Supplemental information for:

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

Zachary Watson[1], Can Li[2], Sean W. Freeman[1], Fei Liu[2,3], Huanxin Zhang[4], Jun Wang[4], Shan-Hu Lee[1]

[1]Department of Atmospheric and Earth Science, University of Alabama in Huntsville, Huntsville, AL 35758
[2]NASA Goddard Space Flight Center, Greenbelt, MD 20771
[3]Goddard Earth Sciences Technology and Research (GESTAR) II, Morgan State University, Baltimore, MD 21251
[4]Department of Chemical and Biochemical Engineering, Center for Global & Regional Environmental Research, and Iowa Technology Institute, The University of Iowa, Iowa City, Iowa 52240

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

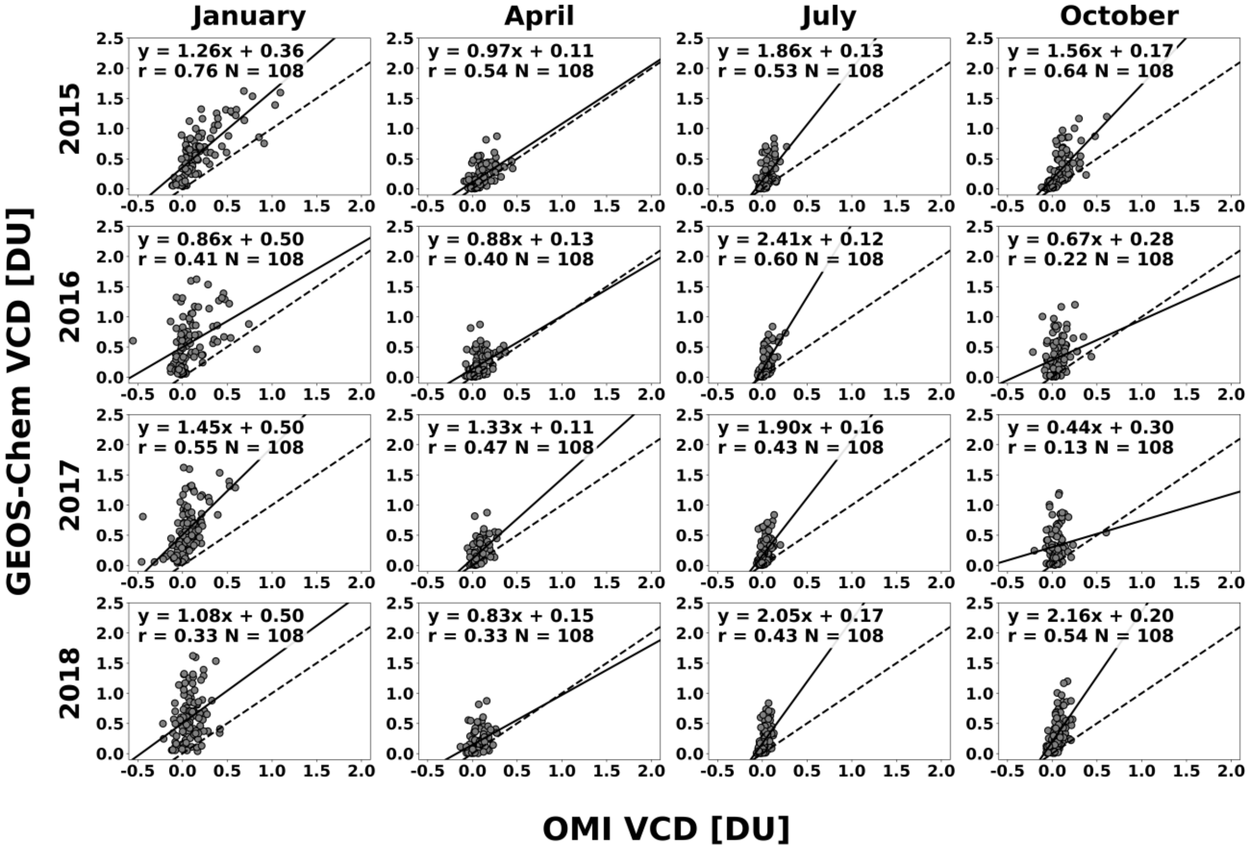


Figure S1: Scatterplots between monthly mean GEOS-Chem SO_2 VCDs and OMI SO_2 VCDs gridded to the model resolution (2.5° x 2.0°). Each column represents a different monthly simulation (from left to right: January, April, July, and October). Each row represents a year of the study period (from top to bottom: 2015, 2016, 2017, and 2018). For each month, the GEOS-Chem SO_2 VCDs stay constant with values from the 2015 simulations and the OMI VCDs change with time. Note: the linear regression fit for October 2016 and 2017 are not statistically significant at the 90% significance level due to a lack of data availability during that year, likely due more days with cloud cover during the OMI overpass.

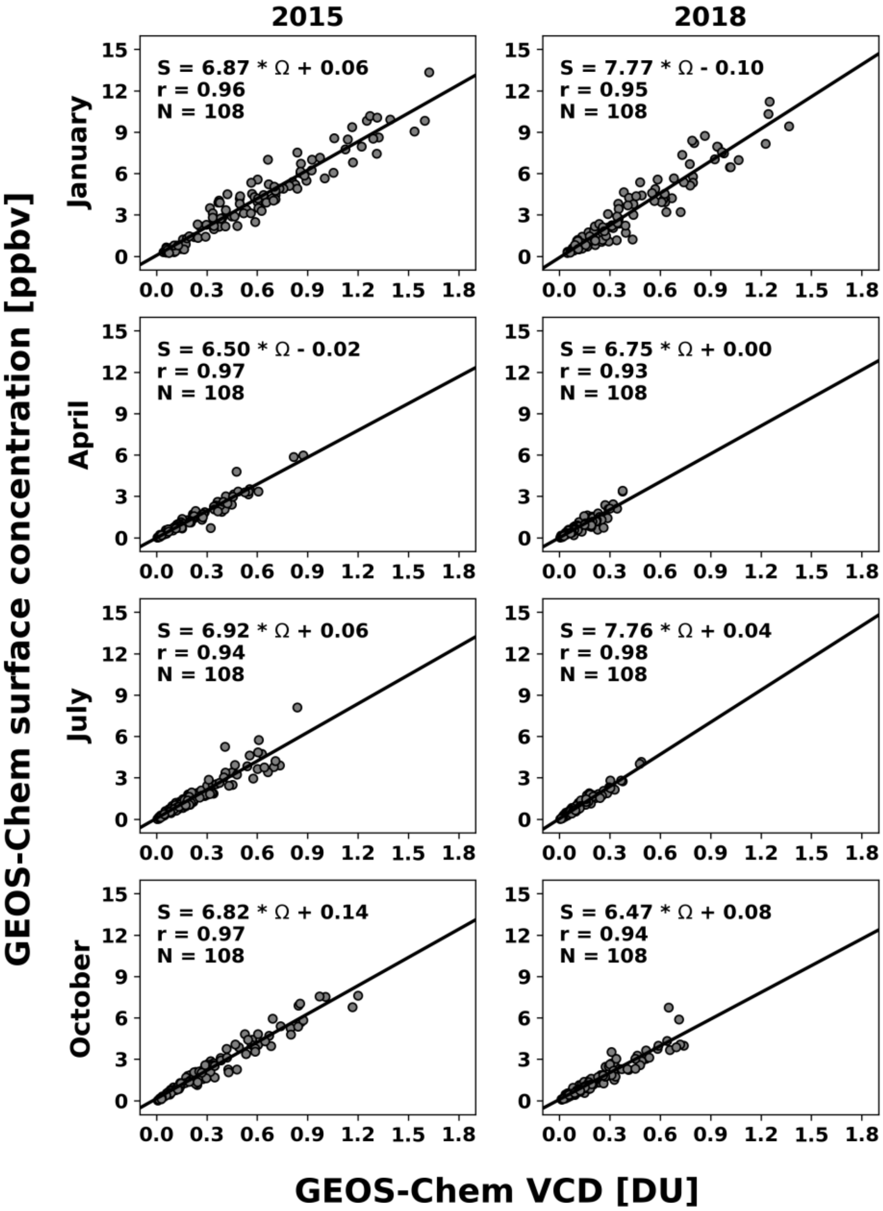


Figure S2: Scatterplots between the monthly mean simulated surface SO_2 concentrations and SO_2 VCDs from the 2015 and 2018 GEOS-Chem simulations. Each panel contains a linear regression analysis with the best fit line (solid line) and number of grid cells used in the analysis. The slope represents the simulations surface-to-VCD ratio (SVR).

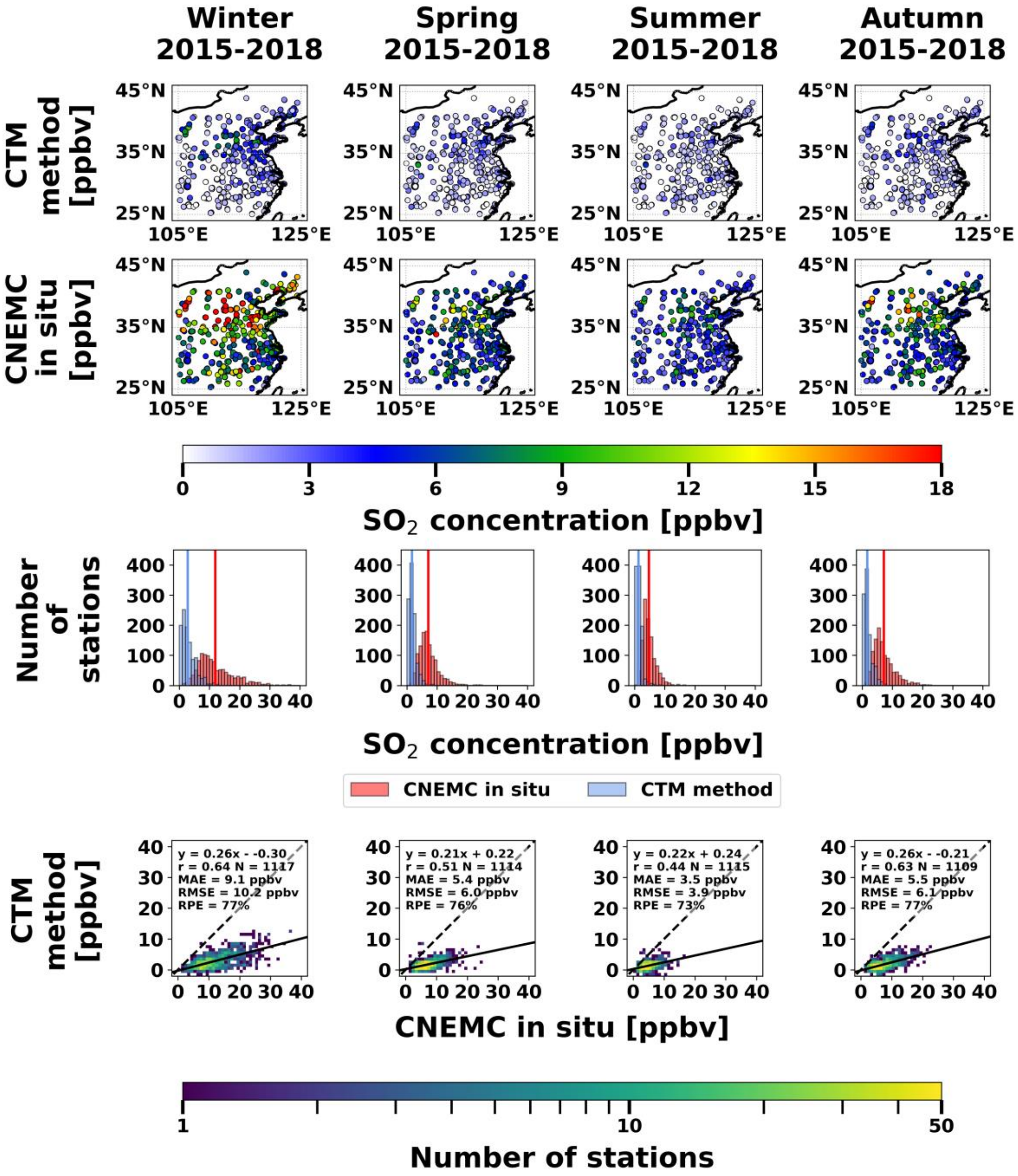


Figure S3: Spatial distributions of the seasonal surface SO_2 concentrations averaged from 2015-2018 for the CTM-based method (top row) and CNEMC in-situ measurements (second row), histograms of the surface concentrations from each dataset with vertical bars representing the means (third row), and scatterplots 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 lines), 1:1 line (black dashed line), MAE, RMSE, and RPE.

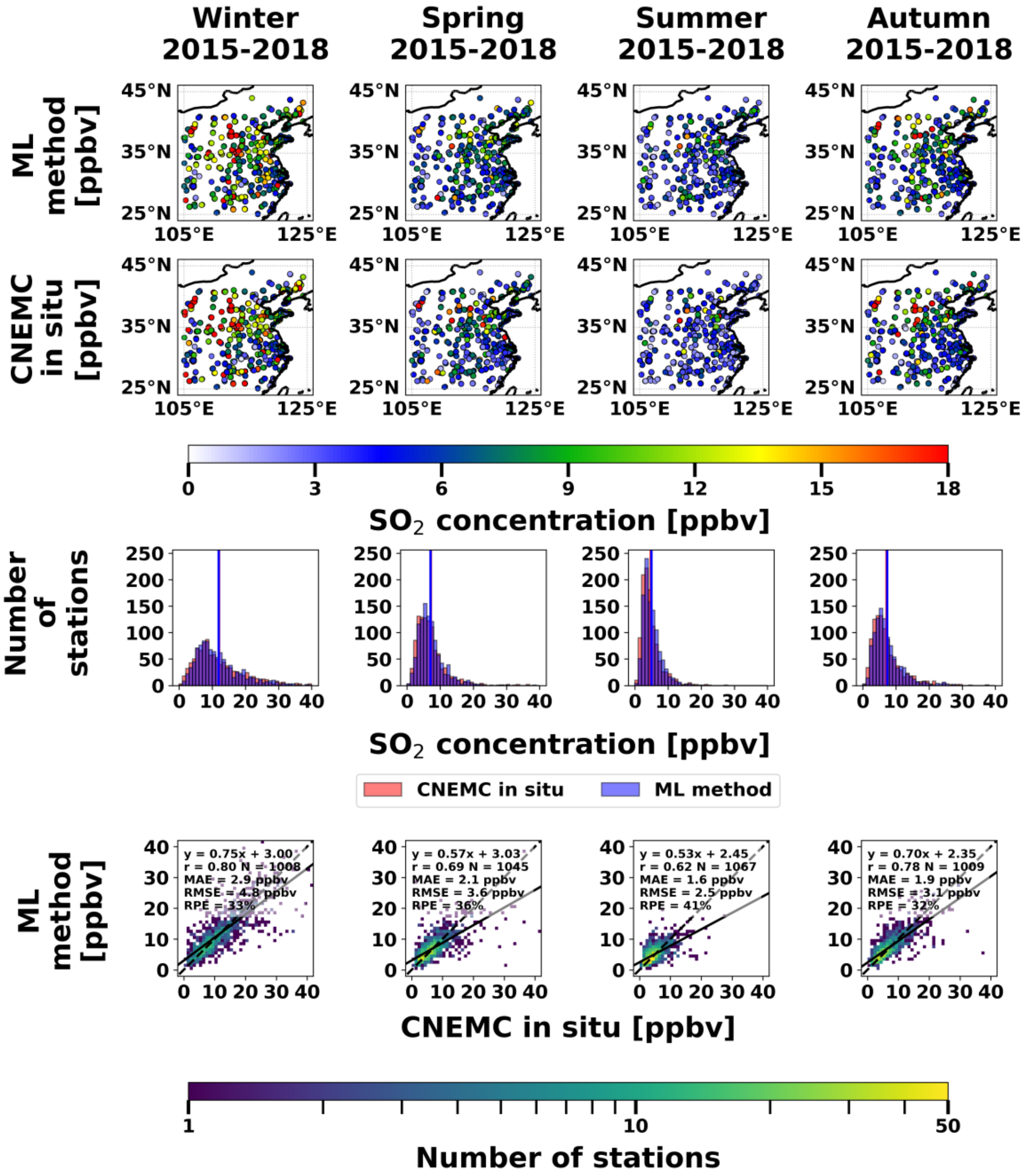


Figure S4: Spatial distributions of the seasonal surface SO$_2$ concentrations averaged from 2015-2018 for the ML-based method (top row) and CNEMC in-situ measurements (second row), histograms of the surface concentrations from each dataset with vertical bars representing the means (third row), and scatterplots 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 lines), 1:1 line (black dashed line), MAE, RMSE, and RPE.

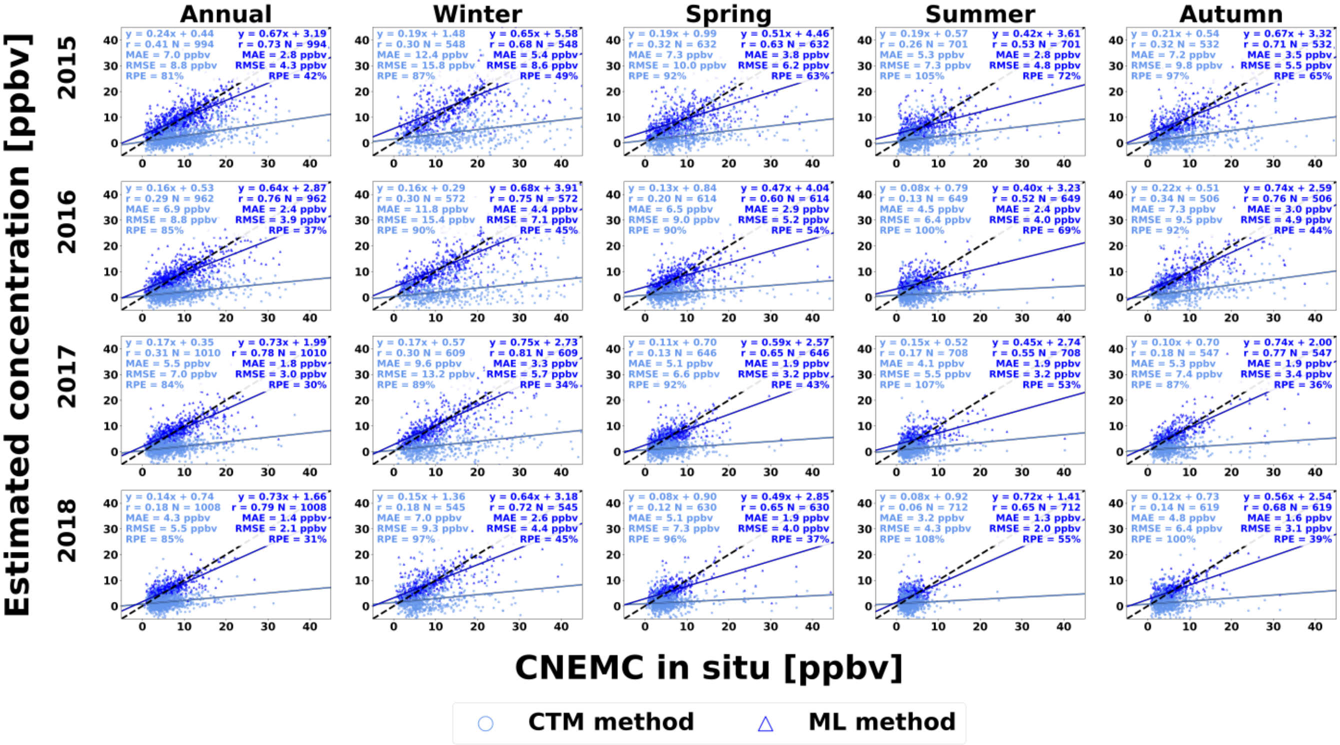


Figure S5: Scatterplots showing the estimated surface SO_2 concentrations from the CTM method (light blue circles) and ML method (dark blue triangles) against the in-situ measurements for individual years and seasons during the study period. Each column represents a different averaging period (from left to right: annual, winter, spring, summer, and autumn), and each row represents a different year of the study period (from top to bottom: 2015, 2016, 2017, and 2018). Each scatterplot includes a linear regression analysis with the best fit line (solid lines), 1:1 line (black dashed line), MAE, RMSE, and RPE.

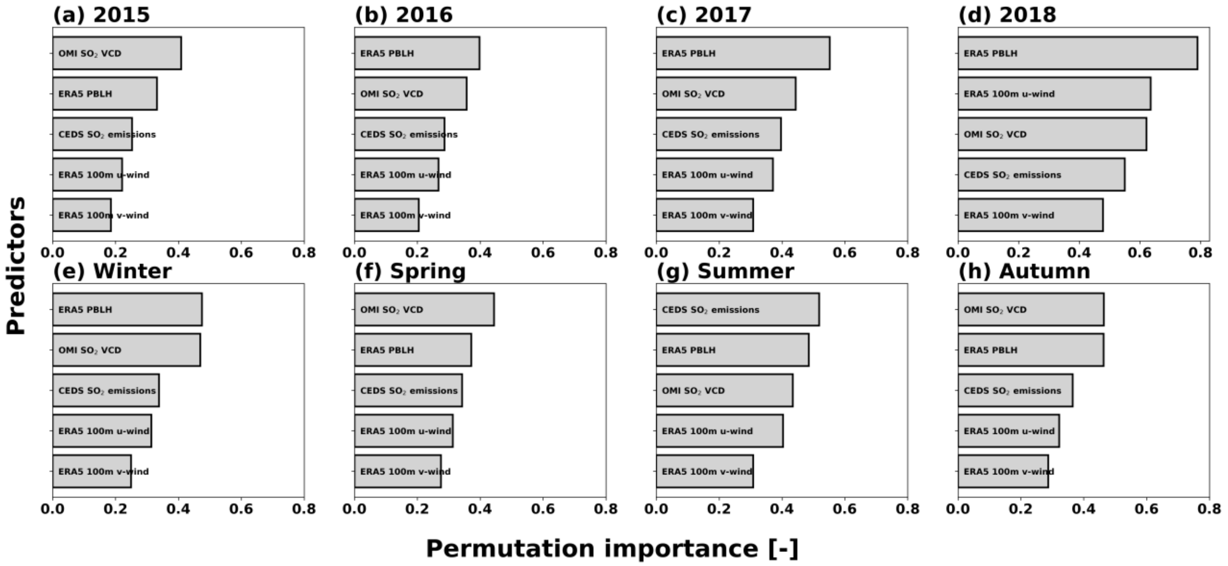
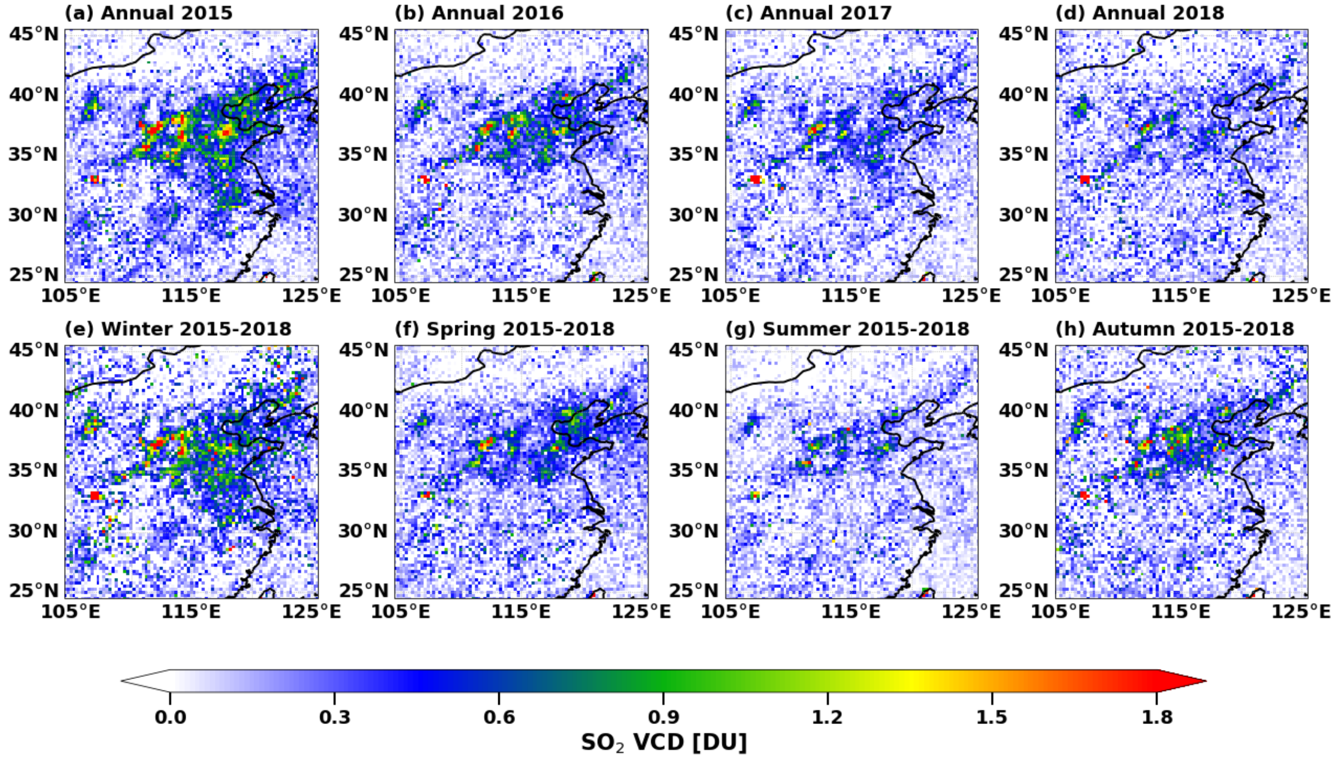


Figure S6: Permutation importance analyses to see how the ML model changed its predictions as a function of (a-d) year and (e-f) season. Longer bars indicate a larger impact on the results.



55 **Figure S7: 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° 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 the top row represents the seasonal means averaged from 2015-2018 for each season.**

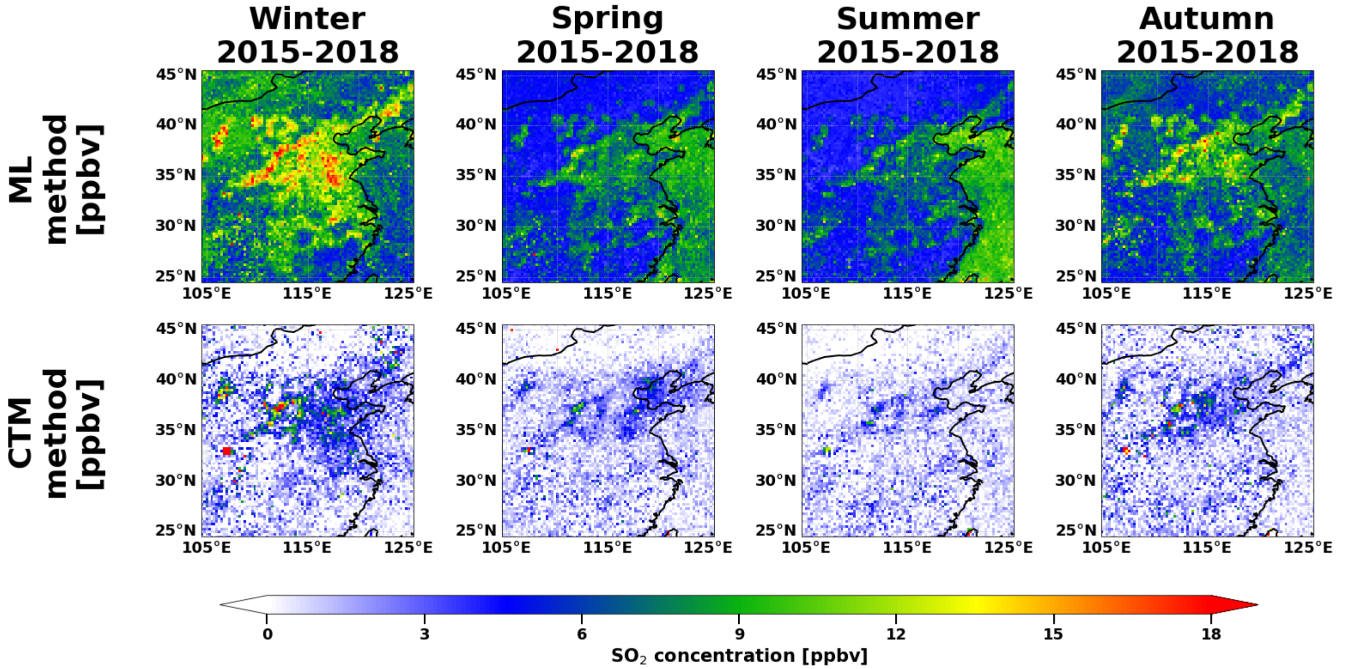


Figure S8: Maps of the annual mean surface SO_2 concentrations in ppbv from the ML method (top row) and CTM method (bottom row) over the study area at 0.25° horizontal resolution. Each column represents a different season averaged from 2015-2018 (from left to right: winter, spring, summer, and autumn).