Response to Reviewers for "Airborne observation during KORUS-AQ show aerosol optical depth are more spatially self-consistent than aerosol intensive properties"

We appreciate comments from Referee #1 (Andrew Sayer) and the push to enhance this manuscript's quality. Please see attached the responses to each comment in blue italic. We have added discussions spanning most of the Referee's comments, with emphasis on descriptions of the expected variability in metrics other than autocorrelation (including a new table) and enhancement to the statistical comparisons using more appropriate bivariate fitting routines. We have also adjusted multiple figures with the Referee's comments in mind. In line with the spirit of the comments, we have enhanced the writing quality throughout the manuscript. See below for more details.

Referee #1 (Andrew Sayer)

The topic is important and within scope for ACP. The quality of writing and presentation is high (though I have a few suggestions for changes to Figures). My overall recommendation is for minor revisions. I would be willing to review the revision if the Editor would like. My specific comments and suggestions for revision are as follows:

Thank you.

- The main metric used to quantify the spatial scale of variation is the distance at which the autocorrelation drops off to 85% of its value in the smallest distance bin. I was wondering why 85% was chosen? I would have thought it more common to state in terms of an e-folding distance – unless the autocorrelation profile doesn't look like exponential decay (which some of them might not). Either way I'd appreciate some (brief – not repeating the whole analysis) discussion in the paper of why this particular threshold was chosen and if results qualitatively change if a different metric is used – for example the e-folding distance, or an autocorrelation drop to e.g. 70% of the max rather than 85% (given a correlation of 0.7 corresponds to about 50% of the variance in the field). Looking at curves my guess is in most cases the picture would be the same, but as thresholds are a bit arbitrary it is good to check sensitivity to them.

  This is a good question. The purpose of this 85% metric is to catch the inflection point of the autocorrelation, with less uncertainty. We tested multiple different metrics, including the e-folding and at 90%. While the overall trends with respect to decay did not vary much, their magnitude did. We included a new paragraph in section 4.4 and a new table, to describe the distances at various metrics. See here: "The 15% decrease metric is used to identify where there is an inflection point in autocorrelation, however distances where autocorrelation decays by 10% or by 1/e show similar trends (see Table 2 for examples). For AOD, the mean distance where there is 10% decrease in autocorrelation occurs at 26.8 km, roughly 1/2 the distance of 15% decrease, while an 1/e decrease (~37% reduction), the mean distance is 167.5 km. Because of the larger dependence to samples (showing larger spread in autocorrelation at longer distances), the 1/e results in a larger spread of distances (standard

deviation of 54 km). Similarly for AE, where the largest standard deviation and spread is found from the 1/e decrease level."
And table 2:

Relative auto-correlation	AOD_{500}					AE				
	4STAR	MERRA-2	GOCI	in situ (LARGE)	**Average**	4STAR	MERRA-2	GOCI	in situ (LARGE)	**Average**
90%										
distance [km]	25 [10, 35]	65 [35,65]	10 [3,160]	7.5 [7.5, 7.5]	**26.88**	10 [10, 15]	65 [35, 100]	0.6 [0.27,1.35]	0.27 [0.2, 0.27]	**18.34**
mean of difference	0.003	-0.0015	0.008	0.013*	**0.003**	-0.0036	-0.0161	-0.00098	0.0015	**-0.005**
standard deviation	0.1364	0.0504	0.0985	0.292*	**0.095**	0.1465	0.206	0.3072	0.1776	**0.209**
85%										
distance [km]	35 [25, 35]	100 [35,100]	65 [5, 160]	10 [10,10]	**52.5**	15 [15,25]	65 [35,160]	7.5 [5,7.5]	0.9 [0.6,0.9]	**22.1**
mean of difference	-0.0072	-0.0073	0.0185	0.0173*	**0.001**	-0.0039	-0.0161	0.0027	0.0048	**-0.003**
standard deviation	0.1436	0.0682	0.1442	0.318*	**0.119**	0.162	0.2066	0.408	0.217	**0.248**
1/e (63%)										
distance [km]	160 [35,160]	250 [35,250]	160 [65, 160]	100 [35, 100]	**167.5**	100 [100,100]	160 [100, 160]	100 [25, 100]	65 [65, 100]	**106.25**
mean of difference	-0.0073	-0.0985	-0.033	0.0057*	**-0.046**	0.037	-0.065	0.028	0.037	**0.009**
standard deviation	0.1557	0.1357	0.217	0.389*	**0.169**	0.247	0.354	0.507	0.371	**0.370**

Table 2 *- Distance bins at which different relative autocorrelation is reached for AOD and AE from 4STAR, MERRA-2, GOCI, and in situ (LARGE). The range in distance (square brackets) is obtained from the autocorrelations that are varied by one standard deviation of the 50-member Monte Carlo ensemble of flight segments. The differences in AOD and AE from all flight segments at the distance bins are reported by their mean and standard deviation. The Average AOD mean and standard deviation of the difference is averaged from 4STAR, MERRA-2, and GOCI, while AE also includes in situ (LARGE). *The in situ extinction coefficient mean difference and standard deviation are multiplied by 2.5 km for easier comparison to the AOD mean and standard deviation values.*

- Related to the above, it would be interesting to quantify at a couple of places what the typical variation in the field is for these autocorrelation drops (e.g. at the distance of 85% autocorrelation, what is the variance of the difference between AOD or FMF or AE at that point and at zero lag). This helps give an idea of how numerically important some of these variations are (with the understanding that these magnitudes might not be transferable to other regions or seasons). For example at the 22.7 km distance where AE autocorrelation has dropped to 85% is the AE difference about 0.1 or 0.3 or?

To answer this question, and the previous we added a new table which describes the distances at various relative autocorrelation values, and the mean difference in the distance lags. The means are as expected very near zero - implying a somewhat even distribution of the sampling. The standard deviation of the difference between values measured at changing distances does vary, and increase with longer distances. See table 2 above and this new paragraph at the end of section 4.4:

"The distances at which autocorrelation varies can also be understood through the expected variation of the aerosol properties (AOD or AE). Since we use the combination of all level flight legs, the difference between AOD or AE between measurements binned by their lag distance has a mean and median very close to zero, while the standard deviation grows with distance. The near-zero mean difference in AOD or AE at varying distances implies an even distribution of measurements. Table 2 showcases the values of the mean, median, and standard deviation for AOD and AE at distances with different autocorrelation reduction."

- KORUS-AQ also included a dense deployment of ground based AERONET sites (mostly around Seoul). I wonder if these could be used as an additional data source for the autocorrelation analysis to see if the overall picture of relative scales of variation holds as for the 20 DC-8 flights. While they would not be spatiotemporally collocated with the other data sources used, the data have low uncertainty and good temporal sampling. I am not sure if the inter-site spacings are sufficiently varied to fill out the autocorrelation distance profile, but it could be worth looking at the distance pairings to see if this could be a useful addition.

  While not exactly used for autocorrelation, Choi et al., 2021 examined the distribution of AERONET measurements during KORUS-AQ and the correlation to each other based on distance from different sites. It is more difficult to resolve with stationary measurements, a distance of 50 km roughly translates to a correlation of 85% for fine mode AOD. In contrast the slope of correlation with distance for Fine Mode Fraction is shallower than fine and coarse mode AERONET AOD, suggesting that the aerosol size distribution was the most spatially homogeneous (Fig. 3 in Choi et al., 2021). But the low accuracy of a linear fit to this metric, and the high rate of decrease within the first 100 km of the correlation of one AERONET site to another, suggest a more complex dependence than a linear decrease, as observed in this work (see fig 9). An abbreviated example is presented in section 4.4:

  "These results contrast to those reported by ground-based observations from AERONET during KORUS-AQ, as presented by Choi et al., (2021), which shows smaller changes in FMF than AOD for coarse or fine mode as a function of distance between the AERONET ground sites (0.11/100 km for FMF, 0.16/100 km, and 0.14/100 km for AOD fine and coarse mode). However, Choi et al. (2021) also shows a lower correlation in FMF, and arguably, a non-linear relationship, particularly at distances shorter than 100 km. This non-linear relationship is presented here in Fig. 10. "

  Reference:

Choi, Y., Ghim, Y. S., Rozenhaimer, M. S., Redemann, J., LeBlanc, S. E., Flynn, C. J., Johnson, R. J., Lee, Y., Lee, T., Park, T., Schwarz, J. P., Lamb, K. D. and Perring, A. E.: Temporal and spatial variations of aerosol optical properties over the Korean peninsula during KORUS-AQ, Atmos. Environ., (February), 118301, doi:10.1016/j.atmosenv.2021.118301, 2021.

- Line 355: the Abstract highlights average and variability of AOD/AE for flights below 500 m but the text here highlights those numbers for flights below 1000 m. Later in the paper there’s some discussion of profiles below/above 500 m but the main results here are all framed relative to 1000 m. I thought I’d mention as I’m not sure whether this difference in reporting altitude between the Abstract and main text was intentional.

  We changed the abstract to reflect the values framed relative to 1000 m to be consistent with the values presented in section 4.1. Other changes throughout the manuscript are intentional.

- Figures 3, 6: if I understand correctly the spectral plots are means and standard deviations. The data are shown on a log scale so the lower tails of the standard deviations often go down to the y axis. I think it could be more meaningful to plot geometric means (i.e. mean and standard deviation of log(AOD)) or else median and interquartile range (or central 68% of points). These, especially the latter, would be informative of the shape of the AOD distribution at each wavelength.

  Both figures already showcase the range and interquartile extents for the distribution, but that seems to not be clear. The figure caption and figures (Fig 6 is now Fig 7) themselves have been adjusted and clarified. See here:

  “

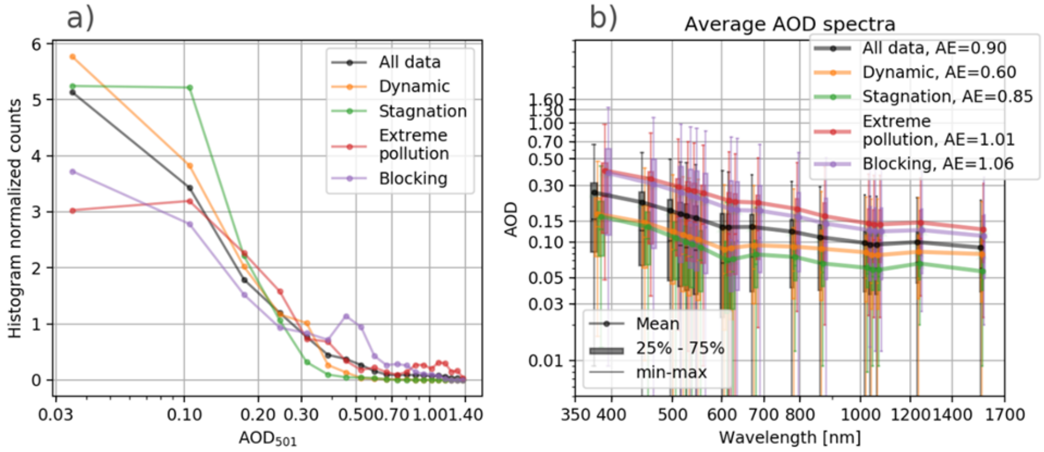

Figure 3 *- (a) Histogram of AOD at 501 nm measured by 4STAR distribution from KORUS-AQ, separated by meteorological periods. (b) Corresponding AOD average spectra for each meteorological period, with the error bars denoting the range of AOD (excluding outliers) during that time period, with the thicker bars denoting the interquartile range. The square symbols and error bars are slightly shifted from each other for clarity. The AE in b) is calculated from the average spectra of each respective meteorological period from 453 nm to 870 nm.*

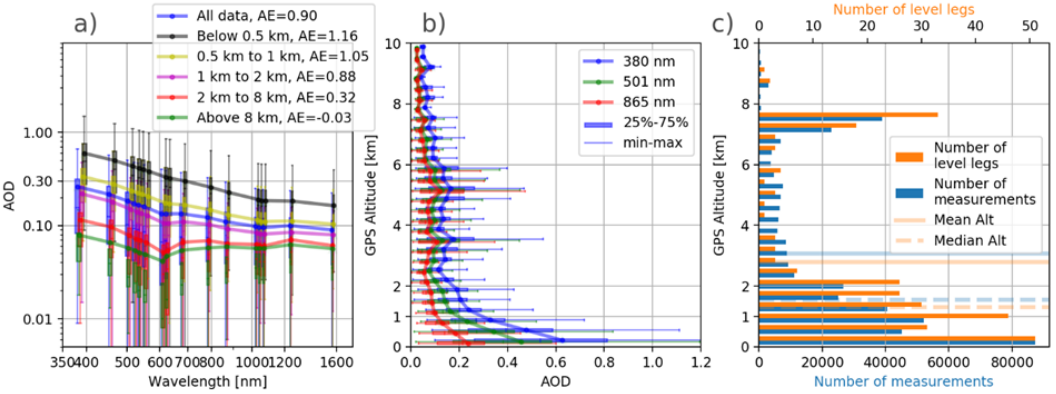

__Figure 7__ - Aggregated AODs observed during KORUS-AQ as a function of observation altitude, for (a) with average AOD spectra, and (b) binned vertically for a subset of wavelengths. The range in binned values are presented by the error bars, while the thicker bar denotes interquartile range (25%-75%). The number of spectra per height bins in a) is 64 736, 41 821, 63 130, 121 569, and 31 076 from lowest to highest respectively. (c) The histogram of the altitude by number of data points (bottom axis) and by number of level legs (top axis), with the mean and median altitudes indicated by solid and dashed lines with the respective colors. "

- Figures B1, B2, B3 and lines 444-448: I am assuming that the regressions here are ordinary least squares (OLS) linear (unless I missed it, it's not explicit). They should really be removed because this technique is inappropriate for these types of data. Some assumptions required for the validity of OLS linear regression include (a) an underlying linear relationship; (b) independent samples; (c) a single underlying (ideally Gaussian) distribution, (d) negligible uncertainty on the independent variable, and (e) equal variance of the dependent variable across the range of the independent variable. Looking at the clouds of points, the linearity assumption appears invalid for B1(b) and B2(b). The independence assumption is likely invalid throughout given the point of this paper shows high levels of correlation across the domain. The distribution shape assumption is likely invalid since AOD tends to be skewed and closer to lognormal, plus the different meteorological fields having different AOD distributions mean we don't have draws from a single distribution but perhaps 4. The independent variable assumption is valid since, as noted, the 4STAR AOD uncertainty in the midvisible is about 0.03, which is not negligible relative to the low AODs commonly found for the bulk of the data. The AE is also uncertain. Note this assumption can be overcome by use of e.g. reduced major axis (RMA) regression accounting for the uncertainty in the independent variable, but this doesn't help with the others. RMA might also be impractical in the present case because my guess is that a non-negligible fraction of the uncertainty in all the data sets here is systematic (e.g. radiometric calibration uncertainty through deployment) so would also be correlated. The equal variance assumption appears to be violated for panel B3(a) and possibly B1(a) (this can also be overcome using weighted regression if pointwise uncertainties are known beforehand). In short, all the data sets violate some of the assumptions, and the numbers and uncertainties presented as regression results are not quantitatively correct. The OLS technique is often used in our field but this does not make it right. I recommend the authors remove the regressions from the plots. In any case I don't think they are really needed to get to the main point about the level of comparability of the data. I think showing R2 is ok (as the collinearity of the data is of interest) but rather than regression equations perhaps some metrics like RMS difference, mean offset, mean absolute difference could be used instead. The discussion in lines 444-448 of the paper should be amended as a result. I don't mean to harp on about this point but since inappropriate regressions are common in our field think it's important to try and stop the practice when I get a chance in peer review.

I've updated the linear fit for all these figures to use a bivariate linear fit, as described by York et al., 2004, and reinforced for use by Cantrell, 2008. These figures also now include the Root-Mean-Square-Error metric for better comparisons and error bars for each x and y coordinate, and are reported in section 4.2. While using this new fit, the corresponding slopes have changed, although the main behavior are unchanged (flatter slope of coarse mode over land than ocean, and better fit for fine mode than coarse mode). The new discussion lines now read:

"In addition to the already known AERONET comparisons over land, we find that the coarse mode AOD from GOCI has a lower RMSE over ocean (RMSE=0.093) than land (RMSE=0.112), albeit with relatively low correlation (R2 of 0.058 and 0.066 respectively). This low correlation is accompanied by nearly flat slope when comparing GOCI to 4STAR coarse mode AOD over ocean (0.26±0.05), and less so over land (0.49±0.12), as estimated using a bivariate linear fit (York et al. 2004). The fine mode AOD is much closer to the expected 1:1 line with slopes of 0.83±0.03 and 0.78±0.09 over ocean and land respectively, and low biases of 0.05 and 0.06 for fine mode AOD (see Fig. B2). "

- Lines 857-859: "Satellite algorithms that assume that aerosol size does not vary as much as aerosol optical depth should be reassessed." I am not aware of data sets produced from algorithms that make assumptions like that, on the scales of tens of km being discussed here – are there any? Most either operate on single pixels (i.e. no spatial constraints) or do multi-pixel processing at a much finer scale than the spatial scales reported on here (e.g. MISR at 4.4 km, GRASP applied to POLDER at 10 km). The VIIRS SOAR ocean algorithm assumes the same fine mode and coarse mode microphysics across 6 km grid cells but AOD and FMF are allowed to vary without spatial constraints for each 750 m pixel within that area. MAIAC used to have some constraints but now retrievals (at 1 km) are spatially independent. It sounds like the GOCI data set used here might (I'm not 100% certain by the way the model selection is described in the paper) but again that's going from 0.5 to 6 km so a lot finer than the scales of variation here. It would be good to either give examples of algorithms here or else delete the comment if there are none using such constraints at the relevant scales.

  Lines 857-859 has been changed to read: "This work showcases that in some regions the spatial scale at which aerosol size varies is smaller than that for aerosol optical depth."

Language comments:

- Title: I am not 100% on this, but I am a bit uneasy about “aerosol optical depth are”. I think it should either be “aerosol optical depths are” or “aerosol optical depth is”.

  Thank you, changed to “aerosol optical depths are”

- “Angstrom” should be typeset as “Ångström” throughout.

  Thank you, changed.

- Line 428: “sporadic aerosols events” should be “sporadic aerosol events”.

  Thank you, changed.