Answers to referee #2 comments, received and published on 16 October 2013, on the manuscript:

"Evaluation of GEOS-5 sulfur dioxide simulations during the Frostburg, MD 2010 field campaign"

We thank the reviewers for providing comments that helped to improve the quality of the paper. The detailed responses to comments are listed below (text in black shows comments from the reviewers, and the text in blue is our answer):

This paper presents a modeling study with the GEOS-5+GOCART global model that focuses on the sensitivity of surface level SO2 and sulfate to anthropogenic SO2 emissions, particularly the injection height of these emissions. Model results are evaluated using surface measurements from a monitoring network in the continental US, and aircraft and surface remote sensing measurements from a short duration field campaign over north-eastern Maryland. The main conclusion is that emitting SO2 from energy sources at 100-500 m rather than in the lowest model layer greatly reduces the strong high bias of the surface SO2 concentrations. However, this change had little impact on surface concentrations of sulfate, which have a weaker high bias, suggesting that sulfate removal may too slow in the model. The subject matter is appropriate for ACP, and the paper is concise and generally well written. It should be acceptable for publication in ACP after some revision.

My main concern is the lack of technical discussion about the appropriate injection height for SO2 emissions from fossil fuel power plants. The elevation range used (100- 500 m) is given without any justification or discussion. Plume rise of emissions from large point sources has been studied for many decades, and it is incorporated in the emission modules of several mesoscale to regional scale models (e.g., EPA CMAQ, WRF-Chem). Although information needed to estimate plume rise and injection height is not available (to my knowledge) for global emissions datasets, it is available for the US. The authors might consider it outside the scope of their study to incorporate such information, although it would certainly strengthen their results. Some technical discussion of injection heights is definitely needed.

GEOS-5 is a global model and the emissions we are using are global gridded emission dataset. In the revised run, EDGAR v4.1 offers the opportunity to partition the sources between the energy sectors on the one hand and non-energy sources on the other. We release the energy SO2 anthropogenic emission above 100 meters mainly because of the averaged stack height for power plants. While there is literature on the plume rise of emissions from large point sources - information that has been incorporated in more traditional air quality models such as CMAQ and WRF-Chem – to our knowledge these strategies have not been applied to global models like ours. The simple vertical partitioning we apply here should be understood as a starting point for our global air quality simulations, and will be subject to further investigation.

Specific Comments

As noted by Referee 1, there are two differences between the control and revised run SO2 emissions: magnitudes and injection heights. If there are appreciable differences in magnitudes, then the authors should perform a third simulation in which only one of these emissions differences was applied. Some of the discussion suggests that the emission magnitudes do not differ substantially. If this is the case, the third run is less important, but the authors should provide some quantitative comparison of the two emissions data sets. E.g., give the annual emissions for the entire globe, for the continental US (or the area in Figs. 1 and 2), and for the portion of the US where most of the monitoring sites (see Fig. 6) are located. Spatial correlation coefficients for the two emissions datasets might also be provided. Also, Figs. 1 and 2 should be revised to allow visual comparison.

Indeed between the control run and the revised run, there are two differences: the emission inventory dataset and the injection height. A table (Table 1) with annual emission rates for 2005 for the 2 datasets for the entire globe, the US and over the eastern US where a lot of power plants are located has been added to the paper. As suggested, Figure 1 and 2 has been replaced and in the new version Figure 1 shows the anthropogenic SO2 emission for the 2 inventories and the differences between both.

STDV statistics. Like Referee 1, I was not sure that I understood how this is calculated. If STDV is simply the standard deviation of a simulation's results (hourly concentrations at the observation sites), then the importance of these statistics to the analysis is unclear. Please clarify both how STDVs are calculated and their importance. Also, in Fig. 6 and 8, would showing RMS rather than STDV results in the middle rows be more useful?

In the paper the standard deviation (STDV) calculated is the STDV of the differences between the modeled and observed values, so this is the variability of the error between the two values. Page 21772, line 5: the text was clarified "the standard deviation of the differences (STDV) and the mean…."
The $(RMS)^2$ can be seen as the sum of an estimate of variance ($STDV^2$) + estimate of $(bias)^2$, this is why we decided to show the STDV in the 2nd row and the bias in the 3rd row.

Section 3. Have the authors looked into seasonal and diurnal differences in the simulated surface SO2 bias? If these differences are small, then this could be stated in one or two sentences. If they are large, more discussion would be informative. Tables 1-2 (number of points) and Figures 10-11 suggest that January 2010 results are missing for the control run. Please explain, and include run duration information (including spin-up) at top of p. 21769). If this is correct, is it appropriate to use different time periods for the Table 1-2 and Figure 5-8 results?

We don't have seasonally varying emissions, following the reviewer comment, we looked at the seasonal differences in the surface SO2 bias and globally we did not find substantial seasonal variation in the surface SO2 bias and sulfate bias. The EPA data that we have are daily means, we did not look at diurnal differences.

January 2010 results are missing for the control run, in order to be consistent in the number of data between the two runs, the SO2 and sulfate comparisons have been redone removing January 2010 for the revised run. The numbers and the statistics in both tables 2-3 have been updated as well as the numbers on the corresponding figures.

P. 21774, L. 1-3. I concur with the comment by P. Castellanos on sulfate removal rate. The authors should at least provide numbers for emissions increases and/or observed surface sulfate decreases from 2005 to 2010, and compare these to the (normalized) model bias for sulfate.

Following P. Castellanos and the reviewer comments, a new comparison has been performed for the year 2005 between ground-based EPA sulfate measurements and GEOS-5 simulated sulfate. A positive bias remains in the comparison but lower than the one observed for the year 2010. The positive bias in sulfate might also be attributed to the overestimated SO2 emissions for 2010. Following this new analysis the text in section 3.2 "Sulfate aerosol" p 21773 has been updated.

The model is global, so why limit evaluation to continental US? European observations (EMEP) could also be included. Also, does the injection height change have much impact on the SO2 and sulfate global lifetimes?

We took the opportunity of the Frostburg campaign in Maryland to evaluate the SO2 simulated by GEOS-5 and we extended our analysis to the US. Evaluate the SO2 over Europe was beyond the scope of this paper and could be done in a future work.
We have computed the SO2 lifetimes from the control run (it would take time to rerun the model with the EDGAR emissions emitted only at the surface for an entire year) and the SO2 lifetime values can be different by about 5-10% (In the control run, the height is different but also the emissions).

Section 4.1. Since Piney Run Station is "in a mountain valley", it is not ideal for evaluation of 25 km resolution model results, which probably do not resolve flow details at scales below about 50 km. Add some discussion about the topography here. What are the valley bottom and ridge top elevations and the valley width? Is the actual terrain much more complex than that used by the model? Also, consider showing surface SO2 on Nov. 8-9 separately from Fig. 11, with an expanded time axis. Section 4.2 Add more discussion of the meteorology on and preceding Nov. 8 and 9.
Was the weather generally clear, or were there low clouds or precipitation in the area (which would strongly affect SO2 concentrations), and if so, how well did the model simulate them? What were the wind speeds and direction below 1 km (where most of the SO2 was found) at the MFDOAS site on these days, and were there any shifts on Nov. 9 that might explain the MFDOAS downward trend?

The Frostburg campaign was a good opportunity to evaluate the model because of all measurements made during the campaign, even if the model was run at a nominal 25 km horizontal resolution that is coarser than the mountain/valley terrain. The topography has been added to the map Figure 8.
To answer on how well GEOS-5 simulate the winds at 500 mb and precipitation over the

US, we have made comparisons with the National Centers for Environmental Prediction (NCEP) data and the comparisons are globally good. Concerning the precipitation we looked also at GPCP (Global Precipitation Climatology Project) data over the US and again the comparison is satisfying for the 2 days. For the weather, according to MODIS RGB images for Nov 8-9, the sky was clear of clouds during these 2 days over the region of the campaign. The MFDOAS was deployed at Frostburg State University and there was no weather station during the campaign.

Section 4.3. Please be more specific about the "major features" that the simulation captures. For example, the model values appear to drop off more rapidly with height (from within to above the PBL) than the measurements, and model is too low at _25- 35 minutes. If the aircraft recorded air temperature, then how did the model's inversion heights compare to the observed. You compare control and revised run results in nearly all the figures, so why not show the control run SO2 in the rightmost (line) plots of Fig. 13, and discuss it in text?

The description of the comparisons between the aircraft and the model in section 4.3 has been completed in the new version.
The model's inversion height might need more investigation in the future, we have compared the GEOS-5 temperature (at the time and space of the aircraft measurement) with the NCEP temperature and the comparisons are good. The comparison with the aircraft measurements is satisfying but we can have some differences explained by the topography in this region and the resolution of the model.

The minor comments have been taken into account in the new version of the paper.

Minor Comments
P. 21766, L. 15-16. I did not see any substantial discussion of "mixing processes in the model".
P. 21767, L. 1-3. SO2 oxidation is quick only in the presence of clouds, and so is highly variable.
P. 21767, L 19. Change to ". . . SO2 losses due to oxidation and dry/wet removal" ?
P. 21768, L. 5. Change to "Representation of Aerosols and Sulfur Gases in . . ." ?
P. 21768, L. 18-21. Please give some information about the model's vertical resolution in the lowest 1 km, such as the number of vertical levels here.
P. 21769, L. 19. Change to ". . . over the US in 2007 (from Streets et al., 2009)" ?
P. 21770, L. 19-23. These two sentences would seem to fit better on P. 21769 after L. 17.
P. 21771, L. 3-4. Mention that Lee et al. results are for 2006.
P. 21771, L. 13-14. Differences in clouds and precipitation in between 2006 and 2010 could also be a factor here, as well as less SO2 dry deposition in the revised run because of the elevated emissions.
P. 21772, L. 5-6. Please state here that throughout the paper, "log" means natural (or base 10) logarithm.
P. 21772-3 or Tables 1-2. Please give the means and STDVs of the observations, as hey help to put the model-observation comparison statistics (those in ppb and ug/m3

units) and model STDVs into perspective.
Figures. Axis labels and numbers could be larger on many of them.
Fig. 3. Choice of contour levels (which leave 80% of globe as white) could be improved.
Fig. 5. In the caption, add some explanation of the coloring (representing the PDF).
Fig. 9. It would be more informative to show topography (elevation contours) on the map rather that state outlines and rivers.