Answers to referee #1 comments, received and published on 1 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):

The manuscript "Evaluation of GEOS-5 sulfur dioxide simulations during the Frostburg, MD 2010 field campaign" by Buchard et al. uses in-situ observations and field campaign measurements to evaluate their global model simulations. The model was driven by MERRA and two different emission inventories. The main conclusion is that the model tends to overestimate the sulfur dioxide concentration at the surface because of (1) the low injection height and (2) the higher emission rates in the 2005 inventory. I do have some minor comments mostly for clarification. I also have some questions regarding the results, but I understand that some of them might be beyond the scope of this study and should be answered in a different paper. Overall, I find this manuscript well-prepared, concise, scientifically sound, and that it fits well with ACP. Therefore, I recommend this manuscript to be published on ACP after a minor revision.

Here are my questions/minor comments:

1. Page 21768, line 19-21: You mentioned that the model was run at various (actually just two) resolutions, but you only showed the results from the 0.23 x 0.315 resolution simulation. So, I suggest that you revise this sentence to something like "The model was run at a horizontal grid spacing of 0.23 x 0.315".

The model was run at 2 different resolutions: the near real time system runs at a nominal 25 km horizontal resolution, so the control run for 2010 was run at a nominal 25 km resolution. After changing the emissions dataset and in order to make some tests, the model was run at a nominal 50 km resolution. For days in November during the Frostburg campaign, the model (revised run) was run at higher resolution (25 km). In the sulfate comparison part, the GEOS-5 aerosol reanalysis was run at a nominal 50 km resolution. The sentence p 21768 19-21 was updated.

2. Page 21768, line 24-26: What fields are provided to constrain the model's meteorology?

The MERRA fields are provided to constrain the model's meteorology. It is specified p 21 769 l 1-4: "For this study GEOS-5 is run in replay-mode using the GMAO atmospheric analyses from the Modern Era Retrospective analysis for Research and Applications (MERRA) available every six hours."

3. Page 21768, line 26: I do not understand the difference between the GEOS-5 and the CTM (e.g., GEOS-Chem). Could you elaborate?

GEOS-5 and GEOS-Chem are two different global models developed by different team. When GEOS-5 is run in replay mode, the meteorological fields come from the GMAO atmospheric analyses (MERRA), so the same version of the model. This allows GEOS-5 to function like a traditional chemical transport model.
Geos-Chem is driven by assimilated meteorological observations coming from the GMAO atmospheric analyses (GEOS-4, GEOS-5, MERRA).

4. Page 21769-21770: Regarding the difference between CR and RR, there seem to be two differences: (1) they use different emission inventories, and (2) they use different injection heights. Is that correct? If so, how can one separate the effect from one another? For example, in Figure 10 and 11, you mentioned that the different injection heights between CR and RR is responsible for the different surface SO_2 concentration. Is this because the two emission inventories have roughly the same SO_2 emission rates? In any event, I suggest that you replace Figure 1 and 2 with contour plots of SO_2 emission rates (column integrated) from the two inventories, and then show the difference between the two inventories.

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 SO_2 emission for the 2 inventories and the differences between both.

5. Figure 4, 10, 11, 12, 13: Please consider to change the colors, with observations in black (with gray shading) and simulations in red and blue.

We would like to keep the colors originally chosen.

6. Page 21771, line 3: Did Lee et al. (2011) use the same emission inventory?

Lee et al. (2011) performed some simulations using GEOS-Chem and different emission inventory. They used regional inventories over the US.

7. Page 21772, line 5: I do not understand how you define the standard deviation of the two datasets. We usually view standard deviation of a dataset as its variability. What does standard deviation of two datasets mean? Please explain.

In the paper the standard deviation (STDV) calculated is the STDV of the differences between the observed and modeled 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…."

8. Page 21772, line 8: You mentioned "the scatter . . . is significant". I do not understand this statement. Do you mean the observations and simulations are significantly different

(or similar)? Please explain the significance test you used to support this statement (i.e., what significance test, what null hypothesis, at what significance level, etc).

We agree that the word "significant" was not appropriate in the sentence, so the p 21772, l 8 the text was changed: "In both plots, there is considerable scatter between modeled and observed daily means with correlation coefficients…"

9. Figure 5, 7: It seems that the model overestimates the SO2 concentration in general (as you stated), but especially when the observed values are very low. Do you have an explanation?

We agree that the SO2 surface concentrations are overestimated and especially when the observed values are very low, this is a topic that we are investigating. For low SO2 concentration values the absolute value of error is higher for ground-based measurements.

10. Figure 6, middle column, top and middle plots: it seems that the model has different behaviors between the northeast US and the mid-west US. Do you have an explanation?

When looking at the new figure 1 showing the SO2 emissions, the new emissions in the revised run went down over the East coast and coupled with the injection height, the revised run is less in agreement with the observations than the control run over this region.

11. Figure 6 and 8: Why did you reverse the color bar for the middle and right plots on the top panel?

The color bar has been reversed in order to be consistent with the color coding used for the STDV and bias (middle and bottom panel). For the 3 panels the color blue indicates improvement relative to the control run (greater correlation coefficient (r) values for the top panel, lower STDV and lower bias for the middle and bottom panel).

12. Regarding the comparison between model simulations and aircraft measurements (Section 4.3): What is the frequency of the measurement? What is the model output frequency? What is the model time step?

The aircraft measurements were every 10 seconds, the flights lasted about 2 hours. The model outputs were every 3 hours. The model has been interpolated in space and time in order to perform the comparisons.