Response to Reviewer #2

We thank Reviewer #2 (R2) for the time spent reviewing our manuscript. We feel that it has improved significantly from incorporating his/her thoughts and suggestions. We respond to the reviewer criticism as follows (Review comments italicized):

1) *To start with, an evaluation of the model performance in correctly simulating the global mean TOA SW anomalies and global mean temperature anomalies needs to be shown so that to prove the model skills is correctly capturing the first order forcing and temperature response to the historical eruptions. Same remark for the South American precipitation mean seasonal cycle at least in DJF (selecting two levels of precipitation contours as on figure 6 won't just do the trick). This should be a first order sanity check for the South American Monsoon mean climatology and for the L20 eruptions of the historical periods for which observation are available.*

Following the comments of multiple reviewers, we do plan to revisit the historical/validation section of our paper. This includes an improved figure for the seasonal cycle in South American precipitation, in addition to the isotopes that we have already done.

The GISS climate model has a long history of making comparisons of the 1991 eruption of Mt. Pinatubo to observations (Hansen et al., 1996). Global temperatures are reduced by (on order) half a degree in the months following the Mt. Pinatubo eruption. "Zero order" analyses of this sort for ModelE2-R have been performed in a number of other studies and it cannot be the point of this paper to repeat all these previous analyses. Instead we will include a discussion of these papers and include appropriate references (e.g., see list provided at end of response to reviewer 1).

The question is complicated for TOA SW. In fact, this question posed to another modeling group would be mute – some groups for CMIP5 represent volcanic eruptions *exactly* as a TOA SW forcing. The implementation in the GISS code is more complicated – see Lacis et al 1992. – such that the TOA SW anomaly is influenced by not only the AOD of the sulfate aerosols, but also their size distribution. Although Mt. Pinatubo may widely be regarded as "well-observed" there is still considerable uncertainty regarding its forcing. The SAGE II instrument was saturated during the 1991 eruption, and the maximum AOD and size distribution have considerable uncertainty. More recent analyses of the Pinatubo aerosol forcing (e.g., Santer et al., 2014; Schmidt et al., 2014) have come to new conclusions that will lead to substantially reduced Aerosol Optical Depth and differences in particle size between CMIP5 and CMIP6. For CMIP6 (including PMIP4), the AOD of Mt. Pinatubo will likely be reduced.

The experiments presented here followed the CMIP5 / PMIP3 protocol for forcing of AOD and size distribution. For the historical eruptions, we would pass the 'first order' sanity check. For a more in depth analysis of signal-to-noise of the GISS model (and comparing to CCSM) for temperature and precipitation, please see (Marvel et al., 2015).

In general, it is worth pointing out that the main historical eruptions (e.g., Mt. Pinatubo) do not represent a useful validation target in our context. The spatial pattern of the post-eruption response is dominated by internal variability (e.g., ENSO). We did attempt to remove ENSO in our late 20th century (L20) analysis, but its expression over South America in particular is non-linear. ENSO and additional unforced variability mask the volcanic forcing at the regional level in L20 eruptions. Thus, the fact that the model does not "look like" observations following a given L20 event is not a reasonable criticism, as free-running GCM's are not built for this purpose.

For all these reasons we intended to shift focus to the larger LM composite in this study, which features a larger sample of events (larger signal-to-noise). We will improve this segway and motivation.

2) *My second comment concerns the method used to build the super-posed epoch and composite analysis. The authors compute anomalies respectively to the period three years before and 5 years after each eruption for both temperature and precipitations in observations for El Chichon and Pinatubo eruptions. By doing so the authors remove part of the volcanic signal. Why choosing this period? GISTEMP anomalies are based on the 1961-1990 climatology. Did you check the consistency between the two anomalies? I'd suggest removing the 1961-1990 climatology, for precipitation and temperatures so that to avoid removing the climatology with part of the climate response to volcanic forcing.*

We will re-visit the superposed epoch analysis, including addition of statistical analysis to improve the presentation.

With respect to the choice of base period, we will do a much better job of describing this in the manuscript. The choice we use simply shifts the entire curve up or down relative to the suggestion by Reviewer #2, but does not influence its temporal structure. What we actually did in the case of temperature was to use GISTEMP land-ocean temperature index, which is already provided as deviations from the 1951-1980 period (not 1961-1990) and the model data, where anomalies were then computed using that same long-term climatology.

However, we also subtract a constant in order to force the data in Figure 3 to have zero mean. Even though Mt. Pinatubo results in global cooling, large-scale tropical mean anomalies in the late 1980s and 1990s are still positive relative to the 1951-1980 climatology, due to the long-term warming trend. Since it may be awkward to display a plot of this sort with all positive anomalies, we subtract off the mean anomaly during years -3 to 5 from each data point. It is true that this is equivalent to using those years as our reference period. However, we do not view this as "removing the volcanic signal." The Hansen et al. (1996) publication makes a nice stack of 5 historical eruptions. Their method averages the 12-months prior to the eruption. This is the method that is used to make many epoch analysis stacks.

To better illustrate this, the following two figures show a comparison of Figure 3 in our discussion paper (only for Mt. Pinatubo here), in both cases with the same monthly-mean anomalies (fill color) using our choice of reference period.

The lines represent moving averages of the instrumental data (black), model ensemble mean (orange), and the individual ensemble members (grey dashed). The bottom figure shows how the data would look if we retained the 1951-1980 climatology. Regardless of choice, this will not affect our interpretation of the post-volcanic signal in the epoch analysis.

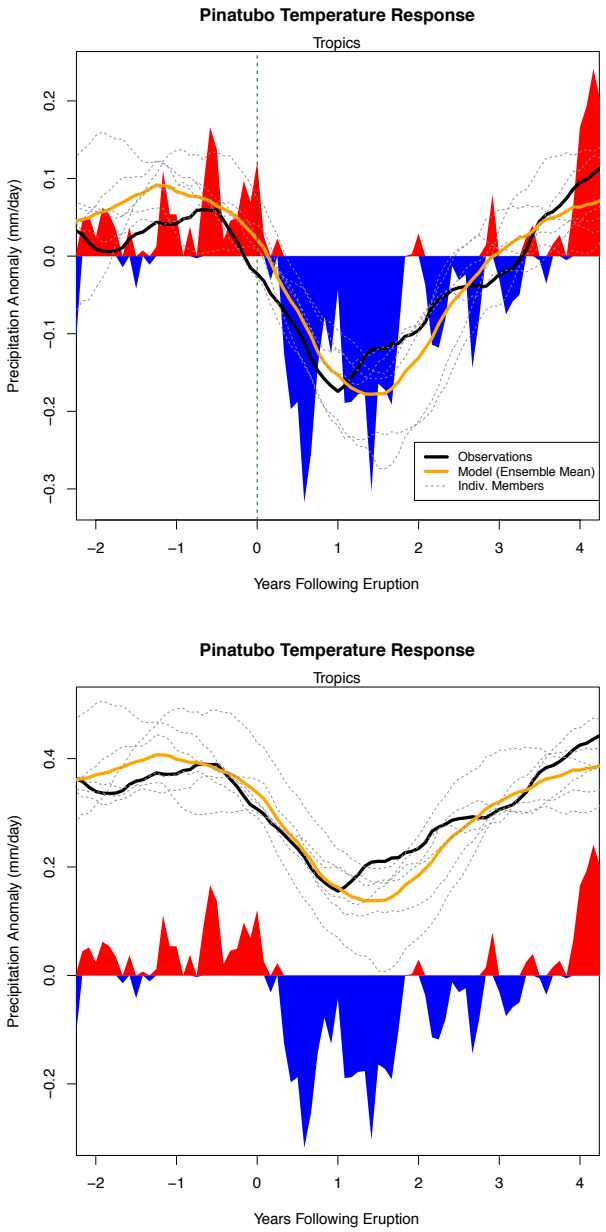

Figure 1: Tropical-mean temperature from years -3 to 5 for the Mt. Pinatubo eruption. Both panels use monthly-mean data (fill color) from GISTEMP Land-Ocean Temperature Index base-lined to give a mean of zero over displayed period. The 18 month running average in observations (solid black line), ModelE2-R ensemble mean (solid orange), and six individual ensemble members (dashed grey) are shown for the same reference period (top) and 1951-198- reference period (bottom).

3) A general comment for all the analyses displayed in the manuscript is the absence of statistical significance evaluation on each figure or plot. I suspect that two eruptions only, is not enough and most of the signal (which is very small) shown on the first figures is within the interval of internal variability. This needs to be evaluated with appropriate statistical methods used to extract the signal from the noise. Tropical South America temperature and precipitation interannual variability is high and the authors should discuss the results respectively to the background noise. No statistical confidence levels are shown. For example on Figure 4 and Figure 8, the colors map is built to be white between +/-0.1 C (mm.day-1). I really doubt that this is a real measure of significance applicable for the whole globe. The authors should address this matter seriously so that they can discuss in a convincing way the signal attributable to the volcanic forcing.

We agree that we could be more rigorous and transparent in our statistical testing. We will stress and revise for clarity in our revised manuscript that in the LM composites (Figure 7,8,10) we did actually test for significance and set any non-significant result to zero. In fact all non-significant areas in these plots were masked white, regardless of the amplitude of their signal. But in addition we also masked all areas with a very low signal (inside the -0.1 to 0.1 range) as white, which may have caused the appearance of only masking areas between -0.1 and 0.1 white. To increase clarity we will consider re-plotting these figures showing all data and use stippling for significance instead.

We will also add statistical significance levels to the historical section of our paper where appropriate (following up on point #1). We do agree that the continental-scale anomalous response is well within the bounds of natural variability, which relates to our concern on the utility of the historical analog for model validation.

4) *Why the authors did run only 6 model members for the L20 eruptions? How were they built? ENSO might be the dominant factor in the simulate response over South America so I would suggest to increase the ensemble size and sample initial states so that in the ensemble mean, the volcanic signal could be extracted from internal unforced variability without any bias toward any ENSO phase. As it is now, we can't really trust the model results as no discussions or diagnostics are shown concerning the appropriateness of the model ensemble to detect the volcanic forcing.*

Six is the number of ensemble members (with volcanic forcing) that are available as continuations of the "past1000" set of experiments with ModelE2-R (http://data.giss.nasa.gov/modelE/ar5/). Until very recently, GISS was the only model that had multiple ensemble members for the last millennium. As with any work dealing with the more complex end in the hierarchy of climate modeling, there are practical limitations in how many simulations have been performed by different modeling groups.

While it is true that averaging over a larger ensemble would improve detection of a forced signal, this does not imply that it would facilitate the ability to

validate the model with observations (which itself is only one realization of an ensemble of possible realities, and largely influenced by unforced variability). Enhancing the ensemble size also would do nothing to address the issue of systematic biases in the historical forcing. It may increase the probability that a given realization of the ensemble better mimics the initial state of the atmosphere prior to observed historical eruptions, but a detailed exploration of this specific aspect is beyond the scope and intended purpose of this paper. Instead we aimed at addressing this issue by averaging over a larger sample of events (with improved signal-to-noise ratio) by focusing on the LM composite. We believe this to be the better approach than increasing the ensemble size for the L20 composite for which the average forcing is much smaller.

5) *The model results displayed on both Figure 4 and 5 show absolutely no agreement with observations (temperatures and precipitations) while the estimated robust signal attributable to any of these volcanic eruptions is not shown (signal to noise ratio). Same remark as above using a color map built to have white shade at a fixed contour is not a measure of significance. The authors can't state based on these figures that the model is able to reproduce the temperature or the precipitation responses, as the spatial patterns and amplitude are not consistent with observations. So far figure 4 and 5 suggest that the model is not able to reproduce any post-eruption signal and is not appropriate to evaluating the impact of volcanic forcing on the South American Monsoon*

The argument made that "the model results displayed on both Figure 4 and 5 show absolutely no agreement with observations," relates back to our response in #1 above, on whether one can realistically expect the model to agree with observations. Please also note that the displayed model results represent an average over several ensemble members, while observations are by definition just one realization, so we cannot compare these usefully. In general this may not have been the best choice of presentation, and we will re-visit these figures and how we can best convey the relevant information.

Below, we show the six individual ensemble members for the post El-Chichón DJF temperature anomaly (relative to the previous five years, as was done in the discussion paper). Although we showed the ensemble mean in the paper (Figure 4, in the fourth column and second row) there is still considerable spread among the ensemble members. This limits our ability to confidently validate the model by comparing the post-eruption model and instrumental response.

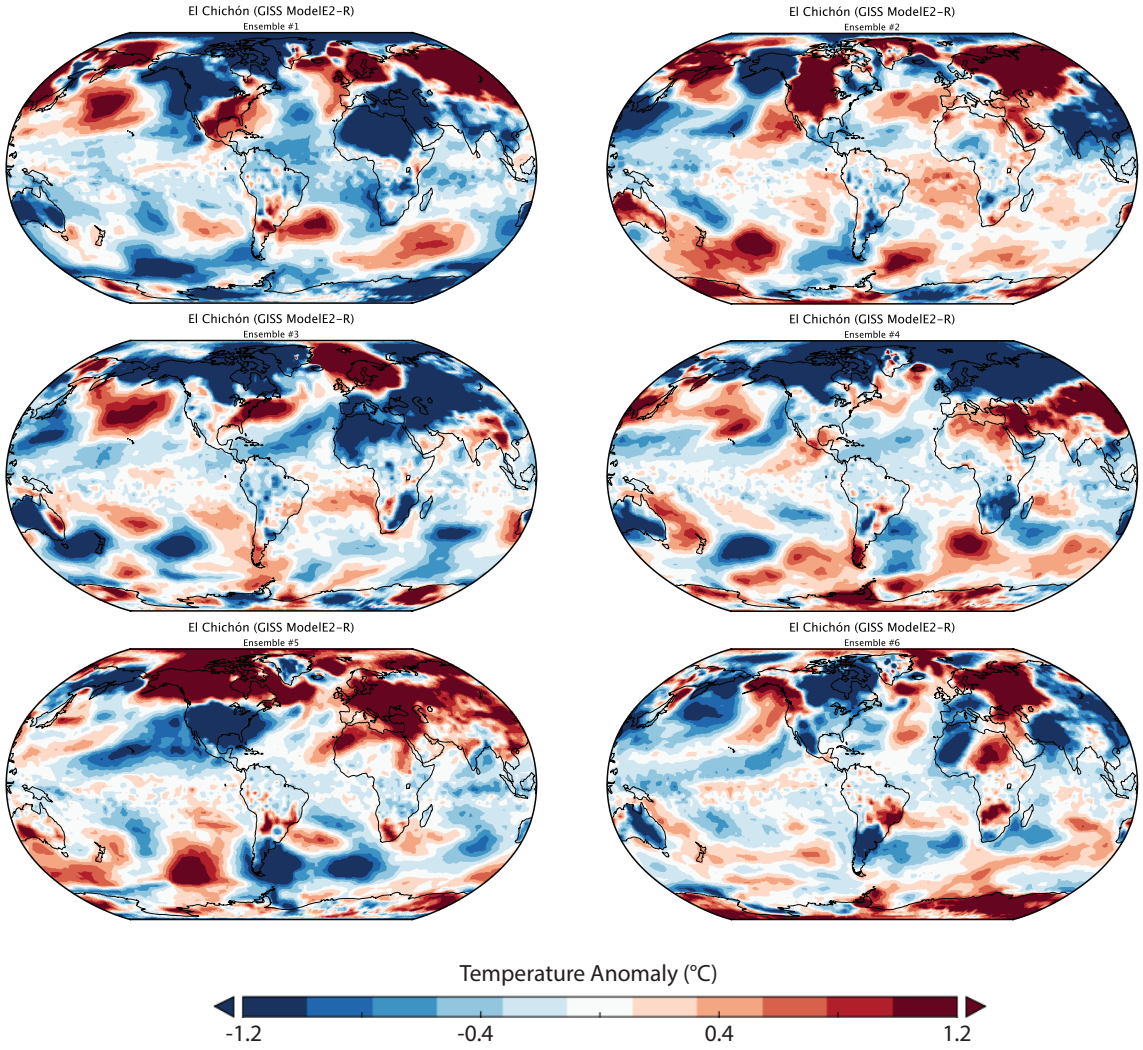

Figure 2: Temperature anomaly (°C) following El-Chichón for the six NASA GISS ModelE2-R ensemble members. Results for DJF.

6*) Last paragraph of page 3387: The authors should clarify what is the mis-scaling of the Gao forcing and why for the model composites covering the L20 eruptions, it is not an issue.*

The code to implement the Gao-derived aerosol loading (given in Tg) and convert to Aerosol Optical Depth and effective Radius (that is prescribed in ModelE2-R) did not include a constant, and thus was mis-scaled by a factor of ~0.51 and results in too large a radiative forcing. After 1850, the volcanic forcing is based on the Sato index and is correctly scaled. Thus, we omit the Gao ensemble members for the pre-industrial component of our study.

7) *First two paragraph page 3388: The authors state that the volcanic forcing should dominate the response in the LM composite. This is a very strong statement as*

different solar forcing scenarios have been used not to mention the two different land-use forcing scenarios (especially over South America) employed in the different LM member. The authors can't make such statement without providing detection-attribution analyses and other diagnostics over South America showing that the various land-use and solar irradiance forcings didn't have any impact on the post-eruption mean response (temperature and precipitations) and ensemble spread for each selected LM eruptions. Addressing this issue is not trivial and it shouldn't be overlooked. As it is, the LM composites can't be used to address specifically the volcanic response as other forcings are at play and may very well contribute significantly to the simulate response.

R2 is concerned with the use of a mixed-forcings ensemble. We first note that if a volcanic-only last millennium GISS ensemble were available in the CMIP5/PMIP3 generation, we would have used it.

We believe R2's point would certainly be important if our analysis focused on the decadal-to-centennial timescale, where volcanic forcing is competing with many other forcings during the Last Millennium (although note that Atwood et al. (2015) document that volcanic forcing dominates even at the centennial timescale, at least at the global scale).

However, we stand by the argument that since the analysis focuses on changes within just a couple of years following pinpointed eruptions (relative to surrounding years), the presence of other "slow" and much smaller-amplitude forcings simply do not matter. For example, suppose we constructed a Pinatubo composite by averaging over 48 realizations of Pinatubo, and focused on the immediate 1-3 year response in the historical simulations – no one would reasonably argue that the CO_2 increase or solar cycle coincident with that change to be an important confounding influence. The same holds here, and most of the events averaged in our paper are even larger than Pinatubo.

To highlight this aspect more clearly, we show results from several ModelE2-R experiments that had differences in the imposed solar reconstruction, but without volcanic forcing included (rows 2-3). We create composites by averaging over the same 16 eruptions (i.e., the same dates as used to create the volcanic composite). This is done separately for both the DJF and JJA season. The mean of these 2 different solar forcing composites, featuring 32 events, are shown in row 4. Results from the control simulation with no forcing are also included (row 5). The top row of this plot includes the volcanic forcing, as in the manuscript (results not masked for significance) based on 16 eruptions and three ensemble members (48 events).

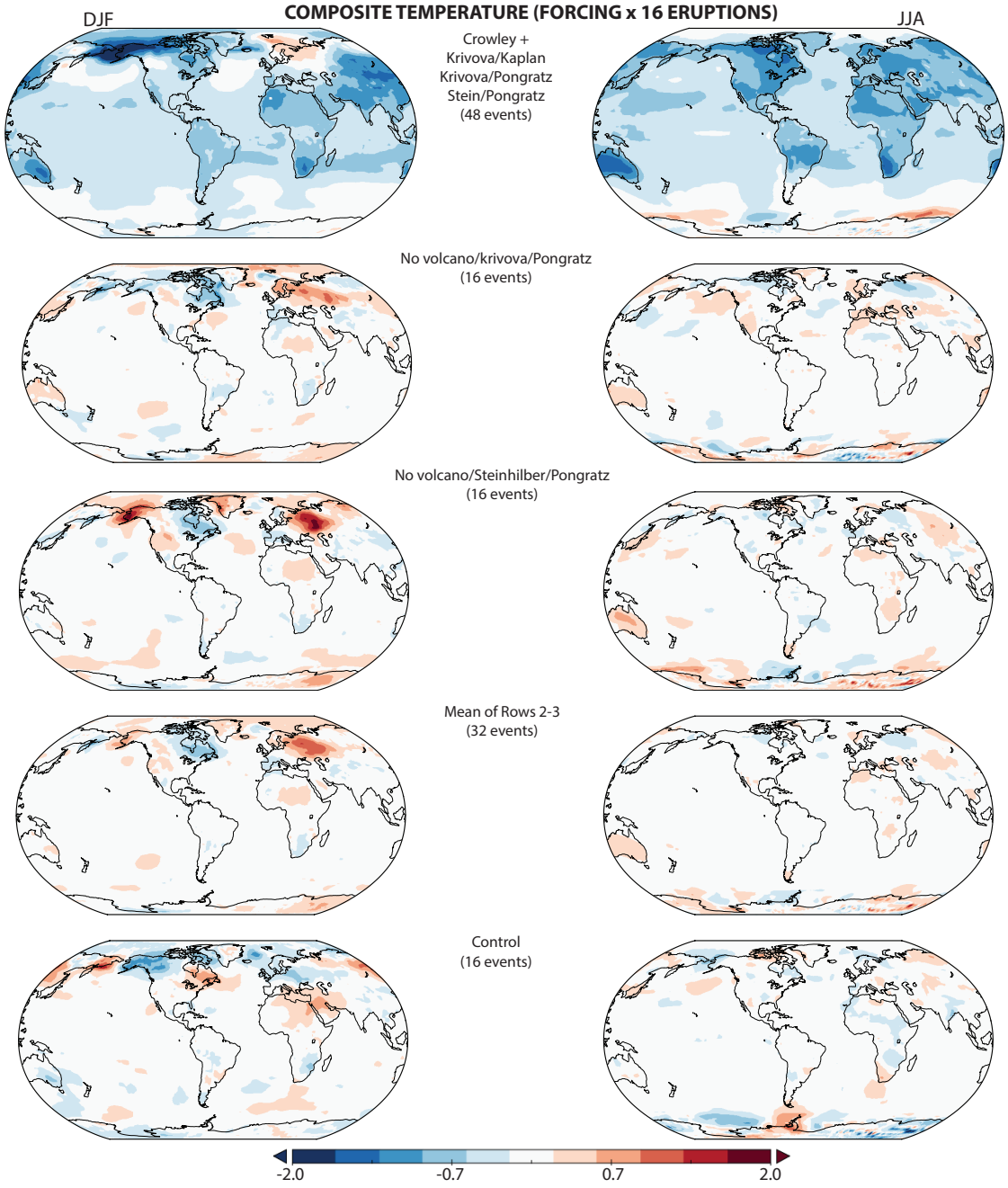

Figure 3: Composite temperature anomaly (°C) for 16 events (multiplied by number of ensembles) using the methodology in discussion paper. Row 1 with volcanic forcing (48 events) , row 2-3 with no volcanic forcing but differences in solar forcing (16 events each), row 4 is the ensemble mean of rows 2-3 (32 events). Row 5 is the 16 events from the control simulation. Results for DJF (left column) and JJA (right column).

The volcanic response stands out clearly in the ensemble, both over South America and on a global scale, even after averaging over 48 realizations of internal variability. The solar signal is just too small. There is no evidence of a coherent forced response to solar forcing, when compositing over such short random time periods (see rows 2-4). The variability in the wintertime high-latitudes still stands

out, but there is no solar signal on this time scale in the tropics. Repeating this analysis for land use forcing or looking at the precipitation response instead of temperature does not change these conclusions. Hence land-use and solar irradiance forcing do not significantly affect the post-eruption mean response of temperature or precipitation in the tropics.

We will consider adding a supplemental figure to stratify our composite by times during high or low solar forcing. We expect them to look the same.

8. Section 3.2.1 first paragraph: Is -/+0.1 C statistically significant as shown on Figure 7 or is it again a color map choice? Does not look right owing to the high SST variability over land and ocean in these regions. I'd ask the author to verify this.

We have addressed this point regarding statistical significance in our response under item 3.

9) *It is difficult to believe based on the results displayed that in in the case of volcanic forcing it appears that the amplitude of the temperature-response to volcanic eruptions over tropical South America is much larger than the rather weak and spatially incoherent precipitation signal. The forcing used (Gao and Crowley) for the LM simulations are well known now to have been largely overestimated as the temperature response in CMIP5 LM simulations while the good performance of the model used in this study against Pinatubo eruption (for which plenty observation are available) for the forcing and response has not been shown. Same for the South American mean climatology.*

If we understand R2 correctly, he/she is surprised at the relative coherence of the temperature response when compared to precipitation. We have argued above that we have sampled more than enough volcanic events in our composite to isolate the signal, and the anomalous temperature field is almost always the "simplest" climate response to any global forcing. The anticipated response to elevated CO_2, for example, is relatively smooth and well characterized by a single number, with the usual caveats of polar/land amplification and minima in sub-polar regions. Precipitation is much more heterogeneous.

We are sympathetic to R2's point that the Gao/Crowley forcing datasets are virtually certain to be "wrong." However, those are the datasets that are available and which have been used in the forcing of CMIP5/PMIP3 generation last millennium simulations, and even in post-CMIP5 efforts (e.g., the CESM Last Millennium ensemble). It is not obvious that the temperature response in CMIP5 models is overly sensitive; they just may be seeing too large a forcing (e.g., because the aerosol size distribution is incorrect). The displayed temperature and precipitation (and isotope) patterns are arising from the same forcing.

Unfortunately, we are shackled to the current state of the science on paleo-volcanic forcing. Newer reconstructions such those provided by Sigl et al. (2015) have not yet been implemented in fully coupled GCMs, but even here any estimated forcing will just be a simple historical scaling and likely not robust (though Arfeuille et al., (2014) do a better calculation back to 1600 C.E.). We further would like to

note that errors in the timing of the eruptions, which exist in the Gao/Crowley forcing datasets and pointed out by Sigl et al. (2015), are not relevant in our context because we know exactly when the model is forced and build our composites accordingly. Errors in the amplitude or spatial structure will potentially matter, and we have used the dataset that actually has "smaller" events. We do show a scaling against AOD for key variables in Figure 9, to lend insight into how the typical response may change if we scale the mean forcing differently.

We do agree that these mismatches have inspired proxy testing, improvements in volcanic forcing, and development of volcanic implementation in climate models. It is an exciting – and new – area of research. But, not the focus of the paper here, which looks specifically at the last millennium simulations.

Aside from this, a fully consistent emissions-based estimate of the aerosol loading, growth of particles, interaction with chemistry and clouds, release of other substances (halogens, water vapor, etc.) is at the frontier of this field and not well implemented by any group. So the model response of course needs to be viewed as a slave to the imposed forcing.

Minor comments:

- Page 3377, line 27-28 and page 3378 line 1-2. The authors state Sulfate aerosols from the Mt. Pinatubo eruption had an effective radius of up to 0.5–0.8, comparable in size to a visible wavelength and strongly scattering to incoming solar radiation. Unless the particles can reach sizes larger than 1–2, this scattering more than offsets the small increase in infrared opacity from the aerosols, and results in a cooling of Earth's surface (Turco et al., 1982; Lacis et al., 1992).

- I'd replace "of up to 0.5–0.8" by "ranging between 0.2 and 0.8 with unimodal size distribution mean radius of 0.5" As for the statement "larger than 1–2", according to theoretical calculation (Lacis et al 1992) the LW forcing would dominate for particles larger than 2.2.

We will improve our description, thank you.

-Page 3380, last paragraph: The continent spans a vast meridional extent (from 10N to 55S), contains the world's largest rainforest (the Amazon), in addition to a rather Mars-like desert (Atacama) that competes only with the dry valleys of Antarctica for the driest location on Earth. What is a "Mars-like" desert? Not really scientifically meaningful. I'd rather give the amount of precipitation per year. As for the comparison to Antarctica for the driest location on Earth, is it proven? If yes the reference is missing.

We agree that this paragraph was not well written. We have removed the anecdotal reference to Mars and Antarctica and now use actual precipitation amounts to discuss the spatial precipitation variability over the South American continent.

- *Methodology section: Line 14: The authors need to clearly define how the ensembles were built, in terms of forcings and initial conditions. How many members and how they differ exactly from each other? A table summarizing this is needed.*

We will include another table in the manuscript to discuss the forcings (and other model details) for the three different ensemble members used in the LM composites, and six for L20.

- *Page 3385, line 19: GPCCv6 is better and is actually what you show in the supplementary material. Please clarify.*

We did use a merged satellite-land precipitation product (GPCP v2.1). We will include the most recent version (v2.2) that became available after the analysis was done, though there are no differences over South America following the eruption events targeted in our paper. GPCC v6 is a land-gauge product only. It is true that we did include in Figure S1 a representative set of examples for the number of observations that are available around each eruption point, which was a readily accessible diagnostic in the GPCC netCDF files online (but not GPCP). GPCP v2.1 uses GPCC precipitation gauge analysis as a key input. We will clarify this in the Figure S1 caption.

References cited:

Arfeuille, F., Weisenstein, D., Mack, H., Rozanov, E., Peter, T., and Brönnimann, S., 2014: Volcanic forcing for climate modeling: a new microphysics-based data set covering years 1600–present, Clim. Past, 10, 359–375, doi: 10.5194/cp-10-359-2014.

Atwood, et al. 2015: Quantifying climate forcings and feedbacks over the last millennium in CMIP5/PMIP3 models. *J. Climate* (in press).

Hansen, J., M. Sato, R. Ruedy, A. Lacis, K. Asamoah, S. Borenstein, E. Brown, B. Cairns, G. Caliri, M. Campbell, B. Curran, S. de Castro, L. Druyan, M. Fox, C. Johnson, J. Lerner, M.P. McCormick, R.L. Miller, P. Minnis, A. Morrison, L. Pandolfo, I. Ramberran, F. Zaucker, M. Robinson, P. Russell, K. Shah, P. Stone, I. Tegen, L. Thomason, J. Wilder, and H. Wilson, 1996: A Pinatubo climate modeling investigation. In *The Mount Pinatubo Eruption: Effects on the Atmosphere and Climate*, NATO ASI Series Vol. I 42. G. Fiocco, D. Fua, and G. Visconti, Eds. Springer-Verlag, 233-272.

Lacis, A., J. Hansen, and M. Sato, 1992: Climate forcing by stratospheric aerosols. *Geophys. Res. Lett.*, **19**, 1607-1610, doi:10.1029/92GL01620.

Marvel, K., G.A. Schmidt, D. Shindell, C. Bonfils, A.N. LeGrande, L. Nazarenko, and K. Tsigaridis, 2015: Do responses to different anthropogenic forcings add linearly in

climate models? *Environ. Res. Lett.*, **10**, no. 10, 104010, doi:10.1088/1748-9326/10/10/104010.

Santer, B.D., et al., 2014: Volcanic contribution to decadal changes in tropospheric temperature. *Nature Geosci.*, **7**, no. 3, 185-189, doi:10.1038/ngeo2098.

Schmidt, G.A., D.T. Shindell, and K. Tsigaridis, 2014: Reconciling warming trends. *Nature Geosci.*, **7**, no. 3, 158-160, doi:10.1038/ngeo2105.

Sigl, M., et al. 2015: Timing and climate forcing of volcanic eruptions for the past 2,500 years, *Nature*, **523**, 543-549, doi:10.1038/nature14565.