Response to Anonymous Referee #1:

We are currently working to address all of the general and specific comments made by Anonymous Referee #1 (AR1). We appreciate all of the detailed comments provided in AR1's review. We agree that the paper should be presented more concisely and more discussion connecting the topic to other related work should be included.

General Comments:

We agree that the paper is in need of a major revision so that our work is more clearly stated and connected to other related work in the community. We will include some examples of the wildfire events in more recent history, and discuss the broader significance, as you have suggested.

Why 2007-8?

AIRPACT-3's lifetime was from 2006 to 2013, and this work was funded by a proposal written in 2011, that agreed to look at AIRPACT, OMI, and other results from 2007-2008. That period was chosen because it is before the OMI row anomaly occurred, which greatly reduced the number of available NO2 retrievals in 2009. The period was also of interest because there were especially large fire emissions those years.

Why AIRS and not other CO?

We chose to use satellites that were part of the A-Train, so that all useful and available data taken at the same time could be assessed together. In our preliminary analysis we also included MOPITT CO and MISR plume top heights. However, the morning-time overpass and limited spatial coverage were not ideal for a wildfire analysis.

Why not AERONET?

There are studies that validate MODIS retrievals using AERONET, but we did not choose to do so ourselves. The large spatial coverage offered by MODIS was of more interest than the limited number of AERONET sites in our model domain. We feel that MODIS AOD has enough accuracy for the scope of this project.

Why not O3 from OMI satellite?

We have used research-grade OMI tropospheric ozone and operational AIRS tropospheric ozone to assess model performance of long-range transport events. Determining tropospheric ozone from space is a notoriously difficult retrieval to make with very large uncertainties. Furthermore, the large concentration of aerosols during wildfire events severely impacts the ability of OMI to retrieve useful and accurate information about tropospheric ozone.

Any ideas for future use of VIIRS, GOES-R, or assimilation?

We feel that AOD retrievals from space are relatively reliable, compared to trace-gas retrievals. The satellite community will have a lot of options for AOD moving forward. We include MODIS AOD results on our forecast website, and will likely add VIIRS soon.

With the inherent time-lag of obtaining satellite data, operational forecasts with assimilation can only be done for "yesterday's simulation." We realize that some performance benefit may be gained by rerunning the previous day's simulation with assimilation, providing better initial conditions for the true forecast. However, at this time, the AIRPACT community would rather spend any extra computational time on forecasting further into the future, rather than rerunning the past. When reliable geostationary satellite retrievals of air quality become available in the future, investments to include satellite assimilation into the AIRPACT forecast will be of much greater performance value.

The uncertainty in satellite retrievals needs to be factored into the overall philosophy.

We agree, and will include more details about the satellite retrieval uncertainties in the revised manuscript.

It should be clarified if the model runs discussed here are the original forecast (it seems not) or a reanalysis with improved fire data that became available later (my current understanding).

These simulations use the finalized fire data that comes out a few days after the fact.

Is the meteorology the original forecast or was the model run again with the "actual" meteorology?

These simulations use the original meteorological forecasts.

Specific Comments:

(Items highlighted with no response were accepted into the revised manuscript)

P2, L3: "a suite"

P2, L25-27: Are these "biases" significant given reasonable estimated uncertainty in the remote sensing products?

This will be discussed in the next manuscript revision.

P3, L5: Wildfires are not just forest fires. Much or most of the PNW is grassland, which also has large fires.

Changed to "rural landscape".

P3, L7: change "respiratory" to "health" since cardiovascular impacts actually dominate.

P3, L7: I don't think the goal is to "alert" people that the AQ is bad, but too forecast bad AQ ahead of time.

Changed to "Informing the public about upcoming poor air quality expected from fires …"

P3, L14: Maybe change "potential health" to "air quality" - the actual health impacts from a given air quality adds another much larger layer of uncertainty.

P3, L15: not just PM so maybe the text after the comma should just be: "but the task is challenging."

P3, L16, Column measurements from space are useful to compare with models, but they have uncertainties and because they are column measurements there is really no such thing as satellite retrievals of AQ yet.

See Crumeyrolle, S., Chen, G., Ziemba, L., Beyersdorf, A., Thornhill, L., Winstead, E., Moore, R. H., Shook, M. A., Hudgins, C., and Anderson, B. E.: Factors that influence surface PM2.5 values inferred from satellite observations: perspective gained for the US Baltimore–Washington metropolitan area during DISCOVER-AQ, Atmos. Chem. Phys., 14, 2139-2153, doi:10.5194/acp-14-2139-2014, 2014.

Surface measurements are where the people live, but the column satellite data is useful to "connect the dots" between surface observations and evaluate overall model performance. I'd maybe express this as something like: "Satellite-based column measurements enhance the coverage available from surface networks and are useful to evaluate model performance."

Changed to "air quality indicators".

P4, L5: change "led to" to "combined with" don't think a dry spring causes a summer Drought

The severity of summer droughts is definitely connected to lack of precipitation in earlier seasons, since the soil is not recharged with moisture before the hot weather ensues. Fuels are much drier as well, so fire seasons can start much sooner. Changed to "led into".

P4, L18: What is meant by "the south" and should ID/MT also be exceptions given the text on L23?

Changed to "the southern U.S."

P4, L20: Would ARCTAS CARB data be of any value in AIRPACT evaluation for 2008?

This recommendation is not within the scope of this project.

P5, top: It's not necessary to name all the fires here or in Figure 1.

We feel that the timing of these events are of interest to some readers. This paragraph will be shortened, but we do not agree that removing labels from Figure 1 (which corresponds to Table 1) is necessary.

P5, L15: I would just show all the burned area in Fig 1 with no names since several smaller fires could be just as important as one big one.

We have given the names of large fire complexes, which include multiple smaller fires. We have chosen to keep the fire complex name labels, since they add linkage to the rest of the paper.

P5, L24: Why project to 2005 instead of 2007/8? Could you evaluate the EGAS software by projecting and then comparing to the 2011 NEI?

AIRPACT undergoes periodic emissions inventory updates in collaboration with local, tribal, state, federal, and international agencies. This project is not intended to evaluate anthropogenic emissions and this recommendation is not within the scope of this project.

P5, L26: change "over" to "from" or say "Canadian anthropogenic emissions are : : :"

P6, L2-3: Maybe a word or two to clarify what is meant by "processing" emissions?

Changed to "spatially and temporally allocated".

P6, L28: change (jargon) "ICS-209" to "fire"

P7, L2: clarify "well"

This will be clarified in the next manuscript revision.

P7, L4-6: Here and in general. This sounds like a partial re-analysis – in other words, not testing the original forecast, but testing an improved forecast using updated fire info, but still with the old meteorology? It should be clear what was done and justified why. It would be of interest to know the accuracy of the original operational forecast. From the broader perspective how does actual vs original fire change in magnitude, location, timing, and how does that impact the modeled results? Also, how are fires forecast? In other words SMARTFIRE incompletely tabulates past fires if I understand right. Is that partial fire activity assumed to persist to generate a forecast? A sentence could clarify this.

This will be discussed in the next manuscript revision.

P7, L7-16: It's my understanding that none of these models have ever been validated, but in any case, the extent to which they have should be provided. For instance, on line 14, combustion "phases" are referred to, which don't actually exist on real fires that burn with a mix of flaming and smoldering.

This will be discussed in the next manuscript revision.

P7, L16: "short-lived" fuels makes no sense.

Changed to "fast-burning".

P7, L19-20: How about just saying the 60% is fixed in the model, but real fuel consumption can vary about the nominal value?

This will be clarified in the next manuscript revision.

P7, L27-8: It doesn't seem to make sense to release all smoldering emissions into surface layer when it is well-known that smoldering emissions are entrained into convection columns and can go to any altitude the column does. I guess the paper sort of verifies that, so OK.

We understand that the FEPS plume-rise scenario is not ideal in this sense. However, it is one of the two available methods to simulate fire plume rise using the SMOKE processor. Therefore, we feel it provides value to the paper to include the results of both standard model pathways.

P8, L10-11: Change "most" to "much" Aqua retrievals are useful, but they are only offered in areas with no clouds and not so much smoke that the cloud mask thinks it is a cloud. Retreivals with estimated uncertainty above a threshold are rejected, but the remaining ones are known to be biased low compared to AERONET and MISR: T. F. Eck, B. N. Holben, J. S. Reid, M. M. Mukelabai, S. J. Piketh, O. Torres, H. T. Jethva, E. J. Hyer, D. E. Ward, O. Dubovik, A. Sinyuk, J. S. Schafer, D. M. Giles, M. Sorokin, A. Smirnov and I. Slutsker, A seasonal trend of single scattering albedo in southern African biomass-burning particles: Implications for satellite products and estimates of emissions for the world's largest biomass-burning source, Journal of Geophysical Research: Atmospheres, Volume 118, Issue 12, 27 June 2013, Pages: 6414–6432, DOI: 10.1002/jgrd.50500

This will be clarified in the next manuscript revision.

P8, L19: I'm not questioning the decision to re-grid by grabbing closest value instead of e.g. weighted averaging or more complex re-mapping, but if possible one should estimate the additional contribution of this step to overall uncertainty?

The MODIS L2 AOD gridding is slightly higher resolution than AIRPACT-3. Weighted averaging or other re-mapping schemes are not necessary here, since weighted averaging gives nearly the same result. This will be clarified in the next manuscript revision.

P8, L25: Is it clear that between the uncertainty in re-gridded MODIS AOD and the model calculated AOD that a statistically significant comparison results? It is of interest and value to report and discuss comparisons even if they are not technically "significant," but it would be helpful to be able to compare uncertainties to biases etc.

The re-gridding of MODIS AOD to the AIRPACT-3 grid does not really introduce any noticeable uncertainty. Model calculated AOD does have large uncertainties. This will be clarified in the next manuscript revision.

P9, L7: Section 2.3 If AIRS CO is one of the many CO products with low sensitivity to boundary layer CO that should be mentioned. One could consult the Kopacz et al 2010 ACP paper for an idea of the accuracy of individual CO products as opposed to combining them all.

Nearly every satellite retrieval of air quality indicators has low sensitivity to the boundary layer. We are not combining any CO satellite products other than AIRS for evaluation, and MOPITT for assimilation into the MOZART-4 simulations (used for boundary conditions). AIRS CO sensitivity will be further clarified in the next manuscript revision.

P9-10: Section 2.4 does well to mention that uncertainty occurs due to a potentially inappropriate a-priori vertical profile and air mass factor effects for the OMI NO2. In addition, the effect of all the data massaging, and the fact that NO2 is rapidly converted to PAN and nitrate in fire plumes, which may not be represented correctly at all times in CMAQ could be mentioned
http://www.atmos-chem-phys.net/12/1397/2012/acp-12-1397-2012.html
http://www.atmos-chem-phys.net/10/9739/2010/acp-10-9739-2010.html

The AIRPACT-3 uncertainties in NO2 comparisons will be clarified in the next manuscript revision.

P10, L27-28: I think the aerosol typing depends on depolarization ratio, but not sure about the attenuated backscatter, altitude, location, and surface type. In general, for the remote sensing instruments, there is too much basic info on things like launch date and principle of measurement and not enough on accuracy and coverage.

This information was obtained from the CALIOP references. More information on accuracy and coverage will be clarified in the next manuscript revision.

P11, L24: It makes sense to compare the plume rise above the deresolved elevation in AIRPACT to the more accurate plume rise a.g.l. measured by CALIOP, but even this is tricky if the terrain is not flat or for whatever reason the actual plume height (a.g.l. or m.s.l) is not constant. But I suspect the CALIOP plume heights are one of the more exact comparisons possible though some vertical uncertainty could be estimated.

CALIOP uncertainty will be discussed in the next manuscript revision.

P12, Sect 2.6: If archived, the GOES visible and "fire channel" is very helpful for understanding fire timing or the temporal profile of emissions.

Unfortunately, GOES archives are not always readily available. Furthermore, this study does not attempt to assess the standard temporal profile used by SMOKE/CMAQ modelers.

P12, L24: Is it better to say Canada is outside the AIRPACT domain rather than AIRPACT has no fire emissions in Canada? Canadian fire emission can impact the US AIRPACT domain and are ostensibly provided by MOZART in the boundary conditions.

Part of Canada is within the AIRPACT domain, for which there are no emissions in this project. Fire impacts from MOZART are largely represented in the CO values, but not well for aerosols. This will be discussed further in the next manuscript revision.

P12, L25: If AIRPACT simulated a doubling or more of PM2.5 and a data set showed no increase during "fire events" - this seems like a case that is important to include in the comparison. It also seems like these events are discussed later in the paper? Or is that only at sites where these events were occasional rather than universal?

As noted in the manuscript: "The ground-site analysis presented here uses combinations of 140 U.S. surface monitor locations where AIRPACT-3 predicted more than double the normal surface PM2.5, as a result of wildfire emissions." All surface monitor locations chosen showed at least some evidence of fire impact in the observations. This will be clarified in the next manuscript revision.

P13, L1-2: This doesn't make any sense to me. Aren't you comparing the model produced column to the satellite product column (in Fig 2-7) and why would that be restricted to places with surface sites?

As noted in the manuscript: "The primary analysis of AOD, tropospheric column NO_2, and total column CO includes all 140 site locations. A secondary rural-sites-only subset includes 43 locations with no influence from transported urban pollution in the remote sensing record." For the purpose of generating model performance statistics, we decided to assess model performance at the discrete site locations rather than across the entire domain. This was done so that surface monitor observations and satellite retrievals could be compared more consistently, and so that the randomness of the location of usable retrievals did not skew our results spatially or with urban signatures. This will be clarified in the next manuscript revision.

P13, L5-8: Interesting and useful idea to select and compare separately only the cases where both model and satellite are strongly elevated. The restriction could conceivably inflate the degree of agreement, but it also potentially selects for higher S:N and lower uncertainty in the satellite product!

P14, L3: Here is one of many places where I wonder if AIRPACT really "underpredicted" or was AIRPACT actually similar within combined uncertainties, or if the satellite over-predicted, etc. It seems better to just consistently refer to differences (like in the figures) or offsets rather than imply value judgments, except maybe against CALIOP? Also there should be some definition here that is relevant throughout the text that specifies what you mean by "agreed well" as opposed to under/over-predicted? E.g. is within +/- 20% OK? Good work, but deserving of more precise terminology.

The remote-sensing daily log lumps things into very broad bins (e.g. +/- 40%). This was done by manually comparing maps so that differences due to location and obvious satellite errors would be avoided. This will be clarified in the next manuscript revision.

P14, L5-7: This is a good example of a difference that is not an "over-prediction" by the model, since the modeled fires really happened. Thus, this work is also simultaneously evaluating the remote sensing products.

We are not intending to evaluate the remote sensing product here.

P14, L9: In light of above; "performance" might be better as "agreement" and over/under prediction as "differences."

The daily remote sensing log is meant to serve as a qualitative assessment of when the model did ok, missed fires, or had some level of under or over-prediction. We hesitate to use quantifiable terms such as "differences" for such a qualitative assessment.

P14, L9-10: Were fire locations predicted? Or were they modeled retrospectively? Not sure what is being done here. Were the previous day's fires from SMARTFIRE assumed to persist? If the fire popped up after the satellite, how was it predicted despite the satellite missing it? It seems like you are referring to a model run done after the fires using the updated fire information.

As noted in the manuscript: "The fire reports used in this analysis are from the final SMARTFIRE archive, as distinct from the information reported in near real-time, which can often be incomplete." This is the "updated fire information" with no spin-up emissions and no persistence used in BlueSky.

L10 What is meant by "intensity"?

Replaced with "air quality impacts".

P14, L15-17: So this is interesting once you've defined what you mean by: 1) "event," 2) how you compare to an "event," and 3) "well". And "large" could be replaced with a ">X km" definition? It seems 100 km is adopted later?

This will be clarified in the next manuscript revision.

P14, L26-27: High model NO2 values could result from the model lifetime being too long in some plumes?

In our experience, the high NO2 values seen in the model results are largely due to the affect the averaging kernel has.

L28 "under-biased" = "biased low"?

P15, L1-3: Use fractions or percentages, but don't mix in same sentence; and what is the significance of "140 sites" for column and model data??

"Fractional bias", a standard air quality model performance indicator, is represented as a percentage, as stipulated in Table 2. As noted in the methods section, the original site locations used started with 140 sites where AIRPACT showed at least double PM2.5 during a fire event.

P15, L3: If NO2 was 39% low on average despite being "over-predicted" 48% of time, then it seems there must have been a few massive "under-prediction" events?

Comparing the qualitative remote sensing log results to the quantitative statistics may have some value and will be discussed in the next revision of the manuscript.

Table 1a footnote: misspelled "source". Regional totals more useful that USA totals.

This was somehow changed during typesetting. We include USA totals so that individual state totals can be put into context of the entire nation's fire year. We do not have totals of fires strictly in the AIRPACT domain from the NIFC, as state-level reporting is the finest available from this source.

Table 1b title L4: "approximate" to "approximation" and, in general, what is meaning of a range of ignition dates?

Some fire complexes were not ignited on one single day (e.g. a storm system with lightning strikes may occur over a period of a few days), as was the case for many of the ID/MT fires that ignited in 2007.

P15, L6-9: Noting that the mismatches for AOD and NO2 are larger when there is more "signal to noise" but CO still seems to agree "pretty well."

This will be clarified in the next manuscript revision.

P18, L7: Fires can entrain surface soil, dust, and ash into the convection column along with the smoke. L11: Missing a word?

Discussion of VFM results will be omitted from the revised manuscript.

P18, L19: Low O3 due to the CMAQ-SAPRC chemical mechanism might be expected (Alvarado and Prinn 2009 in JGR).

This will be discussed in the next manuscript revision.

Page 20-21: In comparing to MBO PM several factors suggest the AIRPACT-modeled SOA is too low (e.g. the usual lower modeled-AOD in plumes longer than 100 km). However, the MBO PM is based on scattering and scattering can sometimes increase without an increase in mass, most likely due to a change in the size distribution (Akagi et al 2012 link given earlier). In general SOA is highly variable and poorly understood (Vakkari et al., 2014 in GRL).

This will be discussed in the next manuscript revision.

P21, L21-22: I don't expect the global model to capture spikes. Is it really possible to differentiate between how well the fire emissions are represented in MOZART and how transport, the chemistry mechanism, or resolution effect the comparison? If a problem with the MOZART emissions can be demonstrated it should be described at least semi-quantitatively as a problem with amount, timing, or whatever it is rather than saying "poor."

This will be discussed in the next manuscript revision.

P22, L15: It would be interesting to try some model runs with the plume rise offset consistent with the CALIOP-measured underestimate. The CALIOP/AIRPACT plume height comparison is not highly correlated, but this could still be tried. Are there plans to fix plume rise by scaling, etc?

AIRPACT-3 has been retired with no plans of further model runs for this project.

P22, L27: This is the first mention of overestimation of plume height a.g.l. Is this referring to just a handful of cases?

More information about these cases will be mentioned in Section 3.3 and readers will be directed again to Figure 8.

P23, L12: change "when detecting" to "near many" since OMI doesn't "detect fires" and some fires do inject emissions near the surface.

Changed to "over".

P23, L11-16: A number of recent papers take the OMI NO2 retrievals in smoke as an accurate basis for global NO2 emissions estimates and so is this accuracy being disputed?

While the NO2 column over clouds or high albedo smoke does enhance the signal to OMI, the resulting retrieval reflects conditions above the plume, not within the plume where most of the pollution exists. This will be discussed in the next manuscript revision.

Also is the AIRS a-priori CO profile any better suited for fire conditions? AIRS-CO seems consistent with AIRPACT.

CO is a much easier pollutant to track from source, since it is relatively long lived and can travel further than the high density aerosols. This will be discussed in the next manuscript revision.

P23, L18-19: Does "but there were often similar estimates of column CO over active fire regions" mean good agreement on column CO "often" occurred between AIRPACT and AIRS above fire locations. Also, are there any useful surface observations of CO?

In general, AIRPACT CO performed well when compared to AIRS CO and surface co at MBO. The details of CO performance will be further discussed in the revised manuscript.

P23, L19: By "The AIRS retrieval is not sensitive to the surface" do you mean literally that it is not affected by land cover type, or that it has low sensitivity in the boundary layer, or something else?

Changed to "CO near the surface".

P23, L27-28: Probably good to cite some papers relying on more advanced measurements of PM that find SOA is highly variable: e.g. from none at all to a factor of four (Jolleys et al. 2012 in ES&T; Yokelson et al. 2009 in ACP; Vakkari et al. 2014 in GRL).

This will be discussed in the next manuscript revision.

P24, L5-9: This seems like an important result and should probably be developed/ integrated into text and tables more fully, rather than appearing almost as an afterthought at the end of the paper.

This iteration of model results (physically allocating all smoldering emissions into the plume) was a test to see if our high over-prediction spikes would be solved, which it did, but it is not a supported model development method and does not treat plumes accurately (buoyancy not constrained). This will be discussed in the next manuscript revision.

P24, L15: Do you actually mean that some of the fires in the historical SMARTFIRE database don't exist?

Changed "completely absent" to "were missed" to indicate that SMARTFIRE misses some fires that occurred.

P24, L20-24: How would complex terrain or cloud cover cause SMARTFIRE to miss fires when it includes the ICS-209s? Maybe wilderness fires that no report is filed on?

ICS-209 reports can be missing some details during large fire seasons when firemen are busy in the field. SMARTFIRE is highly supplemented by HMS, which does a good job of detecting fires, but finite satellite resources cannot detect all fires in all conditions. As such, it is one source of uncertainty, especially since HMS detects are given a default fire size. The revised manuscript will include some literature review on this subject.

How would a lack of dead woody fuels cause a fuel loading underestimate? Maybe change "that completely lack dead woody fuels" to "for which dead woody fuels are omitted"

Grassland and shrubland in FCCS/BlueSky (which include vast areas within the AIRPACT domain) allocate ~2.5 tons per acre of grass fuels, and no dead woody fuels. However, the terrain in some of these areas is not completely void of dead woody fuel. In short, the FCCS map and classifications of fuel loading are not a completely accurate representation of fuels. Changed to "have sparse woody fuels but are classified with zero dead woody fuels in the FCCS".

P24, L27-P25, L1: By under-predicted emissions, do you mean total emissions as opposed to certain species? Why would emissions scale with plume heights? Is buoyancy assumed proportional to amount of fuel burned?

The heat content of a fire location is directly proportional to the total fuel consumed. At least, that is how it is modeled in BlueSky. This will be clarified in the next manuscript revision.

P25, L3-20: This reads like confident conclusions about the benefits of model changes that were not discussed in the paper or tested except for number 4, and it omits the thing you did demonstrate needs fixing: the plume height. Suggest presenting this as a list of additional (in addition to plume height) future avenues to explore for potential improvement.

This list will be clarified and separated into two distinct contexts: recent revisions to the BlueSky framework that address some of these issues and lessons learned from the work discussed previously in the paper.

Table 2: Shouldn't the formulas for percentages require a 100 instead of a 1 as first number?

We could insert a " x 100%" after these respective formulas, if deemed necessary.

Why is the bias and error sometimes computed with respect to the observation and sometimes with respect to the mean of the model and observation? If the model and observation are equally valid then the concept of model performance or "under or over" prediction throughout the text seems less meaningful.

These standard model performance statistics are used in many air quality model evaluation studies.

In Table 2, the normalized quantities are defined as percentages, but then not used as percentages in Table 3.

Table 3 and other similar tables will be updated to report normalized quantities as percentages.

Table 4, Title: I thought both the satellite and AIRPACT data are rural only in this comparison. How about instead of "performance" in sentence one and including sentence two, just say "Summary of matched threshold comparison limited to polluted rural sites for 3 July : : :."?

The surface monitor results include all of the originally listed sites. The satellite results included in the matched-threshold analysis are a "rural" subset.

Figure 1: could be better without fire names.

We would prefer to keep the labels for reference, but we have pasted the figure without labels below.

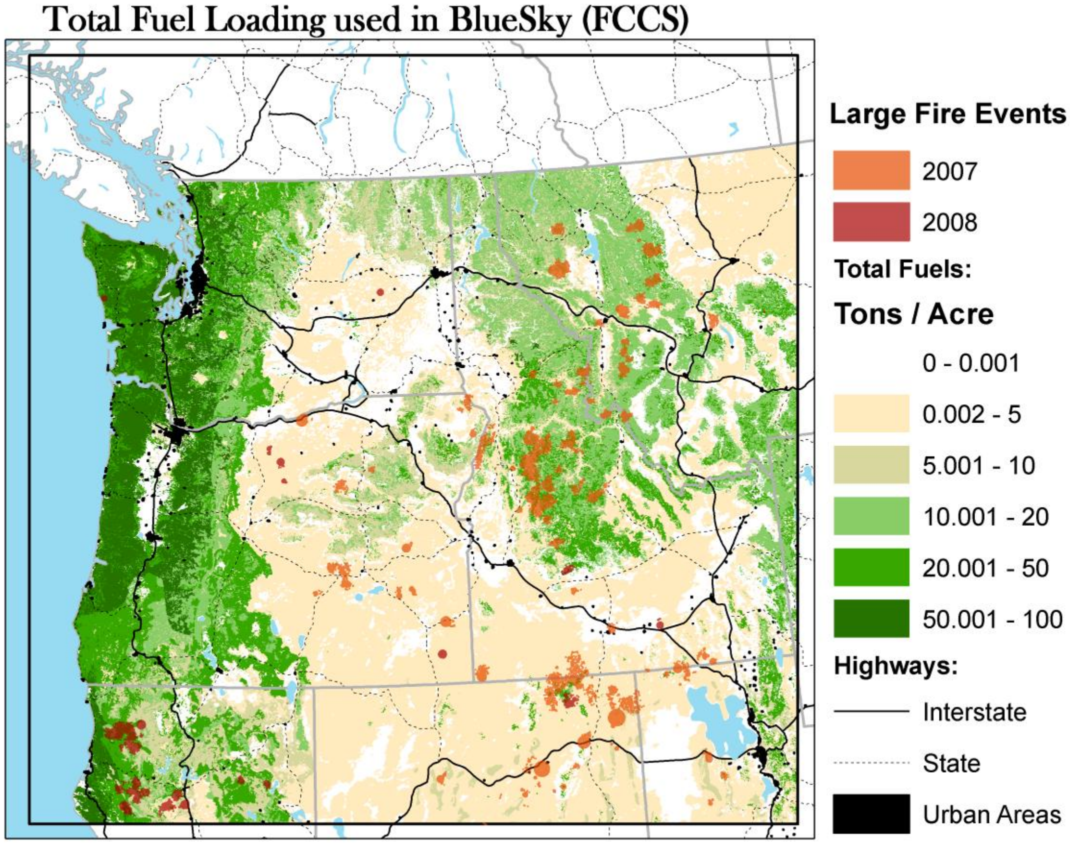

Figure 2 caption, L4: “and” before “exclusion (also in rest of similar figures).

Figure 9: SMOKE model seems to reduce false positive events. I thought this was showing a single site at first and now wondering if the PM spikes that occur even when averaging over all the sites are due to massive modeled impacts at a few sites closer to fires?

As noted in the caption, this figure of daily 24-hr averaged PM2.5 (and ozone) is averaged across all sites. Yes the spikes do occur when the model makes very large over-predictions in a general fire impact area. The SMOKE plume rise algorithm mitigates this since smoldering emissions are allocated across a few layers close to the surface. This is in contrast to the FEPS plume rise algorithm which puts all the smoldering emissions in model layer 1 and can sometimes result in unrealistically large surface concentrations.