Authors' Response to Referee Comments (egusphere-2025-1735)

The authors would like to thank the editor and the two reviewers for taking the time to review our manuscript and provide critical, constructive feedback on our work. The suggestions by the reviewers helped to improve the novelty of the work and provide deeper insights to support our conclusions.

To address suggestions from reviewer comment 1 (RC1), we first performed bootstrapping to estimate individual sources of uncertainty. These were combined with uncertainties derived from assumptions in the methodology provide a summation of the uncertainties through error propagation for the chemical transport model (CTM) based method and bootstrapping-based sensitivity tests for the machine learning (ML) based method. We also provided additional insights about our model simulations to address the consistent underestimation of the surface sulfur dioxide (SO_2) concentrations from the CTM-based method and improve the interpretation of the ML-based method, specifically related to the permutation importance analysis. Additionally, we provided evidence that support methodological decisions including using a 40 km averaging radius for the CTM-based method and the split between training and independent testing datasets for the ML-based method. Finally, we incorporated all minor comments to improve the clarity of the text.

To address suggestions from reviewer comment 2 (RC2), we used a full year of archived GEOS-CF simulations, which have higher spatial and temporal resolution than our GEOS-Chem simulations, to address the impacts of spatial resolution and temporal representativeness of model simulations on the results of the CTM-based method. We also performed additional sensitivity tests for the eXtreme Gradient Boosting (XGBoost) machine learning to show how changing the architecture, as well as changing the source and resolution of the meteorology dataset, impacts the predicted surface concentrations from the ML-based method. Finally, we also showed that the XGBoost model performs better than other ML model types and has several benefits of using a small number of predictors despite a slight reduction in performance compared to previous studies.

In this document, we responded point-by-point to the referee comments with the original referee comments shown in blue, the authors' response and discussion for each comment shown in black, and additions or changes to the manuscript text shown in red.

Referee Comment 1 (RC1):

This study examines the ability of two methods to translate satellite column measurements of SO_2 into surface concentrations. The authors focus their study over eastern China where there are substantial point sources of SO_2. They compare and contrast the abilities of these two methods – one involving the GEOS-Chem model and the other a machine learning model – to reproduce in situ surface measurements of SO_2 across their study region. They find that the machine learning model is generally better at reproducing observed spatial and temporal variation, although the GEOS-Chem model approach also did a good job. They highlight that the GEOS-Chem model approach is typically better in regions when in situ data are absent.

The following comments are intended to improve the value of the study to a wider readership.

1. The authors mention methodological uncertainties throughout the paper but this reviewer didn't see any summation of these uncertainties reported alongside the surface SO_2 concentration estimates. This kind of information would help any potential user to assess the usefulness of the reported data. This reviewer recommends that the authors explicitly state the origin and magnitude of each source of uncertainty. For example, using one year of model output to interpret multiple years of satellite data introduces an uncertainty that the authors have reported.

Response:

We agree that it is useful to summarize all sources of error and determine the combined uncertainty of the estimated surface SO_2 concentrations using both the CTM- and ML-based methods. In fact, this information has never been reported in previous studies for either method. First, we estimated the uncertainties in the input variables including Ozone Monitoring Instrument (OMI) SO_2 vertical column densities (VCDs), GEOS-Chem surface-to-VCD ratios (SVRs), and European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) meteorology using a moving-block bootstrapping technique (Flyvbjerg & Petersen, 1989). We randomly sampled a longitude, latitude, and date with replacement and used a five day "block" on each side of the random date. The standard deviation was calculated for each temporal "block" for 10000 random samples. The average standard deviation over all the bootstraps was then defined as the uncertainty for that variable.

The summation of error was the CTM-based method was determined using error propagation since it is based on an analytical relationship between the OMI SO_2 VCDs and GEOS-Chem SVRs. The bootstrapped uncertainty for the OMI SO_2 VCDs was determined to be 0.67 Dobson Units (DU; 1 DU = 2.69 x 10^{16} molecules cm^{-2}), and the bootstrapped uncertainty in the GEOS-Chem SVRs was 1.4 ppbv DU^{-1}. We also needed to quantify the uncertainties due to assumptions made regarding the temporal sampling and representativeness of the GEOS-Chem SVRs. The first assumption was only using the SVRs from the month in the middle of each season to convert the OMI VCDs for days within that particular season (quasi-seasonal assumption), and the second assumption was only using 1 year of simulated SVRs to convert four years of OMI VCDs into surface concentrations (single-year assumption). The uncertainty for the quasi-seasonal assumption was calculated using the average difference in the SVR between the month in the middle of the season and the other months in the seasons. We ran an additional GEOS-Chem simulation spanning March, April, and May (MAM) 2015 to see how the monthly average SVR changed within a full season. We also performed this analysis using a full year of archived GEOS-CF data for all seasons. For more information regarding the additional GEOS-Chem simulation, GEOS-CF data, and the impact of temporal representativeness on the CTM-based method, see RC2 General Comment 2. In short, the difference in SVR for the GEOS-Chem MAM simulation (0.61 ppbv DU^{-1}) was similar to that of MAM from GEOS-CF (0.57 ppbv DU^{-1}). Therefore, we calculated the uncertainty for the quasi-seasonal assumption from the full year of GEOS-CF to include the other seasons. This uncertainty for the entire year was determined to be 0.8 ppbv DU^{-1}. The uncertainty for the single-year assumption was calculated using the average difference in the GEOS-Chem SVR for each month in the 2015 and 2018 GEOS-Chem simulations. This uncertainty was determined to be 0.6 ppbv DU^{-1}. The bootstrapped uncertainty and the uncertainties from the methodological assumptions were combined into a single value by assuming they are independent and can be combined using the sum of the squares of each uncertainty, which results in a total uncertainty of 1.7 ppbv DU^{-1}. To complete the error propagation for the CTM-based method, we first simplified Eqn. 1 from the text:

$$S_{OMI} = (SVR_{GC}) \times \Omega_{OMI}, \qquad \text{(R1)}$$

where S_{OMI} is the estimated surface concentration, SVR_{GC} is the SVR from GEOS-Chem, and Ω_{OMI} is the OMI SO_2 VCD. We made this simplification because we determined that the sub-grid variability terms from Eqn. 1 mainly cancel each other out due to the negligible free-tropospheric component of the VCD (see RC2 Specific Comment 4 for more information). Next, we used Eqn. R1 in the general error propagation formula to derive Eqn. R2:

$$\sigma_{S_{OMI}} = \sqrt{{\sigma_{SVR_{GC}}}^2 (\Omega_{OMI})^2 + {\sigma_{\Omega_{OMI}}}^2 (SVR_{GC})^2}, \qquad \text{(R2)}$$

where σ is the uncertainty and the subscripts and other terms correspond with those in Eqn. R1. Plugging in the uncertainties from bootstrapping and the assumptions, the average propagated error in the surface SO_2 concentrations from the CTM-based method is 4.9 ppbv, which is very high compared to the concentrations obtained from this method, which are typically between 0 - 4 ppbv, as seen in Fig. 3 from the text. All sources and summation of uncertainty for the CTM-based method is summarized in Table R1.

Table R1: Sources and magnitudes of uncertainty for the CTM-based method. Uncertainties for the OMI SO_2 VCDs and GEOS-Chem SO_2 SVRs were determined using moving-block bootstrapping. The uncertainty for the quasi-seasonal SVR assumption was determined using GEOS-CF data. The single-year SVR assumption was determined using the 2015 and 2018 GEOS-Chem simulations. The overall uncertainty for the CTM-based method was determined using error propagation.

Variable	Uncertainty
OMI SO_2 VCDs	± 0.67 DU
GEOS-Chem SO_2 SVR	± 1.4 ppbv DU^{-1}
Quasi-Seasonal SVR Assumption	± 0.8 ppbv DU^{-1}
Single-Year SVR Assumption	± 0.6 ppbv DU^{-1}
Overall Uncertainty	± 4.9 ppbv

Determining the summation of error is much more difficult for the ML-based method. While moving-block bootstrapping can still be used to determine the uncertainty of the model inputs, since machine learning models act like a "black box," the relationships between the variables cannot be expressed analytically and error propagation cannot be completed. Instead, we estimated the overall uncertainty of the ML-based method by using a traditional bootstrapping technique to resample the training dataset with replacement and train many XGBoost models with the same architecture to see how the predictions of the model change with different input data.

We used the same moving-block bootstrapping technique as the CTM-based method to estimate uncertainties for the ERA5 meteorological variables. The uncertainties in the U- and V-wind speeds were each 1.9 m s-1, and the uncertainty in the boundary layer height was 326 m (Table R2). It is important to note that the uncertainty in the CEDS emission inventory could not be quantified here. The temporal resolution (monthly) is too coarse for moving-block bootstrapping, there is no "truth" dataset to validate them, and there have not been any reported uncertainties of the CEDS inventory in the literature (e.g., Hoesly et al., 2018; McDuffie et al., 2020).

To calculate the overall uncertainty, we resampled the training dataset from the text with replacement to build new, slightly different training datasets. These were then used to train an ensemble of XGBoost models with the same architecture to show how changes to the training dataset may impact the predicted surface concentrations. The models were assessed using the same independent testing dataset from the text to ensure the evaluation was consistent between the different models. The standard deviations of the estimated surface concentrations were calculated for each location and time in the independent testing dataset using each member of the ensemble. The overall uncertainty was then taken as the averaged standard deviation over all locations and times in the independent testing dataset. For the ML-based method, the overall uncertainty was estimated to be 2.0 ppbv. This is improved from the CTM-based method, especially since the typical surface SO_2 concentrations from the ML-based method ranged between 5 – 10 ppbv, as seen in Fig. 4 from the main text. It is worth noting that the overall uncertainty for the ML-based method does not necessarily account for uncertainty in the model inputs, but since traditional error propagation and summation of uncertainties are not possible for machine learning, this is our best estimate at how the training data can impact the predictions from the model. All sources and summation of uncertainty for the ML-based method is summarized in Table R2.

Since the discussion of methodological uncertainties is important for the reader and has not been extensively discussed in previous papers for either method, we decided to add a new subsection to the paper titled "Methodological uncertainties" under the "Data and methods" section to briefly summarize these findings and help contextualize each methodology for the reader. Note that in the response below, Table 1 and Table 2 correspond with Table R1 and Table R2, respectively:

Table R2: Sources and magnitudes of uncertainty for the ML-based method. Uncertainties for the OMI SO_2 VCDs and ERA5 meteorology were determined using moving-block bootstrapping. The overall uncertainty for the ML-based method was determined using bootstrapping on the training dataset and retraining multiple XGBoost models to estimate the uncertainty in the model training. The uncertainty for the CEDS inventory was not able to be quantified (NQ).

Variable	Uncertainty
OMI SO_2 VCDs	± 0.67 DU
ERA5 U-Wind	± 1.9 m s^{-1}
ERA5 V-Wind	± 1.9 m s^{-1}
ERA5 Boundary Layer Height	± 326 m
CEDS Emissions	NQ
Overall Uncertainty	± 2.0 ppbv

“2.6 Methodological uncertainties

This study provides the first detailed discussion of the individual sources and summation of uncertainties for either methodology. To estimate the uncertainty of the input variables for both methods, we performed moving-block bootstrapping with 10000 iterations on the daily gridded data. For each bootstrap, a horizontal coordinate and date was randomly sampled with replacement. For each random sample, a temporal block of five days in each direction from the randomly sampled day was used to calculate the standard deviation. After all bootstraps were completed, the uncertainty was defined as the average of the standard deviations calculated from each iteration. This was done for the OMI SO_2 VCDs, GEOS-Chem SVRs, and ERA5 meteorology. The uncertainty in the CEDS emission inventory was not included due to the monthly temporal resolution, and a lack of uncertainty quantification in previous literature (e.g., Hoesly et al., 2018; McDuffie et al., 2020).

For the CTM-based method, the summation of error was determined using error propagation. For the error propagation, Eq. 1 was simplified such that:

$$S_{OMI} = (SVR_{GC}) \times \Omega_{OMI}, \tag{4}$$

where SVR_{GC} is the monthly averaged SVR from the GEOS-Chem simulations. Equation 4 was used in the error propagation formula to obtain Eq. 5:

$$\sigma_{S_{OMI}} = \sqrt{\sigma_{SVR_{GC}}{}^2(\Omega_{OMI})^2 + \sigma_{\Omega_{OMI}}{}^2(SVR_{GC})^2} \tag{5}$$

where $\sigma_{S_{OMI}}$ is the propagated error of the CTM-derived concentration, $\sigma_{SVR_{GC}}$ is the uncertainty in the GEOS-Chem SVR, and $\sigma_{\Omega_{OMI}}$ is the uncertainty in the OMI SO_2 VCD. The uncertainty in the GEOS-Chem SVR was initially calculated with bootstrapping, but also needs to account for the uncertainties of the quasi-seasonal and single-year assumptions in the CTM-based methodology. The quasi-seasonal and single-year assumptions were defined and quantified in Section 2.4. These three sources of GEOS-Chem SVR uncertainty were assumed to be independent of each other and were combined using the sum of the squares of each term. The results of the bootstrapping and error propagation are shown in Table 1. Ultimately, the methodological uncertainty of the CTM-based method is ±4.9 ppbv when considering the uncertainties of the OMI SO_2 VCDs (±0.67 ppbv DU^{-1}) and GEOS-Chem SVRs (±1.7 ppbv DU^{-1}). The OMI SO_2 VCD uncertainty has a relative standard deviation of 136%, which is comparable to the reported uncertainty of 60 – 120% for moderately polluted areas from Li et al. (2020).

It is much less straightforward to propagate error through a ML model since it effectively acts as a “black box,” so analytical error propagation methods cannot be used. First, uncertainties of the ERA5 meteorological fields were calculated using the moving-block bootstrapping approach. To obtain the overall uncertainty, we used traditional bootstrapping techniques to resample the training dataset with replacement and train an ensemble of XGBoost models to obtain an uncertainty in the model output based on changes in the training data given to the model. To maintain consistency, the same independent testing dataset was used to make the model predictions for each bootstrap. The standard deviation was calculated for each station and day across the different models and was then averaged over space and time to obtain the overall uncertainty. The uncertainties of the ML inputs and overall uncertainty from the retraining analysis are shown in Table 2. For the ML-based method, the overall uncertainty was estimated to be ±2.0 ppbv, which is lower than the propagated error for the CTM-based method. The overall uncertainty for the ML-based method does not directly account for uncertainty in the model inputs, but since traditional error propagation and summation of uncertainties are not possible for ML, this is our best estimate at how the training data can impact the predictions from the model." (Lines 239-285).

2. Line 127: increasing the time steps by 50%. Please assure this reader that this adjustment does not violate the CFL condition.

Response:

Philip et al. (2016) performed sensitivity tests comparing the different internal timestep lengths on the accuracy of GEOS-Chem. This study does not mention the Courant number (C) or CFL condition, so we will calculate it ourselves based on the resolution. In our simulations, the transport timestep was lengthened from 600s to 900 s, the chemistry timestep was lengthened from 1200 s to 1800 s, and the radiation timestep was left at the default 10800 s. The Courant number was using Eqn. R3:

$$C = U \frac{\Delta t}{\Delta x}, \qquad \text{(R3)}$$

where C is the Courant number, U is the velocity of the fluid, Δt is the model timestep, and Δx is the grid spacing. We solved for a maximum value of C using a characteristic horizontal wind speed (U) of 10 m s^{-1}, transport timestep (Δt) of 900 s, and a horizontal grid spacing (Δx) of 2.0° (222,222 m), the Courant number for these values is 0.041, indicating numerical stability and non-violation of the CFL condition since C is much less than 1. We added the Courant number to the text to ensure future readers that we are not violating the CFL condition:

"The internal time steps for the chemistry and advection calculations in the model were lengthened by 50% from the default values to reduce simulation times while minimizing errors following Philip et al. (2016). Despite the longer internal timesteps, the Courant-Friedrichs-Lewy condition is maintained with a Courant number of 0.041, indicating numerical stability of the simulations." (Lines 136-139).

3. In the description of the GEOS-Chem technique, this reviewer was curious about SO_2 retrievals with little or no sensitivity to the surface, perhaps due to elevated aerosols over industrialised regions. In those cases, perhaps the retrievals are removed from further analysis but the authors might also be misallocating an SO_2 column with elevated values in the free troposphere to changes in the surface. This might help to explain the results shown in Figure 3. This reviewer is also wondering whether it might also explain why boundary layer height is the single most important predictor for the machine learning model (Figure 5). At least some discussion is needed about this point.

Response:

Previous studies show that aerosols over eastern Asia (including China) are typically from local, anthropogenic sources, primarily due to fossil fuel combustion (Yang et al., 2024). Measurements from satellites and ground-based sun photometers indicate that aerosols over eastern China typically have SSA values around 0.90 (Devi & Satheesh, 2022; Jiang et al., 2024; Zhang and Li, 2019; Zheng et al., 2021) and have higher SSA values during high aerosol loading (Zheng et al., 2021). This indicates that these aerosols considered to be slightly to moderately absorbing based on definitions from Devi & Satheesh (2022) and Zheng et al. (2021) and are also more reflective at high concentrations. The OMI SO_2 retrieval algorithm does not directly account for the radiational effects of aerosols in either the spectral retrieval or air mass factor calculation to convert from slant column densities to vertical column densities (Li et al., 2013; Li et al., 2020). However, since the SSA values from urban aerosols over China are large, columns with high aerosol loading may be removed by the cloud screening algorithm. Quantitatively investigating the impacts of aerosols on the OMI SO_2 retrieval is far beyond the scope of this work and would require its own study.

One of the fundamental assumptions of the CTM-based method is that the model has the correct profile shape and is able to correctly attribute SO_2 to the proper layer of the atmosphere, as explained in Liu et al. (2004). Figure R1 shows the vertical SO_2 profiles from the GEOS-Chem for each of the simulations at several diverse locations in the study region including the North China Plain, Qin Mountains, southeastern China, northeastern China, and Inner Mongolia. The profiles consistently show the highest concentrations at the surface and within the boundary layer, with rapidly decreasing concentrations into the free troposphere and above. The profiles from our GEOS-Chem simulations have similar concentrations and shapes compared to those from aircraft-based measurements (e.g., Li et al., 2012; Norman et al., 2025; Shan et al., 2025; Xue et al., 2010) and nested domain GEOS-Chem simulations (Norman et al., 2025) over China, especially in the lowest 5 km of the atmosphere. Additionally, the standard deviations of the vertical profile are small in the free troposphere as indicated by the error bars, suggesting that the SO_2 concentrations are consistently low and the model does not regularly capture elevated plumes (Fig. R1). This highlights a potential limitation of the CTM-based method. If OMI detected an elevated plume of SO_2, this would increase the OMI VCD, but the CTM-based method would incorrectly attribute this to the surface, resulting in overestimated surface concentrations rather than the consistent underestimation seen in Fig. 3 from the main text. Figure R2 shows scatterplots of the seasonal average CNEMC surface SO_2 concentrations against the OMI VCDs. The moderate correlation between the two suggests that on relatively long timescales, such as the seasonal and annual averaging periods investigated in this work, OMI VCDs are generally representative of near surface SO_2 with higher VCDs in locations with higher surface concentrations.

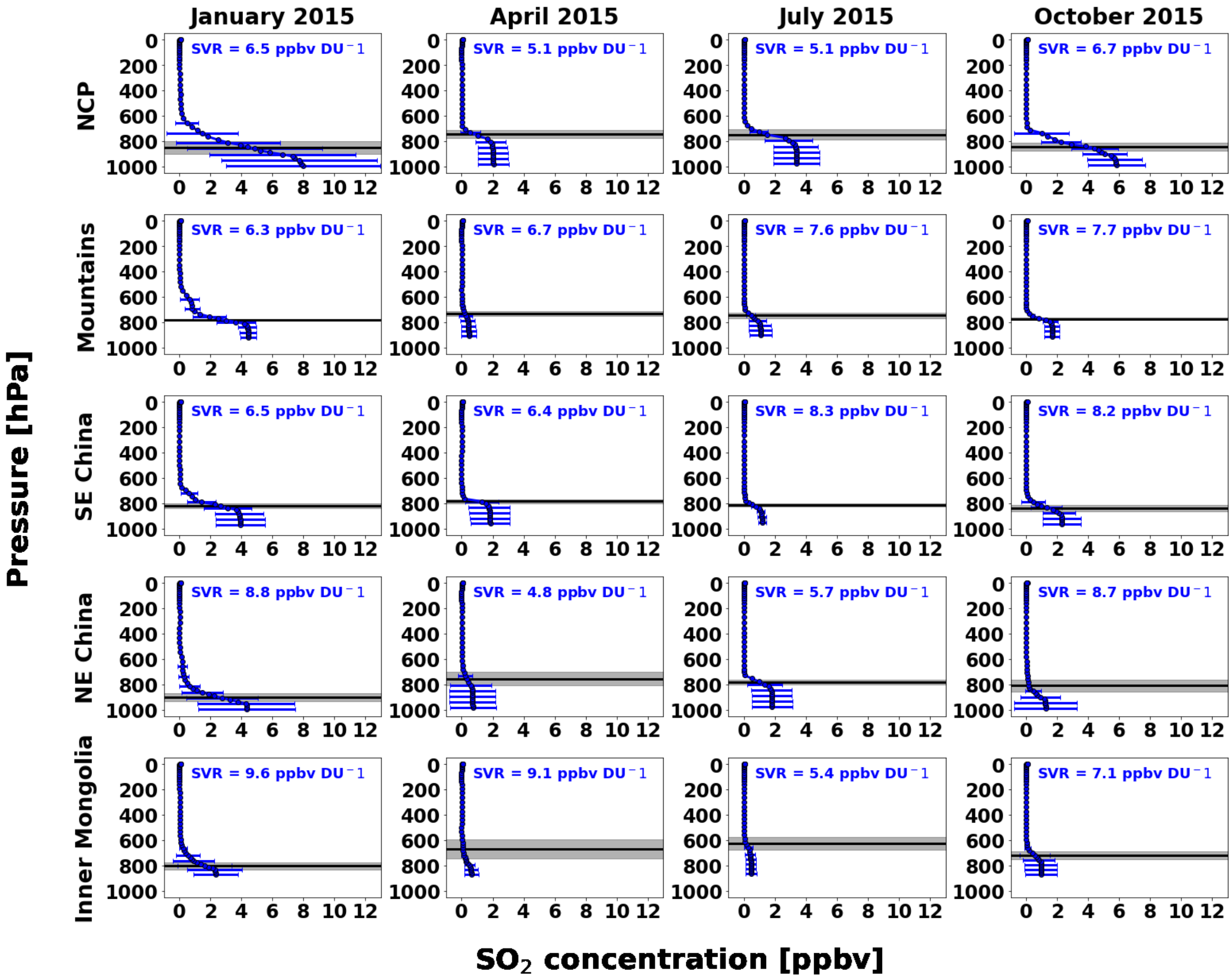

Figure R1: Monthly averaged vertical SO_2 profiles (blue lines) from each of the 2015 GEOS-Chem simulations (from left to right: January, April, July, and October) with 1 standard deviation error bars, GEOS-Chem SVRs, and GEOS-FP boundary layer heights (black line with 1 standard deviation shading) at 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). This figure was added as Fig. S1 in the supplementary material.

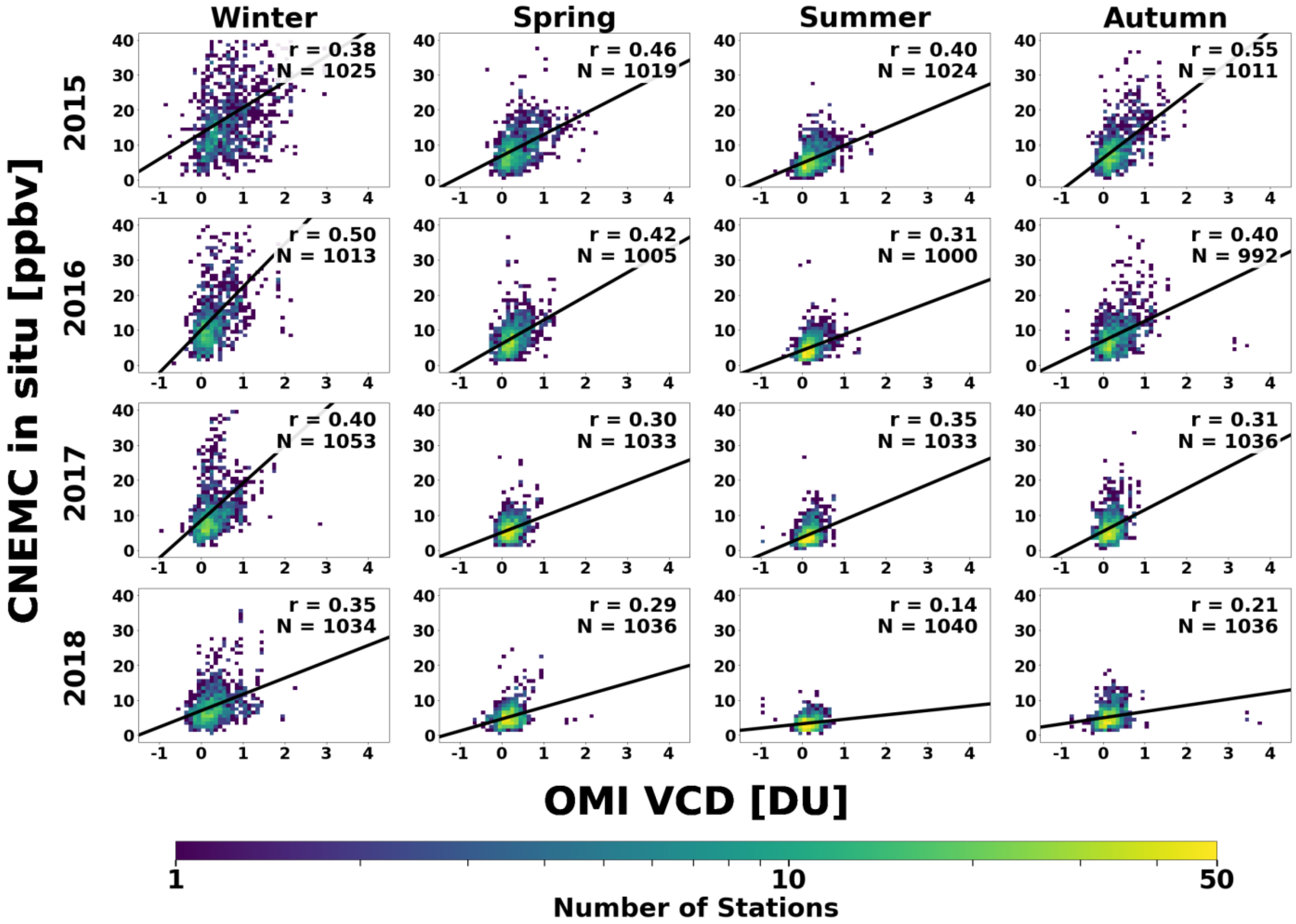

Figure R2: Scatterplots between the seasonal average CNEMC in situ surface SO_2 concentrations and OMI SO_2 VCDs. Each column represents a different season, and each row represents a different year of the study period. Each scatterplot is colored by the number of stations in each bin (every 1 ppbv for surface SO_2 concentrations, and every 0.1 DU for OMI SO_2 VCDs). Each scatterplot contains a linear regression analysis with the best fit line (solid line), with the correlation coefficient and number of stations.

Rather than misattributed OMI SO_2 columns, the consistent underestimation is likely from an underestimated SVR in the model. Figure R3 shows boxplots of the monthly average SVRs from observations (calculated from CNEMC surface concentrations and OMI VCDs) and the GEOS-Chem simulations. The boxplots show that the SVRs in GEOS-Chem are significantly lower than the observed, nearly by a factor of five. This may be due to an incorrect profile shape, which is dictated primarily by the magnitude of the surface concentrations and boundary layer height. Due to the coarse resolution, the average surface SO_2 concentrations in GEOS-Chem for January, April, July, and October 2015 were 6.4 ppbv, 2.3 ppbv, 2.8 ppbv, and 3.8 ppbv, respectively, which is much lower than the CNEMC surface concentrations of 23.7 ppbv, 9.9 ppbv, 6.0 ppbv, and 9.2 ppbv, respectively. The underestimated surface SO_2 concentrations in GEOS-Chem may result in an unrepresentative profile shape, which could affect the accuracy of the SVR. Additionally, since most of the SO_2 in the vertical column is located near the surface, the boundary layer height also plays an important role in defining the profile shape in the lower part of the atmosphere (Lin & McElroy, 2010). Both of these factors may contribute to inaccuracies in the SVR for the CTM-based method.

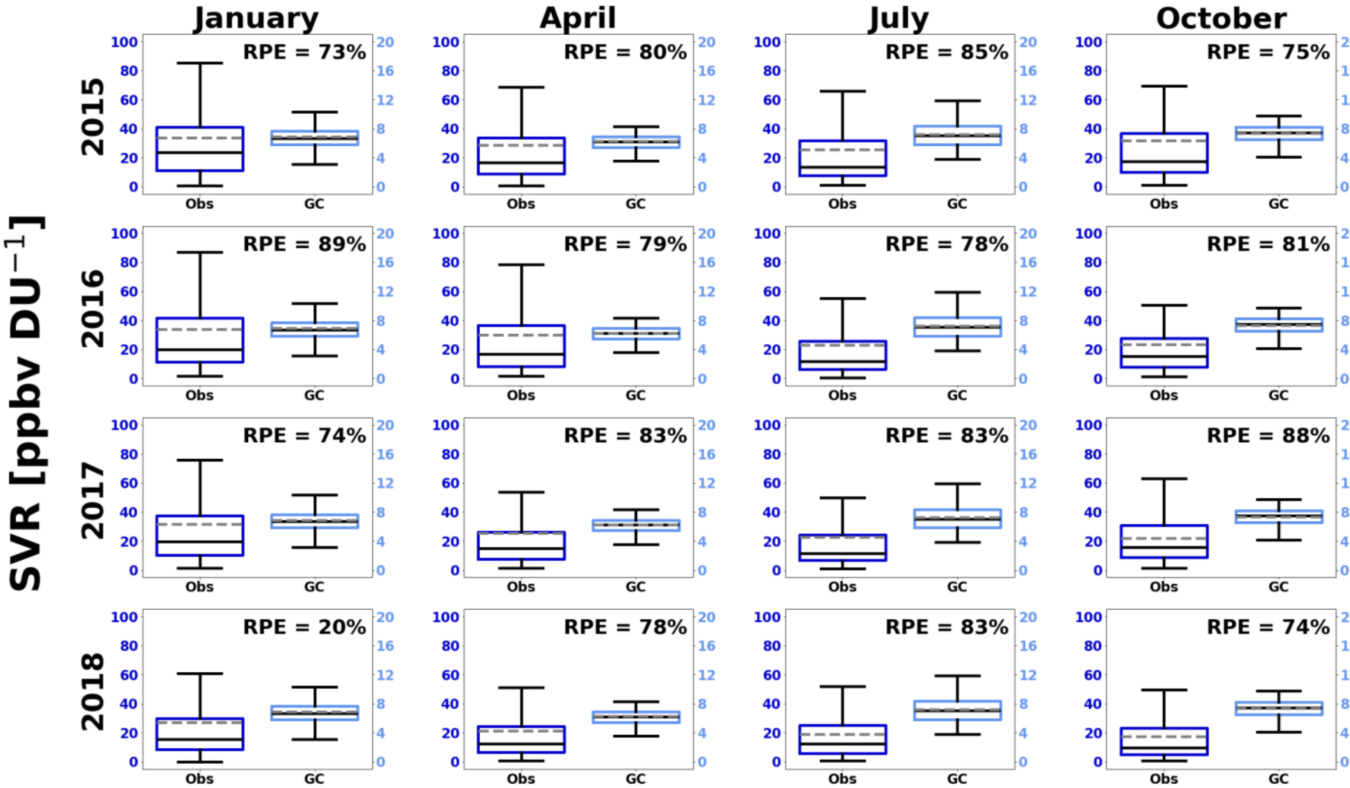

Figure R3: Boxplots showing the monthly averaged observed surface-to-VCD ratio (SVR) from CNEMC in situ surface SO_2 concentrations and OMI SO_2 VCDs (Obs; dark blue) and SVRs from the GEOS-Chem model (GC; light blue). Each column represents a different monthly average (from left to right: January, April, 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. Note that the y-axis 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. This figure was added as Fig. S6 in the supplementary material.

For the ML model, the boundary layer heights may be the most important predictor since it has a pronounced seasonality that is inversely related to the surface concentrations through known physical processes. During periods of cold temperatures and low solar radiation (i.e., winter months), the boundary layer heights remain low, which traps SO_2 near the surface with very little convective dilution, leading to high concentrations. The opposite is true during periods of warm temperatures with intense solar radiation (i.e., summer months), leading to abundant convective mixing, which raises the boundary layer height and allows for SO_2 to dilute and disperse, leading to lower concentrations. Maps of the seasonal mean ERA5 boundary layer heights are shown in Fig. R4. The matching seasonality due to the physical link between boundary layer heights and surface concentrations is likely why it is the most important predictor in the model. This is especially true since most of the SO_2 is located in the near surface layer, as shown in Figure R1, so it makes sense that the concentrations are highly influenced by the boundary layer height. Additionally, based on the bootstrapped uncertainties from RC1 General Comment 1, there short-term variations in the boundary layer heights are much smaller than the OMI VCDs with relative standard deviations of 20% and 136%, respectively. This may allow the ML model to learn the relationships between boundary layer heights and the observed surface SO_2 concentrations more easily than between the surface concentrations and OMI VCDs.

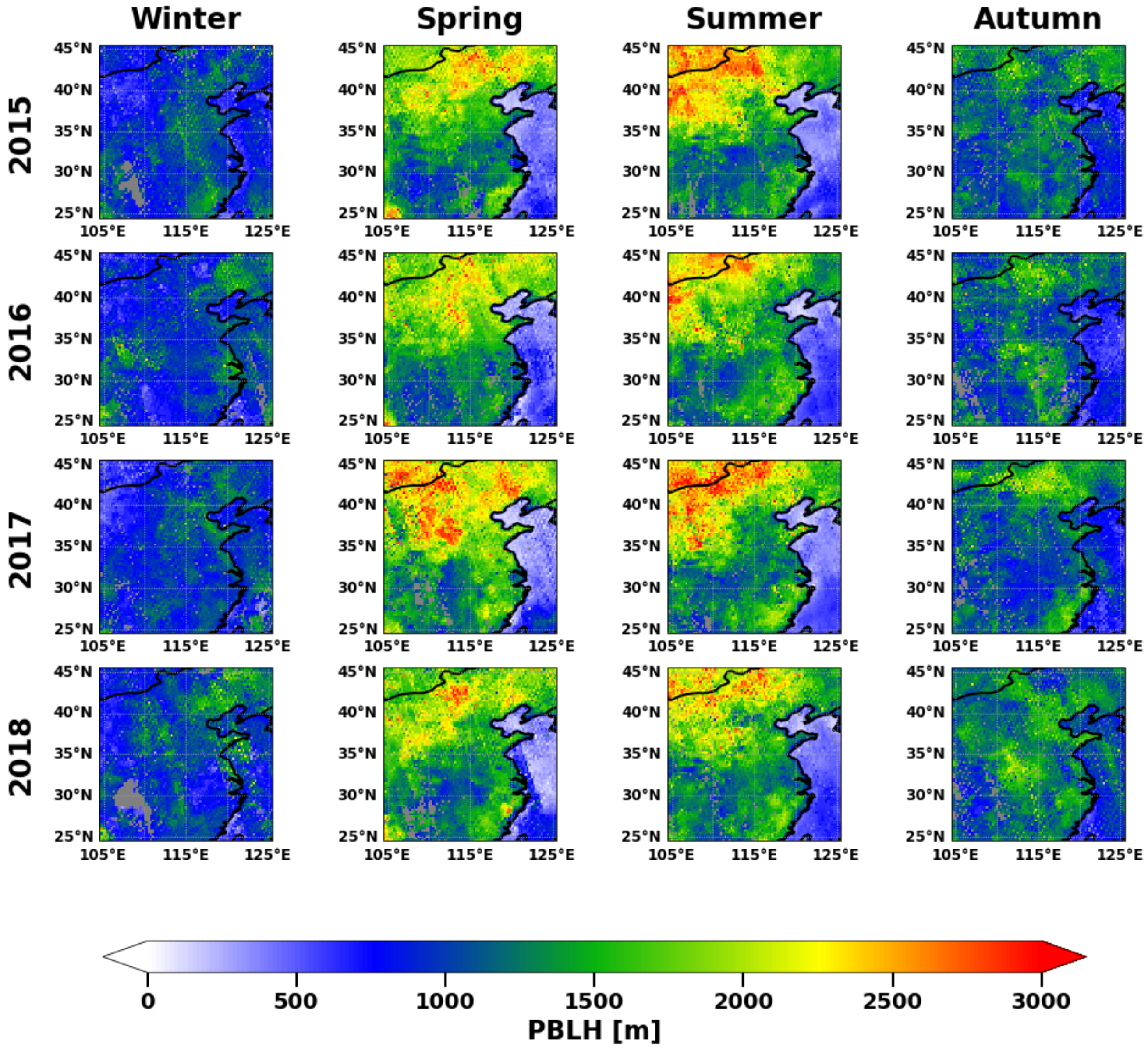

Figure R4: Maps of the seasonal mean ERA5 boundary layer heights in m. Each column represents a different season, and each row represents a different year of the study period. This figure was added as Fig. S15 in the supplementary material.

In the text, we added literature support for the representativeness of our GEOS-Chem simulated SO_2 profiles and briefly discussed their limitation for elevated SO_2:

"The approach from Lee et al. (2011) was used to infer surface SO_2 concentrations from OMI VCDs and simulated vertical SO_2 profiles from GEOS-Chem (GC). Lee et al. (2011) showed that the CTM-based method provided accurate results even with CTM resolutions that are much coarser than the satellite data. The monthly averaged profiles and SVRs from GEOS-Chem are shown in Fig. S1. The profiles indicate that most of the SO_2 within the vertical column is located near the surface and within the boundary layer (Fig. S1). The concentrations then drop to near zero in the free troposphere and have small variations, indicating a lack of elevated SO_2 plumes (Fig. S1). The profiles from the GEOS-Chem simulations are similar to those from aircraft observations (e.g., Li et al., 2012; Norman et al., 2025; Shan et al., 2025; Xue et al., 2010) and higher resolution simulations (Norman et al., 2025) over China." (Lines 143-150).

We also included additional information about the role of the boundary layer height in the machine learning results and the comparison of the gridded products since the oceans provided very clear evidence about the influence of boundary layer heights on predicted surface SO_2 concentrations:

"We performed a permutation importance analysis to assess how each predictor impacted the model predictions. Figure 5a indicates that the boundary layer heights and OMI SO_2 VCDs are the two most influential predictors followed by emissions and wind speeds. It is also worth noting that all of the predictors contribute toward the estimated surface concentrations with all permutation importance scores falling between 0.2 and 0.5 with none being unused. The boundary layer heights have a much smaller variation on short timescales compared to the OMI SO_2 VCDs. Based on the bootstrapped uncertainties from Table 2, the relative standard deviations are 20% for boundary layer heights, and 136% for OMI VCDs. As a result, the ML model is likely able to learn the relationship between boundary layer height and surface SO_2 concentrations more easily than the OMI SO_2 VCDs. Scatterplots between each ML predictor variable and the ML estimated surface SO_2 concentrations with Spearman rank coefficients (r_s) are shown in Figs. 5b-f. The ML-derived SO_2 concentrations increase with larger SO_2 VCDs and emissions, as well as decrease with increasing boundary layer heights and wind speeds (Figs. 5b-f). The surface concentrations and boundary layer heights each have a strong, inverse seasonality, as shown in Fig. S14 and Fig. S15, respectively, so the strong temporal correlations between them also likely lead to a high permutation importance in the model. The behavior of the ML predictions is consistent with the expected physical relationships between each predictor and the surface SO_2 concentrations. Large OMI VCDs and emissions indicate areas of high SO_2 loading, and large boundary layer heights and wind speeds lead to mixing and the dilution of SO_2. The magnitudes of the r_s values are small, indicating that the model may be making predictions based on the interactions between variables rather than any individual predictor. The small number of predictors used in our model allows us to link the ML predictions to known atmospheric processes, adding confidence to the model in its ability to accurately estimate the surface concentrations." (Lines 405-422).

"Since the ML predictions are significantly affected by boundary layer heights (Fig. 5), the model is likely incorrectly associating the low marine boundary layer with areas elevated SO_2, as suggested by the seasonally averaged ERA5 boundary layer heights in Fig. S15. Inaccuracies over the oceans have also been reported in Kang et al. (2021) where ML was used to estimate surface concentrations of NO_2 and ozone and was attributed to a lack of training data for the ML model in these locations. Since the ML model was only trained for conditions over land, it learned the relationship between high surface SO_2 concentrations and low continental boundary layer heights during the winter months but could not accurately apply this knowledge over the oceans." (Lines 513-519).

4. Line 140: Explain to the reader why 40 km was chosen? Do alternative values significantly change the results?

Response:

A radius of 40 km (i.e., nearest 3 x 3 grid of OMI pixels centered on the station) was chosen for several reasons. First, it is similar to values used in previous studies (e.g., Kharol et al., 2017). Additionally, a 40 km radius maximizes both the slope and correlation compared to radii of 10 (i.e., nearest OMI pixel) and 70 km (i.e., nearest 5 x 5 grid of OMI pixels centered on the station), as seen in Fig. R5. Since this was an important methodological decision, we incorperated this supporting evidence to the text and added Fig. R5 into the supplementary material as Fig. S2:

"To compare the estimated surface concentrations to the in situ surface monitoring data, we used a 40 km averaging radius around each station to increase the amount of usable data and further reduce the noise in the OMI data. This is similar to previous studies (i.e., Kharol et al., 2017) and maximizes both the slope and correlation compared to other radii, as shown in Fig. S2." (Lines 160-163).

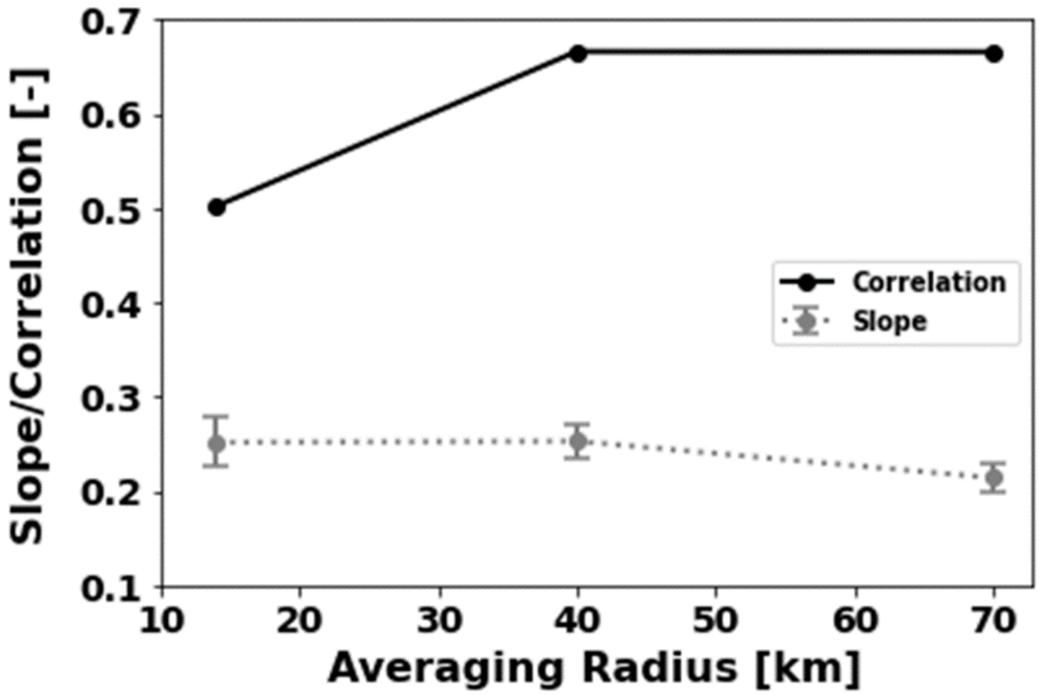

Figure R5: Change in correlation and slope of OMI-derived concentrations compared to CNEMC in-situ measurements 2015 as a function of the averaging radius around each CNEMC site. Error bars represent the 95% confidence interval in slope based on the standard error of the linear regression fit. Radii 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). This figure was added as Fig. S2 in the supplementary material.

5. Line 174: Is it normal practice to use so much data for training? Later in this paragraph the authors mention the resulting machine learning model overfitting the data. Have the authors considered using fewer data to train and more data to test the resulting model?

Response:

The split of 90% of the data for testing and 10% of the data for validation was used following previous studies including Zhang et al. (2022), Yang et al. (2023a), and Yang et al. (2023b) as stated in the text. Early in our development of the machine learning model, we initially used 80% of the data for training and 20% for validation, but we changed it to maintain consistency with the previous studies. After changing the split, we also saw that the performance of the independent testing dataset slightly improved.

To statistically show how the size of the training dataset impacts the validation results, we trained models using the same architecture as in the paper, but with different training dataset sizes varying between 10% and 90%. Figure R6 shows that as the amount of training data increases, the performance of the independent dataset becomes better, slowly leveling off for large training datasets. Figure R5 also shows as the amount of training data increases, the performance of the predictions based on the training dataset decreases, which also levels off for large training datasets. The model is still technically overfitting because of the notable gap in performance between the training and testing datasets, but this indicates that using a large training dataset actually yields the most comparable performance between them. Figure R6 was added to the supplemental material as Figure S9:

"This split of the training and independent testing datasets was used by previous studies (e.g., Zhang et al., 2022; Yang et al., 2023a; Yang et al., 2023b) and was shown to have the best performance for the independent testing dataset for our model as shown in Fig. S12." (Lines 225-227).

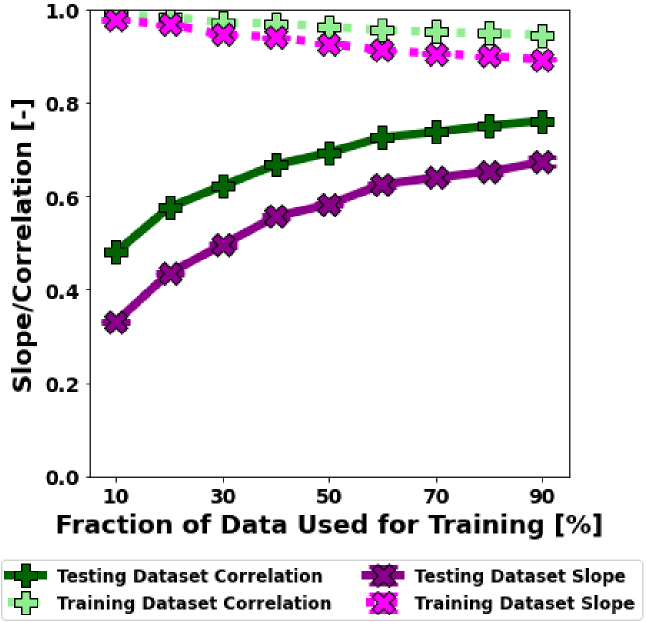

Figure R6: Slopes (magenta crosses) and correlations (green crosses) of XGBoost predicted surface SO_2 concentrations validated against CNEMC in situ measurements as a function of the amount of data used to train the model for both the training (light shades) and independent testing datasets (dark shades). This figure was added as Fig. S12 in the supplementary material.

6. Line 420: This reviewer may have missed this point in the manuscript, but I didn't see any evidence that the GEOS-Chem approach reproduced temporal distribution observed by the CNEMC in situ data. Figure 7 shows a muted seasonal cycle with a correlation coefficient of typically less than 0.4 so the model only captures at most 20% of the observed variation.

Response:

Due to the underprediction of the surface concentrations from the CTM-based method, its temporal variation has a much smaller amplitude compared to the CNEMC in situ measurements. The CNEMC measurements had changes of around 5–9 ppbv between the winter and summer seasons in each year while the CTM-based method had changes of around 1.5–2.5 ppbv. Despite the much smaller seasonal variations for the CTM-based method, the relative change between winter and summer were quite similar compared to the CNEMC observations and ML-based method. The CTM-based method had a relative seasonal fluctuation in the surface concentrations by a factor of 2.4 compared to a factor of 1.9 for the CNEMC measurements. Figure 7 is not able to effectively show this because of the significantly smaller magnitudes of the CTM-based surface concentrations.

Table R3: Annual and seasonal average surface SO_2 concentrations from 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

To make this result clearer to the readers, all data displayed in Fig. 7 can be found in Table R3, which was also added to the supplemental material as Table S1. We also clarified in the text that there was agreement in the relative temporal variations, but not the absolute temporal variations:

"The ML-based method also captured the same temporal variations as the in situ measurements each with a 44% decrease in concentrations from 2015-2018, and an average seasonal fluctuation by a factor of 1.9 between the winter and summer seasons (Figs. 7a-b). The CTM-based method also had good agreement in the temporal trends of the in situ measurements but was not as good as the ML-based method with a 36% decrease from 2015-2018 and a seasonal fluctuation by a factor of 2.4 (Figs. 7a-b). Since the CTM-based surface SO_2 concentrations were underestimated, the magnitude of the temporal trends is much smaller than the observations and ML-based method, but the relative change was similar, as shown by Table S1. The decrease in SO_2 from 2015-2018 detected by both methods is consistent with previous studies that showed a reduction in emissions over China shown using satellite VCDs (Li et al., 2017; Wang et al., 2020a), satellite-derived emissions (Fioletov et al., 2023), and surface concentrations (Wei et al., 2023; Zhang et al., 2021). Despite the similarities in the year-to-year and season-to-season variations, the greatest difference between the time series of the two methods was the magnitude of the concentrations." (Lines 468-478).

Minor points

1. Line 108: Regridding does not result in good data quality.

Response:

Agreed. We moved the regridding information to earlier in the paragraph to keep it separate from the data quality screening:

"Aura flies in a sun-synchronous polar orbit, and OMI is used to retrieve SO_2 VCDs with daily global coverage and a spatial resolution of 13 km x 24 km at nadir, a significant improvement from previous satellite-based instruments. The VCDs were gridded to a horizontal resolution of 0.25° x 0.25° to decrease noise in the SO_2 retrieval without significantly coarsening it from the native measurement resolution." (Lines 105-109).

Now, the information about data quality is limited to cloud fraction, solar zenith angle, cross-track position, and row anomaly and is more representative of the reference:

"To ensure good data quality, we screened out measurements with cloud fractions greater than 0.3, solar zenith angles greater than 65°, located in the outer ten cross-track positions, or affected by the row anomaly (NASA, 2020)." (Lines 116-118).

2. Line 120: The current version of GEOS-Chem bears little resemblance to the model described by Bey et al, 2001. Strongly suggest using a more updated reference.

Response:

Agreed. We replaced the outdated reference with the Zenodo site for the specific version of the model that we used. Most recently published studies utilizing GEOS-Chem either simply provide a link to the website (e.g., Keller et al., 2021) or to the Zenodo site for the version used (e.g., Norman et al., 2025) rather than citing published studies (unless discussing a particular scheme or module of the model).

"We used simulated SVRs from the GEOS-Chem model (version 14.2.2; The International GEOS-Chem User Community, 2023) to convert the OMI VCDs into surface concentrations for the CTM-based method." (Lines 128-129).

"The International GEOS-Chem User Community.: geoschem/GCClassic: GCClassic 14.2.2 (14.2.2), Zenodo [code], doi:10.5281/zenodo.10034814, 2023." (Lines 650-651).

3. Line 125: When stating horizontal resolution, this reviewer suggests you label which of the values represents latitude and longitude.

Response:

Agreed. This change is now reflected in the text:

"The model was run at a horizontal resolution of 2.5° (longitude) x 2.0° (latitude) with 47 vertical layers and was driven by assimilated GEOS-FP meteorology (Lucchesi, 2018) and the Community Emissions Data System (CEDS) anthropogenic emission inventory (Hoesly et al., 2018)." (Lines 133-135).

4. Line 136: difference in (horizontal) resolution…

Response:

This clarification was made in the text:

"Since there is a significant difference in horizontal resolution between the satellite and model data, OMI VCDs were used to provide sub-model grid variability (ν) using Eqn. 2:…" (Lines 155-157).

Referee Comment 2 (RC2):

This study tried to compare the CTM-based and ML-based approaches to estimate surface sulfur dioxide concentrations from satellite vertical column density from OMI over eastern China. However, my strongest concern is about the novelty of this paper that simply combined results from two cases that were rather basic in each approach and the analyses and conclusions did not offer new insights and almost identical as the combination from prior studies using only CTM-based or ML-based approach. In addition, for each approach, the treatment was oversimplified. For example, the CTM-based approach was conducted using a rather coarse-resolution simulation over only 4 months in a single year to compare with observations of annual means in 4 years, with poor performance compared to recent studies indicated by the correlation coefficients. The ML-based approach was conducted using only 5 input variables with poor performance compared to recent studies as well. The paper would benefit from better identifying the novelty of this manuscript and more science-based refinements on each approach.

Response:

Even though we used previously developed methodologies to estimate the surface concentrations using the CTM- and ML-based methods, there are several novel aspects of this work that expand upon what was accomplished by previous studies. First, the CTM- and ML-based methods have never been used over the same locations and times as one another. Furthermore, the accuracy of both methods has never been directly compared using the same truth dataset for validation. While the results from different studies using the CTM- and ML-based methods could be compared to one another, there was infrequently any overlap in the study locations, time periods, and truth datasets, so only general conclusions could be drawn regarding their relative performance. This work is also the first to present estimated surface SO_2 concentrations from the CTM-based method on a timescale shorter than an annual average. Our incorporation of seasonal means highlights the ability of the CTM-based method to capture temporal trends both within a year and over a period of several years for the first time. Finally, this work also presents a quantification of uncertainties for both methods that go far beyond the discussion of previous studies, including addressing the impacts of spatial resolution and temporal representativeness of CTM simulations on the estimated surface concentrations (as suggested by RC1 General Comment 1). Throughout the manuscript, we more directly highlighted where our work expanded on the findings from previous literature:

"For the first time, we quantified methodological uncertainties for both methods, directly compared their performance on the same truth dataset, and validated the CTM-based method on a sub-annual timescale." (Lines 17-19).

"Although the CTM- and ML-based methods have each been used to estimate surface SO_2 concentrations from satellite retrievals, there is a lack of direct comparisons between them. Here, we estimated surface SO_2 concentrations using OMI SO_2 VCDs over eastern China (105-125°E, 25-45°N) from 2015-2018 to directly compare the two methods. First, we quantified methodological uncertainties for each method for the first time. Next, we used simulated SVRs from the GEOS-Chem model to estimate the surface SO_2 concentrations from OMI using the CTM-based method. Then, we used a ML model to predict surface SO_2 concentrations from OMI VCDs, meteorological variables, and an emission inventory, which are all physically relevant to the spatial distribution or lifetime of SO_2. The results from each method were validated against ground-based in situ measurements from the China National Environmental Monitoring Centre (CNEMC) air quality monitoring network on annual and seasonal mean timescales, the latter of which has never been done for the CTM-based method. Finally, we compared the performance of each method on the same truth dataset over the same times and locations for the first time to gain insights on their abilities and limitations to accurately estimate the surface SO_2 concentrations from satellite data." (Lines 75-85).

"This study provides the first detailed discussion of the individual sources and summation of uncertainties for either methodology." (Lines 240-241).

"The surface concentrations from the CTM-based method were also separated by season, averaged from 2015-2018, and validated against in situ measurements for the first time. As shown in Fig. S13, the CTM-based method was able to accurately capture the spatial distribution (r = 0.56) and seasonality of the in situ measurements with higher concentrations in the winter and lower concentrations in the summer but still suffered from underestimation (slope = 0.24; RPE = 76%)." (Lines 314-318).

"Here, the CTM- and ML-based methods were directly compared using the same truth dataset over the same locations and study period for the first time. Each method was resampled to match the independent testing dataset (i.e., data retained from ML training) and the performance of each method was assessed given identically sampled data." (Lines 435-437).

"The novelty of this study includes a first time investigation of quantifying methodological uncertainties for both the CTM- and ML-based techniques, a validation of seasonal mean surface concentrations from the CTM-based method, and a direct comparison between the two methods on the same truth dataset." (Lines 541-544).

General Comments

1. Justification of the study region over eastern China. This study was for eastern China only, and should be reflected in the title. What is the rational of restricting the study area to eastern China only?

Response:

We chose to restrict our study to eastern China because it is a region of abundant SO_2 emissions, and we wanted to investigate a region that has a sufficient SO_2 signal to overcome the relatively small signal-to-noise ratio of the OMI SO_2 product. The justification of focusing on eastern China incorporated into the title and clarified in the text:

"Estimating surface sulfur dioxide concentrations from satellite data over eastern China: Using chemical transport models vs. machine learning" (Lines 1-3).

"Eastern China has abundant anthropogenic SO_2 emissions and thus is a region with elevated surface concentrations. Satellite SO_2 retrievals typically have a low signal-to-noise ratio due to interfering absorbers (Li et al., 2020), so regions with large SO_2 emissions and pollution, such as eastern China, are required to obtain sufficient signals from the spectrometer and provide more reliable retrievals compared to less polluted regions." (Lines 88-91).

2. For CTM-based approach, it is recommended to conduct full-year simulations over each year to be more temporally representative.

Response:

We agree that using a complete simulation over the four year study period would result in more temporally representative simulations; however, we do not have the computational resources to run GEOS-Chem for the full study period. Some previous studies have used CTM simulations that were not representative of the entire study period. For example, Kharol et al. (2017) only used two years of archived air quality forecasts to estimate surface SO_2 concentrations from 10 years of OMI data using the CTM-based method, although there was no indication on how those two years of simulations were used to convert the OMI VCDs into surface concentrations (i.e., monthly mean SVRs, annual mean SVRs, etc.). Since we only had simulations for January, April, July, and October 2015 to convert OMI VCDs into surface concentrations using the CTM-based method, we quantified the uncertainty of our assumptions regarding the temporal representativeness (as suggested by RC1 General Comment 1), as well as tested the impacts of temporal representativeness on the accuracy of the estimated surface concentrations. The first assumption is using the monthly average SVRs from January, April, July, and October to calculate daily surface concentrations within the winter, spring, summer, and autumn seasons, respectively, which is referred to as the quasi-seasonal assumption. The other assumption is using only 2015 SVRs to calculate surface concentrations over the full four year study period from 2015-2018, which referred to as the "single-year assumption."

To address the quasi-seasonal assumption, we first performed an additional GEOS-Chem simulation to quantify how much the monthly average SVR varies within a particular season compared to the month in the middle of the season. The simulation was run for spring (March, April, May; MAM) 2015. Boxplots showing the SVRs from all available simulations can be seen in Fig. R7a. Based on this simulation, we found the difference in the monthly mean SVR for April compared to March and May were 5% and 9%, respectively, suggesting a small amount of intraseasonal variability in the GEOS-Chem SVRs during the spring season. The average difference in SVR was 0.6 ppbv DU^{-1} for March and May; however, since the new GEOS-Chem simulation only covered one season, we also employed a full year of archived GEOS-CF (NASA GMAO, 2023) model data, which has higher spatial (0.25° x 0.25°) and temporal (hourly) resolution than the GEOS-Chem simulations. The GEOS-CF model is described in detail in Keller et al. (2021) and was found to produce similar results to GEOS-Chem due to using the same chemistry module. The first full year of archived GEOS-CF data is 2018, which still falls in our study period. The GEOS-CF data was coarsened to match the GEOS-Chem resolution (2.5° x 2.0°) to provide a more direct comparison. The SVRs from GEOS-CF at both native and coarsened resolution for the full year are shown in Fig. R7b. To maintain consistency with GEOS-Chem, we used the coarsened GEOS-CF data to define the uncertainty in the quasi-seasonal assumption. In general, we found that the average intraseasonal variability for MAM from GEOS-CF (7% and 10% for March and May, respectively) matched well with GEOS-Chem (5% and 9% for March and May, respectively). From the GEOS-CF data, the average discrepancy for all the full year was around 12%, or 0.8 ppbv DU^{-1}, as shown in Table S4. The uncertainty derived from GEOS-CF was used as the uncertainty of the quasi-seasonal assumption rather than the single season from GEOS-Chem since they were similar for MAM, but GEOS-CF covers all four seasons. Fig. R7 also indicates that SVRs tend to remain consistent over the course of the year with slightly lower values during the summer months and slightly higher values during the winter.

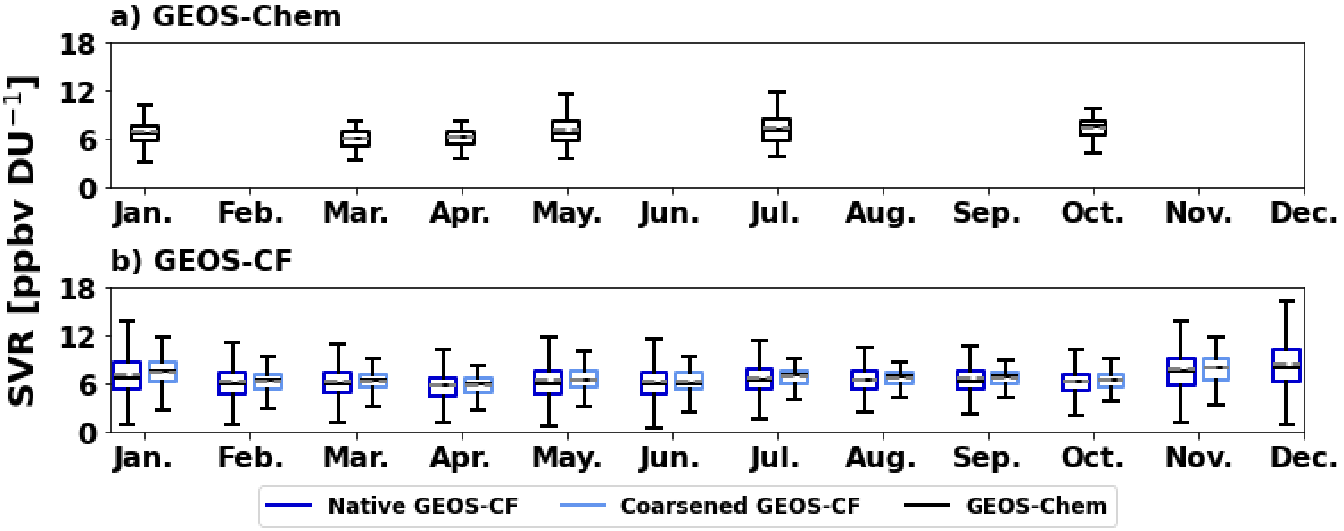

Figure R7: Boxplots showing monthly average a) 2018 GEOS-CF SVRs at native (0.25° x 0.25°; dark blue) and coarsened (2.5° x 2.0°; light blue) resolution, and b) 2015 GEOS-Chem SVRs at 2.5° x 2.0° resolution (black). GEOS-Chem simulations were available for January, March, April, May, July, and October. Months without a boxplot were not simulated. This figure was added as Fig. S3 in the supplementary material.

Table R4: Relative and absolute discrepancies between the SVRs from the month in the middle of each season (January, April, July, and October for winter, spring, summer, and autumn, respectively.) compared to the other months in that season to address the quasi-seasonal assumption. GEOS-Chem data is only available for spring 2015 and GEOS-CF data is available for all seasons of 2018.

	Relative SVR Discrepancy [%]		Absolute SVR Discrepancy [ppbv DU^{-1}]	
Comparison	GEOS-Chem	GEOS-CF	GEOS-Chem	GEOS-CF
Dec. – Jan.	-	8	-	1.0
Feb. – Jan.	-	17	-	1.0
Mar. – Apr.	5	7	0.3	0.4
May – Apr.	9	10	0.9	0.7
Jun. – Jul.	-	15	-	0.8
Aug. – Jul.	-	4	-	0.3
Sep. – Oct.	-	8	-	0.4
Nov. – Oct.	-	25	-	1.6

Next, we addressed the assumption of only using a single year of GEOS-Chem simulations to calculate surface SO_2 concentrations for the entire four year period. The justification that we initially included in the text comparing the OMI and GEOS-Chem VCDs (Lines 143-151) did not directly account for the impacts on the SVRs since it neglected to compare the surface concentrations between the CNEMC in situ measurements and GEOS-Chem. Here, we aim to provide a much clearer quantification of temporal representativeness by looking at SVRs directly rather than VCDs alone. First, we compared the SVRs from the 2015 and 2018 GEOS-Chem simulations, which can be shown in Fig. R8. For each simulation, the difference in the SVRs between the 2015 and 2018 simulations is only around 9%, or 0.6 ppbv DU^{-1}. This value was used as the uncertainty for the single-year assumption in the error propagation. We also wanted to compare the GEOS-Chem SVRs to those from observations to see if the difference between them changes over time. The observed SVRs were calculated using surface concentrations from the CNEMC in situ measurements gridded to 0.25° x 0.25° resolution and VCDs from OMI. The monthly average SVRs from observations and GEOS-Chem are shown in Fig. R3 (see RC1 General Comment 3). In short, the difference between the observed and GEOS-Chem SVRs were large (around a factor of 5), but consistent from year-to-year and season-to-season. We believe that this way of showing consistency in the SVRs over time is more representative than just looking at the VCDs since they account for SO_2 both at the surface and in the atmospheric column, which have equal impacts on the SVR. The consistent underestimation in the SVRs compared to observations may also address the consistent underestimation of the CTM-based method seen in Fig. 3 from the text.

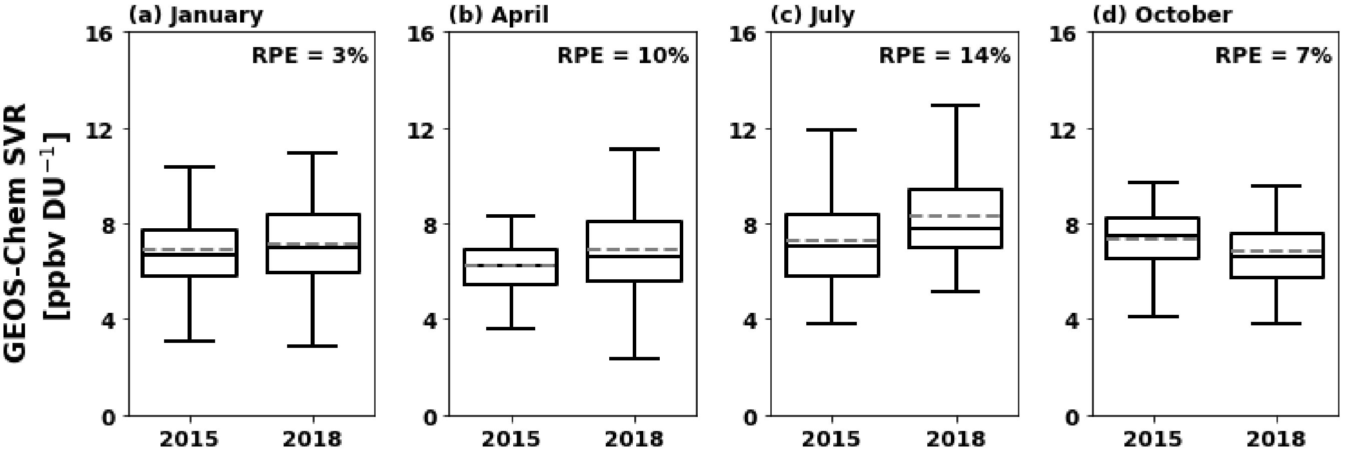

Figure R8: Boxplots showing the monthly averaged SVR from the 2015 and 2018 GEOS-Chem simulations for a) January, b) April, c) July, and d) October. Each panel contains the relative percent error between the two simulations. This figure was added as Fig. S7 in the supplementary material.

In addition to quantifying our assumptions regarding the temporal representativeness, we also investigated the impact of the temporal sampling of the CTM simulations on the accuracy of the predicted surface concentrations. To keep consistency between GEOS-Chem and GEOS-CF, we are limiting this investigation to only include OMI VCDs, GEOS-Chem simulations, and GEOS-CF simulations from 2018. The different temporal sampling methods we used include “real time” which means using the daily SVR from the model is used to calculate the daily surface concentrations, “monthly,” which means using the monthly average SVR to calculate the daily surface concentrations within that month, “quasi-seasonal,” which means using the monthly average SVR from January, April, July, and October to calculate the daily surface concentrations within their respective seasons (winter, spring, summer, and autumn, respectively), “seasonal,” which uses the seasonal mean SVR to calculate the daily surface concentrations within that seasons, and “annual,” which uses the annual average SVR to calculate the daily surface concentrations within that year. The daily surface concentrations from the CTM-based method for each of the GEOS-CF temporal sampling conditions are shown in Fig. R9b-f, with the annual mean concentrations in Fig. R10b-f. Surprisingly, Fig. R9b-f and Fig. R10b-f show that there is not a significant difference in the accuracy of the estimated surface concentrations with different temporal sampling. In Fig. R9b-f, the daily surface concentrations have small ranges in the slopes and correlations at 0.18 - 0.20 and 0.20 - 0.23, respectively with very little change in the MAE, RMSE, and RPE. In Fig. R10b-f, we see a similar trend with slopes only ranging from 0.23 - 0.29, correlations ranging from 0.33 - 0.40. Even though the daily profiles from the CTMs should be more temporally representative of the SO_2 loading, they do not appear to provide better results using the CTM-based method. This may be due to an inability to accurately capture short term changes in SO_2 profiles, and thus, the SVRs. This suggests that our use of the quasi-seasonal temporal sampling should not significantly affect the accuracy of our results.

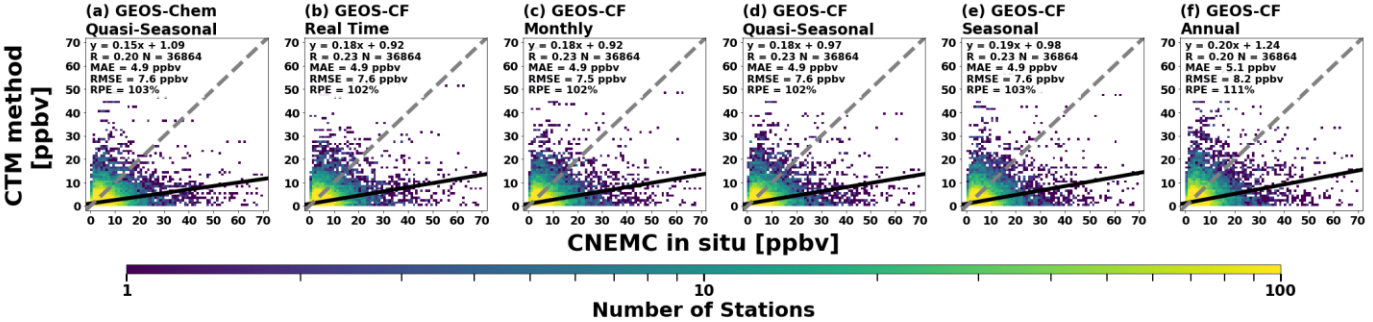

Figure R9: Scatterplots of the 2018 daily surface SO_2 concentrations from the CTM-based method against the CNEMC in situ measurements for different models including a) GEOS-Chem (2.5° x 2.0° horizontal resolution) and b-f) GEOS-CF (0.25° x 0.25° horizontal resolution) for different temporal sampling of the CTM SVRs when combined with daily OMI data to calculate daily surface SO_2 concentrations. "Real time" sampling indicates that daily SVRs were used to calculate daily surface concentrations. "Monthly" sampling indicates that monthly averaged SVRs were used to calculate daily surface concentrations within that given month. "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. "Seasonal" sampling indicates that SVRs averaged over DJF, MAM, JJA, and SON were used to calculate daily surface concentrations within that given season. "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.

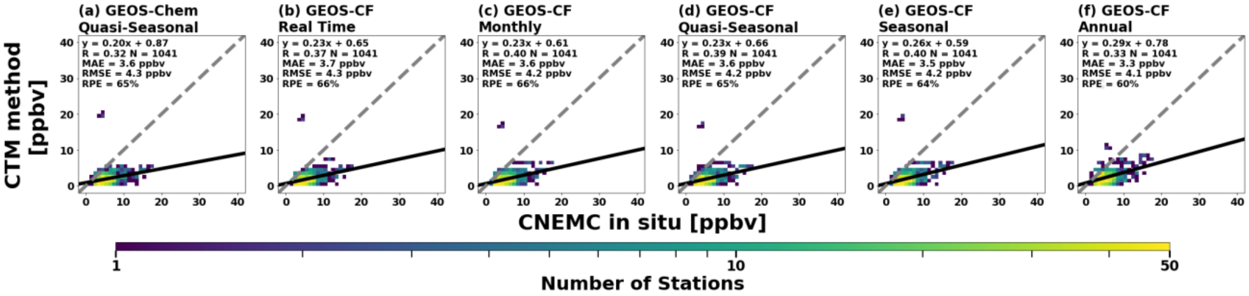

Figure R10: Same as in Fig. R9 but for 2018 annual average surface SO_2 concentrations. This figure was added as Fig. S4 in the supplementary material.

In the text, we provided a full description of these sensitivity tests and quantification of uncertainty regarding the temporal representativeness of the simulations:

"Since only simulations for January, April, July, and October 2015 were available to provide SVRs, there are two inherent assumptions regarding the temporal representativeness of the SVRs. The first assumption was using quasi-seasonal temporal sampling for the SVRs and calculating the estimated surface concentrations. To test the impact of temporal representativeness on the estimated surface concentrations, we ran an additional GEOS-Chem simulation to cover all of spring (MAM) 2015. We also employed a full year of archived 2018 GEOS-CF data (NASA GMAO, 2023), which has improved temporal (hourly) and spatial (0.25° x 0.25°) resolution compared to GEOS-Chem and uses the same chemistry module, so they tend to produce similar results (Keller et al., 2021). We found that the intraseasonal variability in the SVR was 0.6 ppbv DU^{-1} for MAM in both GEOS-Chem and GEOS-CF, as shown by Fig. S3. Therefore, we used the GEOS-CF data to estimate this uncertainty for the entire year. We found that the average intraseasonal variability in the SVRs for the full year was 0.8 ppbv DU^{-1} (Fig. S3). We also used the full year of GEOS-CF data to test the impact of temporal representativeness on the annual mean surface SO_2 concentrations. Figures S4b-f show that there was no significant difference in the accuracy of the annual average surface SO_2 concentrations among the different temporal sampling techniques ranging from daily to annual mean SVRs. The slopes and correlations of the surface concentrations were consistent only ranging between 0.23 – 0.29 and 0.33 – 0.40, respectively (Fig. S4b-f)." (Lines 165-178).

"The other assumption was only using a single year of simulations to convert four years of OMI data into surface concentrations. Kharol et al. (2017) did not have simulations that spanned their entire analysis period, but the implications of this were never discussed. To address this, we first compared the monthly averaged SVRs from observations (calculated using CNEMC surface concentrations and OMI VCDs) for each year in the study period to the 2015 GEOS-Chem simulations to ensure there is no significant changes over time. Figure S6 shows boxplots of the observed and GEOS-Chem SVRs with the percent difference between them. In general, the differences between the observed and GEOS-Chem SVRs were consistent across all years of the study period, typically ranging from 73 – 89% (Fig. S6). We also ran additional GEOS-Chem simulations for January, April, July, and October 2018 to assess if the simulated SVRs change over time. Boxplots for these two sets of simulations can be seen in Fig. S7 and indicate that the GEOS-Chem SVRs only changed by 0.8 ppbv DU^{-1}, or 9%, from 2015 to 2018. The implications of these uncertainties on the resultant concentrations is discussed further in Sect. 2.6." (Lines 182-191).

3. Resolution mismatch between CTM and OMI satellite observations. It is recommended to conduct nested simulations over eastern China to have a finer resolution to be consistent with finer OMI observations.

Response:

We agree that there is a large mismatch in resolution between the GEOS-Chem simulations and the OMI VCDs and that it may have an impact on the results. One of the reasons we did not do a nested simulation is that were already being impacted by the computational expense of the simulations at the coarser resolution. Additionally, this method has been proven to be reliable with coarse-resolution model simulations, as shown by Lee et al. (2011) and Zhang et al. (2021). The method that we used from Lee et al. (2011) utilized OMI VCDs to downscale the model simulations using the subgrid variability, as seen in Eq. 1 and Eq. 2 of the main text. These scaling factors were applied to the simulated surface concentrations and boundary layer component of the VCD; however, since the free tropospheric VCD is negligible compared to the boundary layer VCD in GEOS-Chem, as suggested by the profile shapes in Figure R1, the subgrid variability terms in the numerator and denominator of Eq. 1 largely cancel each other out. See RC2 Specific Comment 4 for more information regarding the impact of the subgrid variability terms on the resultant surface concentrations. In short, since the subgrid variability terms have a small effect on the estimated surface concentrations, downscaling may not be an effective technique for resolving the difference in resolution, and this issue does need to be addressed using higher resolution model simulations. As previously stated, we do not have the computational resources to run a nested simulation in a reasonable amount of time, so to test the impacts of the model resolution on the results, we used the archived GEOS-CF data described in RC2 General Comment 2.

First, we compared the monthly averaged profile shape for January, April, July, and October 2018 between the GEOS-Chem and GEOS-CF simulations. We used the same locations as Figure R1 to represent a diverse set of locations across the study region. Fig. R11 shows the vertical SO_2 profiles from GEOS-Chem and GEOS-CF at their native resolutions. The figure shows some differences in the simulated profile with the higher resolution GEOS-CF having higher concentrations in the boundary layer, as well as the free troposphere with occasional elevated plumes. Despite the differences in the profiles, there was no significant change in the SVRs, suggesting that the resultant surface concentrations from the CTM-based method would be similar regardless of using the coarse resolution GEOS-Chem simulations or higher resolution GEOS-CF simulations.

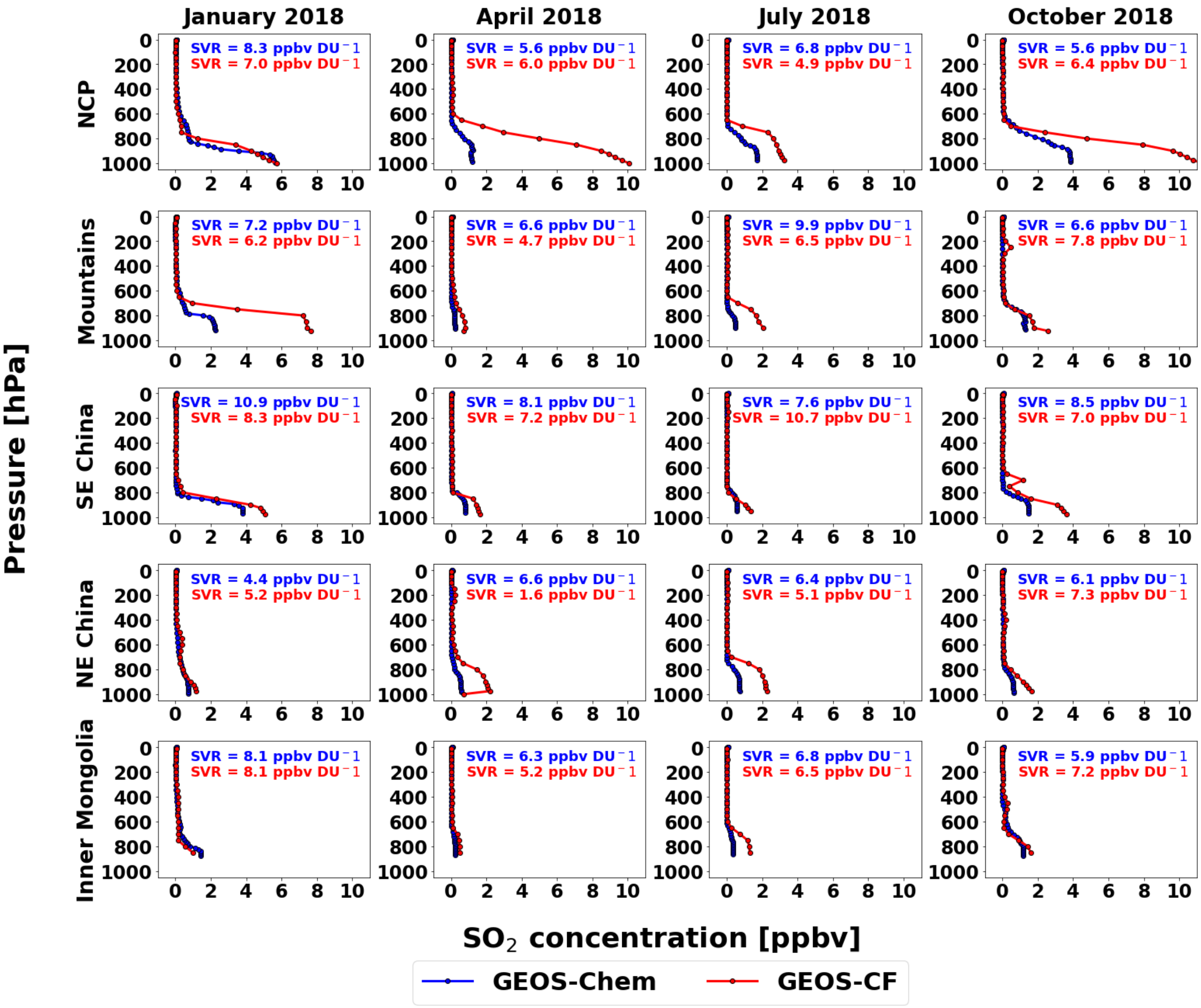

Figure R11: Monthly averaged vertical SO_2 profiles (blue lines) from each of the 2018 GEOS-Chem (blue) and GEOS-CF (red) simulations and their respective SVRs at five locations in different parts of the study region including, from top to bottom, the North China Plain (NCP; 115 °E, 38 °N), the Qin Mountains (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). Each column represents a different month (from left to right: January, April, July, and October).

To test the impacts of spatial resolution of the CTM simulations on the resultant surface concentrations from the CTM-based method, we used both the 2018 GEOS-Chem and GEOS-CF simulations at their native resolutions. Figures R9a and R9d, as well as Figs. R10a and R10d show that there is no significant improvement in the accuracy of the surface concentrations estimated by the CTM-based method for the 2018 data for the same temporal sampling (quasi-seasonal). To investigate this further, we applied the quasi-seasonal temporal sampling from the 2018 GEOS-Chem and GEOS-CF to estimate the surface concentrations for all four years of the study period, similar to what was done in the main text. Scatterplots showing the daily and annual average surface concentrations from the CTM-based method from both GEOS-Chem and GEOS-CF are shown in Figs. R12 and R13, respectively. These two figures further suggest that for each year of the study period, there is no significant improvement in the accuracy of the estimated surface concentrations using a higher resolution CTM simulation compared to a coarse resolution simulation, as indicated by the small changes in slope, correlation, MAE, RMSE, and RPE.

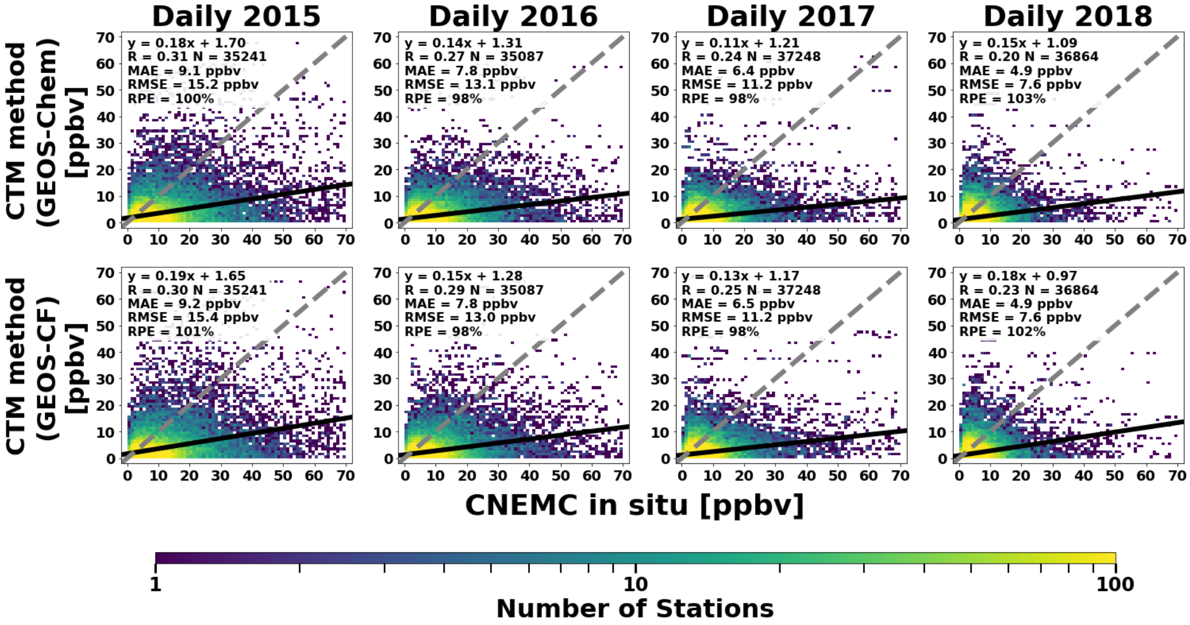

Figure R12: Scatterplots of the daily surface SO_2 concentrations from the CTM-based method against the CNEMC in situ measurements for 2018 GEOS-Chem (top row) and GEOS-CF model simulations. Both models had the same temporal sampling (quasi-seasonal) but different spatial resolutions (2.5° x 2.0° and 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.

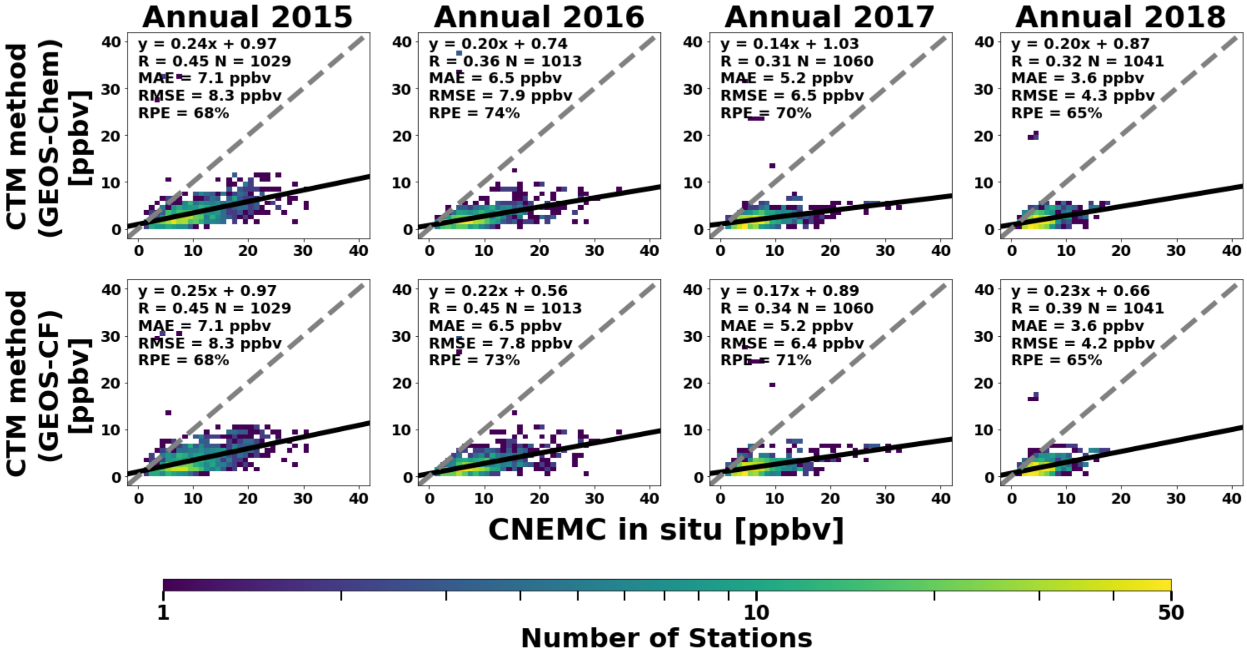

Figure R13: Same as in Fig. R12 but for annual mean surface SO_2 concentrations. This figure was added as Fig. S5 in the supplementary material.

In the text, we addressed our sensitivity tests regarding the spatial resolution in Section 2.4 where the CTM-based method is described:

"The approach from Lee et al. (2011) was used to infer surface SO_2 concentrations from OMI VCDs and simulated vertical SO_2 profiles from GEOS-Chem (GC). Lee et al. (2011) showed that the CTM-based method provided accurate results even with CTM resolutions that are much coarser than the satellite data." (Lines 143-145).

"The sensitivity analysis with GEOS-CF also suggested that improving the spatial resolution of the CTM while maintaining the same temporal sampling of the SVRs did not have a large impact on the accuracy of the estimated surface concentrations despite an improvement in spatial resolution by nearly a factor a 10, as indicated by Figs. S4a and S4d for 2018 data, as well as Fig. S5 for all years of the study period." (Lines 178-181).

4. Why XGBoost ML method specifically? How about other ML techniques?

Response:

We chose to use XGBoost for our machine learning architecture for several reasons. When we were initially developed in the ML component of our work, we tried using an artificial neural network (ANN). The structure of the ANN had an input layer, normalization layer, three dense layers with 16, 8, and 4 neurons each, and an output layer. The model was also compiled with the Adam optimizer, learning rate of 0.0005, and a mean squared error loss function. The ANN was also trained on more variables than our XGBoost model. From the OMI SO2 product, we included SO_2 VCDs, ozone VCDs, 342 nm reflectance, solar zenith angle, solar azimuth angle, viewing zenith angle, viewing azimuth angle, elevation, and aerosol index to account for the detection of SO_2, as well as parameters that may affect the sensitivity of the retrieval to SO_2 located near the surface. We also used meteorological variables from ERA5 including 2 m temperature, 2 m dew point, 100 m u-wind, 100 m v-wind, and boundary layer height in the model training to account for parameters that may affect the lifetime of atmospheric SO_2. We also included SO_2 emissions from the CEDS inventory in the model training to account for source locations. We found that even with the inclusion of these 11 variables, the data was underfitting to the ANN and we could not get good performance from neither the training nor independent testing datasets, as shown by Fig. R14. For a related discussion regarding the impact of the number of predictors on the model training and performance, see RC2 Specific Comment 11.

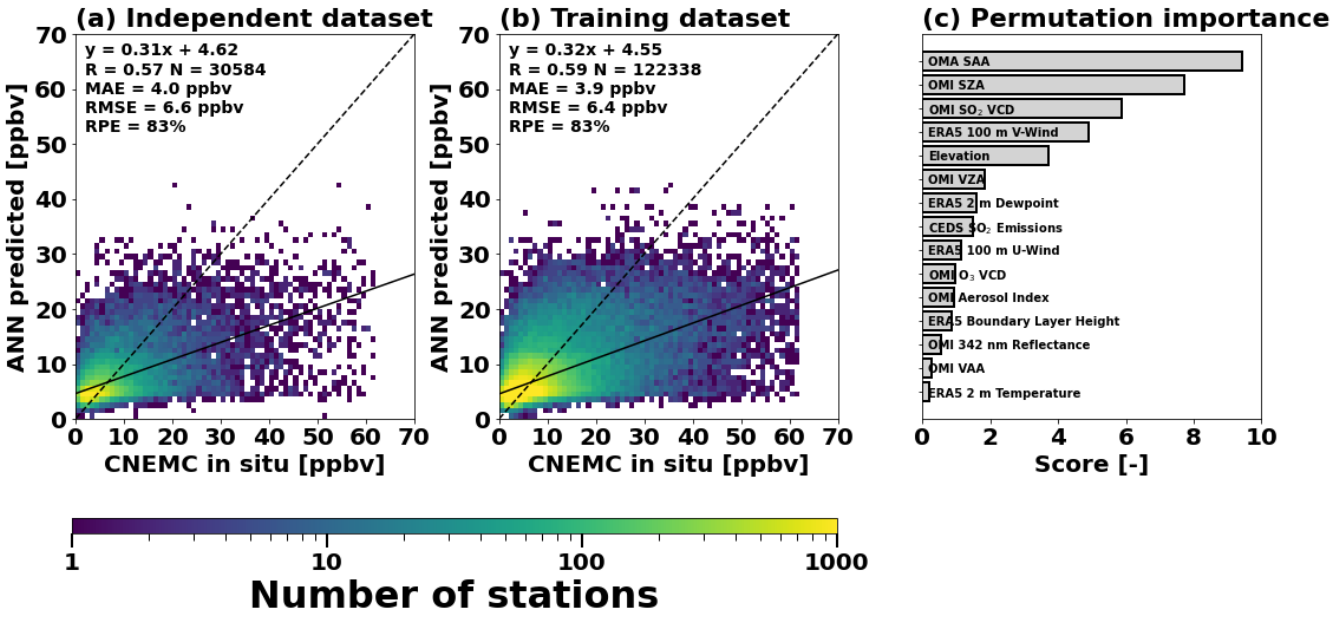

Figure R14: Scatterplots between the daily ANN model predictions and CNEMC in situ measurements for the (a) independent dataset and (b) training dataset, as well as a (c) permutation importance analysis. Each scatterplot includes a linear regression analysis with best fit line (solid line) and discrepancy statistics for the estimated surface SO_2 concentrations compared to in-situ measurements. The scatterplots are binned every 1 ppbv. The dashed line indicates the 1:1 line.

The reason that we switched to an XGBoost model architecture in particular was based on the results from previous studies that performed tests showing which type of model architecture performs better for this application. These studies showed that gradient boosting methods like LightGBM and XGBoost tended to perform better estimating surface concentrations from satellite retrievals and other variables compared to other machine learning architectures. Kang et al. (2021) compared the performance of different model types for estimating surface nitrogen dioxide and ozone concentrations from TROPOMI satellite data and other variables over China, which determined that the XGBoost and LightGBM models had similar performance to one another (r = 0.84, 0.88), as well as better performance than multiple linear regression (r = 0.73, 0.73), support vector machines (r = 0.76, 0.79), and random forest (r = 0.81, 0.86) methods. Additionally, Zhang et al. (2022) used a LightGBM model for estimating surface SO_2 concentrations from OMI data over northern China and was able to achieve accurate results (r = 0.94).

To explain to the readers why we chose the XGBoost machine learning architecture in particular, we provided the supporting information from the previous studies into the text:

"Previous studies have shown that XGBoost and LightGBM models are able to estimate surface concentrations from satellite data more effectively than other ML architectures as shown by Kang et al. (2021) and Zhang et al. (2022)." (Lines 197-199).

Specific Comments

1. Line 23-24: what is the supporting evidence from this study for the statement that CTM-based approach is better than the ML-based approach over areas without monitoring sites.

Response:

Due to the lack of measurements, it is not possible to quantitatively and statistically compare the CTM- and ML-based approaches in these locations. However, we performed a qualitative analysis and saw that the ML-model was predicting high concentrations over the oceans, likely due to the low marine boundary layer as suggested by Fig. R4. Since the model was trained over land, the model likely learned that the low continental boundary layer was associated with high SO_2 concentrations during the wintertime. As a result, when the model is used to predict surface SO_2 concentrations over the ocean, it is likely interpreting the low marine boundary layers as areas of high SO_2 concentrations, especially because boundary layer heights are the most important variable in the model, as seen in Figure 5. This was described in the text in Section 4.2, but we clarified this in the abstract:

"Despite the higher accuracy of the ML-based method at the monitoring sites, the CTM-based method produced more reasonable gridded spatial distributions over areas without monitoring data, such as over the oceans, since its estimations are independent from the in situ measurements." (Lines 23-25).

2. Line 121-122: is one-month spin-up enough for CTM simulations? The ground-based observations are for annual means spanning 4 years. It is recommended to conduct full-year simulations over each year. Also, only one month in each season may not be representative to support seasonality analyses in Figures 6 and 7.

Response:

We decided to do a one-month spin-up following the procedure outlined in Kharol et al. (2015) which used the CTM-based method to estimate surface nitrogen dioxide concentrations from OMI retrievals. Additionally, after running the additional GEOS-Chem simulation for March, April, and May 2015 (see RC2 General Comment 2 for more information), the difference in the SO_2 concentrations between 1 month of spin-up from the original set of simulations and 2 months of spin-up from the additional simulation was only 0.01 ppbv. This clarification has been added to the text:

"We ran simulations for January, April, July, and October 2015. Each simulation was conducted with a one-month spin-up following Kharol et al. (2015)." (Lines 129-131).

The surface concentrations obtained from both the CTM- and ML-based methods were calculated on a daily timescale from the daily OMI VCDs and the monthly average SVRs from the four GEOS-Chem simulations for days within their respective seasons. This means that the surface concentrations from both the CTM- and ML-based methods were estimated once per day for the entire study period. This means that they are directly comparable to the daily CNEMC measurements. The seasonality analysis in Figs. 6-7 were derived from the daily surface concentrations from each dataset including the CTM-based method, ML-based method, and CNEMC measurements. We clarified this information in the methodology.

"To reduce the computational expense, we used the monthly average SVR from each simulation to estimate the daily surface concentrations within the corresponding winter (DJF), spring (MAM), summer (JJA), and autumn (SON) months (referred hereafter as quasi-seasonal temporal sampling) for all years of the study period." (Lines 131-133).

"The surface SO_2 concentrations for the CTM-based method (S_{OMI}) were calculated on a daily basis at 0.25° x 0.25° resolution using the daily OMI VCDs and averaged GEOS-Chem SVRs from the model grid cell that the OMI measurement lies within using Eqn. 1:…" (Lines 150-152).

For more information regarding the temporal representativeness of the CTM simulations on the estimated surface concentrations, see RC2 General Comment 2. In short, we found that the temporal sampling of the SVR from the CTM does not have a significant impact on the results.

3. Line 129: it is suggested from this paper that OMI overpassing time over eastern China is 12:15 pm to 2:45 pm, while here it is suggested simulation outputs are only at 2 pm local time. It is recommended to sample simulation output from 12 pm – 3 pm to be consistent with OMI overpassing time.

Response:

We agree that having hourly model output over the OMI overpass window would be more temporally representative. We limited the GEOS-Chem output to be every 3 hours to speed up our simulations since we were already struggling with computational expense. To address the typical diurnal variation of the simulated SVRs during the OMI overpass window, we used archived GEOS-CF data as described in RC2 General Comment 2. Figure R15 shows how the hourly SVR changes as a function of time. The figure shown that the SVR at 2:00 pm local time is only around 3% lower than the SVR averaged across the overpass window (dashed line). This small discrepancy should not significantly affect the estimated surface concentrations and would not cause the extremely large underestimation observed in Fig. 3.

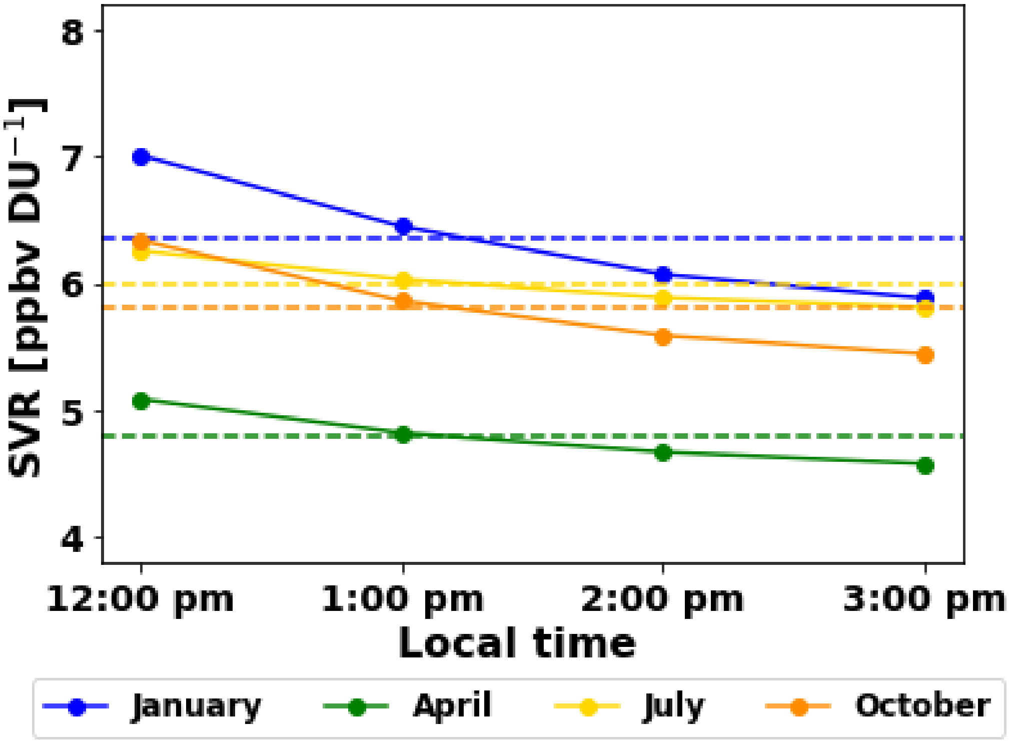

Figure R15: Time series of monthly averaged GEOS-CF SVRs for each hour in the OMI overpass period. Dotted lines represent the average SVR for the OMI overpass window for that month.

4. Line 139: what are the uncertainties for using observations from OMI at 0.25 degree resolution to represent subgrid variability for simulations at 2 degree? At least a test using nested simulation at 0.25 degree over eastern China should be added.

Response:

We investigated the impact of the subgrid variability term on the resultant surface concentrations from the CTM-based method to see if this method of accounting for the difference in resolution is effective. We found that the subgrid variability terms in Eqn. 1 effectively cancel each other out, so the uncertainties associated with them would cancel out as well. The reason for this is because nearly all of the SO_2 in the vertical column comes from the boundary layer component while the free-tropospheric component is negligible, as suggested by the simulated SO_2 profiles in Fig. R1. Table R5 shows the percentage of the total VCD that is attributed to the boundary layer and free-tropospheric components from the profiles in Fig. R1. The boundary layer component of the VCD accounts for at least 97.5% of the total VCD while the free-tropospheric component only accounts for up to 2.5%. To determine the impact of the subgrid variability term on the surface concentrations calculated using the CTM method, sensitivity tests were performed by calculating the surface concentrations with and without the subgrid variability terms. In general, it was found that the subgrid variability terms changed the surface concentrations by less than 3%, as shown in Table R5. Therefore, the impacts of subgrid variability, as well as its uncertainties, should largely cancel out since they have similar impacts on the numerator and denominator of Eqn. 1. As previously stated, we do not have the computational resources to run a nested domain simulation; however, we did find that the accuracy in the CTM-based method was not significantly affected by using archived GEOS-CF simulations with 0.25° x 0.25° resolution compared to our GEOS-Chem simulations with 2.5° x 2.0° resolution. See RC2 General Comment 3 for more information regarding the impacts of CTM resolution on the estimated surface concentrations.

Table R5: Relative components of the GEOS-Chem SO_2 VCDs split between the boundary layer and free troposphere as defined by GEOS-FP boundary layer height, and the relative impact of including the subgrid variability term when calculating the surface SO_2 concentrations using the CTM-based method.

Monthly Averaged GEOS-Chem Simulation	PBL VCD Fraction [%]	FT VCD Fraction [%]	Impact of Subgrid Variability on CTM-Method Surface Concentrations [%]
January 2015	98.8	1.2	0.2
April 2015	97.5	2.5	2.8
July 2015	98.2	1.8	0.6
October 2015	98.5	1.5	1.7

5. Line 145-146: the statement is not consistent with Figure S1. From Figure S1, it shows large variability of the correlation coefficients over different years. For example, for results in January, the correlation coefficient is 0.76 in the year of 2015, compared to 0.33 in the year of 2018. Actually, Figure S1 supports that there is large variability over years, and thus simulations over each year should be conducted instead of using simulations over only 4 months in 2015.

Response:

This is a good point. In fact, we realized that using only the VCDs to assess variability over time is not representative of the CTM-based method since it also highly depends on surface concentrations. Instead, we decided to redo this analysis by comparing the observed SVRs (calculated using CNEMC surface concentrations and OMI VCDs) to the VCDs from the GEOS-Chem simulations. This analysis was completed in response to RC2 General Comment 2. In short, we found that by incorporating the surface concentrations, the percent difference between the observed and 2015 GEOS-Chem SVRs was much more stable over the four year study period compared to the VCDs alone (Fig. R3). For more information, see RC2 General Comment 2.

6. Line 148-149: I view the scatter plot shows different concentration ranges in the year of 2015 and 2018 and thus there could be large spatial variability of sulfur dioxide concentrations. The more comparable slopes only indicate general regional convergence.

Response:

We agree that we misinterpreted what the slope and correlation mean in the scatterplot. Instead, we compared boxplots of the simulated SVRs from 2015 and 2018 instead of plotting the simulated surface concentrations and VCDs against each other (Fig. R8). This analysis was completed in response to RC2 General Comment 2. In short, we found that the SVRs from 2015 and

2018 had similar ranges and an average percent difference of only 9%, which provides a much clearer comparison between the SVRs of the two sets of simulations than the scatterplots from the original text (Fig. R3). For more information, see RC2 General Comment 2.

7. Line 159-160: what are the supporting statistics for using a depth of 15 splits?

Response:

Figure R16 shows the performance of the model by using the slope and correlations of a linear regression analysis between the observed and predicted surface concentrations for both the training and independent testing datasets as a function of maximum tree depth. We also performed this analysis for the number of trees in the ensemble, as shown in Fig. R17. In both plots, a depth of 15 splits and ensemble of 500 trees occur in regions where there is no further improvement in the performance of the model, indicating that these values should produce effective results (Figs. R16-R17).

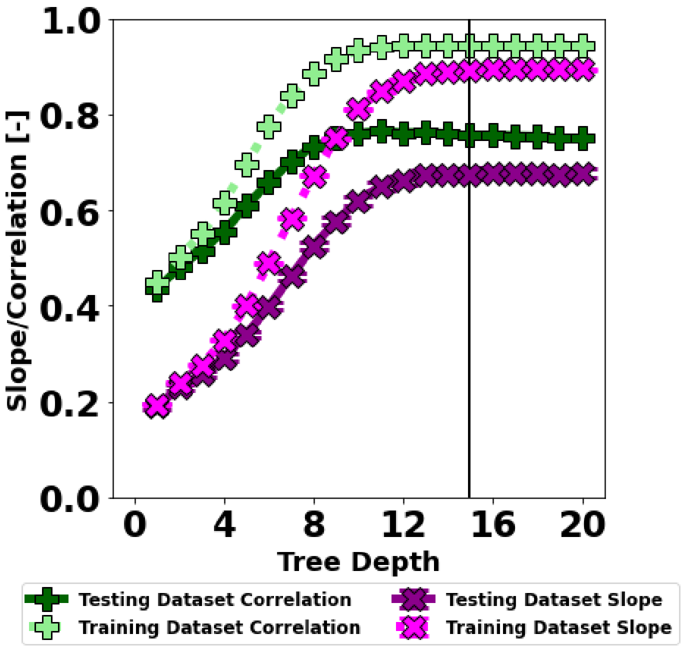

Figure R16: Performance of the XGBoost model measured by the slope and correlation of a linear regression analysis between the daily ML-predicted and CNEMC in situ surface SO_2 concentrations as a function of tree depth. Increasing the tree depth significantly improved the performance of the model until stabilizing around a value of 15 (black line), so this value was used for the model. This figure was added as Fig. S8 in the supplementary material.

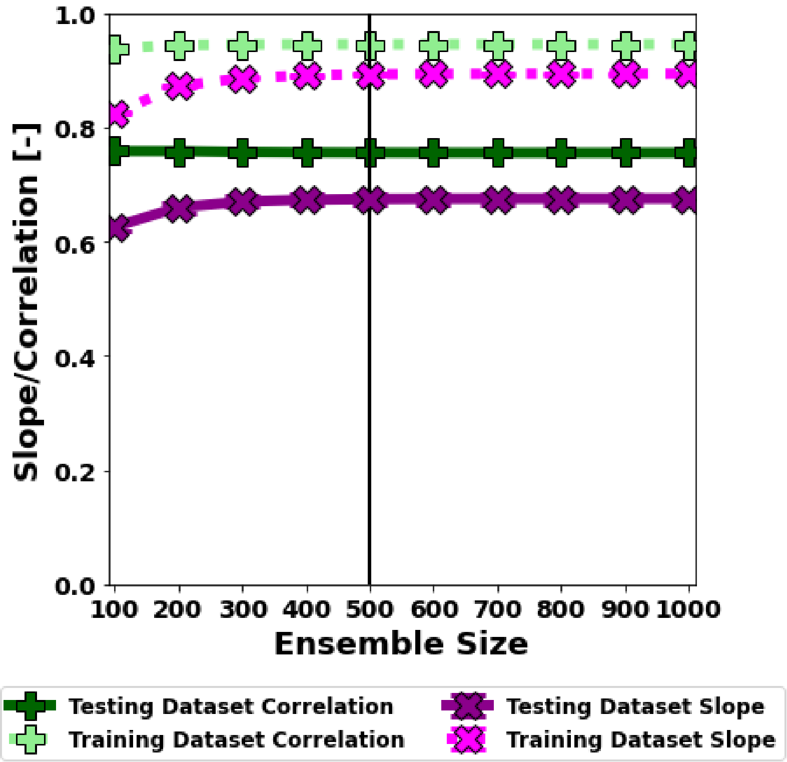

Figure R17: Performance of the XGBoost model measured by the slope and correlation of a linear regression analysis between the daily ML-predicted and CNEMC in situ surface SO_2 concentrations as a function of ensemble size. Increasing the ensemble size slightly improved the performance of the model until stabilizing around a value of 500 (black line), so this value was used for the model. This figure was added as Fig. S9 in the supplementary material.

Figures R16 and R17 were added to the supplemental information as Figs. S8 and S9, respectively and briefly described in the main text:

"We trained our XGBoost model with an ensemble of 500 trees, a maximum tree depth of 15 splits, and a learning rate of 0.15 on a mean squared error loss function. Neither a larger ensemble nor deeper trees improved the performance of the model, as shown by Fig. S8 and Fig. S9, respectively." (Lines 199-201).

8. Line 164-167: why not keep the inputs the same between the CTM-based and ML-based approaches? In other words, why not choosing GEOS-FP meteorological fields as used in the CTM, as the inputs for training ML model? How would the difference between meteorological fields contribute to the differences of CTM-based and ML-based estimates?

Response:

The main reason we wanted to use ERA5 rather than GEOS-FP for the ML work was due to its better spatial and temporal resolution. ERA5 has all meteorological variables that we used in our ML models available at 0.25° x 0.25° spatial resolution and instantaneous values at hourly temporal resolution. The GEOS-FP version we used to drive the GEOS-Chem simulations is at a much coarser 2.5° x 2.0° horizontal resolution with hourly averaged boundary layer heights and 3-hourly instantaneous wind speeds. GEOS-FP does offer a higher resolution product for running nested domain simulations at 0.25° x 0.3125° horizontal resolution with the same temporal resolution as the coarser version. To address this comment, we compared the performance of the XGBoost model with ERA5 meteorology from the main text to XGBoost models with identical architectures, but use GEOS-FP meteorology instead of ERA5. We trained two new XGBoost models using different GEOS-FP products: one at the coarse resolution and one at the nested resolution. For simplicity, we regridded the high-resolution GEOS-FP meteorology using interpolation to match the ERA5 resolution (0.25° x 0.25°).

Figures R18 and R19 show the performance of the ML model using the coarse-resolution (2.5° x 2.0°) and nested resolution (0.25° x 0.25°) GEOS-FP meteorology as inputs in place of ERA5. Despite using a different dataset, the permutation importance did not change between Fig. 5 and Figs. R18-R19 with the boundary layer height being the most important followed by OMI SO_2 VCDs, CEDS SO_2 emissions, and wind speeds. The performance of the machine learning models were also similar between the ERA5 meteorology (slope = 0.67; r = 0.76; Fig. 2), the coarse-resolution GEOS-FP meteorology (slope = 0.69; r = 0.77; Fig. R18), and the high-resolution GEOS-FP meteorology (slope = 0.67; r = 0.76; Fig. R19). It is an interesting result that the XGBoost model using coarse resolution GEOS-FP meteorology slightly outperforms the models using both ERA5 and nested GEOS-FP meteorology datasets, although this difference is quite small. It appears that the meteorology dataset used does not have a significant impact on the model results. For this reason, we will maintain our use of ERA5 in our work due to it having higher spatial resolution and has better temporal representativeness of the OMI overpass window.

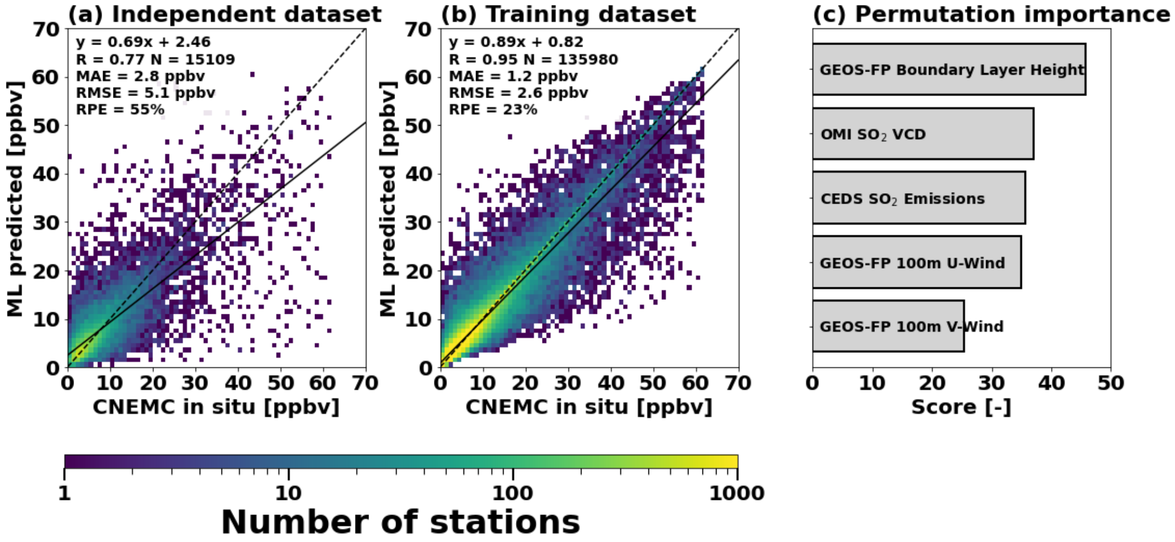

Figure R18: Scatterplots between the daily ML model predictions and CNEMC in situ measurements for the (a) independent dataset and (b) training dataset using coarse resolution (2.5° x 2.0°) GEOS-FP meteorology for model training and predictions. Each panel includes a linear regression analysis with best fit line (solid line) and discrepancy statistics for the estimated surface SO_2 concentrations compared to in-situ measurements. The scatterplots are binned every 1 ppbv. The dashed line indicates the 1:1 line.

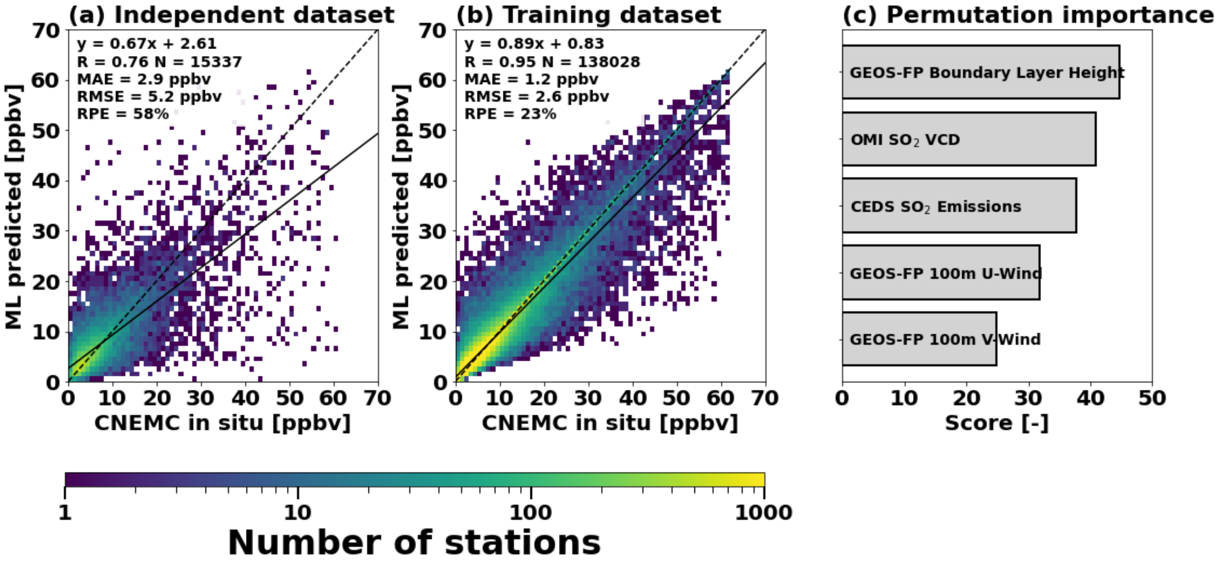

Figure R19: Same as in Fig. R18 but using high resolution (0.25° x 0.25°) GEOS-FP meteorology for model training and predictions.

9. Line 203: the simulation is not for the full year and thus it is not annual mean actually. It is recommended to conduct full year simulations over each year for temporal representativeness and to be consistent with the ground-based observations.

Response:

The annual and seasonal mean surface concentrations were calculated from daily surface concentrations from the CTM-based method, ML-based method, and CNEMC in situ measurements that were then averaged over time. The SVRs obtained from the GEOS-Chem simulations were the only values that were not fully representative of the full year or season, but they were used to calculate daily surface concentrations from the OMI VCDs. Additionally, we conducted tests with archived GEOS-CF data with higher spatial and temporal resolution that showed that the impact of temporal representativeness on the SVRs was relatively small. More details can be found in RC2 General Comment 2.

10. Figure 3: are the comparisons against all ground-based sites or over the same independent testing sites used in ML-based approach? Also, how are the simulation-observation pairs sampled? As the simulation is at very coarse resolution of 2 degree, while monitoring sites are single points. Are ground-based observations averaged over each grid box?

Response:

The results from Section 3 (CTM-based method) use all of the available ground based sites. The results in Sections 4 (ML-based method) and 5 (direct comparisons between the methods) each only use the sites from the independent testing dataset.

We did not directly compare the simulation output to the observations. We used the SVRs from the simulations to convert the OMI VCDs within each GEOS-Chem grid cell into the surface concentrations at the OMI resolution. We attempted to account for subgrid variability from GEOS-Chem using OMI data to reduce the impacts of coarseness on the estimated surface SO_2 concentrations from the CTM-based method, but this sub-grid variability term was likely canceled out as described in RC2 Specific Comment 4.

To compare the gridded surface concentrations (at 0.25° x 0.25° resolution) to the surface stations, we averaged the satellite-derived surface concentrations in a 40 km radius around each CNEMC site, a similar technique as Kharol et al. (2017). A 40 km radius is equivalent to averaging a 3 x 3 grid of OMI data points with the CNEMC site located in the center pixel. We used a radius of 40 km since it maximizes the slope and correlation compared to other radii, as shown by Fig. R6. Since we used the averaging radius technique from previous studies, we did not need to average the CNEMC sites within each grid box. This was explained in Section 2.4 from the text, but was rephrased for clarity:

"To compare the estimated surface concentrations to the in situ surface monitoring data, we used a 40 km averaging radius around each station to increase the amount of usable data and further reduce the noise in the OMI data. This is similar to previous studies (i.e., Kharol et al., 2017) and maximizes both the slope and correlation compared to other radii, as shown in Fig. S2." (Lines 160-163).

11. Line 282: what are the predictors used in Yang et al. (2023b) and Zhang et al. (2022)? What is the supporting evidence to restrict predictors to the specific 5 predictors used in this study? Are there any other physical predictors that need to be included in this study?

Response:

Yang et al. (2023b) trained their deep neural network using latitude, longitude, day of year, sun azimuth, sun elevation, as well as reflectance data from band 2 (0.45-0.51 µm), band 4 (0.64-0.67 µm), band 7 (2.11-2.29 µm), and band 10 (11.50-12.51 µm) from Landsat-8.

Zhang et al. (2022) developed a three-component machine learning model. The first component was trained using OMI satellite data, meteorology, emissions, and land use types (BASE). The second component of the model was trained using the same variables to estimate a scaling factor for back-extrapolation (RBE). The final component of the model used both the estimated surface concentrations from BASE and scaling factor from RBE to obtain a new estimate of the surface SO_2 concentrations. To train the models, Zhang et al. (2022) used OMI SO_2 VCDs, boundary layer height, zero plane displacement height, surface pressure, 10 m specific humidity, 2 m specific humidity, sea level pressure, 10 m air temperature, 2 m air temperature, total column ozone, total column odd oxygen, total precipitable ice water, total precipitable liquid water, total precipitable water vapor, tropopause pressure based on blended estimate, tropopause pressure based on thermal estimate, tropopause pressure based on Ertel potential vorticity (EPV) estimate, tropopause specific humidity using blended tropopause pressure estimate, temperature using blended tropopause pressure estimate, surface skin temperature, 10 m eastward wind, 2 m eastward wind, 50 m eastward wind, 10 m northward wind, 2 m northward wind, 50 m northward wind, SO_2 surface mass concentration, land use types (5 types of cropland, 4 types of grassland, 14 types of forest land, 6 types of shrubland, 3 types of bare areas, urban areas, water areas, and snow/ice areas), population density, elevation, normalized difference vegetation index, emissions from energy, industry, residential, and transportation sectors, as well as geospatial data including day of year, longitude, latitude, and the product of longitude times latitude.

One of the reasons we wanted to limit the number of predictors used for model training was highlighted in Zhang et al. (2022). The paper specifically mentions that they identified areas that were affected by non-physical predictions due to interactions between meteorological variables and land use types. Second, we wanted to use only physical quantities that relate to the emission or lifetime of SO_2. We selected the five variables in our model due to known physical relationships with surface SO_2 concentrations. OMI SO_2 VCDs describe where near-surface SO_2 may be located on a particular day. The ERA5 boundary layer heights and wind speeds represent the ability of SO_2 to either accumulate near the surface (low boundary layer heights and wind speeds) or mix out (high boundary layer heights and wind speeds). Boundary layer heights also tend to be correlated with temperature and solar radiation at the surface, so it also has a strong seasonal signal. The CEDS inventory also accounts for the locations where SO_2 is emitted into the atmosphere. Finally, we also wanted to use as few spatial and temporal proxies as possible so that the model is only dependent on measurable quantities. We wanted the model to be able to use the OMI observations and meteorological data to learn the spatial distribution and seasonality of SO_2 based on measurements rather than learning the latitudes, longitudes, and/or times of year where SO_2 is observed at the surface by the monitoring network. By avoiding these proxies, the model can then be used to estimate surface SO_2 concentrations in other locations both within the study region, as done in Section 4.2, and outside of the study region in potential future work.

Another reason we chose to restrict the number of predictors in our model is because in previous versions of our work, they did not contribute to the predictions, caused poor results, or did not have a known physical relationship to the lifetime or distribution of surface SO_2 concentrations. For example, temperature and humidity were included in previous versions of the model, although these were found to not significantly contribute to the ML predictions based on a permutation analysis, as seen in Fig. R20. In earlier versions of the model, we also included solar geometry (solar zenith/azimuth angles) since these variables can be used as a proxy for OH radicals that react with SO_2 in the atmosphere (Burnett, C.R. & Burnett, E.B., 1981), as well as affect the sensitivity of the satellite retrieval to near-surface SO_2; however, these caused an unrealistic spatial distribution of SO_2 due to the strong latitudinal gradients in the solar geometry, especially in the winter months (Fig. R21). A permutation importance analysis also revealed that the solar geometry variables were highly influential in the model (Fig. R20). Earlier versions of the ML model also attempted to account for interfering absorbers at similar wavelengths to SO_2 such as total column ozone and aerosol index, but these were also found to be relatively unimportant in the model predictions (Fig. R20).

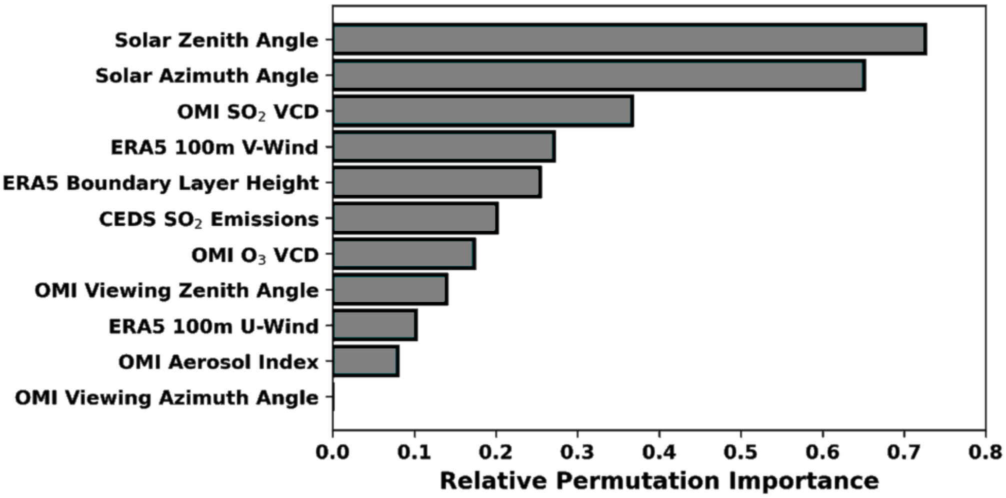

Figure R20: Permutation importance analysis performed on a previous version of the model with 11 predictors including SO_2 VCDs, ozone VCDs, aerosol index, solar zenith angle, solar azimuth angle, viewing zenith angle, and viewing azimuth angle from the OMI product, 100 m u-wind, 100 m v-wind, 2 m temperature, 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 observed in Fig. R21. This figure was added as Fig. S11 in the supplementary material.

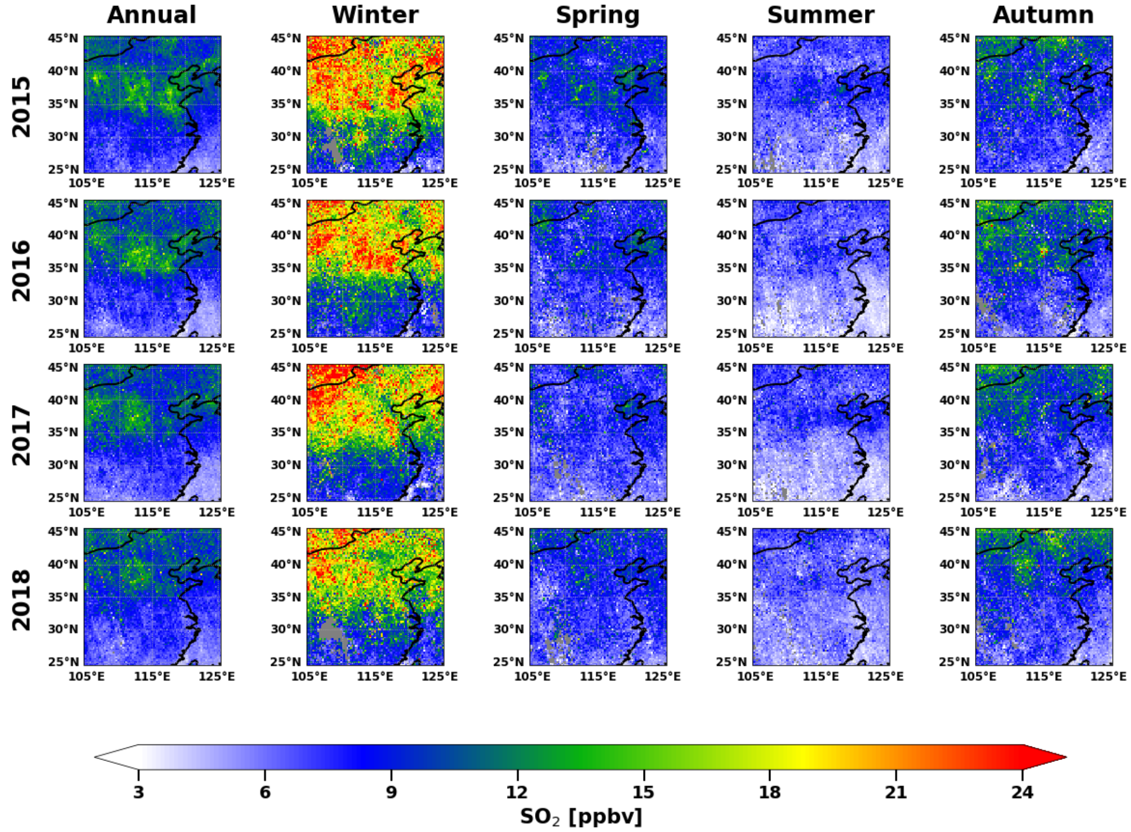

Figure R21: Maps of surface SO_2 concentrations from a previous version of the model with 11 predictors including SO_2 VCDs, ozone VCDs, aerosol index, solar zenith angle, solar azimuth angle, viewing zenith angle, and viewing azimuth angle from the OMI product, 100 m u-wind, 100 m v-wind, 2 m temperature, 2 m dew point temperature, and boundary layer height from ERA5, and SO_2 emissions from the CEDS emission inventory. The 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. This figure was added as Fig. S10 in the supplementary material.

We incorporated a brief discussion about how reducing the number of predictors improved our results in the text when describing the architecture and training variables of our XGBoost model:

"Earlier versions of the model were trained on 11 predictors, but the predicted surface concentrations produced an unrealistic spatial distribution of SO_2, as shown in Fig. S10. Additionally, some of the predictors were shown to be relatively unimportant to the model output, as indicated by the permutation importance in Fig. S11. The reduction of predictors from 11 down to five led to an improvement in the statistical performance and spatial distribution of the estimated surface concentrations, suggesting that utilizing known physical relationships between variables is more beneficial than the number of predictors in a ML model." (Lines 215-220).

12. Figures 6 and 7: only one month in each season in a single year of 2015 is not representative for seasonality analyses over 4 years. It is recommended to conduct complete 4-year simulations.

Response:

The annual and seasonal mean surface concentrations (including those from Figs. 6-7) were calculated from daily surface concentrations from the CTM-based method, ML-based method, and CNEMC in situ measurements that were then averaged over time. The SVRs obtained from the GEOS-Chem simulations were the only values that were not fully representative of the full year or season, but they were used to calculate daily surface concentrations from the OMI VCDs. Additionally, we conducted tests with archived GEOS-CF data with higher spatial and temporal resolution that showed that the impact of temporal representativeness on the SVRs was relatively small. More details can be found in RC2 General Comment 2.

References

Burnett, C. R., and Burnett, E. B.: Spectroscopic Measurments of the Vertical Column Abundance of Hydroxyl (OH) in the Earth’s Atmosphere, J. Geophys. Res., 86, C6, 5185-5202, doi:10.1029/JC086iC06p05185, 1981.

Devi, A. and Satheesh, S. K.: Global maps of aerosol single scattering albedo using combined CERES-MODIS retrieval, Atmos. Chem. Phys., 22, 5365-5376, doi:10.5194/acp-22-5365-2022, 2022.

Flyvbjerg, H. and Petersen, H. G.: Error estimates on averages of correlated data, J. Chem. Phys., 91, 461, doi:10.1063/1.457480, 1989.

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., Seibert, J. J., Vu, L., Andres, R. J., Bolt, R. M., Bond, T. C., Dawidowski, L., Kholod, N., Kurokawa, J.-I., Li, M., Liu, L., Lu, Z., Moura, M. C. P., O'Rourke, P. R., and Zhang, Q.: Historical (1750–2014) anthropogenic emissions of reactive gases and aerosols from the Community Emissions Data System (CEDS), Geosci. Model Dev., 11, 369–408, doi:10.5194/gmd-11-369-2018, 2018.

Jiang, X., Xue, Y., de Leeuw, G., Jin, C., Zhang, S., Sun, Y., and Wu, S.: Retrieval of hourly aerosol single scattering albedo over land using geostationary satellite data, npj Clim. Atmos. Sci., 7, 157, doi:10.1038/s41612-024-00690-6, 2024.

Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.-K., and Kim, S.: Estimation of surface-level NO2 and O3 concentrations using TROPOMI data and machine learning over East Asia, Environ. Pollut., 288, 117711, doi:10.1016/j.envpol.2021.117711, 2021.

Keller, C. A., Knowland, K. E., Duncan, B. N., Liu, J., Anderson, D. C., Das, S., Lucchesi, R. A., Lundgren, E. W., Nicely, J. M., Nielsen, E., Ott, L. E., Saunders, E., Strode, S. A., Wales, P. A., Jacob, D. J., and Pawson, S.: Description of the NASA GEOS Composition Forecast Modeling System GEOS-CF v1.0, Journal of Advances in Earth Modeling Systems, 13, e2020MS002413, doi:10.1029/2020MS002413, 2021.

Kharol S. K., Martin, R. V., Philip, S., Boys, B., Lamsal, L. N., Jerrett, M., Brauer, M., Crouse, D. L., McLinden C., and Burnett, R. T.: Assessment of the magnitude and recent trends in satellite-derived ground-level nitrogen dioxide over North America, Atmospheric Environment, 118, 236-245, doi:10.1016/j.atmosenv.2015.08.011, 2015.

Kharol, S. K., McLinden, C. A., Sioris, C. E., Shephard, M. W., Fioletov, V., van Donkelaar, A., Philip, S., and Martin, R. V.: OMI satellite observations of decadal changes in ground-level sulfur dioxide over North America, Atmos. Chem. Phys., 17, 5921–5929, doi:10.5194/acp-17-5921-2017, 2017.

Lee, C., Martin, R. V., van Donkelaar, A., Lee, H., Dickerson, R. R., Hains, J. C., Krotkov, N., Richter, A., Vinnikov, K., and Schwab, J. J.: SO2 emissions and lifetimes: Estimates from inverse modeling using in situ and global, space-based (SCIAMACHY and OMI) observations, J. Geophys. Res., 116, D06304, doi:10.1029/2010JD014758, 2011.

Li, C., Joiner, J., Krotkov, N. A., and Bhartia, P. K.: A fast and sensitive new satellite SO2 retrieval algorithm based on principal component analysis: Application to the ozone monitoring instrument, Geophys. Res. Lett., 40, 6314–6318, doi:10.1002/2013GL058134, 2013.

Li, C., Stehr, J. W., Marufu, L. T., Li, Z., and Dickerson, R. R.: Aircraft measurements of SO2 and aerosols over northeastern China: Vertical profiles and the influence of weather on air quality, Atmospheric Environment, 62, 492-501, doi:10.1016/j.atmosenv.2012.07.076, 2012.

Li, C., Krotkov, N. A., Leonard, P. J. T., Carn, S., Joiner, J., Spurr, R. J. D., and Vasilkov, A.: Version 2 Ozone Monitoring Instrument SO2 product (OMSO2 V2): New anthropogenic SO2 vertical column density dataset, Atmos. Meas. Tech., 13, 6175–6191, doi:10.5194/amt-13-6175-2020, 2020.

Lin, J.-T. and McElroy, M. B.: Impacts of boundary layer mixing on pollutant vertical profiles in the lower troposphere: Implications to satellite remote sensing, Atmospheric Environment, 44, 1726-1739, doi:10.1016/j.atmosenv.2010.02.009, 2010.

Liu, Y., Park, R. J., Jacob, D. J., Li, Q., Kilaru, V., and Sarnat, J. A.: Mapping annual mean ground-level PM2.5 concentrations using Multiangle Imaging Spectroradiometer aerosol optical thickness over the contiguous United States, J. Geophys. Res., 109, D22206, doi:10.1029/2004JD005025, 2004.

McDuffie, E. E., Smith, S. J., O’Rourke, P., Tibrewal, K., Venkataraman, C., Marais, E. A., Zheng, B., Crippa, M., Brauer, M., and Martin, R. V.: A global anthropogenic emission inventory of atmospheric pollutants from sector- and fuel-specific sources

(1970-2017): an application of the Community Emissions Data System (CEDS), Earth Syst. Sci. Data, 12, 3413-3442, doi:10.5194/essd-12-3413-2020, 2020.

National Aeronautics and Space Administration Global Modeling and Assimilation Office (NASA GMAO): GEOS Composition Forecast (GEOS-CF), NASA Center for Climate Simulation, NASA Goddard Space Flight Center [Dataset], available at https://portal.nccs.nasa.gov/datashare/gmao/geos-cf/v1/ana/.

Norman, O. G., Heald, C. L., Bililign, S., Campuzano-Jost, P., Coe, H., Fiddler, M. N., Green, J. R., Jimenez, J. L., Kaiser, K., Liao, J., Middlebrook, A. M., Nault, B. A., Nowak, J. B., Schneider, J., and Welti, A.: Exploring the processes controlling secondary inorganic aerosol: evaluating the global GEOS-Chem simulation using a suite of aircraft campaigns, Atmos. Chem. Phys., 25, 771-795, doi:10.5194/acp-25-771-2015, 2025.

Philip, S., Martin, R. V., and Keller, C. A.: Sensitivity of chemistry-transport model simulations to the duration of chemical and transport operators: A case study with GEOS-Chem v10-01, Geosci. Model Dev., 9, 1683–1695, doi:10.5194/gmd-9-1683-2016, 2016.

Shan, Y., Zhu, Y., Sui, H., Zhao, N., Li, H., Wen, L., Chen, T., Qi, Y., Qi, W., Wang, X., Zhang, Y., Xue, L., and Wang, W.: Vertical distribution and regional transport of air pollution over Northeast China: Insights from an intensive aircraft study, Atmospheric Environment, 358, 121327, doi:10.1016/j.atmosenv.2025.121327, 2025.

Xue, L., Ding, A., Gao, J., Wang, T., Wang, W., Wang, X., Lei, H., Jin, D., and Qi, Y.: Aircraft measurements of the vertical distribution of sulfur dioxide and aerosol scattering coefficient in China, Atmospheric Environment, 44, 278-282, doi:10.1016/j.atmosenv.2009.10.026, 2010.

Yang, Q., Kim, J., Cho, Y., Lee, W.-J., Lee, D.-W., Yuan, Q., Wang, F., Zhou, C., Zhang, X., Xiao, X., Guo, M., Guo, Y., Carmichael, G. R., and Gao, M.: A synchronized estimation of hourly surface concentrations of six criteria air pollutants with GEMS data, npj Clim. Atmos. Sci., 6, 94, doi:10.1038/s41612-023-00407-1, 2023a.

Yang, Q., Yuan, Q., Gao, M., and Li, T.: A new perspective to satellite-based retrieval of ground-level air pollution: Simultaneous estimation of multiple pollutants based on physics-informed multi-task learning, Sci. Total Environ., 857, 159542, doi:10.1016/j.scitotenv.2022.159542, 2023b.

Yang, Y., Mou, S., Wang, H., Wang, P., Li, B., and Liao, H.: Global source apportionment of aerosols into major emission regions and sectors over 1850-2017, Atmos. Chem. Phys., 24, 6509-6523, doi:10.5194/acp-24-6509-2024, 2024.

Zhang, L. and Li, J.: Variability of Major Aerosol Types in China Classified Using AERONET Measurements, Remote Sens., 11, 20, 2334, doi:10.3390/rs11202334, 2019.

Zhang, S., Mi, T., Wu, Q., Luo, Y., Grieneisen, M. L., Shi, G., Yang, F., and Zhan, Y.: A data-augmentation approach to deriving long-term surface SO2 across Northern China: Implications for interpretable machine learning, Sci. Total Environ., 827, 154278, doi:10.1016/j.scitotenv.2022.154278, 2022.

Zhang, X., Wang, Z., Cheng, M., Wu, X., Zhan, N., and Xu, J.: Long-term ambient SO2 concentration and its exposure risk across China inferred from OMI observations from 2005 to 2018, Atmos. Res., 247, 105150, doi:10.1016/j.atmosres.2020.105150, 2021.

Zheng, Y., Che, H., Xia, X., Wang, Y., Yang, L., Chen, J., Wang, H., Zhao, H., Li, L., Zhang, L., Gui, K., Yang, X., Liang, Y., and Zhang, X.: Aerosol optical properties and its type classification based on multiyear joint observation campaign in north China plain megalopolis, Chemosphere, 273, 128560, doi:10.1016/j.chemosphere.2020.128560, 2021.