In this study, the authors compared mineral dust and combustion aerosol distributions from CALIPSO, MERRA-2, and FLEXPART model in the Arctic during 2008-2015. They found that both MERRA-2 and FLEXPART predict the vertical distribution of combustion aerosols with relatively high confidence, as does FLEXPART for mineral dust. MERRA-2 and FLEXPART have substantially higher false negative rates for mineral dust in conditions favouring diamond dust formation. All three products predicted that wintertime dust and combustion aerosols occur most frequently over the same Siberian regions where diamond dust is most common in the winter. The manuscript is well written, and results are clearly presented. I have a few comments for the authors to consider.

General Comments:

The dust size in diameter in MERRA-2 (GOCART aerosol model) should be 0.2-2, 2-3.6, 3.6-6, 6-12, and 12-20 μm. Many papers have cited incorrectly. Please make it clear whether the size ranges are for radius or diameter (both MERRA-2 and FLEXPART), and it is better to be consistent to use diameter as FLEXPART (the numbers appear to be for diameter). For the grouping of the bins, bin 1 contains not only submicron but also supermicron dust. If you want to exclude dust with diameter larger than 10 μm, you may need to use bins 2-4. Or you can use all bins. MERRA-2 has the size up to 20 μm as FLEXPART does. Please make sure that the size bins for MERRA-2 are correctly used and referred throughout the paper.

The authors focused on qualitative comparisons of aerosol presence information between CALIPSO and MERRA-2/FLEXPART. I'm a little confused about the methodology, especially for the introduction/explanation of Table 2. It would be nice if the authors could explain more in details. What does the tested range of minimum concentrations in Table 2 mean? Ranges like "BC: >41 to >100 ng m^{-3}" are confusing. Different vertical levels have different threshold values? Does that minimum range correspond to 67-92.5% quantile? What does 67-92.5% quantile mean? Larger than 67% of the values to 92.5% of the values? Why MERRA-2 and FLEXPART use the same quantile range? How are the range of minimum concentrations used for the calculations of FN and FP rates? How would you determine the amount of dust, BC, and OC to be detectable with CALIPSO? Is it ideal to use a simulator? "Polluted dust" was included in both "dust" and "pollution" group. How much could this affect the FP and FN rates?

Specific comments:

Line 128-129, the assimilation of aerosol data in MERRA-2 outside the Arctic may greatly increase the transport. For mineral dust, the lower to middle troposphere may be affected more by local sources. The middle to upper troposphere may be affected more by the transport.

Line 204-205, it is a little confusing. Is it CALIPSO vs MERRA-2/FLEXPART or dust vs combustion aerosol? Consider rephrasing it?

Figure 1, the panels seem to be small comparing to the large white space between them. Please consider adjusting them. Try to make the panels larger and compact.

Figure 3, "Both models can predict the … that near-surface dust sources may be underestimated" These should better be discussed in the main text instead of in the figure caption.

Figure 4, there seems to be too much white space, like Figure 1, try to make the panels larger and compact. For the colorbar, usually one color indicates a value range. The red and blue color in the colorbar indicate >100% and <0% values? Please change it if necessary.