Using Machine Learning to Estimate Surface-Level Sulfur Dioxide Concentrations from Satellite-Based Measurements

Zachary Watson[1], Sean W. Freeman[1], Can Li[2], Joanna Joiner[2], and 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

Abstract

Sulfur dioxide (SO_2) is a criteria air pollutant due to its contributions to aerosol formation, rainfall acidification, and harm to human health. The placement of air quality monitoring sites is typically biased towards urban areas, leaving large areas with very limited monitoring data. The Ozone Monitoring Instrument (OMI) has been used to provide estimates of SO_2 vertical column densities (VCDs) globally at spatial resolution of 10s of kms once per day. OMI SO_2 VCDs have been previously used to estimate surface SO_2 concentrations using chemical transport model (CTM) simulations. The CTMs use estimated emissions and assimilated meteorological data, and simulate the chemical and physical processes that determine the vertical profile of SO_2, which can be used to derive a ratio between the surface concentrations and VCDs. These models are complex, computationally expensive, and have large uncertainties in the simulated surface-to-VCD ratio due to biases in emissions and relatively coarse resolution. Machine learning techniques are comparatively easier to use, much less computationally expensive to use after training, and can produce more accurate estimations of surface concentrations than the CTM-based method. The interpretation of machine learning models often poses challenges, and in some cases, non-physical variables unrelated to SO_2 are used as predictors. In this work, we create an artificial neural network (ANN) to relate OMI retrievals and archived GEOS-FP boundary layer heights to surface SO_2 concentrations from the ChinaHighAirPollutants ChinaHighSO2 dataset (CHAP; Wei et al., 2023) on a seasonal average timescale from 2013-2018. Our model only utilizes five variables that are directly relevant to the satellite retrieval, lifetime, and spatial distribution of SO_2. The model was trained on 16 seasons (four of each) with independent validation (one of each season) and testing datasets (one of each season) to avoid overfitting. Our ANN generates surface SO_2 concentrations that are sensitive (slope = 0.51) and consistent ($r = 0.74$) with the CHAP data, but are underpredicted by an average of 1.2 ppbv with a mean absolute error of 2.2 ppbv. These results are better than recent studies utilizing the CTM method. To our knowledge, this is the best performing machine learning model that only uses physical variables to predict surface SO_2. Our work demonstrates that a carefully constructed, simple ML model can accurately estimate surface-based SO_2 concentrations from satellite VCD measurements, and this technique has future promise to expend to newer, higher resolution satellites and other air pollutants.