Response to Anonymous Referee #2:

We appreciate the review offered by Anonymous Referee #2 (AR2) and currently we are working to address all of AR2's general and specific comments. We agree that the paper is too long, and should be presented more concisely. We also will add more specific discussion in the manuscript that offers recommendations for related modeling and satellite efforts in the future. For instance, we plan to: include better recommendations for wildfire VOC emission factors used in BlueSky; discuss the differences expected if the newest version of SMARTFIRE and BlueSky were used; recommend air quality satellite products that could be developed for the community, to better refine detection of wildfire events; and offer insight as to what might be gained from future satellite programs such as TROPOMI.

General Comments

We will include a flow chart to describe the model operations.

We will summarize the results more succinctly in text, table, and graphical format. The number of tables will be cut nearly in half and we also plan to reduce the number of total graphics in the main manuscript. Results with similar patterns will be summarized together (e.g. similarities across the two years).

We will better explain the methods used to determine the categories:
observed but not predicted; under-predicted; predicted well; over-predicted;
and predicted but not observed. This part of the analysis used manual review of air quality comparison maps (e.g. between model and satellite) on a day by day basis, which was necessary as part of the QA/QC process, since satellite data can sometimes have erroneous data that passes automated checks.

Specific Comments

"Abstract- It would be nice to see a sentence or two on how the modeling could be improved to better simulate wildfires in the future."

The revised abstract will include recommendations for better wildfire VOC emission factors, based on recent literature.

Lines 225-226. What were the criteria for deciding that the MODIS retrievals were "high quality"?

All MODIS AOD retrievals come with a quality assurance flag that splits retrievals into quality categores. We used the combined "Land and Ocean" product which utilizes AOT at 0.55 micron for both ocean (best) and land (corrected) with best quality data (Quality flag=3). This will be explained in the revised manuscript.

Line 311. It is not clear what a "VFM curtain" is, please elaborate.

VMF curtains are explained in the previous paragraph. Part of our analysis included vertical feature mask plots, but we subsequently determined that the CALIOP algorithm does not accurately predict aerosol subtype. Thus we did not include the aerosol subtype vertical feature mask plots in the manuscript. The VFM product, though, was subsequently used to derive

aerosol plume top height comparisons. We will constrain our discussion of VFM curtains in the methods section, removing unneeded details, and making sure to succinctly revise the plume top height comparison methods. This should help to remove any confusion about the VFM product and how it was used in our final analysis.

Lines 346-347. Often negative values, while not physically possible, tell us how precise a measurement is. I assume that "screened" means that negative values were discarded. Does this skew the comparison?

We calculated statistics with a variety of screening methods (including keeping the negative values) and found that the only the fractional statistics were affected due to this method used. This study is focused on relatively short-term pollution events, so we feel it is appropriate to simply discard negative MODIS values. We will further justify this choice in the revised manuscript. Considering information on the MODIS aerosol website, discarding the negative values shouldn't have much effect on our results. A relevant quote from the MODIS aerosol site http://modis-atmos.gsfc.nasa.gov/MOD04_L2/format.html : "Note: We are permitting small negative Aerosol Optical Depth values in order to avoid an arbitrary negative bias at the low AOD end in long term statistics. This is because MODIS does not have sensitivity over land to retrieve aerosol to better than +/-0.05. This means in very clean conditions the algorithm cannot determine if the AOD = 0, 0.05 or -0.05. If we eliminate all the negative numbers and keep all the positive numbers, we introduce an artificial bias. Thus, we allow negative retrievals up to -0.05. To interpret these: If you are calculating long-term statistics, simply add the negatives into the mix and don't worry about them. If you were looking at individual retrievals then count negative retrievals as 'very clean'. You could force them to be AOD = 0, for example. It really depends on the application."

Line 420. What does the term "under-biased" mean? It is unclear to me.

The phrase "AIRPACT was under-biased" is unclear. It will be changed to say "AIRPACT was biased low" in the revised manuscript.

Lines 427-428. What is a "matched-threshold analysis"?

The "matched-threshold" analysis is explained in the methods on Lines 359-364. We will make sure to reiterate the thresholds used when reporting these results to make their meaning more clear.

Lines 555-557. This sentence doesn't seem to make any sense, I can't tell what is meant here.

We feel that the way this section was written is confusing and difficult to follow. This will be revised in the next version of the manuscript, including simplification of both the text and table. We intend to replace this section with wording more along the lines of what is written below:

"The AIRPACT-3 daily 24-hr PM2.5 performance was assessed from a policy standpoint for both the daily (35 $\mu g/m^3$) and annual (12 $\mu g/m^3$) National Ambient Air Quality Standards (NAAQS) threshold values. For each site, we calculated the number of days when both the model forecast and the observations showed $PM_{2.5}$ concentrations greater than the NAAQS. We tallied the number of these days during the analysis period, for 67 sites in 2007 and 82 sites in 2008. For the FEPS plume-rise scenario we found: 0.2% of the data pairs were in agreement, with values higher than the daily threshold; 0.3% of the data pairs included observations higher than the daily threshold, with no such model prediction; 1.8% of the data pairs included model predictions higher than the daily threshold, with no such observation; and 97.7% of the data pairs were in agreement, with values less than the daily threshold. The SMOKE plume-rise scenario reduced

the number of model predictions that were higher than the daily threshold, with no such observation, by 27% (or 1.3% of the total data pairs).

For the FEPS plume-rise scenario we also found: 1.8% of the data pairs were in agreement, with values higher than the annual threshold; 4% of the data pairs included observations higher than the annual threshold, with no such model prediction; 3.5% of the data pairs included model predictions higher than the annual threshold, with no such observations; and 90.7% of the data pairs were in agreement, with values less than the annual threshold. The SMOKE plume-rise scenario increased the number of data pairs that were in agreement, with values higher than the annual threshold, by 17% (2.1% of the total data pairs). Further details of the $PM_{2.5}$ NAAQS comparison are in Table 8, including separate details for 2007 and 2008. The location and numbers per site are shown for the FEPS scenario in the supplemental materials."

Lines 644-645. In all the previous section the comparison text has been AOD, NO2, CO. Don't change it here, it will just confuse matters.

We agree, and will make the necessary revisions.

Conclusions and Future work: It would be nice to have the authors opinion on whether current emissions inventory are adequate for regional modeling of wildfire, or whether, and what improvements are needed. The CO data would seem the most applicable for this purpose. Does the consistent under-estimating by the model imply that the inventories are low? Does this problem with low inventories account for some of the under-estimating of particle mass, hence AOD?

We intend to include our opinions that the VOC emissions factors in BlueSky are low (and old), so they should be increased to reflect more current literature values. This will increase the AOD in the model due to SOA. We agree that CO gives a nice estimate of the "emissions inventory" but feel that the small under-estimations in CO results could be due to a variety of things including satellite retrieval errors and inconsistencies in model parameters such as fire locations, fire size, fuel moisture, fuel loading, heat content, or plume rise. Since the AIRS CO retrieval imparts different sensitivities vertically throughout the atmosphere, something as simple as a small adjustment in plume rise parameters can have obvious affects on the bias results. Furthermore, the once daily retrievals of CO do not allow an accurate validation of overall emissions, especially since fires emit most of their pollutants later in the afternoon. Our opinions on these details will be further explained in the next version of the manuscript.

Figures- All the maps (Figures 2-7, and S1-10) should show the location of MBO on at least one panel.

This will be included in the next manuscript version.