Response to Reviewer 1 (S. Rahmstorf)

Thanks to the reviewer for his comments. Responses to specific technical points are discussed below, with original reviewer comments in red.

"The first part uses a simple 1-box energy balance model to test whether a mixed climate signal (consisting of ENSO, solar cycles, two volcanic eruptions and a trend) in this model can be decomposed with linear regression methods.
The first finding is that if the climate response is described by a long e-folding time scale, then the recovery of the signal components by linear regression is poor, while it is much better using an approach that explicitly includes such an e-folding time scale. This is obvious even without performing the test, but it hinges on a huge 'if'. The key question is whether the real climate system is well described by such a single e-folding time scale, and as the author himself shows in the next section with proper climate models, the answer is no. That is why we decided against including a "memory" in form of an e-folding time scale in Foster & Rahmstorf (2011)."

The question of whether the linear regression performs poorly does not depend on whether the real climate system acts on a "single e-folding time scale". This can be seen in the next section, where – as the reviewer suggests – the AOGCM results do not show responses described in the single response time manner, but the linear regression approach still performs poorly in picking up the longer term effect of the volcanic recovery (see the last 2 rows of Table 1, for example). Indeed, the memory-less assumption (per FR11) is suboptimal for any sort of response time longer than a month, and the problem increases depending on the length of actual response time. Note that it is nearly impossible to find any physically realistic effective heat capacity that would cause the full response to manifest within a single month. In addition to the references already given for longer volcanic responses, Wigley et al. (2005) suggest that for the variety of realistic climate sensitivity values, the exponential decay time is likely between 27 and 44 months. The e-folding times tested in my manuscript span an even wider range than this, and still in no case does the memory-less linear regression approach perform particularly well.

The reviewer correctly points out that Method 2, which includes a single e-folding time scale, does require that such a model is a reasonable approximation for the real climate system in order for it to perform well. And as I show in Section 4, such an approximation is likely invalid, at least if the real climate system responds similar to the AOGCMs examined. Note that I am *not* presenting Method 2 as a successful alternative to Method 1; indeed, both methods fail to effectively decompose the temperature response in the AOGCMs, and Method 2 performs poorly when the assumption of an underlying linear increase with time is not met. Rather, the point is that simply including memory in the system via an e-folding time and using a multiple regression approach does *not* solve the problems of adjusting temperatures (particularly when the form of the underlying trend is incorrect). This is why, in the last section, a different method is used to adjust temperatures. Per Reviewer 2's suggestions, I will add some clarifying text about Method 2 and "perfect models" in the revision to help clarify this point.

The second finding of Masters is that the regression performs poorly in that it overestimates the response to solar cycles in the 1-box model. The author tacitly assumes that it overestimates this also for the real world and puts this down to there being "only a few solar cycles" to fit. However, that is very unlikely, because almost the same amplitude in the response to solar cycles was also found by Lean and

Rind (GRL 2008) with data starting in the 1880s, thus including 11 solar cycles. Recovering a quasi-periodic signal from noisy data when 11 full cycles are available is quite robust, and Lean and Rind found that global temperature lags solar forcing by just 1 month, in full agreement with our regression analysis starting in 1979. This rapid response time scale found in the data is incompatible with the model of Masters, where he considered response time scales of 14, 57 and 194 months. By design his model excludes the important rapid response component, which is the one captured by linear regression.
The fallacy of using a single time scale can be easily understood by looking at a slightly more realistic energy balance model, one that consists of one third land (including ice-covered ocean) area with essentially zero heat capacity, and two thirds ocean area with a mixed layer heat capacity. In that case on third of the temperature response would occur immediately and two thirds with the e-folding time scale considered by Masters. For his standard case of a 57-months mixed layer time scale, his model would get 19% of the response in the first year. The model that includes land surface would get 46% of the response in the first year. The one-box model greatly underestimates the initial rapid response of the climate system not only by ignoring the land-covered part of our planet but in addition by ignoring the fraction of radiative forcing absorbed straight in the atmosphere. Of the absorbed incoming solar radiation ~30% is absorbed in the atmosphere before reaching the surface.
The need for including multiple response time scales has often been discussed in the literature, recently e.g. by Caldeira and Myhrvold (ERL 2013) who fitted exponentials to GCM output and found that two or three time scales were needed – even though they ignored the sub-annual response by using annual-mean data in the analysis. They write that "substantial temperature changes over land are observed to occur within days in climate model simulations of step-function changes in radiative forcing. The median value of the shortest time constant in the 3-exp fits was 0.6 years, which is less than the annual resolution used in this analysis."

The reviewer notes that "by design his model excludes the important rapid response component, which is the one captured by the linear regression," but as presented in the results, the linear regression *still* "captures" (misdiagnoses) a large solar response even when such a response is not present in the ground-truth model. This is a major issue with the method regardless of the underlying model, and is easily demonstrated with the idealized one-box model presented. IF the multiple regression technique properly captured the amplitude of the solar response in the model, and the only fault was that it failed to capture the longer-term response (which is impossible based on the memory-less assumption), one could make the case that if the bulk of the real-world solar response was near instantaneous the multiple regression technique might perform well enough. However, the fact that we see a solar response with a much larger amplitude than is actually present suggests that the problem with the method is *not* simply an unrealistic response time in the underlying model.

Obviously, a more complex model (with multiple response times) in itself would not produce results that are easier for the multiple linear regression method to decompose, and anyway the ability to decompose a more complex model (namely, AOGCMs) is tested in the next section. However, the reviewer suggests that a more complex model might have some response times that are shorter (per CM13), and points out that a model including the land surface would get 46% of the response in that first year (compared to the 19% of the 57-month e-folding time). However, note that the manuscript also includes a test with the 14-month e-folding time, in which 58% of the response occurs in the first year, and the linear regression method still substantially overestimates the response.

The reviewer also contends that the Lean and Rind (2008) results, because they consider more solar cycles, would be less apt to encounter such an error. As demonstrated in Rypdal (2012), however, the solar response is over-estimated using the regression technique over that period as well, due to the coincidental correlation with five major volcanic responses, as the volcanic responses are much larger than the solar signal. Additionally, Hansen (2011, ACP), which the reviewer references later, also indicates substantially smaller solar responses (fig 18g), even when using the "intermediate" rather than the "slow" response of AOGCMs.

*"**Part 2.** In the second part, a similar exercise as with the simple box model is repeated with an ensemble of global climate models. The results are strikingly different: as can be seen in Figs. 5 and 6, the volcanic response is captured rather well by linear regression."*

While the regression methods capture the immediate amplitude of the volcanic response well enough, the deficiency here comes from their inability to properly capture the longer-term impact. Such a deficiency is one of the key points of this paper, particularly since it has a huge impact when adjusting temperatures to remove these impacts. This is clear again in the last 2 rows of Table 1 – adjusting using the multiple linear regressions in those two cases suggests the "no-volc-solar-ENSO" trend is 1.08 K/Century in GISS-E2-R and 2.60 K/Century in CNRM-CM5 rather than the ground-truth value 0.21 K/Century. These are huge errors, and clearly even if the linear regression methods *appear* to perform slightly better on AOGCMs in those figures, they are inadequate for performing the temperature adjustments with any sort of accuracy.

"This result invalidates the conclusion of the first part, namely that the box model performs much better than linear regression."

As discussed above, despite the box model performing slightly better in the AOGCM cases as well, the paper is not making the case that this box model performs much better than linear regression. Both methods fail to adequately capture the volcanic and solar influences.

This is despite of the evidence that global climate models significantly overestimate the response to volcanic eruptions both in amplitude and duration (see section below). With this in mind, the linear regression model very likely would perform even better for the real world than for the global climate models.

The linear regression model does not perform well in the AOGCMs examined, and would only perform better in the event of response times being extremely short (the amplitude is less important than the response duration). I discuss more the volcanic response of AOGCMs below.

Remarkably, Masters does not say what e-folding time k was found for the fit shown in Figs. 5 and 6 and how this compares to the values assumed in the previous section, but the proximity of the curves for Method 1 and Method 2 shows it must be much shorter than any of the values considered before.

In this case the diagnosed e-folding time was not included because it was shown the AOGCM temperatures cannot be modeled well using this single response time. Nevertheless, these response times can be added: the value for the GISS run was k=8, and for the CNRM run k=10, both which clearly underestimate the longer term responses. The problem here, as the reviewer notes previously, is that the AOGCMs act as models with multiple response times, some more rapid than others. Method 2 over-weights the short-term response times and thus fails to capture the enduring impact of the volcanic recovery. Method 1 can only detect the rapid response and thus performs even worse.

For ENSO, Masters found in the first part that the regression analysis performs very well, but here again he comes to the opposite conclusion with the GCMs, finding a poor performance (even the wrong sign for post-2000 temperatures) and concluding that "the proxy for ENSO (the Niño3.4 SST in this case) diverges from the actual influence of ENSO". This problem is not relevant for Foster and Rahmstorf (2011), since we did not use the Niño3.4 SST but the multivariate ENSO index as a more comprehensive index measuring the amplitude of El Niño and La Niña events. Also, it needs to be tested whether the linkage between ENSO and global-mean temperature might differ between models and the real world.

I agree that the linkage between ENSO (and in particular, the ENSO indices) may differ between the AOGCMs and the real world, and that the CNRM test is probably not an important result with respect to the efficacy of ENSO removal in the real world. I will include additional text detailing this.

Part 3. *In the third part of the paper Masters proposes a new way of adjusting the global temperature time series. It relies on using the output of climate model simulations in order to subtract the solar and volcanic signals from the data.*

There are two problems with this: (a) such climate model ensembles are typically only available with at least five years delay, so this approach is not useful for operational adjustments of global temperatures on an annual basis or so, and (b) this approach assumes that the models are reliably capturing the temperature response to volcanoes and solar cycles (as the author writes himself). In both cases this is not obvious. Evaluating the performance of this proposed adjustment procedure would require that the author makes an attempt at estimating error bars on the simulated volcanic and solar response, which one would expect from a paper that proposes such a method. Below I give some reasons for caution about the models in this regard.

Then in a second step Masters corrects the global temperature series for ENSO using the Niño3.4 SST-based method which he found to give the wrong sign of response in part 2. This is rather surprising, as one might expect that the method passing the test would have been the basis for its application for a proposed new temperature adjustment. Masters gives no scientific reason for using Niño3.4 instead of the more comprehensive multivariate ENSO index, but only motivates it with the "ease of calculating". Based on these questionable adjustments, Masters then claims a "significant deceleration" of global warming. However, this claim of significance is not based on any systematic uncertainty analysis, and in my view none of the sweeping claims made in the conclusions section are properly supported by sound methods in this paper.

Do climate models properly capture the fast time scale of response to volcanic eruptions?

It is well-known that climate models tend to smear out rapid time scales because of their limited grid resolution, leading to numerical diffusion and generally "over-mixed" models both in the horizontal and vertical. Hansen et al. (Atmos. Chem. Phys. 2011) extensively looked at the response functions of climate models and conclude: "We believe, for several reasons, that the GISS modelE-R [also used by Masters] response function is slower than the climate response function of the real world" and that "the slow response ... is typical of most IPCC models." This conclusion is directly relevant for Masters' statement that his "method relies to some degree on the multi-model-mean approximately matching the actual response time of the real world".

This too-slow model response can be seen clearly for the example of the Pinatubo eruption in IPCC models, see the following figure.

Figure: *Observed vs. modelled temperature, from Box 11.1 of the IPCC AR5 WG1 report. Note that all models are too cold starting from the Pinatubo eruption in 1991 for at least five years.*

More systematic comparisons for volcanic eruptions over the last 400 years, using coral data as reliable high-resolution recorders of sea surface temperature, show that climate models systematically overestimate both the amplitude and duration of the response to the eruptions (Tierney 2013).

The purpose of adding the adjusted temperature series was an experiment to see what the resulting temperature series might look like if the volcanic, solar, and ENSO influences were removed based on physical models rather than linear regressions. However, to give such a method the full treatment of uncertainty analysis and testing requires re-running AOGCMs or developing models to mimic these, which in turn is beyond the scope of this paper and should probably be its own manuscript. As such, I would be happy to remove this latter section and instead keep the current paper focused narrowly on the pitfalls of the multiple regression technique in temperature attribution.

Conclusion

In Foster and Rahmstorf (2011) we deliberately used a simple, straightforward regression analysis to obtain the contributions of ENSO, solar cycles and volcanoes to the global temperature evolution. The idea was to derive those straight from data without the use of models, since numerous modelling studies to interpret the temperature evolution already exist, but the ability of models to correctly reproduce these effects remains disputed. Thus a complementary, purely data-based approach seemed useful. Our regression analysis relies on the assumption that a rapid temperature response without long memory is a reasonable approximation. Masters calls this into question in his manuscript, which is a comment on our paper that is based on an earlier blog post on it. His argument hinges on the existence of "a lingering effect of the Pinatubo recovery into the 21st century", but he provides no evidence for its reality.

The manuscript indeed demonstrates that the memory-less assumption has a large effect on the resulting adjusted temperatures if the assumption is not met. It is impossible to prove conclusively that the assumption is not met in reality (otherwise there would be no need to decompose the volcanic response if we already knew it), but there is substantial evidence presented to suggest a longer effect from the Pinatubo recovery:

1) 6 of the 7 AOGCMs shown in the figure indicate the naturally forced response (volcanic recovery + solar activity) contributes to a positive trend in the 21st century, whereas the multiple linear regression method incorrectly diagnoses a cooling contribution. All 7 show a positive contribution starting in 1996, which due to the short response time would be entirely missed using the multiple regression method. The reviewer notes that these responses might be unrealistic, referencing Hansen et al. (Atmos. Chem. Phys. 2011): that for AOGCMs the "*response function is slower than the climate response function of the real world.*" However, in figure 5 of that reference, it is clear that even if the "slow" response (which suggests only ~45% of the temperature response after 10 years) is incorrect, the "intermediate" (~55% after 10 years) and "fast" (~65% after 10 years) alternatives also leave substantial portions of the response well into the 21st century. Contrast this with the implicit assumption of the multiple regression method, which assumes that 100% of the response would be manifested after 10 years. Moreover, if one looks at the simulated responses in figures 18e and 18f of that paper (which uses the "intermediate" rather than the "slow" response), the effect of the Pinatubo recovery indeed lingers well into the 21st century.
2) Simpler energy balance models also show a likely continued effect of the volcanic recovery. No physical model is going to show a memory-less response, which is why the multiple regression

technique is almost certainly going to underestimate the enduring effect of volcanic eruptions. Unless virtually all the response is instantaneous, this can have a substantial impact on the accuracy of the temperature adjustments. Again, if we examine figure 11 of Rypdal (2012), which expands on the work of Hansen (2011) here and shows 4 different response models to volcanic eruptions, it is clear that all four show effects of the volcanic recovery into the 21st century.

3) When testing the multiple regression method on models, we find a substantial overestimate of the solar response and underestimate of the volcanic response relative to the ground-truth. When applying the multiple regression technique to the actual temperature series, as per FR11, it results in an estimated solar efficacy that is 5-6 times larger than the volcanic efficacy when both are converted to the same units Wm^{-2} (using the conversions discussed in the manuscript). While some difference is to be expected in the efficacies of these forcings, and perhaps different response times, it is hard to argue that a discrepancy of this magnitude could be realistic. Given that this follows the same pattern we observed in the tests, this suggests – albeit circumstantially – that such a diagnoses is indeed an artifact of the method when applied to the actual earth system.

Finally, I will note that the thrust of the argument in this paper does not hinge on a "lingering effect of Pinatubo into the 21st century." Certainly the multiple regression technique will perform worse as the length of the volcanic recovery increases, and the long recovery seems quite likely for the reasons mentioned above. Nevertheless, any sort of memory in the system will hinder the multiple regression technique's effectiveness, as will any incorrect assumptions about the form of the underlying trend. Moreover, in the case of the recent period, the technique clearly has a tendency to overestimate the solar influence.

In the first section he simply uses an inadequate model where such a lingering effect is basically built in from the start due to the inclusion of only one slow response time scale. In the second section he disproves this himself when his e-folding model is fitted to GCM results and can only fit those when a much shorter time scale is used, so that his energy balance model results become almost identical to our regression results – and this despite the fact that also the GCMs likely overestimate the response time scale.

Again, this seems to stem from the confusion that Method 2 is supposed to be shown to be a success. I will attempt to clarify that *both* methods, which ultimately use multiple least squares regressions to find the best fits, do not perform well in adjusting temperatures.

Finally he proposes a way to adjust observational data with the help of model simulations and an ENSO adjustment which earlier he showed to perform poorly, but he does so without proper uncertainty analysis and the whole approach is questionable in the way it conflates models and observational data. A better approach is to use climate models to interpret rather than adjust the data, by careful comparison of model results with data, as is done e.g. in the IPCC reports.

Per the discussion above, I will remove this final section of adjusting temperatures from the current manuscript.

[1] Wigley, T. M. L., Ammann, C. M., Santer, B. D. & Raper, S. C. B. Raper. The effect of climate sensitivity on the response to volcanic forcing. J. Geophys. Res. 110, D09107, doi:10.1029/2004/JD005557 (2005)