Journal of Geophysical Research: Atmospheres

Supporting Information for

Multi-model surface temperature responses to removal of U.S. sulfur dioxide emissions

A. J. Conley[1], D. M. Westervelt[4,6], J.-F. Lamarque[1], A. M. Fiore[4,5], D. Shindell[3], G. Correa[4], G. Faluvegi[6,7], L.W. Horowitz[2]

[1]ACOM Laboratory, National Center for Atmospheric Research, Boulder, Colorado, USA

[2]National Oceanic and Atmospheric Administration, Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA

[3]Nicholas School of the Environment, Duke University. Durham, NC, USA

[4]Lamont–Doherty Earth Observatory, Columbia University, Palisades, NY USA

[5]Department of Earth and Environmental Sciences, Columbia University, Palisades, NY, USA

[6]NASA Goddard Institute for Space Studies, Columbia University, New York, NY USA

[7]Center for Climate Systems Research, Columbia University, New York, NY, USA

Introduction

We summarize the statistical methods used to analyze the models. In particular we describe the regions of analysis in Section S1, how we compensated for underlying drift in Section S2, how we estimated standard errors of the mean, including the effects of time-series autocorrelation in Section S3, the consequences of using more advanced statistical methods in section S4, the robustness of results in section S5, and the applicability of empirical orthogonal functions (EOFs) to reduction of confidence intervals in section S6. In Section S7 we explore the contributions of individual species to total aerosol optical depth. In Section S8 we list various potential contributions to the increase sulfate burden seen in the GFDL simulations. Lastly, we plot surface temperature and dust burden responses in both the forcing and coupled runs in order to diagnose feedbacks in the CESM model over Africa.

S1 Region Definitions

We wish to analyze not only global average responses, but also responses in specific regions. Figure S1 shows the regions over which we compute statistical measures and

Table S1 delineates the latitudinal and longitudinal boundaries of these regions. Constructing averages over these regions is meant to decrease the variability, and hence provide additional power in the statistical tests. Looking at the Southern Australia regional average of temperature change in the CESM model, we see from Figure 11, that the change is marginally significant at the 68% level, however, in Figure 10, most of the land surface in this region is significant at the 95% level. A similar loss of statistical power can be seen in the average over Northern Europe for the GFDL model, as it includes some significant cooling over ocean surfaces. Sometimes the regional average includes regions of opposite sign or little change, diluting the significance of temperature change for a region. Furthermore, there is enough horizontal correlation of temperature changes (for annual averages) that regional averages may not provide much reduction in variance. This usefulness of regional averages for statistics of annual average changes should be reevaluated; however, in this study we maintain these definitions to be consistent with previous works.

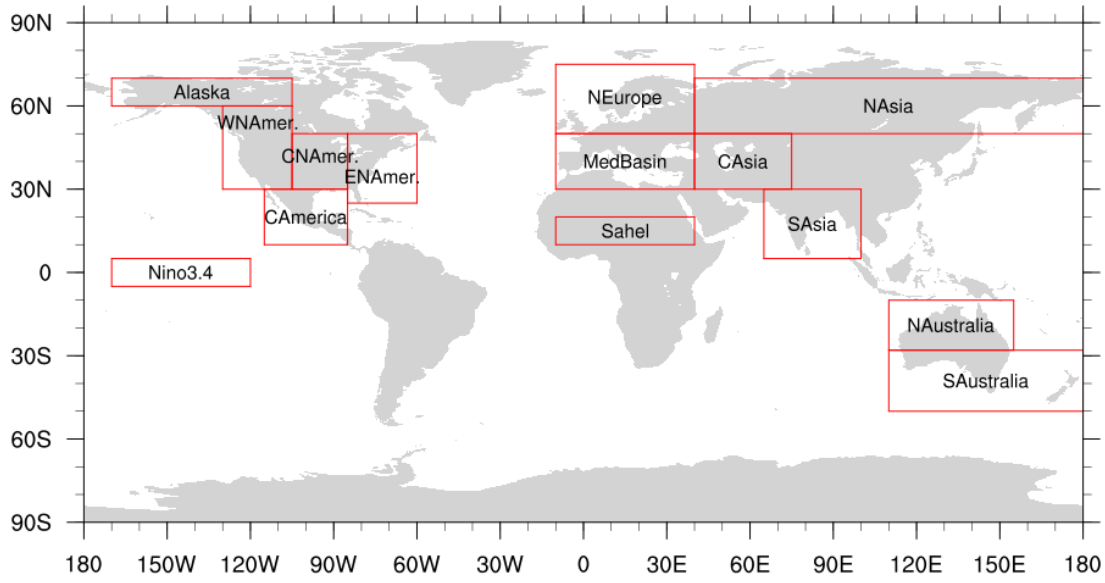

Figure S1. Regional boundaries. In addition to studying global-average responses, we compute average responses in each of the regions above. One additional region is the area north of 60°N. Boundaries are specified in the table below.

South Australia	110E-180E / 50S-28S
North Australia	110E-155W / 28S–10S
Central America	115W-85W / 10N-30N
Western North America	130W-105W / 30N-60N
Central North America	105W-85W / 30N-50N
Eastern North America	85W-60W / 25N-50N
Alaska	170W-105W / 60N-70N
Mediterranean Basin	10W-40E / 30N-50N
Northern Europe	10W-40E / 50N-75N

South Asia	65E-100E / 5N-30N
Central Asia	40E-75E / 30N-50N
North Asia	40E-180E / 50N-70N
Sahel	10W-40E / 10N-20N
N60	60N-90N
Niño 3.4	170W-120W / 5S-5N

Table S1: Regional Boundaries

S2 Temperature Drift

The global average surface temperatures drift upward in the control simulation from all three models (Figure S3). Comparing with Table 3, the magnitude of the drift is comparable to (or larger than) both the size of the temperature responses we wish to detect and the variance of the samples.

Performing a standard t-test comparing the test and control time series without accounting for this drift confounds the drift with the variance in the underlying control and test time series.

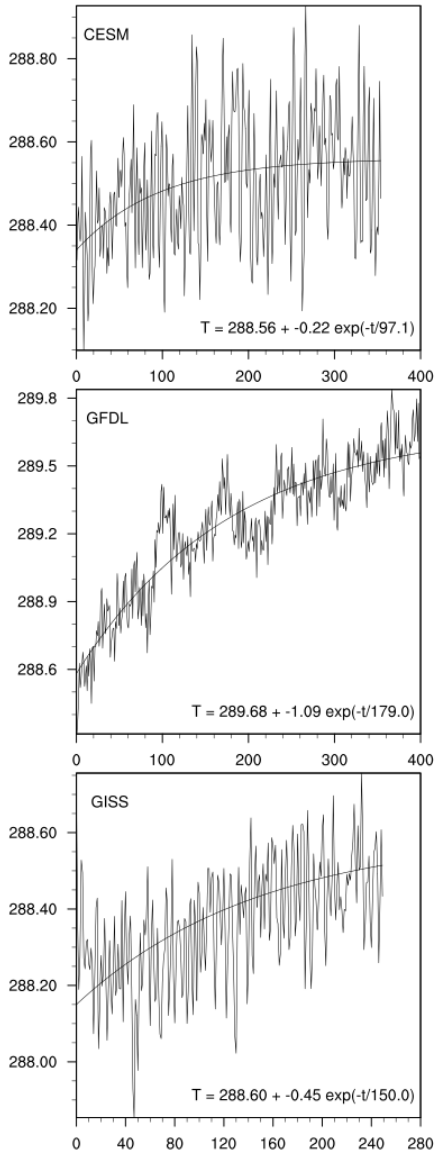

Figure S2: Global Average Temperature drift in control simulations. Curve fits are illustrative only.

Fortunately, the test (perturbation) simulations are constructed by branching at some point in time from the control simulation. By computing statistics on the co-temporal differences between the control and test cases, the statistical tests have much more power. (This is a matched-pair, or paired z-test.) We test to see if the temperature differences can be distinguished from zero at the 1-standard error or 2-standard errors (i.e., 68% or 95% confidence level).

To be explicit, given two co-temporal annual-average time sequences of temperatures, (T_i^c, T_i^t), where the time is indexed by i, c indicates the control time series, and t indicates the test series, we compute the mean temperature difference between the sequences and the standard error of the mean difference as follows.

Defining the annual-mean difference at each co-temporal sample between the control and test case,

$$\Delta T_i = T_i^t - T_i^c$$

we examine the statistics of the mean change in temperature.

$$\overline{\Delta T} = \frac{1}{n}\sum \Delta T_i$$

where the sum is over co-temporal years excluding the first 5 years following the branch point and n is the number of co-temporal years in the sum.

S3 Standard deviations, autocorrelations, standard errors, and z-scores

The sample standard deviation, σ, is defined as

$$\sigma^2 = \frac{1}{n-1}\sum (\Delta T_i - \overline{\Delta T})^2$$

and the lag-k autocorrelation as

$$r_k = \frac{1}{\sigma^2 (n-k)^2}\sum (\Delta T_i - \overline{\Delta T})(\Delta T_{i+k} - \overline{\Delta T})$$

As can be seen in Figure S3, the CESM model shows significant autocorrelation for offsets of 1, 2, 3, 5, and 6 years, indicating that consecutive years are not independent samples. As a result, computation of the standard error, without accounting for the autocorrelation, overestimates the number of independent samples as discussed in *Jones* [1975]. There are a couple of ways to think of this. One is that if neighboring (in time) samples are correlated, there are fewer samples. *Zwiers* [1994] constructs an effective number, n_{eff}, of samples based on the lag-1 autocorrelation:

$$n_{eff} = n\frac{1 - r_1}{1 + r_1}$$

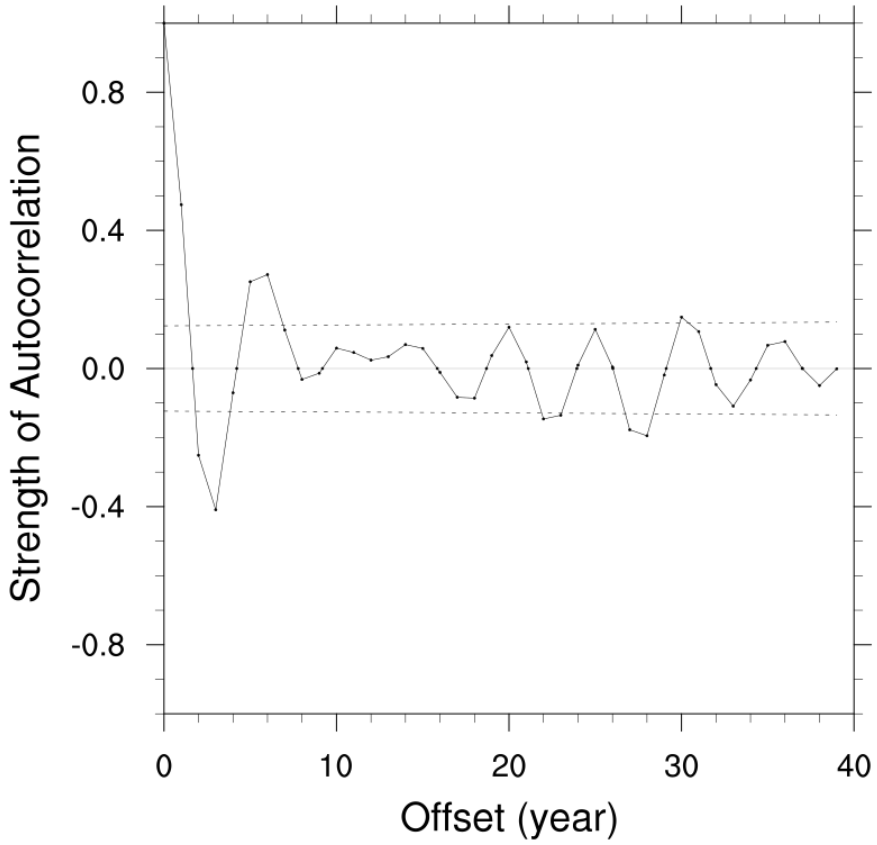

Figure S3: Autocorrelation of global average surface temperature in CESM model. The dashed boundaries are 1-sigma boundaries corresponding to white noise.

Unfortunately, observational studies have identified multi-decadal modes of variability, and these multidecadal modes are not strictly periodic, so computation of autocorrelation using the methods above (which depend upon strict periodicity in order to identify the autocorrelation) may still underestimate the total autocorrelation and therefore overestimate the number of independent samples. Using the autocorrelation function to correct confidence estimates has a limitation for non-periodic internal variability (e.g., PDO or ENSO) in climate models. (E.g., have you sampled enough of the slow oscillations to accurately compute their influence on the mean value?) We see no obvious way to eliminate this limitation to constructing a more valid and complete statistical model. Empirical orthogonal functions (EOFs) are one attempt to identify (and remove) modes of variability in order to construct a more complete and powerful statistical model. We will review the possible use of EOFs for this work below.

For the purposes of this paper, we define (following *Zwiers and von Storch* [1995]) the standard error of the mean as

$$s.e. = \sqrt{\frac{\sigma^2}{n}\frac{1+r_1}{1-r_1}}$$

As shown in Figure S4, the standard error gives an estimate of the confidence in the sample mean. As you can see, using a simulation length of 10 could often lead to an incorrect expectation that the control was warmer than the perturbed run. However, for a sequence of more than 50 years, such an incorrect conclusion would be less likely. As the number of samples increases (say with 200 samples), the confidence interval is small enough to differentiate the means with confidence. The statement that "the result is significant" is a statement that one has enough samples that the confidence intervals do not overlap and thus the means are probably different. Thus, significance is a statement about both the sample size and the distance between the averages. If there is a difference between the test and control averages, enough samples will eventually resolve that

difference; however, the required number of samples can be quite large since the size of the confidence interval shrinks proportionally to $1/\sqrt{n}$.

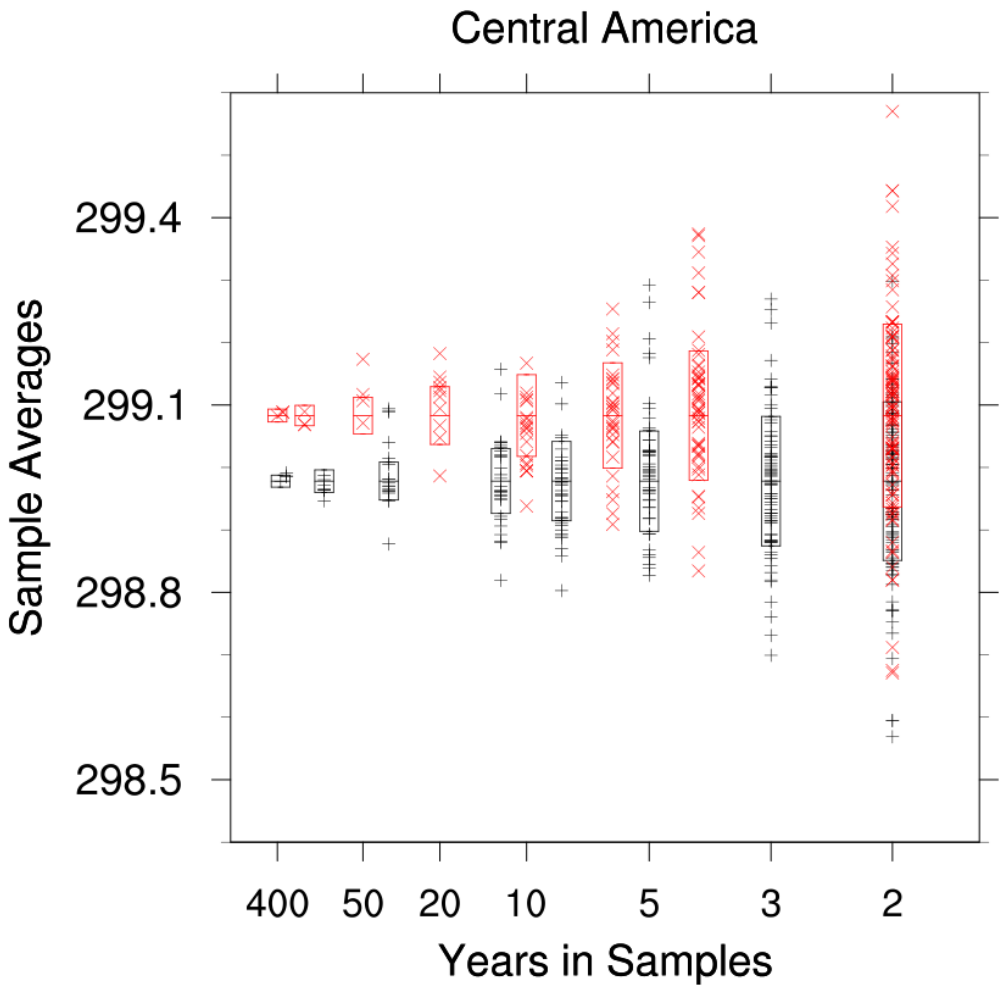

Figure S4: Standard error is the range over which the estimated parameter (in this case the mean) is likely be reside. Abscissa is labeled with the number of samples included in the mean. The abscissa is $1/\sqrt{n}$, since standard error is $\sigma/\sqrt{n}$. Ordinate is the sample average. Black markers are means computed using a sample of (n) consecutive values from the underlying control time sequence. Red markers are the same, but using the perturbed simulation. Boxes indicate the confidence intervals corresponding to 1 standard error for that sample size. The box could be centered on any one of the sample means shown.

As *Crawford and Hale* [1998] demonstrate, the z-score

$$z = \overline{\Delta T}/s.e.$$

is, in practice, sufficiently close to the t-score for more than 50 independent samples. In these simulations, n is larger than 180 and the effective number of samples, as seen in Figure S5, is almost always larger than 100 for these samples, implying that the z-score should be very close to the t-score. Significance at the 68% level is defined as when the z-score is larger than 1,

$$\mathrm{abs}(z) > 1$$

and significance at the 95% level is defined when the z-score is larger than 2.

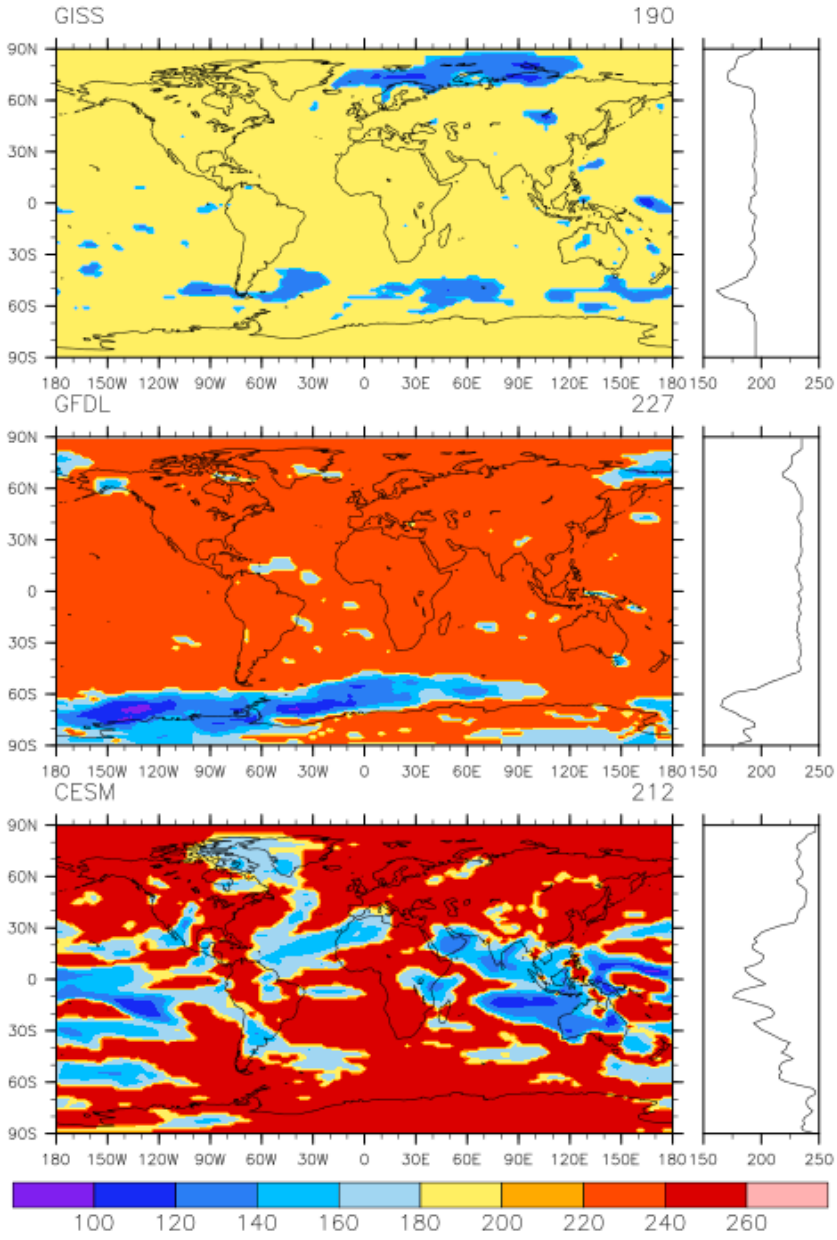

Figure S5: Effective number of samples in each model at each grid point. Zonal means are to the right and global averages are in the upper right. However, the zonal means are not the number of effective samples for the estimates of zonal mean responses. Similarly, the global mean of the point-wise temperatures is not the effective number of samples of the global time series.

S4 Alternative statistical models

We also evaluated the use of more complex statistical models. While an ARIMA (3,0,1) process (three-step autoregressive, one-step moving average) with autoregression weights of (0.3,0.1,0.3) minimizes the *Akaike* [1974] information criterion for the GFDL time series in Figure S1, model estimates of mean and standard error from the method of *Zwiers and von Storch* [1998] and the ARIMA (3,0,1) model are very similar (almost indistinguishable) on global plots and bar and whisker plots. In addition, differences between models dwarf the improvement to estimates provided by the more sophisticated statistical models in our tests of global and regional surface temperature. Furthermore, the optimal statistical model depends on which chemistry-climate model is being analyzed, the variable being analyzed, and even the region being analyzed, raising the question of how to compare statistical results between models when the statistical models are distinct.

S5 Robustness of results

As discussed by *Tebaldi and Knutti* [2007] combining results from multiple models increases skill of predictions. However, narrowing the confidence range also requires comparing results from multiple models where the models have reasonable but distinct physical parameterization [*Knutti et al*., 2017]. It is unclear to what extent these models have sufficiently "distinct physical parameterizations", however the range of sensitivities (SO_4 per SO_2, AOD per SO_4, forcing per AOD, and temperature change per forcing) indicates significant diversity of parameterization. The inclusion of error measures on each of these intermediate sensitivities allows us to know that these models are distinct with some level of confidence. As a result, the multi-model averages of surface temperature change are likely to have more confidence than any one of these models alone; however, the inclusion of only three models makes the construction of an error estimate for the multi-model average difficult. Lacking such an estimate, we simply compute a standard deviation of the multi-model temperature responses as a measure of confidence of the multi-model mean temperature responses seen in Figure 11.

S6 Modes of variability

Each model shows different strengths of autocorrelation at each individual grid point (Figure S5). And the patterns of strong autocorrelation between models are quite distinct. However, those plots show nothing about the horizontal correlation between the time series at different grid points. Performing a singular value decomposition on the time series of all the grid points leads to the empirical orthogonal functions (EOFs) which describe the time series of spatial patterns that describe (orthogonal) spatial patterns and (orthogonal) time series of all the variability in the time series of all grid points [*Preisendorfer and Mobley* 1988, *Bretherton et al*. 1992].

Appealing to the idea that each EOF is a physical mode of variability, one can remove (project out) a number of modes accounting for a significant amount of the variance in the system. We constructed EOFs from the control simulation and projected the first 5 modes from both the control and the perturbed time series of the surface temperatures. We computed means and standard deviations on the residual time series. As can be seen comparing Figures S6 and S7, by projecting out the first 5 EOFs, there is some decrease in the standard deviation of the residual time series over the tropical pacific, perhaps

corresponding to ENSO variability that is partially captured in these EOFs. It is unclear how to incorporate this type of analysis into a refinement of the estimates of the means or a tightening of the confidence intervals. In summary, projecting (removing) a mode of variability from the time series does not improve the confidence in the estimate of the mean.

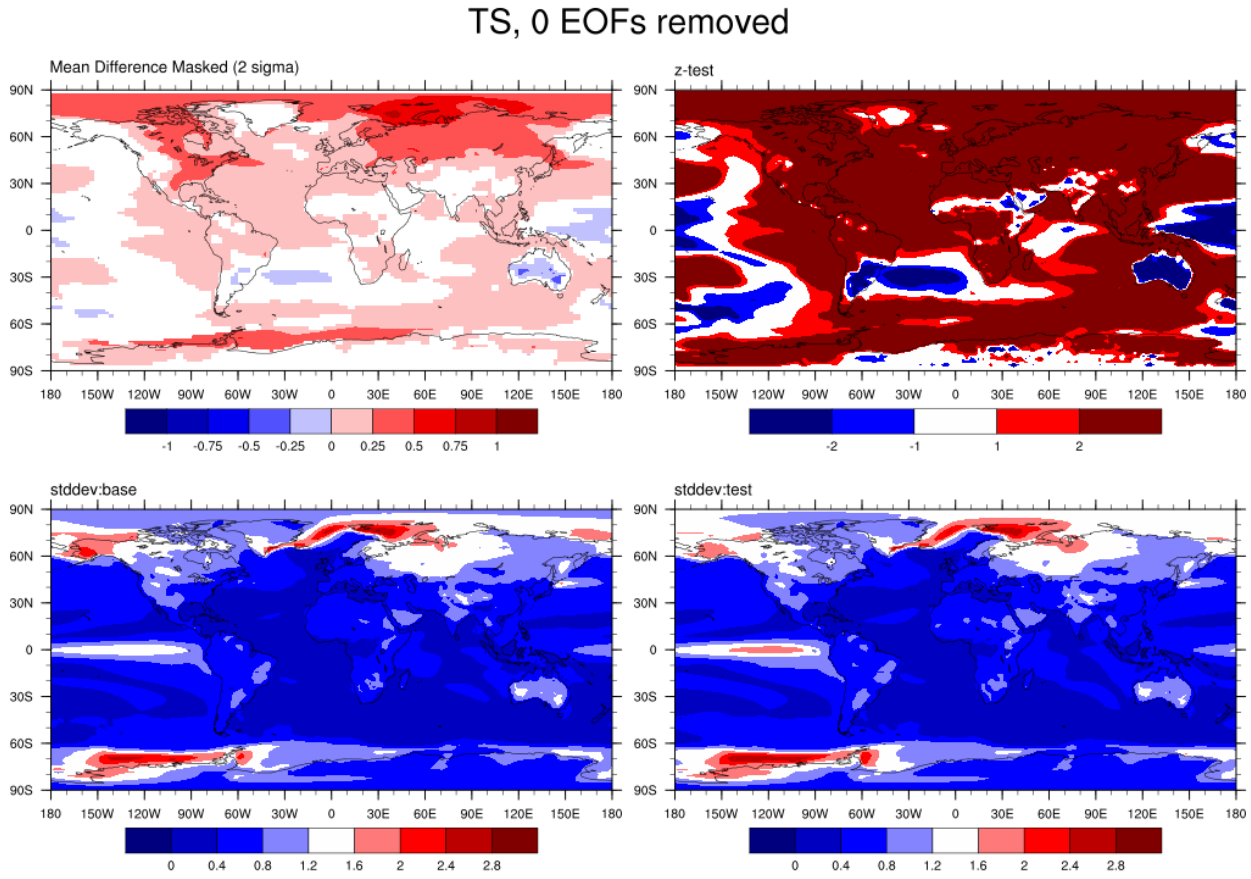

Figure S6: Plot of average grid point temperature difference for the CESM model in the upper left plot, masked to show only regions of significance at the 95% level. The z score is plotted in the upper right. The standard deviation for the control run is plotted in the lower left and the standard deviation for the perturbed (no US SO_2 emissions) in the lower right. In this case no EOFs have been removed from the time series before the statistical analysis; compare with Figure S7.

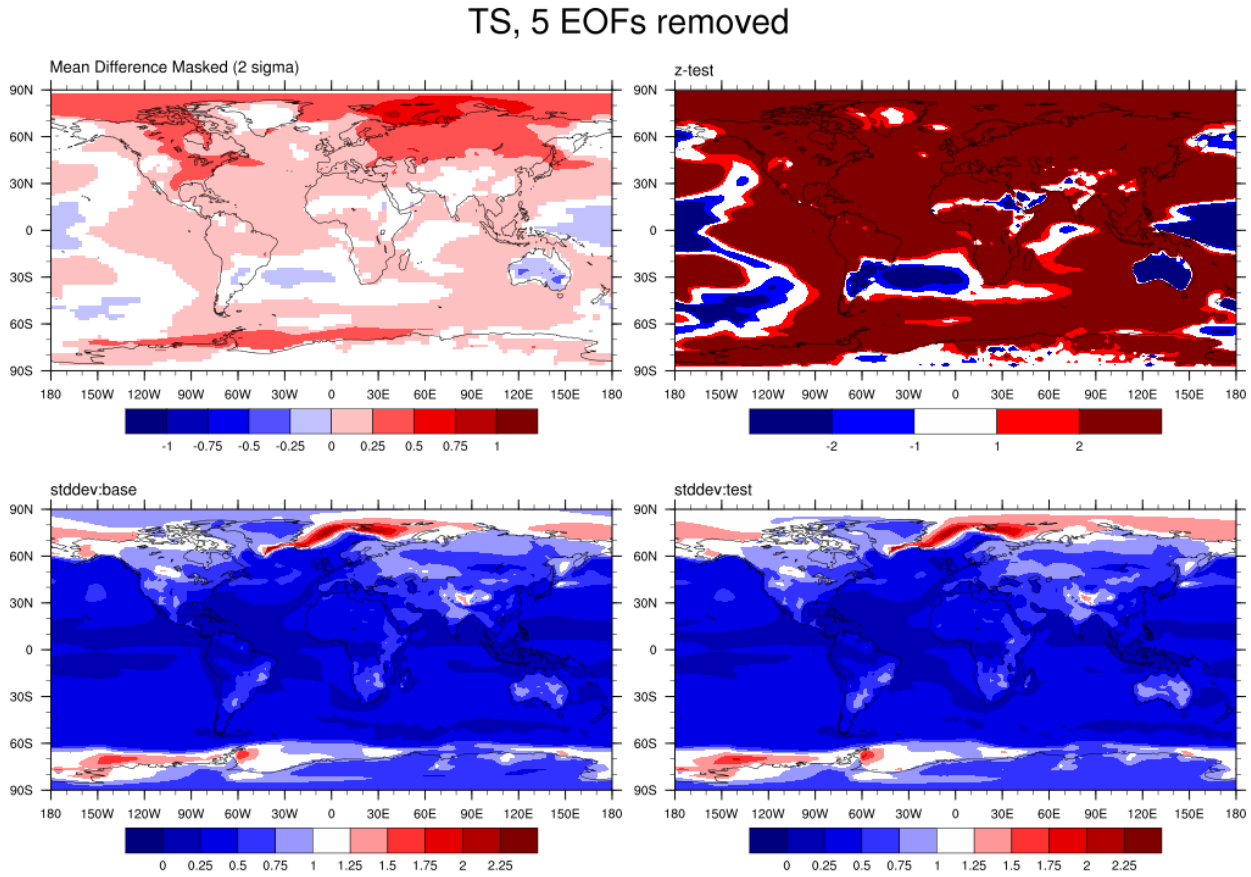

Figure S7: Plot as in Figure S6. In this case the first 5 EOFs have been removed from the time series before the statistical analysis. Note the change in scale for the plots of standard deviation.

S7 Relation of total optical depth to column masses of aerosol species

In order to see how the change in each of the aerosol species affects the total aerosol optical depth we have plotted the time-average change in total optical depth, as well as changes in masses of species contributing to aerosol optical depth. Plotting the change in optical depth against the change in column mass for each lat-lon grid point for the CESM model (Figure S8), we see that contributions to the total optical depth are dominated by both sulfate and dust. There are also contributions at some locations from sea salt and primary organic material (POM). Changes in relative humidity and internal mixing are not considered by this regression model.

We regressed the change in total optical depth against change in column masses,

$$dAOD(\theta,\varphi) = \alpha_S dSulfate(\theta,\varphi) + \alpha_{Bl} dBlackCarbon(\theta,\varphi) + \alpha_{Dust} dDust(\theta,\varphi) + \alpha_{POM} dPOM(\theta,\varphi) + \alpha_{SS} dSeaSalt(\theta,\varphi) + \alpha_{SOA} dSOA(\theta,\varphi) + \alpha_{Residual}$$

where $dAOD(\theta,\varphi)$ is the average change in the total optical depth at each latitude and longitude, $dBlackCarbon(\theta,\varphi)$ is the change in column mass of black carbon as well as

terms for the corresponding dust, primary organic matter, sea salt, and secondary organic aerosols. The residual is not a function of latitude or longitude. As can be seen in Figure S8 below, for the CESM model, the largest contributors to aerosol optical depth are sulfate and dust. After computing the regression coefficients, α_i, (using least squares), we can compute the resulting contributions from each species (e.g., the dust contribution as $\alpha_{Dust} dDust(\theta, \varphi)$) as seen in Figure S9. Similarly, we compute this regression for the GFDL model in Figures S10 and S11. From the scatter plots, it seems that almost all of the change in AOD in the GFDL model is from sulfate. Whereas for CESM, there is a significant contribution from sulfate but also due to the large range of dust mass changes, there is an additional contribution from dust. Note that the α's are effective aerosol extinction cross sections for each of the species. For externally mixed aerosols it is possible to simply divide the species optical depth by the species column mass to get an estimate of the effective extinction cross section, however for internally mixed aerosols, there is no species-specific optical depth; thus, the need to compute these results through some method such as multiple linear regression.

Table S1 Regression coefficients and regressed global contributions of column masses to total aerosol optical depth. For some regression coefficients (marked with an *), the regression coefficients may not even have one significant digit; other coefficients are likely to have 1-2 significant digits.

Regression coefficients (α) [m^2/mg]	**CESM**	**GFDL**
Sulfate	6.5	17.7
Black Carbon	94.4*	29.1*
Dust	.4	.4
Primary Organic Matter	1.7*	11.1
Sea Salt	3.4*	2.8*
Secondary Organic Aerosol	5.7*	-13.6*
intercept	0.111	-0.097
Global average regressed contribution to AOD change		
Sulfate	-1.153	-3.699
Black Carbon	-0.047	0.016
Dust	-0.412	-0.066
Primary Organic Matter	-0.005	0.107
Sea Salt	0.050	0.034
Secondary Organic Aerosol	-0.022	-0.035

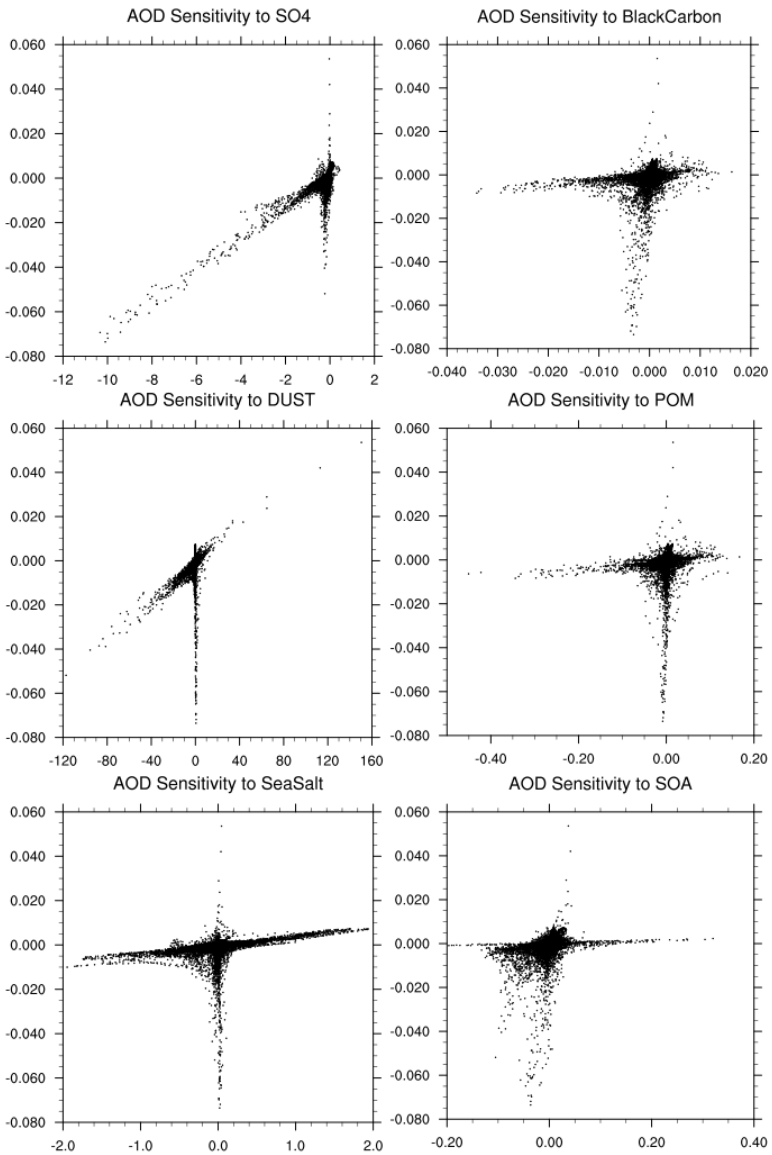

Figure S8. For the CESM model, a scatter plot of change in total aerosol optical depth vs change in column mass for each group of aerosol species. The y-axis is change in optical depth, and x-axis is the change in column mass in mg m^{-2}. Each point represents a particular latitude and longitude point from the time-average of the differenced (zeroed US SO2 – control) simulations. The largest contributors are SO4 and Dust.

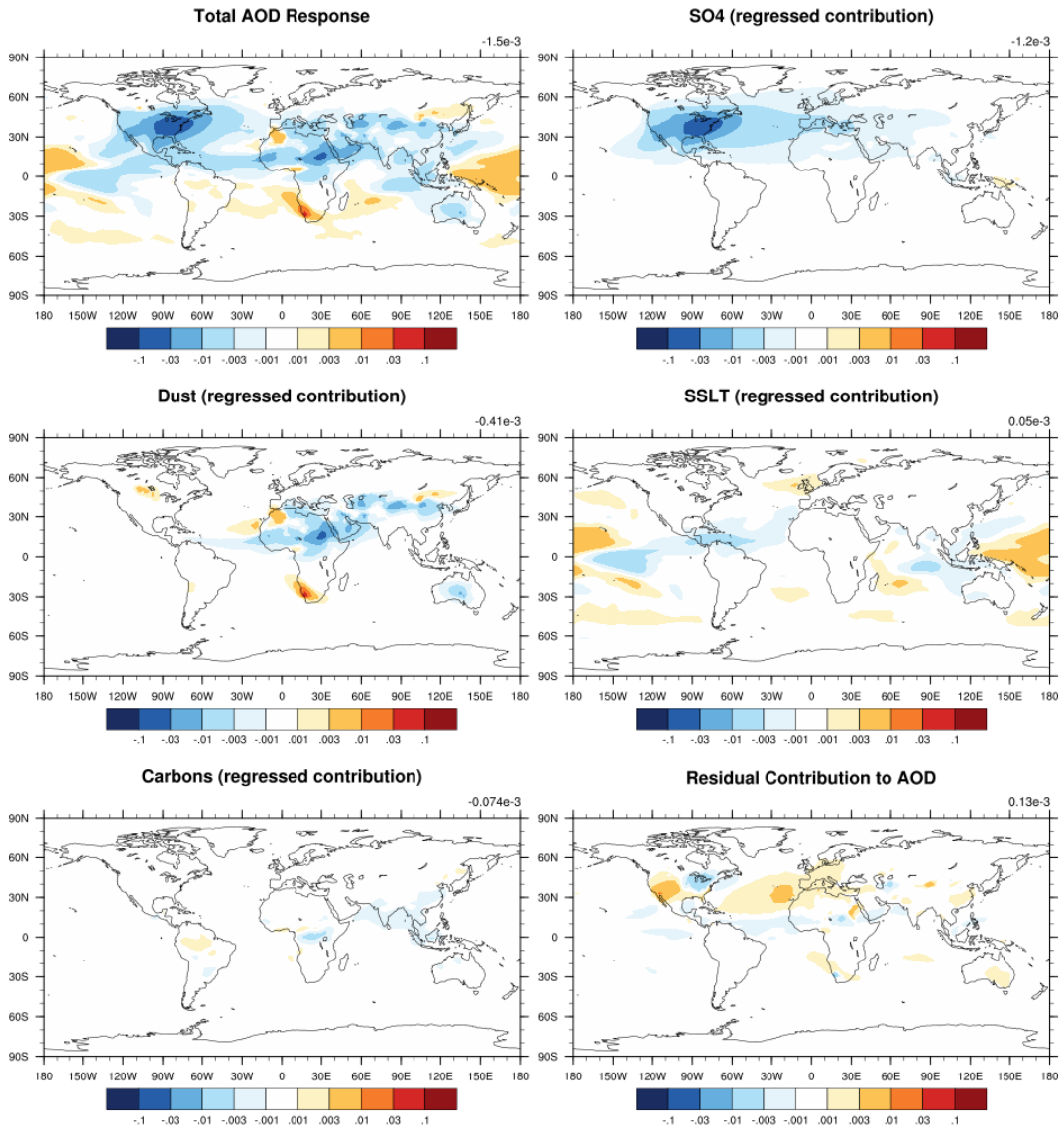

Figure S9. For the CESM model, regressed estimate of the contribution of each species to the total aerosol depth through regression. E.g., the dust contribution is computed as $\alpha_{Dust} dDust(\theta, \varphi)$. The residual is computed for each latitude and longitude as the (total – sum (all species)). Global averages are in the upper right hand of each plot.

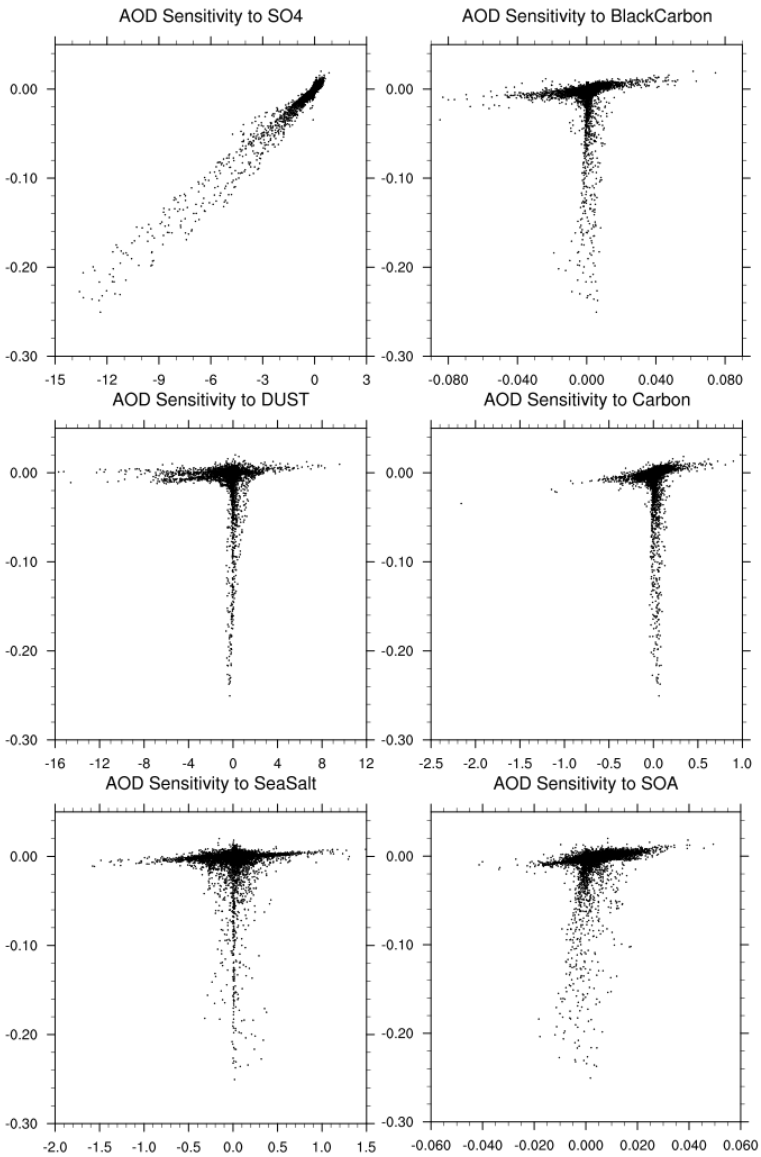

Figure S10. For the GFDL model, a scatter plot of change in total aerosol optical depth vs change in column mass for each group of aerosol species. The y-axis is change in optical depth, and x-axis is the change in column mass in mg m^{-2}. Each point represents a particular latitude and longitude of the time-average of the differenced (zeroed US_SO2 – control) simulations. The largest contributor is SO4.

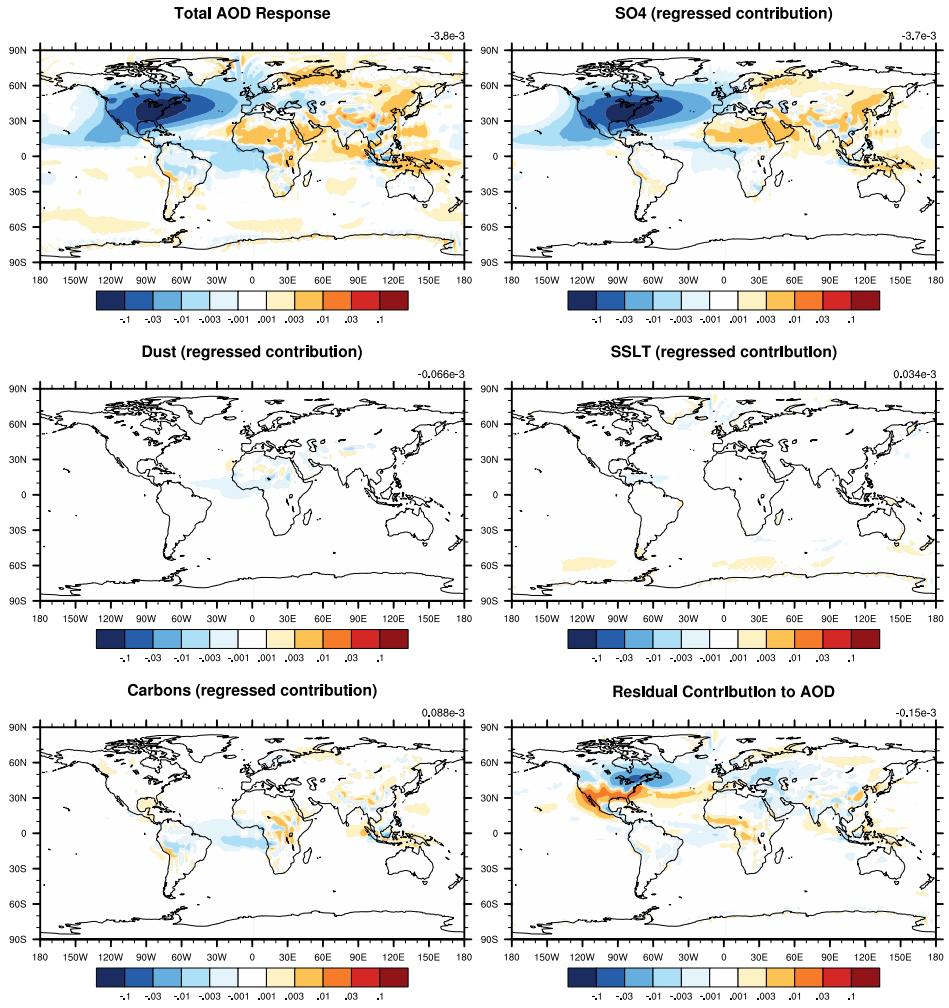

Figure S11 GFDL regressed contributions as in Figure S9.

S8 Remote increase of sulfate in GFDL model over North Africa.

In an attempt to discover the source of increased sulfate burden over Northern Africa and Southern Asia seen in Figure 1, we examined the sulfur dioxide source contributions near

these regions. While the sulfate-burden increases over these regions are an order of magnitude smaller than the sulfate-burden decreases over the source regions, they are still significant. As can be seen in Figure S12, there are no additional sulfur sources near these regions. The remaining possibilities are transport of sulfate or sulfur dioxide to North Africa, or a change in removal processes over this region. As can be seen in Figure S13, the burden changes are very similar in magnitude to the changes in removal processes. The lifetimes are very similar between the simulations; however, it is interesting to see that the lifetime over North Africa is nearly a factor of 15 larger than in extratropical regions. This indicates that if sulfate is transported to this region, it is likely to stay around for a while, and that transport through this region is more likely than in regions with rapid removal processes. This leaves open the possibility that transport from some remote region could be an explanation of this response seen in the GFDL model. Additional explanations could be the changes in photolytic conversion of sulfur dioxide, or changes in oxidation processes. Additional work would need to be done to tease out the possible cause of this small but significant change seen in this particular model.

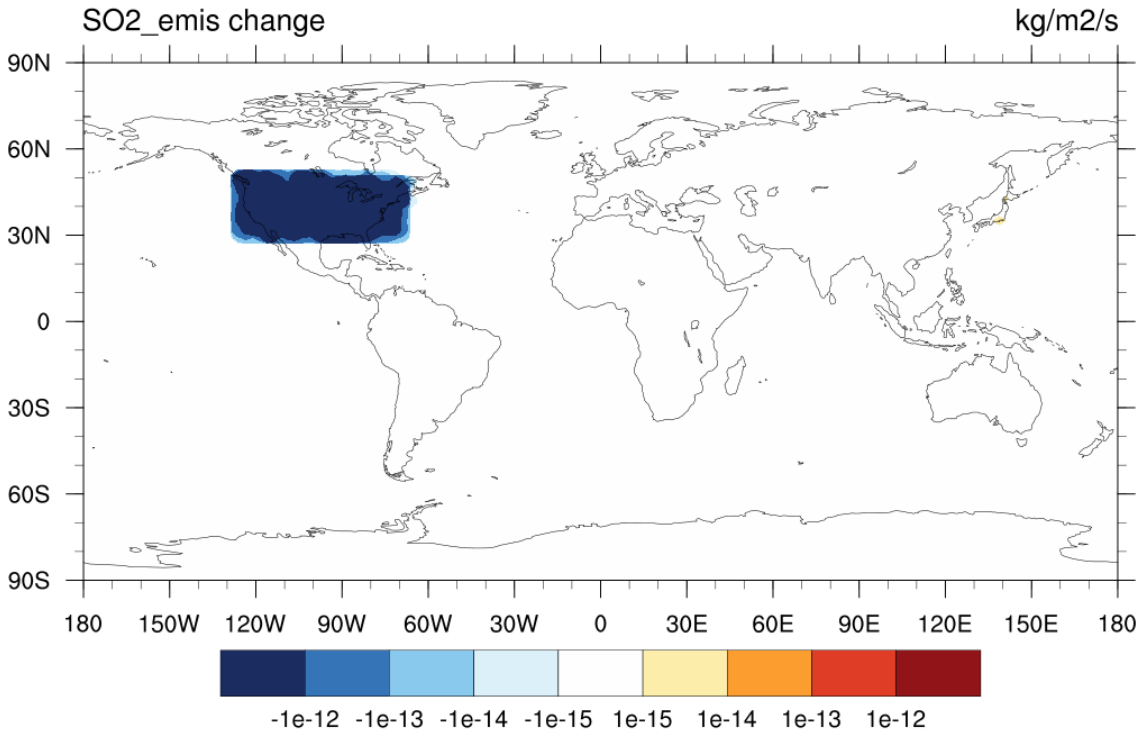

Figure S12 SO2 emissions changes in GFDL model. There are no remote emission changes in SO2. Compare with figure 1, where remote SO4 burden changes in the GFDL model are also more than 2 orders of magnitude smaller than the source region.

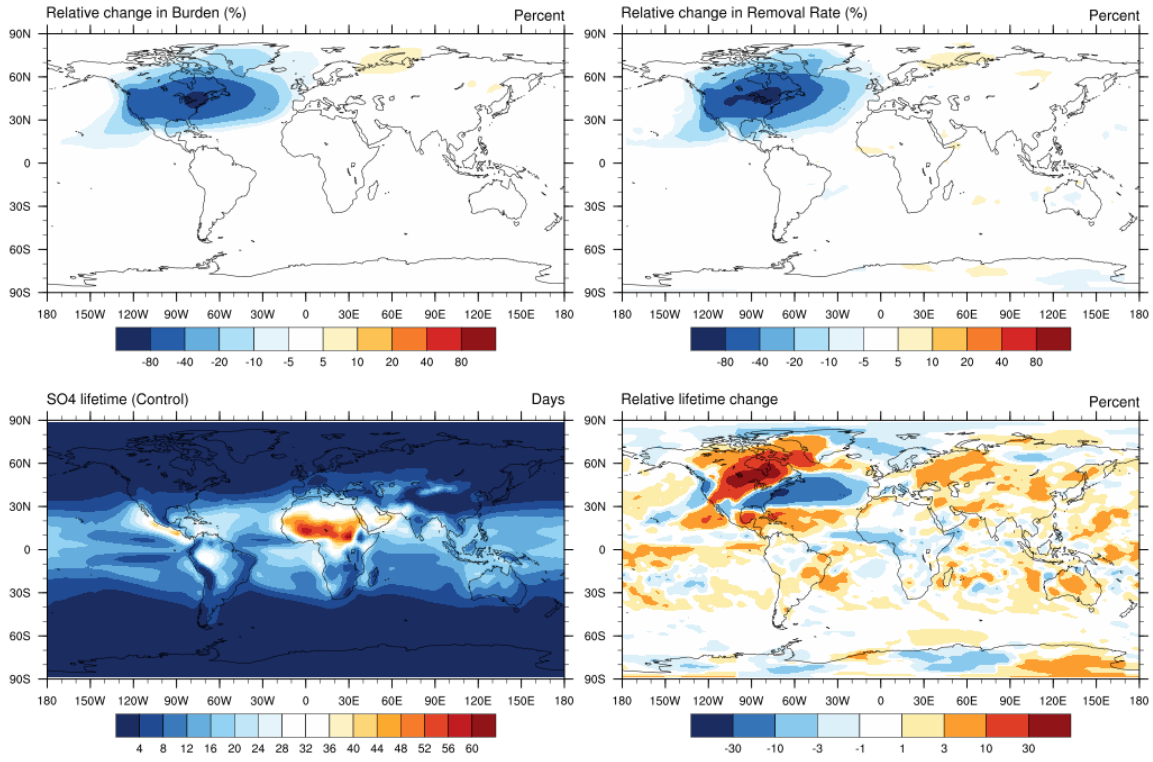

Figure S13 Sulfate burden and removal rate in the GFDL model. The burden is the total column sulfate. The relative change is relative to the control simulation. The removal rate is the combined removal due to wet and dry deposition. The lifetime is burden divided by combined removal rate at each grid point. The lifetime relative change is the difference between the perturbed simulation lifetime and the control simulation lifetimes divided by the control. The dry areas of Africa have a factor of 8-15 longer sulfate lifetimes than extratropical regions; however, there is almost no change in lifetime over any of the dry areas of Africa. There is a small increase in lifetime over much of Asia.

S9 Feedback diagnosed in surface temperature and dust burden responses

The surface temperature responses between the forcing runs and the coupled runs are distinct in some regions. In the CESM model (Figure S14), the pattern of temperature changes from Southern Europe Southward through most of Africa show an opposite pattern. In addition, the temperatures in Eastern Australia are of opposite sign. The responses from Brazil through Mexico are distinct as well. And lastly, the temperature response over Greenland is of opposite sign as well. While for each grid point, the

temperature responses are rarely significant (Figure 10), regional changes (Figure 11) are often significant. A complete analysis would require significantly more work to diagnose the reason for these feedbacks but the difference in patterns of temperature response over Africa indicate that there are temperature feedbacks as well as dust feedbacks in Africa. The dust feedbacks are significant on the global level as seen in Tables 3 and 4.

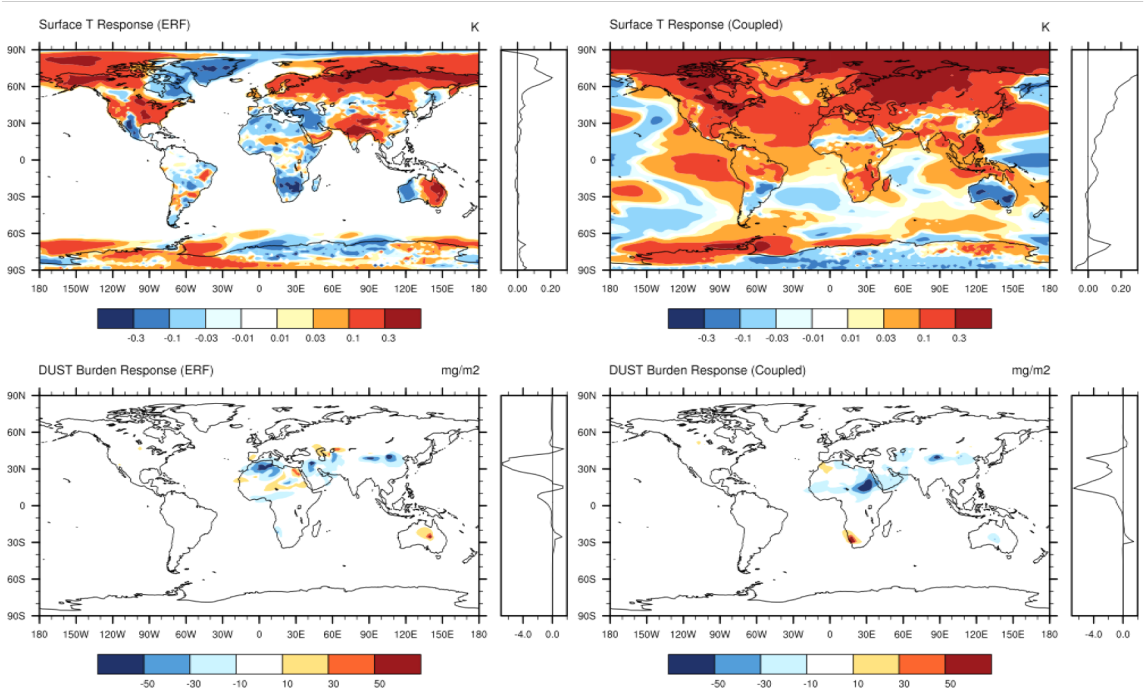

Figure S14 Land surface temperature responses in the forcing and coupled simulations are plotted in the top panels. Similarly, the dust burden response in the forcing and coupled simulations are plotted in the bottom panels. Zonal average responses are seen to the right of each plot.