Response to Reviewer #2:

We thank the reviewer for his/her very useful comments and questions, which helped us improving our manuscript. Below we show the reviewer's comments in roman font and our answers italicized.

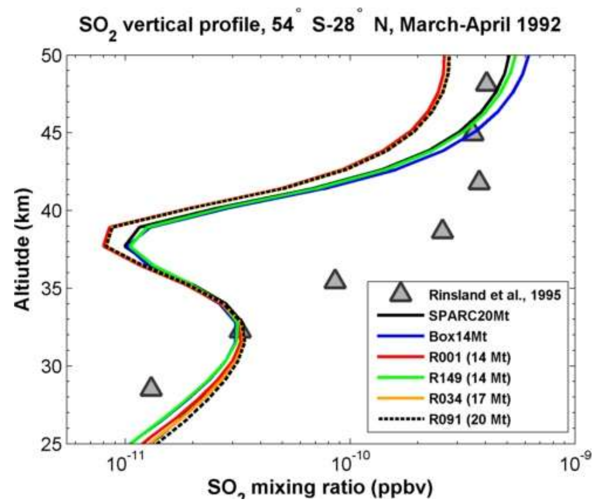
1. General comments

The conclusion that the lowermost estimate of earlier studies for SO2 injected by Pinatubo should be selected is only of limited value because it appears to be significantly perturbed by model artifacts and arbitrary weighting of 'scores' for differences to observations. Giving the MLS SO2 measurements a higher weight and the rather uncertain burdens estimated from SAGE during saturation of the instrument a lower one would completely change the conclusions. From text and figures it is also often not clear how the scores were calculated. In the table and the figures important cases are missing. In the introduction the ATMOS observations of Rinsland et al. (1995) should be cited. They should be included into the scoring scheme too.

We do not agree with the reviewer about the 'model artifacts', which are mentioned here at the start but then not detailed (below OH concentrations and the QBO are mentioned, but these should not be deciding factors). Nor would we agree on the weighting being 'arbitrary'. Actually the choice of weighting is discussed in detail in the text (Section 3.1). Giving the MLS SO2 measurements a higher weight, as suggested by the reviewer, would not be proper due to the very short lifetime of SO2 and the fact that the only available measurements with vertical resolution of SO2 in the stratosphere during the Pinatubo period have been made by MLS, which unfortunately only started its mission only three months after the eruption (as we clearly state on p. 4603, l. 6-9). The burdens are estimated from both HIRS and SAGE. HIRS is applied during the saturation phase of SAGE. The overall uncertainty of the composite of HIRS and SAGE is estimated to be about 20% (as stated at the top of p. 4610).

We are grateful for the suggestion to add the formulas for calculating the scores, which we now do in Section 3.1.

ATMOS observations (ALTAS-1) of Rinsland et al. (1995) are during March 24 and April 2, 1992, which is about 10 months after the Pinatubo eruption (SO_2 lifetime in the stratosphere is about a month), i.e. SO2 concentrations are lower than 10^{-4} of the initial value. Furthermore, ATMOS measurements are in the upper stratosphere (above ~30km), where the source of SO2 mainly comes from photodissociation of H_2SO_4. However, model results above 35 km are very sensitive to the photolysis of H_2SO_4 (see Fig. 2 in Rinsland et al., 1995), which is highly uncertain. Below 35 km, as shown in the attached plot below, the AER model results from different initial injections of SO2 agree reasonably with ATMOS, but do not differ from each other significantly. Therefore, we refrain from including ATMOS observations into the scoring scheme.



2. Specific comments

Section 2.1: Is the pre-calculated OH based on the updated chemistry? Why OH is pre-calculated? Most chemical 2D-models calculate that in interactive mode. Errors in the troposphere due to simplified hydrocarbons are not relevant for Pinatubo. I suppose meteoric dust is not treated explicitly.

OH is pre-calculated because this study uses a version of the AER 2-D model with only sulfate source gases and intermediate products, along with resulting aerosol, interactively calculated. Sulfur gases are sensitive to OH, but OH is not sensitive to sulfur gases at levels below and including the Pinatubo-injected SO_2 amounts (see Bekki, GRL, 22, 913-916, 1995). Omitting the full chemical mechanism that would be used in an ozone simulation makes the model rapid enough for the ensemble calculations presented here. The pre-calculated OH-fields used in the AER 2D model is the same as Notholt et al. (2005). The OH-fields have been calculated with the 3-D chemistry-transport model MATCH-MPIC (Lawrence et al., 2001). Notholt et al. (2005) found that the model results are in reasonable agreement with calculations from 3-D models [Pitari et al., 2002] and with PEM tropics data (Thornton et al., 1999). Furthermore, differences between the AER pre-calculated OH fields and AER full chemistry are within 10-30% throughout the stratosphere (except for very high latitudes), which results in potential discrepancies in the same order as the uncertainty of the rate coefficient of SO_2 + OH. We don't expect this to be significant in comparison with other 2-D model uncertainties.

Meteoric dust is indeed not treated in the AER model, but this is irrelevant under volcanic conditions such as after the Pinatubo eruption, when the meteoritic mass in the stratosphere corresponds to less than 1% of the total aerosol mass.

Section 2.2: Due to the low vertical resolution the 3D model cannot have an internal Quasi-Biennial Oscillation. Is nudging applied? Especially if this is not the case the tropical water vapor tape recorder is artificially fast. Please expand.

QBO in the 3D model is indeed nudged. We add this information in Section 2.2.

Section 3, metrics, paragraph 1: An equation should be given for scores. What is in the denominator? The exact definition is especially important for the extinction which varies on a logarithmic scale. Now the text is rather confusing. Paragraph 2: I cannot follow the arguments for weighting. Before September 1992 SAGE derived burden is very uncertain due to saturation and gap filling. What is month 12 in line 14? Is this December 1991 (Fig. 3) or June 1992? It is not appropriate to give SAGE burdens (and extinctions in lower stratosphere) a large weight from January 1992 to September 1992.

We revised the manuscript and now explicitly provide the equation that we used to calculate the scores in Section 3.1. We changed 'month 12' to 'December 1991'. SAGE is most saturated and uncertain in the tropics during first six months after Pinatubo, i.e. before January 1992, see Russell et al. (1996). HIRS is applied during the most saturated phase of SAGE. See our answers in General Comments.

Section 3, scoring table: The case with rank 1 for SO2 should be listed in Table 1 too. The cases with peak emission at 29.5km should be skipped because that is against any observation. The conclusion in line 19 is strongly dependent on the arbitrary weighting and too early. All scenarios having rank 1 in one criterion should be discussed in more detail.

We now list the case with rank 1 for SO2. We still keep some of the cases with peak emission at 29.5 km in order to show some examples for these worst scenarios. We don't agree on "arbitrary weighting". See our answer in General Comments. We now provide more discussion about all scenarios having rank 1 in one criterion in Section 3.3-3.6.

Section 3, matching SO2: Here also the scenario with rank one in this criterion should be discussed.

We added some discussion about the scenario with rank 1 for SO2 in Section 3.3.

Section 3, matching burden: The sulfate mass without water should be given too because there is often confusion in the literature on this. There is the common problem that simulations are too high in the early phase and too low in the second and third year after the eruption. Elaborate more in this. The sentence with 'age of air' is confusing.

We add the information about sulfate mass without water in the text. "Too high in the early phase" could be due to the fact that the early data in SAGE-4λ are either gap-filled and therefore uncertain, or are based on very few original SAGE measurements (while all other measurement situations were opaque), which leads to a low bias in the SAGE-4λ data set. In other words, the measured data may be erroneously low (This is also the reason we use HIRS data in the first six months after the Pinatubo eruption).

Concerning the too low burden or AOD in the 2^{nd} and 3^{rd} year, actually the 2-D simulations show very reasonable agreement with the observed global aerosol burden (Fig. 3) and AOT (Fig. 6, a new figure suggested by the reviewer #1). For the 3D model, too low tropical AOT in the 2^{nd} and 3^{rd} year might be partially explained by the too low age of air in the 3D model (Sheng et al., 2015).

Section 3, matching extinctions: All shown 2D-simulations overestimate extinction above about 23 km and underestimate it in the lowermost stratosphere 1 year after the eruption. All scores are rather poor in this criterion but have a large weight.

Extinction scores in the lower stratosphere (18-23km) are reasonable and have a much larger weight than those above 23 km and in the lowermost stratosphere, because extinctions at 525 nm and 1020 nm at 18-23 km after the Pinatubo eruption (see Fig. 5) are one to several orders of magnitude larger than those above 23 km and in the lowermost stratosphere (15-16 km). We calculated the score by the relative Euclidean norm, therefore the scores above 23 km and in the lowermost stratosphere have a very small weight. We now clarify this in the text (Section 3.1).

3. Technical corrections

Page 4605, line 12: Typo?

"couple" changed to "coupled".

Page 4606, lines 16ff and Table 1: μ is a bad choice for an altitude parameter since it is normally used in atmospheric sciences for totally different quantities. Better use for example z0.

*μ is a standard notation for the skewed normal distribution. As it is properly defined on p4606-4607 we do not think that it causes any confusion in the text. Therefore, we prefer to keep the present notation.*Page 4608, lines 16f: improve structure, sentence is confusing.

We added a formula and reworded the sentence.

Page 4608, lines 25ff: the numbers should go also into the table caption

Done.

Introduce subsections in section 3.

We introduced subsections in Section 3.

Figure 1: Include all relevant simulation numbers of Table 1 in legend.

Done.