We appreciate your time to read our manuscript and give us these comments. Below please find our reply.

Anonymous Referee #1

1 General Comments

This manuscript presents an evaluation of several aerosol-optical depth products derived from satellite sensor data with ground-based observations in a region of East Asia. The text is well-structured and mostly well-written. However, in my view, two major items remain to be addressed in a revision:

- Several choices in study design are not fully explained and require additional justification (see below).
- It appeared to me that not all numerical preconditions for correlation and regression analyses, both central to the presented study, were met in all situations. Also, statistical significance of regression was not tested for. Details below.

Response: According to your comments in the "individual issues/questions" section, we added some description and explanation of our data processing and analyzing method. We hope this information can help you evaluate our study.

2 Individual Issues/Questions

- 20710-24 (henceforth "10-24" etc.): Is the bias systematic?

Response: There is evidence that this positive bias includes both random error and systematic error due to improper characterization of surface reflectance, uncertainties in the assumed aerosol model, and cloud masking. The 3 km MODIS products sample fewer reflectance pixels to retrieve aerosol pixels relative to the 10 km products, introducing sporadic extreme values of AOD that are avoided more successfully by the 10 km products. Previous studies also indicated that this positive bias in urban areas resulted from improper characterization of bright urban surfaces, a known difficult situation for the Dark Target algorithm (Munchak et al., 2013;Remer et al., 2013). The VIIRS IP product is retrieved at the reflectance pixel level without aggregation, so it is expected to include more noise. Moreover, VIIRS IP is also affected by factors that impede the Dark Target algorithm; thus, this positive bias is due to both random error and algorithm issues. We added the following sentences to explain this on page 21, line 25- page 22, line 6: "There is evidence that this positive bias includes systematic errors due to improper characterization of surface reflectance, uncertainties in the assumed aerosol model, and cloud masking. The 3 km MODIS products sample fewer reflectance pixels to retrieve aerosol pixels relative to the 10 km products, introducing sporadic unrealistic high AOD retrievals that are avoided more successfully by the 10 km products (Munchak et al., 2013). Previous studies also reported that improper characterization of bright urban surfaces, a known difficult situation for

the Dark Target algorithm, led to positive bias in urban/suburban regions (Munchak et al., 2013; Remer et al., 2013). The VIIRS IP product is retrieved at the reflectance pixel level without aggregation, thus it is expected to include more noise."

• 13-12: what does a value of -0.1 indicate? This should probably not be referred to as a "value".

Response: Since the MODIS Collection 5 algorithm, negative retrieval values have been allowed. In this study, both MODIS C6 3 km and GOCI aerosol products have negative retrievals. Though such a negative AOD value is not physically possible, it statistically represents small positive AOD values in the overall data distribution. In other words, removing these negative values would in fact truncate the lower tail of the AOD distribution. Most previous evaluation studies include these negative retrievals as valid values (Levy et al., 2013;Munchak et al., 2013;Remer et al., 2013); thus, we also included these negative values in our analyses.

• 14-7: All AERONET observations are point observations. In evaluating the accuracy of the satellite products, why would there be a need for spatially continuous ground-based observations? I would expect a multi-temporal evaluation using a wide range of AERONET stations to allow for a fairly representative assessment of product quality. Or do you expect distinct spatial patterns in the satellite products? As this point is the central motivation for this study as I understand it, I suggest that you elaborate your argument in this respect.

Response: Thank you for this suggestion. Since we aim to evaluate the performance of high-resolution satellite aerosol products, we need intensive ground observations as "ground truth". For example, the spatial resolution of MODIS C6 3 km products is 3 km, thus in theory the variation in aerosol loading at two locations that are 3 km apart can be detected by these products. However, if the ground stations are 10 km away from each other, we cannot validate this 3 km product at their designed resolution; even if these products can detect the aerosol loading variability across stations separated by 10 km, they may not perform well at a 3-km resolution. The DRAGON-Asia Campaign provides intensive ground measurements and makes it possible to validate these high-resolution satellite aerosol products. To make the motivation of this study clear, we modified these sentences as follows on page 6, line 20–25: "Evaluation of satellite aerosol products' ability to track small-scale aerosol spatial variability is limited due to lack of intensive ground observations of AOD: the permanent AERONET stations can be tens or even hundreds of kilometers apart, leading to insufficient information of the small-scale horizontal distribution of aerosol loading that is required for a precise evaluation at high resolution."

• 14-8: A point observation at the ground does not have a 'spatial resolution' at all. You may be referring to the distance between observations. Please clarify, and change the terminology here and elsewhere.

Response: We modified the sentences as follows on page 6, line 25–28: "In response to the lack of intensive ground AOD observations, AERONET conducted several campaigns, which deployed additional temporary sunphotometers in selected regions and provided valuable information of

small-scale AOD distribution."

- 15-14: What does "high quality" refer to in EDR/IP?

Response: There are several quality assurance steps in the retrieval process and both EDR and IP are assigned quality flags of "high", "degraded", or "low", indicating the confidence of retrievals. The "high" quality AOD is suggested for scientific research and applications by the VIIRS aerosol products team. We added the following sentence in the revised manuscript on page 5, line 9–10: "Detailed description of the quality assurance of VIIRS aerosol products is documented by Liu et al. (2014)." We also modified the sentence as follows on page 8, line 3-4: "Thus, only EDR and IP pixels from May 2012 to June 2013 with high quality (Quality Flag = "high") were processed."

- 16-3: "AERONET stations" do not "measure AOD". Please increase precision of statement.

Response: Thank you for this suggestion. We modified this statement through the manuscript, e.g. we changed "AOD measurements" to "AOD observations".

- 16-9: How can anyone "assure" the quality? Did you perchance mean "quality assessed"?

Response: The phrase "quality assured" is developed by the AERONET science team and is widely used in related articles. Quality-assured data (Level 2.0) have both pre- and post- deployment calibration, leading to uncertainty of about 0.01–0.02. We added the following sentence on page 8, line 23–26: "The Level 2.0 (quality assured) AOD data have both pre- and post-deployment calibration, leading to an uncertainty of about 0.01–0.02 while the Level 1.5 AOD data are cloud-screened but not quality-assured (Otter et al., 2002)."

- 16-27: Why did you reproject the data? This will certainly lead to sampling induced errors!

Response: The original satellite aerosol products are in a geographic coordinate system with latitude and longitude information; however, to build a 3-km/6-km fixed grid, we need to covert the latitude/longitude coordinate system to a projected coordinate system. To clarify the data processing method, we modified the sentence as follows on page 9, line 18: "All the data were converted to the JGD_2000_UTM_Zone_52N coordination system."

- 16-27: Please give details of the averaging/pixel combination method used in the reprojection process.

Response: In the reprojection process, we did not conduct any averaging or pixel combination. This process was basically conducted using ArcGIS to convert satellite and ground AOD data from a latitude/longitude geographic coordinate system to a projected coordinate system, thus allowing us to match satellite retrievals and ground stations based on the distance between them.

• 16-28: What do you mean by "data integration"?

Response: The data integration here means to spatially join satellite retrievals with ground stations based on their locations and to develop collections of coincident satellite–ground AOD pairs for following comparisons. We modified the sentences as follows in this section to clarify this data processing step: "All the data were converted to the JGD_2000_UTM_Zone_52N coordination system. For the matchup process, a 6-km grid and a 3-km grid covering the whole study domain were constructed, corresponding to the spatial resolution of each satellite product. Satellite aerosol data from different sensors were mapped and spatially joined to this 6-km grid (for VIIRS EDR and GOCI products) or 3 km grid (for VIIRS IP and MODIS C6 3 km products) to construct coincident satellite-ground AOD pairs."

• 16-28: Why would "data integration" be necessary? Why not leave all data at their original aspects and resolutions and compare them based on location alone?

Response: The data integration, or the matchup process was necessary because it provided satellite–ground AOD pairs for following comparisons. This process did not change the original aspects or resolutions of the data, and our comparisons were based on location and time.

• 17-12: "maximum sample size" - in what respect?

Response: Since the DRAGON-Asia Campaign and the Beijing sampling experiment were conducted in different time periods, if we used the overlapping periods of these two experiments, we would lose many ground observations, leading to an insufficient sample size. To include the maximum number of ground observations, we allowed the spatial comparisons in Beijing and the Japan – South Korea region to differ in time periods. To clarify this, we modified the sentence as follows on page 10, line 12–16: "Temporal comparisons and spatial comparisons differ in study periods (Table 2): the temporal comparison period was the longest overlap period covered by all five satellite products and the spatial comparison periods in Beijing and the Japan–South Korea region are different in order to include the maximum number of ground observations."

• 17-17: Why did you average 3x3 grid cell environments if your main aim was to assess the quality of high-spatial-resolution data?

Response: We developed two comparison methods in this study: the temporal comparison and the spatial comparison. For the spatial comparison, the intensive ground observations from the DRAGON-Asia Campaign and the Beijing Sampling Experiment provided sufficient satellite–ground AOD pairs to validate satellite aerosol products performance at their designed resolution. For the temporal comparison, since we aimed to validate the ability of satellite aerosol products to track the day-to-day variation of aerosol to improve coverage and benefit from a collection of AOD retrievals (Ichoku et al., 2002), we used average AOD from the 3x3 grid cell buffer. This average method is widely used in previous evaluation studies.

• 18-6: How do you choose a 4x4 pixel window? Do you use the coordinates of the point between the four central pixels for comparison with other data sets?

Response: We used the 3-km grid for comparisons with VIIRS IP data because we did not have intensive ground sampling data to create a 0.75-km grid. We did not choose a 4x4 pixel window, but the 3-km grid cell sampling buffer cover a roughly 4x4 pixel window. To make this clear, we modified sentences as follows on page 11, line 21–25: "For the temporal comparison, we averaged valid IP AOD retrievals falling in the 3-km grid cell centered at each ground AERONET station and the mean and median CV were 0.33 and 0.25, respectively, within the 3 km grid cell buffer. This sampling buffer roughly covered a 4 × 4 pixel group."

• 18-8: "due to the lack of..." - I don't understand this argument. What do you mean by fine-resolution ground-based observations here? What would you ideal ground-based comparison data look like?

Response: The ideal ground-based observation for validation of the VIIRS IP aerosol product should be distributed roughly 0.75 km apart, thus we can test if the VIIRS IP AOD can detect variations in AOD at its 0.75-km nominal resolution. We modified this sentence as follows on page 11, line 26–28: "In the spatial comparison of VIIRS IP, we also used the 3-km sampling buffer due to a lack of more intensive ground AOD observations."

• 18-25: If the distortion towards the fringes of the pass impedes study results, why not use a dynamic spatial averaging approach that takes pixel size into account and tries to keep averaging area approximately constant, regardless of location and satellite system?

Response: Thank you for this suggestion. The major effect of the distortion towards the edge of the scan is that we may lose some ground–satellite AOD pairs because a satellite pixel at the edge of the swath covers a larger area than the nadir pixel size. However, there is no evidence that such missing will bias the comparison results. Using a dynamic spatial averaging approach, like kriging or interpolation, may introduce new error. Moreover, previous evaluation studies rarely used dynamic spatial averaging approach to fill the missing data due to the stretch. To make our results comparable with previous studies, we decided to use the method similar to previous studies.

• 19-9: "grid cell centered on the ground stations" - how does this apply to the 4x4 pixel averaging described above?

Response: This did not apply to IP data because we did not create a 0.75-km grid. To explain this, we added the following sentence in the revised manuscript on page 13, line 4–6: "Since we did not create a 750 -m grid for the VIIRS IP product, VIIRS IP-ground AOD pairs were assigned either "High Quality" or "Low Quality"."

• 19-28: Your figure 5 suggests that the data were used 'as-is'. A correlation analysis assumes normally distributed data, so in the case of AOD a logarithmic transformation would be required.

Did you perform this? If not, what is the rationale?

Response: There are a few important issues that go against log-transformation in the context of this study. First, due to the existence of valid negative AOD values, log transformation cannot be applied to MODIS and GOCI products directly. One solution is to add a fixed small positive number—e.g. 0.05—to both satellite retrievals and AERONET values; however, doing so changed the reference range of EE (± 0.05 ± 0.15 AOD) and made the evaluation metrics incomparable across different satellite aerosol products. Moreover, with log-transformation, linear regression intercepts and slopes lack clear physical meanings. Second, the distributions of AOD values from different sensors, as shown in Figure 4, were not significantly skewed. Due to the existence of small positive AOD values, log-transformation actually introduced slight skews to the left. Third, previous evaluation studies rarely used log-transformation (Levy et al., 2013;Liu et al., 2014;Munchak et al., 2013). Since one of the objectives of this study is to compare the performance of these emerging finer-resolution products in urban regions to their global evaluation results, log-transformation made the evaluation metrics incomparable with previous studies. All things considered, we decided to use the original data in this analysis.

• 21-25: I suggest moving this sentence to the discussion/conclusions section Results/analysis section: You analyse regression slopes and intercepts. I see two potential problems:

1. Like correlation, regression analysis assumes normally distributed date. If no log transformation of the AOD data was performed this condition is probably not met, statistically invalidating the analysis.
2. In regression analysis, a p value is always computed, indicating the probability that the results were purely due to random variation. It is commonly accepted practice to set a significance level before the analysis (e.g. 90%, 95% etc. probability of the relationship NOT being random) and then to discard all relationships outside that frame (p value gt; 0.1, 0.05 etc.) as not statistically significant. A slope and intercept could be the result of random variation in your data set, or they could be statistically significant. Without ap value, no one can tell.

Response: Regarding your first comment that the non-normal distribution of AOD data violated the assumption of linear regression, as we explained in a previous response, there are a few important issues that go against log-transformation in the context of this study, including the existence of valid negative AOD values and inconsistency with previous evaluation studies. Thus, we decided to use the original data in this analysis.

Regarding your second comment, we added an indicator of significance level based on the p-values of the regression slopes and intercepts in Table 3 and Table 4 in the revised manuscript.

• 25-20: "cautious" - how?

Response: Researchers need to calibrate these high-resolution satellite aerosol products in their study regions before applying them. Researchers may need to develop specific methods to process these data: for example, filtering AOD retrievals based on land use information. We

modified the sentence as follows on page 19, line 8–12: "In general, these finer resolution aerosol products included larger bias relative to lower resolution products and researchers must be cautious when applying them, e.g. calibrate these high resolution satellite aerosol products in specified study regions and implement appropriate data filtering strategies."

- Tables 3 and 4: Why are no p values given?

Response: We added an indicator of significance level based on p-value for linear regression slopes and intercepts in Tables 3 and Table 4.

- Figure 5: Since AOD is not normally distributed, it should be shown on a log scale or another suitable transformation.

Response: As explained in our previous response, there are a few important issues that go against log-transformation in the context of this study, including the existence of valid negative AOD values and inconsistence with previous evaluation studies. Thus, we decided to use the original data in this analysis.

3 Technical Details

- 11-15: ground-based
- 12-1 and 12-16: different time formats. Please harmonize throughout manuscript in accordance with journal requirements.
- 13-8 replace "that were" by a comma
- 13-12 remove "range"
- 14-5: small-scale
- 14-7: remove "required"
- 15-3: The size/extent etc. of the study area...
- 15-21: Ground-based measurements (here and elsewhere)
- 15-25: were/are distributed
- 16-2: approximately 10km apart -> with an averate distance of about 10km between two stations (surely 10 km isn't the distance between Osaka and Seoul...)
- 16-2: which can be... check wording
- 16-6: in THE Japan-South Korea region
- 16-17: "that distributed" -> selected sites roughly 6km apart from each other along
- 17-14: cells -> cell
- 20-5: metrics -> metric
- 21-5: results ... suggest

- 21-6: among -> between
- 23-8: over THE Japan-...
- 23-10: DRAGON
- Tables 3 and 4: The "Spatial Comparison" part should be more clearly visually distinct from the "Temporal Comparison" part.
- Figure 3, line 3: observations -> observation
- Figure 3, line 4: retrievals -> retrieval
- Figure 3: red and green are hard to impossible to distinguish for a of humanity (including me :). I suggest using a different pair of colors (e.g. red and blue)
- Figure 5: In their current form, the individual figures seem too small.
- Figure 5: in dash line -> as a dashed line
- Figure 5: in gray solid -> as gray solid

Response: Thank you for these suggestions/corrections, we changed the words and modified the figures in the revised manuscript accordingly.

References:

Ichoku, C., Chu, D. A., Mattoo, S., Kaufman, Y. J., Remer, L. A., Tanré, D., Slutsker, I., and Holben, B. N.: A spatio‐temporal approach for global validation and analysis of MODIS aerosol products, Geophysical Research Letters, 29, MOD1-1-MOD1-4, 2002.

Levy, R., Mattoo, S., Munchak, L., Remer, L., Sayer, A., Patadia, F., and Hsu, N.: The Collection 6 MODIS aerosol products over land and ocean, Atmospheric Measurement Techniques, 6, 2989-3034, 2013.

Liu, H., Remer, L. A., Huang, J., Huang, H. C., Kondragunta, S., Laszlo, I., Oo, M., and Jackson, J. M.: Preliminary evaluation of S‐NPP VIIRS aerosol optical thickness, Journal of Geophysical Research: Atmospheres, 119, 3942-3962, 2014.

Munchak, L., Levy, R., Mattoo, S., Remer, L., Holben, B., Schafer, J., Hostetler, C., and Ferrare, R.: MODIS 3 km aerosol product: applications over land in an urban/suburban region, Atmospheric Measurement Techniques Discussions, 6, 1683-1716, 2013.

Remer, L., Mattoo, S., Levy, R., and Munchak, L.: MODIS 3 km aerosol product: algorithm and global perspective, Atmospheric Measurement Techniques Discussions, 6, 69-112, 2013.

Anonymous Referee #2

This work studies the spatial and temporal characteristics of satellite remote sensing of aerosol products against ground measurements of AERONET, the DRAGON-Asia campaign, and data from a mobile sunphotometer sampling campaign in Beijing. Five emerging satellite aerosol products from three different platforms (i.e. MODIS, VIIRS, GOCI) are evaluated over East Asia in 2012-2013.

In general, the manuscript is well written and organized in a clear and logical way. This manuscript is, as far as I know, the first to compare these five satellite AOD products in one study. Moreover, the VIIRS and GOCI products are rather new and have not yet explored in depth. As such, this study adds knowledge to the atmospheric research community and could be published after addressing the following comments:

Major Comments:

âA˘ c The authors use VIIRS products and comment in page 20712, lines 16-17 that "The VIIRS aerosol product reached validated maturity level in January 2013". In the NASA LAADS website it is written in relation to the use of VIIRS products that "All Suomi NPP VIIRS EDRs are currently beta quality (with known problems) and are not intended for scientific use". A clarification is therefore needed as the data sources for VIIRS and GOCI satellite products are missing.

Response: The VIIRS aerosol product science team published a global evaluation study, reporting that the VIIRS AOD at the provisional maturity level is validated. The provisional maturity level is defined as: "product quality may not be optimal" but it is "ready for operational evaluation". To make this clear, we cited this study and added the following sentence on page 4, line 25-28: "The VIIRS aerosol product reached provisional maturity level in January 2013, which means the "product quality may not be optimal" but it is "ready for operational evaluation" (Liu et al., 2014)."

The GOCI science team recently submitted an evaluation paper of the GOCI aerosol product and it has been published on Atmospheric Measurement Techniques Discussions (Choi et al., 2015). In addition, the GOCI science team published a study about monitoring transboundary particulate pollution using the GOCI aerosol product, indicating that this product can be used for quantitative studies (Park et al., 2014). We hope our evaluation study can contribute to the validation of the GOCI aerosol product. To make this clear, we added the following sentence on page 5, line 28-page 6, line 2: "A recently published evaluation study reported that from March to May 2012, the GOCI AOD had a linear relationship with AERONET AOD with a slope of 1.09 and an intercept of -0.04 (Choi et al., 2015)."

âA˘ c This work presents data from sources with very different temporal and spatial resolutions including a changing footprint (e.g. MODIS) compared to a fixed footprint (i.e. GOCI). It is not clear how these differences have been taken into account? How has data fusion to one grid been done?

Response: We compared the satellite data with ground observations using sampling buffers with

respect to satellite products' resolutions. Both satellite and ground observations were fused to a fixed 3-km/10-km grid based on their locations, processed with ArcGIS. In addition, for polar orbit sensors (VIIRS and MODIS) that provide one observation per day, we used the 1-h time window (± 30 min of satellite pass-over time) for comparisons; for the geostationary orbit sensor (GOCI) that provides multiple observations per day, we conducted comparisons during the 1-h window around 13:30 that overlaid with other sensors, as well as during each of its 8 hourly observation periods.

âAˇ c This In page 20717, line 2 the authors write that the data was "remapped". A detailed explanation in the text of the remapping methodology is missing. I find it an important stage of the work and a detailed explanation will able the reader to understand and reproduce the methodology in a future work. Furthermore, is the remapping a daily procedure? What is the possible bias due to the remapping procedure?

Response: The remapping process here means spatially joining the satellite data with the fixed grid in a projected coordination system. To avoid any confusion, we modified this sentence as follows on page 9, line 21–23: "Satellite aerosol data from different sensors were mapped and spatially joined to this 6-km grid (for VIIRS EDR and GOCI products) or 3-km grid (for VIIRS IP and MODIS C6 3 km products) with respect to their spatial resolution." This process was conducted at daily level. Due to the stretch of MODIS and VIIRS pixels toward the edge of the scan, joining the satellite pixels with the fixed grid may lead to some missing satellite − ground AOD pairs, but there is no evidence that this missing will introduce a significant systematic bias.

âAˇ c I suggest to put more emphasis in the conclusion (and abstract) and throughout the manuscript on the better performance of satellite aerosol products in tracking the day to-day variability than in tracking/representing the spatial variability at high resolution. For example, in the Conclusion the authors claim that small scale variability and point sources can be detected. Unless point source has the size of 3-6 km I do not see how this claim is supported by the results in this manuscript. Also, individual exposure is mentioned on line 10 of p. 20729 – individual exposure estimation in urban areas may be obtained if we assume uniform exposure for all the people that live in a 3-6 km grid cells. If this is what the authors mean this needs to be clarified. Otherwise, I suggest to reduce expectations rather than increase them based on the reported MS results.

Response: We modified these sentences as follows on page 4, line 12-16: "The variability of aerosol loading at local scales in urban areas with complex land surface and meteorological conditions are expected to be greater (Li et al., 2005). Accurately characterizing local-scale $PM_{2.5}$ heterogeneity is critical for assessing population PM exposure, detecting air pollution sources, and monitoring air quality." and on page 23, lines 27-29: "High-resolution satellite aerosol products provide valuable information for the spatial and temporal characterization of $PM_{2.5}$ at local scales."

âAˇ c Sections 3.2, 3.3 – it will be very valuable to show performance metrics for the different

satellite aerosol products after they were calibrated against ground measurements. Namely, once these products are calibrated it is very interesting to know which in fact performs better. Clearly, the calibration should be based on a complete leave one-out cross validation process, such that the model parameters are "optimal" in the sense that they represent all the data but not overfitting the data. Model parameterization should be developed on a regional (spatial) scale and then applied locally on AOD measurements, such that the spatial variability is still evident.

Response: We conducted 10-fold cross-validation analyses for temporal comparisons of VIIRS and GOCI data in the Japan–South Korea region and the regression statistics are similar to the original regression statistics. Due to the small sample size, cross-validation was not conducted for MODIS products. Since we aimed to evaluate rather than calibrate these satellite aerosol products, we did not create a table showing the performance metrics for the calibrated AOD. We added the following sentences on page 19, line 24-28: "Ten-fold cross validation was conducted for the comparison of VIIRS and GOCI products to detect overfitting. The linear regression statistics of cross validation did not change significantly relative to the statistics of comparisons. The cross validation R^2 values of VIIRS EDR, VIIRS IP, GOCI at 13:00, and GOCI 8 observations data were 0.73, 0.51, 0.78, and 0.82, respectively."

All of these satellite aerosol products have their own advantages and disadvantages and are suitable for different research objectives, thus it is hard to say which one performed the best. We added the following sentences on page 23, line 19-26: "These satellite aerosol products have their own advantages and disadvantages. For example, the GOCI aerosol product provides high accuracy AOD retrievals eight times per day, but it only covers East Asia; the VIIRS EDR product provides high accuracy AOD retrievals and global coverage once per day, but its 6 km resolution is relatively low; the MODIS C6 3 km products provide high resolution AOD retrievals with global coverage, but have positive bias in urban regions. Researchers need to apply these aerosol products according to specified research objectives and study design."

Using GOCI 8 observations per day data, we applied the regionally developed linear regression parameters to individual station data in the Japan–South Korea region. The linear regressions with the satellite AOD as a dependent variable and the fitted AOD from a regional model as an independent variable have an R^2 greater than 0.75 at all sites except the AERONET site 'Nara' and 'Osaka', two stations located in Osaka. Limited by sample size, we cannot apply this method to other aerosol products. However, since the spatial distribution of satellite aerosol products from different sensors are similar in this region, we believe that parameters from regional datasets were also valid locally. We added the following sentences on page 19, line 28-page 20, line 7: "In addition, to detect the spatial variability of the satellite retrieval performance, we applied the regionally developed linear regression parameters of GOCI 8 observations data to individual AERONET station in the Japan–South Korea region. The linear regressions with the satellite AOD as the dependent variable and the fitted AOD from a regional model as the independent variable yielded R^2 larger than 0.75 at all sites except the AERONET sites 'Nara' and 'Osaka', two stations located in Osaka. This result indicated that parameters from the regional dataset were valid locally. Limited by sample size, we did not apply this method to other aerosol products."

Minor Comments:

âAˇ c Figure S1 presents the spatial distribution of the stations with the different buffers. (a) The size of the ground station symbols is not proportional and I recommend to reduce the symbol size. (b) I recommend using a scale bar of 3-6-9 km, which is more relevant, instead of 5-10-20 km. (c) The different sample size boxes are not very clear: 3x3, 4x4, 6x6, 9x9? An additional table at the bottom of the figure with an explanation in the manuscript and next to each cell size can possibly make this clearer.

Response: We modified this figure according to your suggestion and added the following explanation in the captions: "The temporal comparison figure (left) shows the buffer of 3 x 3 grid cells for MODIS (pink), VIIRS EDR and GOCI products (blue), as well as the single grid cell buffer for VIIRS IP product (green); the spatial comparison figure (right) shows the single grid cell buffer for each sensor."

âAˇ c Table S2- How was the number of observations (N) from each data source taken into account? Show that the results are affected/not affected by this parameter (N).

Response: Since coverage and accuracy are two major metrics used to evaluate the performance of satellite aerosol products, the number of coincident satellite-ground AOD pairs in this table was aimed to reflect the coverage of each satellite aerosol product. Since the estimated slopes and intercepts were significant, the sample size was sufficient and the results were not affected by N.

âAˇ c The standard deviation within the 3x3 cells isn't reported. I think it is important to report it before averaging the cells in order to study/observe the distance between values within the 3x3 boxes is low.

Response: We calculated the coefficient of variation (CV), which is the standard deviation divided by the mean, of AOD retrievals in the temporal-comparison buffer from various sensors. To avoid effects from large within-buffer variation in aerosol loading, we removed satellite pixels with CV outside the range of ± 1.0. Doing so led to less than 10% missing data and the regression statistics remained almost the same. We reported CVs of AOD retrievals from each sensor in section 2.4 of the revised manuscript and we added the following sentences on page 10, line 16-27: "The coefficients of variation (CV), which is standard deviation divided by mean of AOD retrievals, from various sensors in temporal-comparison sampling buffers were calculated and reported below to assess the homogeneity of aerosol loading within buffers. The mean CV from various aerosol products ranged between 0.18 and 0.35, indicating that, as expected, certain heterogeneity in aerosol loading existed within the temporal-comparison buffer. This relatively small heterogeneity should not be a detriment to the temporal comparison, however; some extremely large CV values that were probably due to very small mean AOD values were observed. In order to avoid potentially large variations in aerosol loading within buffers, we removed satellite pixels with CVs outside the range of ± 1.0

(Liu et al., 2007) in temporal comparisons. Moreover, the existing heterogeneity of AOD loading encouraged us to conduct spatial comparisons implementing smaller sampling buffers."

âA˘ c P 20720, lines 2-4. "slope is the slop e of the linear regression with satellite retrievals as the dependent variable and ground AOD measurements as the independent variable;" it should be exactly the opposite. We want to predict ground PM by AOD so satellite AOD should be the independent variable and ground measurements (hear ground AOD) be the dependent variable. This way the satellite AOD will be consistently used as the independent variable.

Response: Since ground AOD is considered as "true value", we used ground AOD as the independent variable and the satellite retrievals as the dependent variable. In this study, we did not want to estimate the "true" AOD from satellite retrievals; in contrast, we wanted to validate satellite retrievals with ground truth and tested by how much the satellite retrievals deviated from the ground truth; thus, the satellite AOD was the dependent variable. In most previous evaluation studies, the satellite AOD was the dependent variable and the AERONET AOD was the independent variable. When predicting ground-level PM concentrations using satellite AOD, satellite AOD—together with other parameters—are independent variables, but the objectives and interpretations of these two kinds of studies are different.

âA˘ c Page 20720, lines 10-16. Consider moving these lines to the introduction and method sections.

Response: We moved these sentences to the introduction and method section.

âA˘ c p 20721 line 3. Figure 2b shows the site specific average AOD with the regional average AOD subtracted in these three cities – how was the background calculated?

Also, please explain what is the meaning of 0.01 increase in AOD as represented by different colors in Figure 2(b). Moreover, the manuscript (page 20721, line 20) refers to a difference of AOD of 0.4 between stations, a value not represented in the figure.

Response: The regional average (background) AOD was calculated as the average of AOD from all the ground stations located in this region. The background color, mainly green, denotes the elevation of this region with the same color scale as in Figure 1. To clarify this, we changed the color scale of AOD in Figure 2(b) and added the following sentence in the caption of Figure 2: "The background color shows the elevation with the same color scale as in Figure 1." The different colors in Figure 2(b) indicate the difference between AOD from each ground station and the regional average AOD. We added two more colors to this color scale to show that the difference in AOD between two nearby stations in Beijing is about 0.4.

âA˘ c P 20721 lines15-18. I assume that the higher variability in Beijing comes from the (a) poorer performance of the hand held device (e.g. instrument quality), (b) the use of daily average AOD values in DRAGON sites vs. momentarily measurements (in each site-day) in Beijing (e.g. measurement noise, un-representativeness of the measurments in Beijing), and (c) in Beijing the

measurement may have been performed when the devices does not exactly face the sun due to operation errors. I suggest to discuss all these optional sources of errors.

Response: We added the following sentences in the revised manuscript on page 14, line 25-page 15, line 4: "Second, the handheld sunphotometer may introduce larger measurement errors than DRAGON stations, due to both instrument quality and operation errors. Previous evaluation indicates that handheld stability and inaccurate pointing to the Sun significantly affects the accuracy of measurements by Mocrotops II (Ichoku et al., 2002; Morys et al., 2001). Our comparison of Microtops II AOD with nearby AERONET data yielded a slope of ~0.95, a correlation coefficient of ~0.8, and an intercept of 0.16 (Supplemental Material, Text S1), indicating that the handheld sunphotometer AOD are usable."

âAˇ c Page 20722, lines 4-5. Compare the availability of different satellite-based data and AOD from AERONET at 13:00. Terra overpass is at 10:30 local time, it hasn't been mentioned throughout the manuscript if the Terra data was compared to AERONET data at 10:30. One can understand from the text that the Terra observations were compared to AERONET at 13:00. Yet, later in the manuscript, in the first paragraph in page 20724, the overpass time difference of Terra is mentioned. I recommend to either make this clearer or to consider excluding the Terra dataset from this study.

Response: We compared the availability of Terra data with AERONET from 10:00-11:00 am. We kept Terra in this study because its aerosol products are widely used and it provides additional information about aerosol distribution. To make this clear, we added the definition of the 1-h window used for comparisons on page 15, line 21 and page 17, line 14.

âAˇ c As written in page 20724, line 3, the Y-axis in Figure 4 is "relative frequency rather than the total number of retrievals". If the frequency is relative to the number of observations (N) than it (i.e. N) should be specified in the text and/or in the figure. Moreover, as the number of satellite observations has seasonal variation (e.g. due to clouds), I suggest to add the number of observations per satellite per month, possibly in a separate figure/table.

Response: The frequency is relative to the total number of matched AOD retrievals from the corresponding sensor. We modified the sentence on page 17, line 19-23: "This histogram is plotted with the frequency of AOD retrievals from each sensor relative to the total number of matched AOD retrievals from the corresponding sensor rather than the count of AOD retrievals because these aerosol products differ in sampling strategies, leading to different total numbers of coincident satellite–ground AOD pairs." We also added the following sentence to clarify this in the caption of Figure 4: "The x-axis shows AOD values and the y-axis shows the frequency of AOD observations from each sensor relative to the total number of matched AOD observations from the corresponding sensor."

This figure compared the distribution of AOD from each satellite dataset to AOD from AERONET, the ground truth. Thus, we can detect systematic bias. The variation in the number of observations (N) across satellite aerosol products due to differences in the aerosol products'

resolutions and masking strategies does not necessarily lead to different retrieval quality, so we did not specify N in this figure. Since we used AERONET AOD as ground truth and showed the distribution of AOD from matched satellite-ground AOD pairs, this figure indicated distribution of AOD retrievals in cloud-free conditions. The seasonal missing pattern of each AOD dataset due to cloud and weather conditions is out of the scope of this figure. We added the following sentence to clarify this on page 17, line 11-15: "It is notable that the seasonal missing pattern due to cloud cover and weather conditions may vary across these satellite aerosol products. However, since we did not have enough coincident satellite-ground AOD pairs to conduct seasonal evaluation, the seasonal missing patterns and seasonal performance of these satellite aerosol products were not analyzed in this study."

âAˇ c Page 20729 top. Clearly, the conclusion that the 6 km products provide more accurate data than the 3 km products results from the spatiotemporal averaging. This may be useful in some cases but is huge disadvantage in other cases, in particular for environmental health and exposure estimation, which is one of the applications declared by the authors as their interest.

Response: We understand that one major application of aerosol satellite remote sensing is exposure assessment and that's why we introduced quality flags for coincident satellite-ground AOD pairs. There is a trade-off between satellite retrieval coverage and accuracy, and we tried to increase the coverage without significantly decreasing accuracy. We understand that the 3 km products and products at even higher resolution will contribute to fine-scale exposure assessment; however, these products showed higher bias in this and previous evaluations. Researchers need to use these products with caution. We added the following sentence on page 23, line 12-16: "however, VIIRS IP and MODIS C6 3 km products provide additional information about fine-resolution aerosol spatial distribution and will benefit exposure assessments at local scales;"

âAˇ c Figure 6. The color scale should be the same for all figures for a clearer interpretation.

Response: We used the same color scale for all figures with different minimum and maximum values. The minimum value is 0 for VIIRS products and -0.05 for MODIS and GOCI products, and the maximum value is 2.0 for VIIRS products and >2.0 for MODIS and GOCI products. This difference is related to retrieval algorithms and we wanted to indicate this difference in figures, but the fact that these color scales differed in maximum and minimum values did not affect comparisons across these figures.

âAˇ c Table 3. The temporal comparison section and the spatial comparison section should be separated, e.g. by a line above the spatial comparison section.

Response: We modified table 3 and table 4 to make the temporal and spatial comparison sections more visually separated from each other.

âAˇ c Caption to Fig. 2a – what is "Loess curvy" ? Fig. 2b – what is the meaning of the green

background color in non-measurement locations?

Response: The Loess curve is a smooth curve based on a non-parametric regression. We used this curve to show the trend of the correlation coefficient of AOD from two stations with increasing distance. The green background color shows the elevation with the same color scale as in Figure 1. To make this clear, we added the following sentence in the caption of Figure 2: "The background color shows the elevation with the same color scale as in Figure 1."

âAˇ c Fig. 5 is too small and its details cannot be seen. There is a need to improve the presentation of this fig.

Response: We modified the arrangement of Figure 5 and enlarged each of the figures.

References:

Choi, M., Kim, J., Lee, J., Kim, M., Je Park, Y., Jeong, U., Kim, W., Holben, B., Eck, T. F., Lim, J. H., and Song, C. K.: GOCI Yonsei Aerosol Retrieval (YAER) algorithm and validation during DRAGON-NE Asia 2012 campaign, Atmos. Meas. Tech. Discuss., 8, 9565-9609, 10.5194/amtd-8-9565-2015, 2015.

Ichoku, C., Chu, D. A., Mattoo, S., Kaufman, Y. J., Remer, L. A., Tanré, D., Slutsker, I., and Holben, B. N.: A spatio‐temporal approach for global validation and analysis of MODIS aerosol products, Geophysical Research Letters, 29, MOD1-1-MOD1-4, 2002.

Levy, R., Mattoo, S., Munchak, L., Remer, L., Sayer, A., Patadia, F., and Hsu, N.: The Collection 6 MODIS aerosol products over land and ocean, Atmospheric Measurement Techniques, 6, 2989-3034, 2013.

Liu, H., Remer, L. A., Huang, J., Huang, H. C., Kondragunta, S., Laszlo, I., Oo, M., and Jackson, J. M.: Preliminary evaluation of S‐NPP VIIRS aerosol optical thickness, Journal of Geophysical Research: Atmospheres, 119, 3942-3962, 2014.

Munchak, L., Levy, R., Mattoo, S., Remer, L., Holben, B., Schafer, J., Hostetler, C., and Ferrare, R.: MODIS 3 km aerosol product: applications over land in an urban/suburban region, Atmospheric Measurement Techniques Discussions, 6, 1683-1716, 2013.

Park, M., Song, C., Park, R., Lee, J., Kim, J., Lee, S., Woo, J.-H., Carmichael, G., Eck, T. F., and Holben, B. N.: New approach to monitor transboundary particulate pollution over Northeast Asia, Atmospheric Chemistry and Physics, 14, 659-674, 2014.

Remer, L., Mattoo, S., Levy, R., and Munchak, L.: MODIS 3 km aerosol product: algorithm and global perspective, Atmospheric Measurement Techniques Discussions, 6, 69-112, 2013.