Thank you to all reviewer's for your suggestions. We respond to each comment below in italics. We will also submit a version of the article with all the changes marked to the editor.

<u>Response to Andrew Sawyer</u>

This is not a review of the paper, only a short comment about the MODIS aerosol data used.
In this study, the authors use MODIS Collection 6 data for MODIS Aqua, but MODIS Collection 5 data from MODIS Terra. The latest data version, which is expected to remain the standard for at least the next few years, is Collection 6; older versions should generally be considered obsolete and avoided where possible. In case the authors are unaware, I thought I should mention that Collection 6 data are also available for MODIS Terra. In the 'Dark Target' data set which it looks like the authors are using, there were several important bug fixes and updates going from Collection 5 to Collection 6 (see the Levy et al., 2013 paper cited in the manuscript). It may therefore be advisable to repeat the Terra portion of the analysis with Collection 6 data, if this is practical. My expectation is that results for eastern North America may not change too much, but western North America and China may change more significantly. This is consistent with what the authors see in e.g. Figure 6 (Terra/Aqua differences are not consistent with diurnal sampling effects), and so using the latest data version for both platforms may simplify the analysis somewhat. I realise, though, that this may be quite a computational burden to update the data set use at this stage. MODIS Collection 6 also includes Deep Blue aerosol data covering all land surfaces, and therefore may be of additional interest for this type of study. Our experience suggests that the two data sets are quite similar over North America, though, so it might be that not much is gained from using both Deep Blue and Dark Target as model constraints (they would probably have a similar effect on the model over North America).
Thank you for this comment. We started this work using Collection 5, but upon release of Collection 6, we redid all of our analysis with the updated product for Aqua MODIS, but did not reprocess MODIS-Terra due to the timeline of release (and the need for many years of data). However, we kept the discussion of Collection 5 because previous studies using a satellite-based $PM_{2.5}$ *method have relied on Collection 5 and there is substantial difference (as shown) between Collection 5 and Collection 6, which could be a significant source of uncertainty in those previous estimates. We have now removed any discussion of Terra from our results, ensuring that we are using the latest data products in our analysis.*

As a minor unrelated point, I notice that C. A. Pope's surname is typeset as "Pope III" in the paper and reference list. The surname should just read "Pope". If the authors are using LaTeX/BibTeX then I think the correct formatting can be achieved by writing "Pope, III, C. A." rather than "Pope III, C. A.
This is a typesetting issue which we will be sure to catch in the final version.

<u>Response to Anonymous Referee #1</u>

General comments: This paper estimates sources of uncertainty in exposure estimates by analyzing differences in model versus satellite-derived PM2.5 and various concentration-response functions. I think that this is an important contribution to the field because it compares the influence of individual assumptions on estimated mortality and compares results to different studies.

Thank you for your review. Responses are in italics below each comment.

Specific comments:

-Pg. 25333, line 14: Did the five-year population estimates indicate that a linear interpolation was appropriate in China?

Population growth is likely not linear in China over the time period, but it is outside the scope of this paper (beyond acknowledging this as a source of uncertainty) to determine the actual annual change in population given the projections for 2005, 2010, and 2015. Additionally, we calculate the average annual mortality over a 8 year time period, so this should reduce some of the uncertainty.

-Pg. 25334, line 4: How would you expect different spatial resolutions of baseline mortality data to influence your results and comparisons with the studies mentioned?

Spatial resolutions of all variables will impact local results, including the baseline mortality, specifically in regions where grid boxes might straddle two countries or states with very different mortality rates. Many previous studies use a country-wide regional baseline mortality rate; we go one step beyond by using state-specific mortality for the US, however we acknowledge that further spatially resolved mortality data (particularly in China) would provide more accurate local estimates of the burden. To address this concern, we had added discussion throughout the text about resolution and referenced Punger and West (2013) and Thompson et al. (2014) both of which discuss the impact of resolution for estimating health impacts.

-Pg. 25338, line 18: Specify here that satellite-based estimates are gridded at the same spatial resolution as the unconstrained model instead of (or in addition to) further down in this section.

Done.

-Pg. 25339, line 5: Explain why MODIS and MISR were both used (strengths/weakness of each dataset), and why collection 6 for Aqua and 5 for Aqua and Terra. Also, how does this compare with also using SeaWiFS in the more recent van Donkelaar et al. (2014) work?

We used MISR and MODIS to highlight their differences because previous studies have used either or both of these for a combined AOD product. Additionally, MISR is generally biased low and MODIS is biased high (in comparisons with AERONET). This has been noted in other studies, and we address this discrepancy in the text and reference these other studies.

We started this work using Collection 5, but upon release of Collection 6, we redid all of our analysis with the updated product for Aqua MODIS, but did not reprocess MODIS-Terra due to *the timeline of release. However, we retained the discussion of Collection 5 in our submitted text because previous studies using this method have relied on Collection 5 and there is substantial difference (as shown) between Collection 5 and Collection 6, which could be a significant source of uncertainty in those previous estimates. However, given concerns about the data quality of Collection 5 (particularly raised by Andrew Sawyer) and that this comparison is not central to our study, we have now removed all MODIS-Terra data from the analysis to ensure clarity.*

van Donkelaar et al. (2014) use a combined product of MODIS, MISR, and SeaWiFs. Using this product would likely also provide different results because it is a different product. The discussion however would remain the same, that different satellites products have different biases and therefore would result in different estimates. We want to stress that the goal of this paper is not to design a $PM_{2.5}$ product as with the series of papers by van Donkelaar et al. (2010; 2013; 2014; 2015) but to discuss uncertainties in these products and stress the necessity of understanding the data used in assessing health impacts.

-Pg. 25341, line 22: This is mentioned briefly later in the manuscript, but do you have any indication of how MODIS and MISR compare with observations at shorter timescales (daily)? Are the satellites overestimating or underestimating peaks and how could this impact exposure estimates?

The satellites both underestimate and overestimate peaks in AOD depending on the time and location. Compared to AERONET sites, the mean normalized gross error in daily AOD for MODIS is 75% in the western U.S., 50% in western China, 35% in the eastern U.S. and in eastern China. Determining chronic exposure from long-term averages should reduce some of the uncertainty from not capturing daily variability (unless there is a systematic bias). Our discussion in the sensitivity analyses section of using average AODs to compute $PM_{2.5}$ alludes to some of the uncertainty in this, that daily variability can influence the annual means.

-Pg. 25342, line 17: Is there a difference in PM2.5 components between China and the U.S. that could influence results?

Yes, this is why we do some sensitivity tests examining the potential biases in aerosol composition (i.e. the sulfate and BC only sensitivity tests). As to the toxicity of different components, this is still an open area of research (i.e. Chung et al., 2015), therefore we do not estimate or discuss speciated $PM_{2.5}$ estimates here.

Chung Y, Dominici F, Wang Y, Coull BA, Bell ML. 2015. Associations between long-term exposure to chemical constituents of fine particulate matter ($PM_{2.5}$) and mortality in Medicare enrollees in the eastern United States. Environ Health Perspect 123:467–474; http://dx.doi.org/10.1289/ehp.1307549.

-Pg. 25344, line 21: Is there a figure or table with these AERONET results?

We have added a figure with these results.

-Pg. 25344, line 25: Better in the eastern US and eastern China than the western parts of each country? Please clarify.

We have added to the text:"than in the western U.S. and western China"

-Pg. 25348, lines 21-22: What are examples of some of these regional sources?

We have added to the text: "e.g biogenic aerosol in the Southeastern U.S."

-Pg. 25352, line 4: Implications of a study that is smaller and using only white participants?

The implications are that it may not be representative of the larger population that might not have the same demographics.

-Pg. 25353, line 3: Can you comment on how the results of Chen et al. (2013) (or another China-specific study) would impact your results?

We cannot make a quantitative comparison with the Chen et al. (2013) study because they used total suspended particles (TSP) rather than $PM_{2.5}$ (and most epidemiological studies agree that the most harmful constituent is the fine fraction). Pope and Dockery, 2013 do compare the Chen et al. (2013) results with other studies and find that the elevated risk is lower than found in the Laden et al. (2006) and Pope et al. (2002) studies which is in line with Aunan and Pan (2004)as mentioned in the text. We have added the Chen et al. (2013) reference to the text.

-Pg. 25355, line 17: What was the spatial resolution of the Lelieveld et al. (2013) study? As the authors mentioned earlier in the text, spatial resolution might be driving some differences in more populated grid cells.

Lelieveld et al. (2013) uses a model with a horizontal resolution of ~1.1° x 1.1° which is somewhat coarser than our model resolution.

-Conclusions: I think that the discussion of Figure 9 needs to be expanded, which could also include a brief discussion of how different spatial resolutions (among different models and between model and satellite-based estimates), emissions inventories, region-specific health data, etc. impact these estimates. And, if possible, it might be helpful if the authors could give some sort of general recommendations regarding the "best practices" of the factors that are most important for future authors to consider when estimating exposure at either at a global scale or for China and the U.S. specifically.

We have moved Figure 9 to section 4 and expanded the discussion. We have added more comments regarding resolution throughout the paper and added comments about other sources of uncertainty to the conclusion. As stated, the CRF seem to be the most important for attributing mortality, as such, we prefer to leave the recommendation about the "best practices" to experts in that field and instead suggest that a range of results are presented for comparison to other studies which also use a range of different methods.

-Table 1: Would it make sense to include the updated estimates of Lelieveld et al. (2015)? Also, doesn't Evans et al. (2013) provide estimates with different CRFs?

Lelieveld et al., 2015 had not been released at the time of submission but has now been added to the text along with several other estimates. Yes, Evans et al. (2013) does provide estimates with different CRFs which we have now added along with several other studies.

-Table 1: Does the heading mean that some of the U.S. estimates are for all of North America and China for all of Asia/Western Pacific? Please clarify in the caption because it's unclear if these refer to studies that were specific to the countries or to regions. You mention that the Anenberg study is regional in the text but it's unclear about the others. Also, are all of these studies for similar years?

We have clarified the region for the estimate in the table next to the study and added an extra column in Table 1 noting the year of the estimate.

-Figure 4: Anything available for China? It would be helpful to see a plot based on any available data, maybe AERONET as mentioned in the text?

There are now a significant number of surface monitoring sites in China, but there were none with long-term measurements (that were publically accessible) for the period of 2004-2011. We have added a figure with AERONET sites and comparisons (Figure 7).

-Figure 9: Are the previous estimates including only country-based US and China specific studies, or are some regional? This figure is very helpful and I would appreciate a more in-depth discussion.

Yes, there are. We have tried to clarify this in Table 1 and have added to the discussion on Figure 9.

Technical corrections:

-Pg. 25330, line 19: "on the order of.." *fixed*

-Pg. 25344, line 20: Are you referring to Fig. 5a (exposure plot) or to Fig. 6a (AOD plot)? *Figure 6, fixed*

-Pg. 25345, line 2: Missing a period or this is a run-on sentence. *This is a typesetting error (we prefer U.S. to US) which we look for in the final version.*

-Pg. 25345, line 10: Fig. 6b? *yes.*

-Pg. 25346, line 17: "Requires model output," that is unnecessary as written. *Fixed.*

-Pg. 25351, lines 19-23: Much of this is repeated from the introduction and could likely be cut or shortened if the word limit is an issue. *We chose to leave it for clarification.*

-Table 1: Do you mean that Table 4 provides additional information?

Yes, thank you for noting this mistake.

-Table 3: Define threshold abbreviations in caption.

We have altered the table.

-Figure 2: Can you change the font size of the individual studies? This figure is difficult to read, and the information might make more sense in a table.

We have made the font bigger.

-Figure 7: Please clarify that abbreviations are also defined in Table 3.

Done.

-Figure 8: Shouldn't this refer to the last column in Table 4?

Yes. This has been fixed.

References mentioned above: Chen, Y., Ebenstein, A., Greenstone, M. and Li, H.: Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy, Proc. Natl. Acad. Sci. USA, 110(32), 12936–12941, doi:10.1073/pnas.1300018110, 2013. Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. and Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature, 525(7569), 367–371, doi:10.1038/nature15371, 2015.

<u>Response to Anonymous Referee #2</u>

"Exploring the Uncertainty Associated with Satellite-Based Estimates of Premature Mortality due to Exposure to Fine Particulate Matter" by Ford and Heald. This study estimated the premature mortality in US and China by using satellite data and GEOS-Chem model simulations, and quantified the uncertainties of the results caused by different methods and dataset used to derive. The study is useful to constrain the estimated health effect due to increased concentrations of fine particulate matter with satellite-based observations. I have a few major concerns and some specific comments as below. Firstly, the relationship η, which links PM2.5 and AOD, is derived from the GEOS-Chem simulation in this study, although the authors have conducted a couple of sensitivity experiments to understand how much difference would be caused due to the uncertainty in η, I am curious that how these would be different from the real η if directly linking the surface PM2.5 and satellite AOD. Secondly, the relative risk (RR in the paper), which is a key factor to determine the premature mortality due to exposure to PM2.5, differs significantly because the pathophysiological mechanisms are currently unclear. The authors assessed the uncertainty of the estimated mortality rate by using different PM2.5 concentration response function. I wonder is it possible to give us a "better choice" for the study region such as US and China? Finally, the authors have conducted a few sensitivity experiments to test how different factors impact their estimations, which is a good attempt to improve our understanding. The disadvantage is lacking of the detailed explanations and discussions on theses sensitivity results.

For comparison of the actual η vs. model η using ground-based measurements, we refer the reviewer to Snider et al., 2015 and add this reference to the text and for discussion of the satellite η, we refer the reviewer to van Donkelaar et al., 2012. We have added text to the *sensitivity discussion to better address these questions. As the CRF seem to be the most important for attributing mortality, we prefer to leave the recommendation about the "best choice" to experts in that field and instead suggest that a range of results are presented for comparison to other studies which also use a range of different functions.*

Snider, G., Weagle, C. L., Martin, R. V., van Donkelaar, A., Conrad, K., Cunningham, D., Gordon, C., Zwicker, M., Akoshile, C., Artaxo, P., Anh, N. X., Brook, J., Dong, J., Garland, R. M., Greenwald, R., Griffith, D., He, K., Holben, B. N., Kahn, R., Koren, I., Lagrosas, N., Lestari, P., Ma, Z., Vanderlei Martins, J., Quel, E. J., Rudich, Y., Salam, A., Tripathi, S. N., Yu, C., Zhang, Q., Zhang, Y., Brauer, M., Cohen, A., Gibson, M. D., and Liu, Y.: SPARTAN: a global network to evaluate and enhance satellite-based estimates of ground-level particulate matter for global health applications, Atmos. Meas. Tech., 8, 505-521, doi:10.5194/amt-8-505-2015, 2015.

Specific comments:

p.25333, at top of the page, it is difficult to see here how PM2.5 contributes to health from the equations Eq.1 and 2; please add an equation to describe the link between PM2.5 and RR, if possible.

These are standard equations to determine the attributable fraction of mortality due to a specific factor, not just for $PM_{2.5}$ *exposure. Therefore, we prefer to leave the equations as they are in order to align with what is standard in the literature. The concentration response functions are described in Section 2.4 which we refer to in this section.*

p.25333 last paragraph, You use crude death rates, instead respiratory disease, to determine baseline mortalities, which will overestimate the burden of death due to air pollution. Can you find and use the death rates from non-accidental death? In China, it is even cruder as population rather than death rate is used to estimate. Can the authors estimate the biases caused by this?

We do not use overall crude death rates (or "all cause"), but the death rates specific for each disease (respiratory, heart, and lung cancer) for each state and each year. Same for China, these are not crude death rates; they are year-specific age-standardized mortality rates by cause from the WHO for China as described in Section 2.2.

Fig.2. the text is too small to see, I suggest the authors to make this figure bigger.

This may be related largely to ACPD formatting but we have made the font bigger and will ensure that this is legible in the final version.

Table 2: table caption, "… in Eq. (8)…", should be Eq.(7).

Yes. Thank you.

p.25339, line 5-20, You removed the satellite observations with high AOD (>2.0), can you explain how do you decide this threshold? since AOD could be very high (over 2.0) in some cases, e.g. heavy pollution?

We choose this threshold to attempt to take care of cloud contamination as discussed in the methods from our previous work (Ford and Heald, 2012, 2013). We acknowledge this and note that this could remove high pollution events, particularly in China.

Fig.5. How do you compute the values shown in Fig.5? Which field in Eqs corresponds to the results shown here? can you clarify that if the results are P in Eq. (2), or others?

This is a cumulative distribution plot showing the percent of the population exposed to different annually-averaged $PM_{2.5}$ *concentrations calculated using the population (which is P in Equation 2) and the concentration (which is C in the RR equations). It is calculated as a sum of the population in each grid box which has an annual average concentration at or above each concentration on the x-axis. We have added this to the text to clarify.*

p.25344, bottom paragraph. It would be good to give a plot to show the AERONET sites used in the comparisons in both US and China. The quantitative comparison of AOD between satellite and AERONET is not shown in a plot and/or table, and not even given in the text. Please include these comparisons.

A figure has been added (now Figure 7).

p.25347 and Fig.7: As I can see the NMB is apparently largest in Southeastern China from the experiment vertical profile, but there are no explanation in the text. For the test Relative humidity, there are positive NMB in Southeastern and Northeaster China, but negative NMB in western and Central China. The necessary explanations and discussions are needed in this sensitivity tests.

We have added more to this discussion to clarify these results.

Figure 8, figure caption "… in Table 3", should be Table 4.

Fixed.

Figure.9, I suggest to move the Figure 9 and associated text into section 4, rather than last section, i.e. section 5.

Done.

Response to Anonymous Referee #3

The authors present an interesting paper in which they estimate the health burden of PM2.5 in the US and China, compare those estimates with previous studies, and then explore uncertainties in the calculation due to satellite estimates of PM2.5, health function parameters, etc. The paper is unusual in its detailed treatment of atmospheric science and satellite retrievals, as well as concentration-response functions within a single paper. To my mind, this is both a strength of the paper – as different uncertainties are addressed within a

single paper – and a weakenss, as the discussion ranges over a wide body of literature and can be hard to follow at times.
Overall, my sense is that the paper is a worthwhile addition to the existing literature.
Thank you for your review.

However, I feel that the presentation of the complex discussion can be improved and I have some general questions or concerns about the approach:
1) It seems that the main points of the paper are summarized in Figure 9. Differences in health burden are presented when exposure estimates are driven by the model vs. two satellite estimates, and then uncertainty analysis is performed on 3 parameters individually. Given that the uncertainty due to individual parameters has been estimated by the authors, I am surprised that they did not take the next step to estimate an overall uncertainty given uncertainty in those parameters individually. Also, is the uncertainty in CRF in Figure 9 a simple uncertainty given the confidence intervals from a single study, or does it somehow account for uncertainty as illustrated in Figure 8 or Table 4?

Our goal was to show a range of uncertainty due to the specific parameters that we explored. We did not examine every sources of uncertainty, especially with regards to the model. Estimating an overall uncertainty would be a much more complicated process that would likely require a much more thorough examination of the parameter space (along the lines of Lee et al., 2013). For this reason, we do not want to provide an overall uncertainty that might misrepresent the analysis done here. The grey lines show the uncertainty from the confidence intervals of the Krewski et al. study, the colored bar shows the uncertainty from Figure 8 (now Figure 9).

2) The goal as stated p. 25354 is "to explore how mortality burden estimates are made and how choices within this methodology can explain some of [the discrepancies among previous studies]." The authors have succeeded in estimating how different modeling choices or parameters contribute to the overall uncertainty. But as there are many differences among many different studies, I don't know that this paper helps to clarify those differences in results – or it certainly does not explain why a particular study is high or low vs. others. The results shown are not surprising given the current literature, and since previous studies have often included analysis and discussion comparing their results with others, I'm not sure that the authors add something new here. The results are interesting and seem to add to the literature, but I'd encourage the authors to think harder about what is new and present that more clearly.
We have added to the discussion of Figure 9 (now Figure 10) and the conclusions which we believe does a better job clarifying the differences.

3) Related to #2, despite the complexity of the paper and its extensive discussion, I thought the bottom-line messages were rather few. The authors should consider reorganizing in places to reduce repetition, and/or removing excessive discussion.

Thank you for this suggestion. As the reviewer highlighted, this was a study with a wide-ranging discussion, and we have endeavored in this final version to streamline the discussion. In *particular, we believe that moving our discussion of the overall uncertainty from the conclusions into a separate section (4.4) clarifies the conclusions and contributions of the study.*

More specific modeling questions & concerns:
1) I might be wrong, but I'm not aware that anyone uses a linear function as described in equation 3.

It is not commonly used for the reason stated in the text, that it produces very large estimates for high concentrations. However, it is used as an alternative concentration-response curve in Cohen et al. (2004; 2005), and the Hoek et al. 2013 is presented as linear by Pope et al. (2015). Our goal was to start with the simplest form and then explain alternate functions and the impact this has on the estimates. We have tried to clarify this is the text. Additionally, in order to be more in line with recent literature, we have chosen to use the Burnett et al. (2014) as our baseline function and discuss other forms as sensitivity tests.

2) They assume the C0 to be zero. I don't think that there are other studies that use zero as C0, and I am concerned that it requires a significant leap of faith to assume that the same concentration-response relationships hold at concentrations below which we have observations. If the authors keep this assumption, they should do more to discuss and justify this choice.

Thank you for raising this point. Sun et al. (2015) uses a threshold of zero and many studies which estimate the change in mortality from different sensitivity simulations, which use a base case concentration (for changes over time or comparing natural and anthropogenic sources) as the C_0 value are not accounting for a threshold if those values are less than the threshold of the given CRF (e.g.. Anenberg et al., 2010; Silva et al., 2013). Many other studies also test using threshold values below the observations (e.g. Johnston et al., 2012; Evans et al., 2013).

The reason one might not include a threshold is that, as stated in the text, most experts in health impacts of ambient air quality agree that there is no population-level threshold (Roman et al., 2008). Additionally, the literature is full of assumptions about the shape and magnitude of the CRF above which there were no observations in the original studies (it is relatively linear in the range of observations, which is why there is so much discussion on the shape of the CRF at high concentrations) which is even more uncertain than at lower concentrations (see confidence intervals in Burnett et al., 2014).

However, we would like to point out that we are not trying to defend this as the "correct" approach, we instead stress that we explain different assumptions and how this impacts the results. We have tried to clarify this in the text.

More specific comments: - The title focuses on satellite-based estimates, but model estimates are also used here, and uncertainty in concentration-response factors are also a major focus of the analysis.
As stressed in the paper, satellite-based estimates are in many ways also model estimates, they are just constrained model estimates. Additionally, estimating premature mortality due to exposure (as in the title) requires use of a CRF so this is implicit. Furthermore our objective is to

provide context for the interpretation of satellite-derived $PM_{2.5}$ health estimates, and we feel that our title accurately reflects this.

- Given the interest in models, satellites, and ground observations in the paper, I am surprised the authors didn't mention "data fusion" types of approaches such as Brauer et al (EST, 2012), who did data fusion to underlie the Lim et al. global burden estimates. Do data fusion studies reduce these uncertainties? The question may be beyond the scope of the paper, but I thought it deserved at least a little qualitative discussion.
Thank you for this suggestion. Data fusion methods can indeed reduce some of the uncertainty. We have added a reference to Brauer et al., (2012; 2015) and van Donkelaar et al. (2015b) in the conclusion.

p. 25331, l. 8-19 – The discussion here seems to mix up estimates of concentration to drive epidemiological studies vs. concentrations to drive risk assessments. This also seems to be confused a few times later in the paper. I would think that using concentrations to drive risk assessments would be the main purpose here. The last sentence of this paragraph I don't think is true – many epidemiological studies consider health effects for whole populations using monitors as imperfect estimates of concentration (and not as estimates of personal exposure).
Our goal here was not to confuse the two, but to discuss them together as both rely on accurate PM2.5 observations. Risk assessments use CRFs from epidemiological studies, so the difference in how concentration is determined in the original epidemiological study versus how the concentration is determined in a risk assessment using that CRF could be important, and the difference in the population included in the study (gender, age, socioeconomic status, etc.) versus the whole population is likely very important. While monitors can be used to estimate population-level exposure, health effects are still an individual response (specific individuals with specific characteristics died or had an asthma attack) and the available health data may not be representative of the whole population (if there is some sort of bias in the percent of susceptible people or confounding variables not accounted for) but the response is still aggregated to be applicable to the whole population when a relative risk is determined. We have clarified this in the text.

p. 25332, l. 8 – "both" is ambiguous here since you've just talked about monitoring, satellites, and models.
We removed "both of"

p. 25334 top – it's not clear to me whether one value is used for the whole US or if different values are used in different states. If the first, then why is it important to start with state level data and use gridded population?
Different values are used for different states. We have clarified this in the text.

p. 25336 – The authors are correct that different studies use the terms linear and log-linear in different ways. But the discussion here doesn't quite clarify how the authors are using these terms.
These are both log-linear and are now referred to by the equation number later in the text.

p. 25338 top – what is the source for emissions for the rest of the world?
EDGAR and GEIA for anthropogenic emissions, but that is overwritten by regional inventories, such as BRAVO for Mexico, CAC for Canada, EMEP for Europe. We have added these references to the text.

p. 25340 bottom shows model performance compared to IMPROVE and AQS. In contrast, p. 25341 top discusses AERONET AOD, but presents no model performance evaluation.
The goal of satellite-based PM2.5 estimates is to improve surface PM2.5 estimates, not AOD (although inherently the method assumes that the two scale the same). As stated, we use AERONET to discuss the uncertainty in satellite AOD and have added a figure to show this.

p. 25341, l. 23 – what does "initial fraction" mean?
This should be "attributable fraction" and has been changed in the text.

p. 25346, l. 12-15 – This discusses how compensating errors may be hidden by NMB. For that reason, it is common to also present Normalized Mean Error.
Yes, compensating errors can be hidden by the NMB and we have also investigated the NME. However, as discussed in the text, we use NMB here as appropriate for errors in annual mean values (since these are used for chronic exposure).

p. 25352 bottom – this long discussion of low thresholds might seem more appropriate to present in methods (there is some discussion there) or in a discussion sector.
We have removed this paragraph from this section and moved some of it to the methods section.

p. 25354 – Is this the first time Figure 9 is referenced? I find it a little strange to present a new figure in the conclusions section.
We moved this figure and associated text to section 4 following the suggestion of another reviewer.

p.25355, l. 14 – "correctly applying response functions is a determining factor" Are the authors claiming that some previous studies have done this incorrectly? I would think that there may be better or worse choices to make, but that authors may have reasons for choosing the approach that they do. I also wouldn't call these "epidemiological approaches" because these are risk assessment studies, not epidemiology.
Thank you for pointing out this potential misinterpretation of our text! We have removed the words "correctly" and "epidemiological."

p. 25355, l. 17 – "using only populated places" I don't understand what this means. There should be no health effects in unpopulated places.
The populated places data set gives values for a point location rather than a grid and therefore has values for all major cities and town, but only some of the smaller towns in sparsely inhabited regions. We clarify this in the text.

Table 1 – I'm surprised that uncertainty is shown for only one study – certainly at least some of the other studies also estimated uncertainty. Also, Punger & West 2013 estimated US PM2.5 burden. Zheng et al. isn't listed in the references.
We didn't include the confidence intervals, just the different estimates from different CRFs. We have added Punger and West, 2013 and a reference to Zheng et al. 2014.

Figure 2 is pretty small and difficult to read. Is it true that all of these studies are chronic PM2.5?
We believe that this is largely an issue related to ACPD formatting, however we have made the font bigger and will ensure legibility in the final version. Yes, these are all for chronic exposure as stated in the Figure caption.

Figure 8 & 9 – units should be "deaths per year"
Fixed.