Reviewer #1 – This is a very careful, well-thought and reasoned reading of our paper, which has led to many changes for clarity and method. Thank you for your patience and thoughts.

The paper presents comparisons of CALIOP AOD retrievals with NAAPS model results after assimilation of MODIS and MISR AOD data, with focus on night-time observations. Global as well as regional datasets for the entire year of 2007 are discussed. Furthermore, comparisons with ground-based lidar observations in the Caribbean are shown. The authors find significant discrepancies between measured and modeled AOD values as well as between daytime and night-time CALIOP retrievals. The paper is based on a careful data analysis and thus is a valuable contribution to the CALIOP validation efforts. However, the interpretation and discussion of the findings could be significantly improved if the authors went more into the background of CALIOP extinction retrievals, namely the aerosol-typing issue and its consequences. Here, it would also be useful to relate the findings better to other validation and comparison studies, e.g., with ground-based observations at different sites, available in the literature.

MAJOR COMMENTS

1) *The authors compare CALIOP and NAAPS data based on a differentiation of surfaces (ocean, coastal, land) on a global scale and on the selection of specific regions. This approach is probably driven by procedures for passive remote sensing where the surface plays a dominating role in the retrievals. However, as the authors correctly state, CALIOP extinction and AOD retrievals strongly depend on the proper selection of the lidar ratio and thus on the aerosol typing. A respective discussion is missing almost throughout the paper. Only at the end of the manuscript consequences of the aerosol-typing approach are tackled in the specific comparisons with ground-based measurements in the Caribbean (see Fig. 10). To my feeling, a much better reasoning of the obtained discrepancies could be given by rigorous investigation of differences in dependence on aerosol type used in the CALIOP retrievals in conjunction with a critical discussion of the selected lidar ratios (see also comments below).*

Thank you for this critical comment. We agree. However, we ask the Reviewer to recall specifically that this basis for this paper is the nighttime retrieval evaluation, and thus not necessarily a mechanistic study of algorithm engineering as opposed to how robust their performance actually proves. It is absolutely true that for the case of the Caribbean we do quickly talk about the specific impact of the aerosol model on the fractional contributions to total AOD. However, this was done as context for interpreting the vertical day/night profiles themselves relative to MPLNET. We realize that this isn't necessarily a satisfying answer to the reviewer. Please consider, though, a brand new publication that many of us have just collaborated on investigating CALIOP retrievals specifically over Southeast Asia:

http://www.sciencedirect.com/science/article/pii/S0169809512001366

There, in fact, we do break the aerosol models out into regional contributions to total AOD, and identify where there are apparent inconsistencies and correlate them with AOD anomalies depicted in the regional datasets. Within that experimental design, it was much more appropriate to consider the potential effects of aerosol typing. Within the current one, however, it stands

more as “cause” than “effect”, the latter of which being what we are investigating. Ultimately, our interaction with the CALIPSO Science Team on both publications is contributing to a better understanding of the “lidar ratio” impact, which I think we all agree is very important. For that, we hope that the Reviewer will understand that for this publication we believe that this topic, though clearly relevant and noteworthy, falls along the periphery of our experimental design.

Thank you, however, for recognizing this important process, and for raising the question for us to consider.

2) *The authors should put their study in better context to other validation efforts for CALIPSO. In particular, ground-based observations are more or less ignored at the moment. The reference list is dominated by self-citations. Considering the facts that CALIPSO is a joint effort of NASA and CNES and the manuscript is submitted to a European journal, it is somewhat disconcerting to see that only two out of more than 60 citations refer to European research efforts.*

SPECIFIC COMMENTS

1) *Pages 2753-2755: The description of CALIOP data evaluation is difficult to understand and should be improved. In particular, the meaning of the QC flags and of the Extinction_Coefficient_Uncertainty needs better explanation.*

   Some questions in this context:
   Page 2753: What is r?*
   Page 2754, lines 9 and 10: What means 99 without unit?
   Page 2754, line 11: …of all 2010 nighttime… nighttime what?

I think all of the authors here are sympathetic to this comment. Arguably, yes, a reader uneducated in the ways of current satellite datasets and metadata should be able to pick this up and understand the method and narrative. At this junction, the reviewer is likely right that its confusing. We don’t believe, however, that it isn’t *consistent*, though. In deference, the introductory paragraph now reads:

“To ensure that only QA measurements are included, each L2-AProf 5-km extinction profile solved was screened and low-confidence data points excluded before integrating and solving column AOD. Within each data file disseminated by NASA, parameter profiles are included reflecting various diagnostic quality flags for the user to consider for QA evaluation. These terms are defined and described elsewhere [NASA, 2010]. Specifically then for a given 5-km profile, an extinction coefficient value is considered QA and included in the column-integrated sum (i.e., AOD), when at the corresponding range either…”

Thus, we are specifically turning the reader to the Algorithm Theoretical Basis Document, which in the case of the CALIPSO Science Team is ridiculously well written. However, we now specifically say that the data are available in the file, and if you don’t care what the variables truly mean, just apply these standards and you’ll get what we got. Given space limitations and the scope of the paper, again, this is consistent.

*r** is described consistently in the narrative, in order of introduction for each term (page 2754, line 19).

Re:99…great catch! Corrected.

'*Nighttime'* now has 'data' after it…again, good catch, thanks!

2) *Page 2755, discussion on attenuation: …limitation for nadir-aligned profiling… Why specifically for nadir? The approach of omitting profiles when no aerosol is detected close to the surface because of signal attenuation (cut of optically very thick layers) leads to a bias in the dataset, doesn't it? The consequences, e.g., for strong aerosol sources (deserts, fires) should be discussed.*

The sentence that is objected to has been correctly removed. It made no sense in context. Otherwise, to the second point, bias is a strong word. It's a qualified dataset, yet. Frankly, had we not applied this metric, and thus allowed all cases through in spite of aerosol resolution near the surface however, I would in fact argue that the sample is biased from simple undersampling. Therefore, our method actually seems the most consistent and reasonable way for constraining the dataset. In fact, this is what we say in the text in the final sentence. The paragraph has been rewritten, though, in portions to make these points clearer.

3) *Page 2759, line 3: …where each bin value is treated as a single normalized point…What does it mean?*

Agreed, this is difficult to introduce, and we've changed the text here in response. What we're saying here is if we average all points together, period, or treat the 1degX1deg average, in spite of any relative offsets in sample size globally, as single points. So, 4,000,000 data points (give or take) versus 56,400.

4) *Page 2759, line 12: …AOD threshold… threshold for what?*

Changed to "global mean".

5) *Page 2759, lines 22-23: …the west African coast, including the Caribbean, Brazil… is a geographically strange description.*

Agreed! Removed "including"…much more consistent.

6) *Page 2759, line 26 ff.: …Saharan desert dusts, where a-priori assignment of the lidar ratio for extinction retrieval is relatively stable… This statement is contradictory to the ongoing discussion on the correctness of the lidar-ratio assignment for desert dust (see, e.g., Schuster et al., Atmos. Chem. Phys. Discuss., 12, 11641–11697, 2012). The a-priori lidar ratio of dust used in the CALIPSO retrievals is considerably smaller than values measured in pure Saharan dust from ground (see, e.g., Tesche et al., Tellus B, 61, 144-164 2009; Wandinger et al., Geophys. Res. Lett., 37, L10801, 2011; Tesche et al., Tellus B, 63, 649-676, 2011). Thus an underestimation of the global*

dust AOD from CALIOP is very likely. This could be the reason for the discrepancies obtained along the dust belt (Sahara to Gobi) in Fig. 4.

As we describe above, this isn't the paper for investigating these offsets. Further, this wasn't the intention of this statement. The language was poor. It was simply a statement about the temporal variability of the aerosol model chosen, and that along this band aerosol model selection is relatively persistent, which we've now worded more appropriately. However, the Reviewer has a point, and we have added the references and point of contention about the absolute magnitude of the lidar ratio for the dust model in deference.

7) *Page 2759, line 29: …biased high… against what?*

Corrected. "…versus AERONET"

8) *Section 3.1 in general: Most of the discussion of the discrepancies between CALIOP and NAAPS AOD in this section is driven by surface aspects. As mentioned above, CALIOP AOD determination strongly depends on aerosol typing and the respective a-priori selection of the lidar ratio. A discussion of this aspect would be very helpful here.*

We certainly respect this comment. However, we maintain our stance from above: this is not within the scope of what we are trying to do here, which is ultimately evaluate day v. night performance. That said, to properly achieve what the Reviewer is proposing will require studies exactly like this one, and including another one the Lead Author here has just published in Atmos. Res. evaluating CALIOP specifically over SE Asia (referenced above). Again, in that latter paper we go into much more detail on how the aerosol models impact the retrievals, and how the results compare versus NAAPS (high over land, low over water, similar to what is reported here save for the Saharan Dust Belt. Through all of these cumulative pieces of evidence, a true quantitative and global estimate of aerosol model impact can be stitched together. Ultimately, though again, that work is the responsibility of the Langley Development Team/CALIPSO Science Team, and not something that we wish to get into here. The strength of the current paper is the novelty of applying the numerical model as a nighttime validation source. We need to strictly stay within the guise of that experimental design. We do sincerely thank the Reviewer for the comment and the encouragement, and hope that the Atmos. Res. paper begins to answer some of the questions that they are so clearly (and rightly) interested in!

9) *Section 3.2: How and why are these specific regions selected? All of them seem to be somehow influenced by marine conditions. Why are pure continental sites missing? Again, in view of the interpretation of data with respect to aerosol types, it would be very helpful to select regions with relatively "pure" aerosol types as used in the CALIOP retrievals and perform the respective comparisons. Also in Sec. 3.2, the discussion should not only focus on the surface aspects.*

Agreed. There isn't necessarily a clear and easy justification for having chosen these regions. Arguably, and again reiterating the points above, had we really been targeting a correlative study of AOD performance relative to CALIPSO aerosol model type we could have chosen the regions

based on predominance where each of the models are most commonly chosen. But, again, this wasn't the point of the work. For this study, these sites were chosen based on a number of factors, including data points/signal-to-noise (i.e., nearer the equator), availability of ground-based lidar observations for eventual quantitative comparison, etc… The point is well taken, and frankly in the absence of some overarching theme from which to justify the selection, for now we'd just as soon leave the discussion as is rather than hand wave.

10) Section 3.3: Only here, when comparing with ground-based instruments, discussion on the aerosol-type-dependent lidar-ratio assumption is started. However, a clear reasoning for the obtained discrepancies is missing. A discussion on how "polluted dust" is identified from CALIOP level 1 data, what it means for the lidar-ratio selection and thus the extinction calculation, and why this approach fails when dust is mixed into the marine PBL, as it is obviously the case in the Caribbean, would pave the way for some interesting conclusions here.

Thematically assessing the entirety of the Reviewer's comments, the Lead Author genuinely regrets that we got so close to these issues but ultimately left them hanging in this paper. I hope sincerely that the Reviewer will take a look at our SE Asian paper. To be frank, the two papers were originally written within weeks of one another. So, I did this study, wrote this paper first, and then conducted the SE Asian study/wrote that paper second. With respect to many of the thoughts that you're sharing with us here, I definitely was thinking about them after completing this paper and thus dug more deeply into them in the second. Ultimately, though, I don't want to take away from (and hope the Reviewer doesn't discount) what we accomplished here, which I believe exhibits significant merit in assessing the nighttime retrievals. There are some real accomplishments for the Development Team achieved by demonstrating the skill of the Level 2 archive, knowing that the "other half" of the datasets (i.e., the ones you can't validate v. AERONET/MODIS/MISR/VIIRS) are just as good. But, I'm sympathetic to your point. It definitely resonates. So, with that bit of perspective on how this and the other paper were "born", I ask you to keep your comments in mind, take a look at the SE Asian paper and see if we began to meet your expectations there. They are very different papers, ultimately, but with many of the same goals in mind. They were just designed a bit differently, and I think the more "mechanistic" view of the algorithms and their performance is focused on more in the second and will be a much more satisfying read for the Reviewer. Thanks.

OTHER CORRECTIONS
Page 2753, line 27: …avalue…
Page 2757, line24: …at over land…
Page 2761, line 2: …is found the tropical…
Page 2765, line 29: …xceeding…
Page 2766, sentence starting at line 12 is not complete

Thanks. These have all been fixed.

Thank you again for your comments and careful reading!

Reviewer #2 – Thank you for the considerate and supportive reading of our manuscript. We have made the suggested changes, referenced below, and thank you for your constructive thoughts.

> *The manuscript presents an evaluation of the CALIPSO nighttime and daytime cloud-screened 1-degree averaged aerosol product using the NAAPS model with satellite data assimilation. The model is employed to bridge a gap between daytime passive satellite and high quality nighttime lidar observations. The paper is well written; the CALIPSO and NAAPS data qualities are discussed in extensive detail. The potential issues are clearly identified. The paper is clearly relevant and appropriate for publication to AMT. I highly recommend the paper for publication after few general comments are addresses. I do not have much to add to the specific comments of the previous review.*
>
> *GENERAL COMMENTS:*
> *- Authors should clarify the goals of the study. It is not completely clear if the main goal of the paper is to contribute to CALIPSO validation efforts by evaluating the CALIOP cloud-cleared climatology, or to evaluate the applicability of the CALIOP AO climatology for potential assimilation into NAAPS.*

This is a very concerning comment. Thanks for raising the point. We've made multiple changes to bring this emphasis out more clearly within the narrative. In the Abstract, we add as a second sentence:

"In the absence of sunlight, since passive radiometric AOD retrievals rely overwhelmingly on scattered radiances, the model represents one of the few practical global estimates available from which to attempt such a validation."

At the end of the Introduction, we have added:

"The goal of this work is to motivate a practical study of CALIOP nighttime aerosol algorithm performance using one of the few estimates of AOD available globally during darkness."

In the conclusions, we have added to the middle of the first paragraph:

"In the absence of sunlight, since passive radiometric AOD retrievals rely overwhelmingly on scattered radiances, the model represents one of the few practical global estimates available from which to attempt such an evaluation"

> *- Authors should add additional discussions on when CALIPSO-NAAPS differences are likely to be attributed to NAAPS/and satellite retrieval biases (likely for land-water differences) and when differences are due to the CALIPSO retrievals themselves (likely due to aerosol model assumptions as pointed out in the previous review).*

Again, as we discuss in our response to Reviewer #1, much of this discussion would be mechanistic, which was not the purpose of the study. There are points in the discussion presently where we do point out specific offsets and/or biases (i.e., NAAPS being high over Sahara),

particularly when they can be referenced to previous work. The experimental design was to evaluate the nighttime retrievals, investigate representativeness versus the model, and compare when appropriate to daytime skill. Hence, we're studying the "effect" as opposed to the "cause". We are working, again as we talk about above in SE Asia, on mechanistic investigations of CALIOP retrievals. But, this paper had a much more limited focus/design. It would have been very useful to hear the Reviewers comments about specific instances where we could have provided more context within the narrative so as to hopefully address their concern without straying too far from our objectives. Again, there are points in there where a specific point is qualified as a function of something we know about the algorithms. In the absence of these specifics, however, its difficult to really respond. In light of their other comments, which we have definitely tried to remedy, its not clear where to add such discussion?

> *- Authors state on the page 2756 that NAAPS model errors co-vary close with those of satellite data. Is it true for 00-hr NAAPS analysis that was compared to CALIPSO data? NAAPS nighttime data when there is no satellite data is available for assimilation might differ from those of satellites.*

Agreed, this is a very good contextual point. As we introduce that thought there on P2756, its actually out of order. So, to correct this, we've added at that point where the Reviewer raises the point:

"(discussed/introduced below, and as will be depicted in Figs. 1c and 2c)".

At the end of the second paragraph of Section 2.2, we now state:

"Again, as described above, the consistency depicted in these results is a direct response of the model to its assimilation inputs."

At that point, we believe its otherwise clear that there is degradation between the 00-hr and 24-hr states, and thus within the 12-hr forecast that is being relied upon here for validation.

Thanks for these valuable comments, and for helping us emphasize these important points in the paper!