**Response to Nick Schutgens' comment:**

A very useful paper, and it will be interesting to see how to reconcile various measurements. Here I just want to mention the issue of temporal sampling, see also: http://www.atmoschem-phys.net/16/1065/2016/

While this temporal sampling issue is important for model evaluation, it is equally important in comparing different observational datasets. I see two issues relevant to the current paper:

Page 18, Line 732-735: Fig 7 was apparently made with different samplings of the in-situ and AERONET measurements. Is that the case for other figures as well? How might that affect results?

Figures 3-5 were made for matched samplings of the in-situ and AERONET measurements where matched means that the measurements were made within 3 h and 15 km of each other. Given the high correlation in AOD we are reasonably confident that with these sampling constraints the in-situ and remote sensing instruments were measuring in the same atmospheric column. The literature studies cited in Table 4 and included in Figure 6 used similar temporal and spatial matching criteria (see comments column in Table 4), with the exception of the DABEX dust/biomass burning flights (Osborne et al 2008; Johnson et al 2009) which were matched in time but less so in distance (flight profiles were within 100 km of AERONET retrievals).

In contrast, Figures 7 and 8 utilized the multi-year climatological data sets available for each measurement which have different samplings. Your work (e.g., Schutgens et al., 2016) shows there can be large differences when comparing values obtained with different samplings (more than 100% for AOD), particularly when there are high levels of variability in the data. In our manuscript the different temporal samplings are likely one contributor to the relatively small differences observed between the in-situ (red line) and AERONET 1.5 AOD (black line) although other things (e.g., assumptions about aerosol hygroscopicity, missed aerosol (i.e., due to size cut or flight limitations)) will also contribute. The relatively small differences between the in-situ and AERONET 1.5 AOD suggest there may not be much year-to-year variability at these two sites. The long term surface measurements at the site also suggest there is not much year-to year variability. The effects of different sampling are definitely the primary reason for the difference between the AERONET level 2 almucantar values (AOD and AAOD) and the in-situ measurements.

Page 22, Line 925 - 934: The authors suggest better estimates of AAOD may be obtained by using SSA measured at high AOD and applying it to low AOD cases. They mention possible sampling impacts but seem to feel those may not be that important. I'd like to caution against that.

I attach a figure of the difference in yearly SSA, when that SSA is taken at high AOD or at any AOD, for three different models. At least two models allow differences of more than 0.05. (In general, the MIROC-SPRINTARS model agrees best with AERONET Lev 2 SSA while HadGEM-UKCA is often too high and ECHAM-HAM too low.)

We agree that this is an approach to be cautioned against, particularly as systematic variability between loading and SSA has been observed by both in-situ and AERONET measurements at BND, SGP and many other sites (Delene and Ogren, 2002; Andrews et al., 2011b; Schaefer et al., 2014 and our Figure 8). Current work by our group shows this systematic variability is also simulated by many global models. We've re-written the abstract, discussion of Fig 8 and the conclusions to highlight the importance of the systematic variability we've observed and to note that such systematic variability cautions against the use of applying SSA obtained from high loading to obtain AAOD at low loading conditions via the relationship AAOD=SSA*AOD. That said – for the specific case of these two sites, we note that using the monthly median SSA from the high loading retrievals would result in a reasonable monthly median AAOD if the high loading SSA was applied to all AOD values.

Finally, it would be useful if the authors made a suggestion under what conditions AERONET SSA conditions may be used. Is AOD > 0.4 sufficient?
I don't think we can say definitively. Our Figure 6 comparing many field campaign measurements suggests that AOD440>0.25 or 0.3 may be reasonable. Oleg Dubovik (pers. comm. with co-author Stefan Kinne) thinks AOD440>0.4 may be too restrictive but did not suggest a lower alternative. We've added the following text to the discussion of Figure 6: "Figure 6 suggests that AERONET retrievals of SSA could perhaps be used at $AOD_{440}$<0.4, perhaps down to $AOD_{440}$~0.25 or ~0.3 – even at those low AOD values the differences in SSA between AERONET and in-situ still tend to be within the AERONET uncertainty. However, as Figure 6 shows, there are not a lot of direct comparisons to support such a choice."