaged from 2015-2018 for each season.

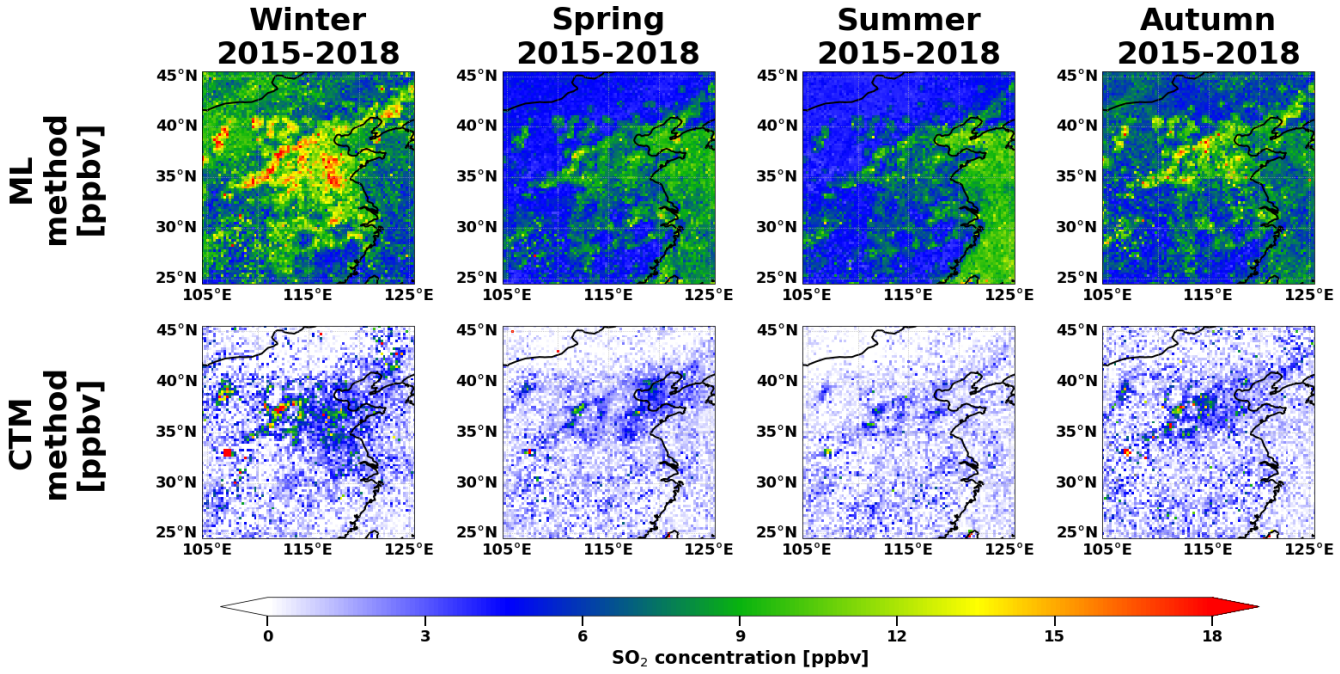


Figure S19: Maps of the annual mean surface SO_2 concentrations in ppbv from the ML method (top row) and CTM method (bottom row) over the study area at 0.25° x 0.25° horizontal resolution. Each column represents a different season averaged from 2015-2018.