## Response to Reviewer 2

We appreciate the reviewer's thoughtful comments and key suggestions. Our responses to his/her comments are highlighted in blue:

- Page 2, lines 56-57: I suggest citing the Witte et al. (2018) paper here, along with implications on this study. (Witte, M. K., Yuan, T., Chuang, P. Y., Platnick, S., Meyer, K. G., Wind, G., & Jonsson, H. H. (2018). MODIS Retrievals of Cloud Effective Radius in Marine Stratocumulus Exhibit No Significant Bias. Geophysical Research Letters, 45(19), 10,656–10,664, doi:10.1029/2018GL079325.)

Thank you for drawing our attention to Witte et al., which presents an interesting perspective on satellite retrieval biases dependent on the cloud probe. In the revised paper, we added the following sentence:

"On the other hand, Witte et al. (2018) found an insignificant bias of MODIS Collection 6 (MODIS Science Team retrievals) relative to in-situ Phase Doppler Interferometer (PDI) observations over the subtropical eastern Pacific. While Witte et al. (2018) point to the importance of counting on in-situ observation that fully capture the droplet size distribution, our study relies on two independent airborne datasets, lending confidence in the satellite assessment."

- Page 3, line 86: Just to clarify, the C-130 only flew 1 to 1.5 hour (60 to 100 minute) flights? Seems short to me, so perhaps I'm misunderstanding.

The 1-1.5h flight duration mainly refers to the period of active in-cloud sampling, however, the total flight duration was 10 hour. NAAMES sampling also includes cloud-free and subcloud observations. Furthermore, the 10 hour flight also includes 4 hours transit. This information was updated in the revised manuscript.

- Page 3, lines 96-97: This statement on overcounting being thought to equally affect all size bins is an assertion without evidence. Was this verified to be the case? It might be an acceptable assumption, but there's nothing here that makes that case.

This operation is typical of the operation of optical probes, in which the concentration of a specific bin is equivalent Conc_bin = Counts_bin / (Sample Area) / (True Air Speed) (Brenguier et al., 1994). Thus, the sample area equally affects each bin of the probe, implying that cloud effective radius remains unaltered.

Brenguier, J. L., Baumgardner, D., & Baker, B. (1994). A Review and Discussion of Processing Algorithms for FSSP Concentration Measurements, *Journal of Atmospheric and Oceanic Technology*, *11*(5), 1409-1414.

- Page 3 line 102: Why is the CIP sampling not the same as the CDP? My understanding is that these two in situ instruments are supposed to complement each other to resolve the full width of the droplet size distribution, and are flown together for that reason.

We agree with the reviewer that, ideally, both cloud probes should be operated in tandem. Unfortunately, due to issues with the CIP measurements, a subset of CIP observations were deemed unreliable.

- Page 4, lines 115-117: Yes, the RSP polarimetric re retrievals may be accurate, at least for the synthetic LES cases considered in Alexandrov. But the key question here, given their use as a benchmark for satellite re retrievals, is whether these retrievals should be consistent with those from total reflectance approaches considering their different vertical weighting functions (e.g., Platnick (2000)). The polarimetric signal is a single-scattering phenomenon and thus is sensitive to the very top of the cloud. Looking at the re profiles in Fig. 3 (and from knowledge of similar profiles from other field campaigns), there is a decrease in re at the very top of the cloud. This decrease may in fact be too small to matter, but the authors don't fully address this other than later in the paper stating that using different tau thresholds (1 and 3) in their averaging of "cloud top" CDP measurements yields only a roughly 0.1μm re change. The single-scattering polarimetric signal may be in large part from the portions of the cloud above even 1 optical depth into the cloud. Please comment on this.

We have added the new sentence in the revised manuscript:

"Satellite $r_e$ larger than its RSP counterpart reflects in part the different sensitivity of each method to the cloud top layer. For instance, in-situ vertical profiles in Figure 3 shows a slight decrease in $r_e$ at the cloud top. Because RSP $r_e$ is more sensitive to the optically thinner layer from the cloud top than those estimated from passive 3.7-μm and 3.9-μm channels, it is expected that even for unbiased retrievals, satellite $r_e$ would be larger than RPS $r_e$. However, this discrepancy should be modest as CDP $r_e$ averaged over an optical depth of 0.4 from the cloud top is only 0.17 μm smaller than that calculated for an optical depth of 2.0."

- Page 4, lines 118-121: Radiometric calibration, and relative radiometry between two imagers, can have a big impact on tau retrievals and their agreement between two sensors (see, e.g., Meyer K, Platnick S, Holz R, Dutcher S, Quinn G, Nagle F. Derivation of Shortwave Radiometric Adjustments for SNPP and NOAA-20 VIIRS for the NASA MODIS-VIIRS Continuity Cloud Products. Remote Sensing. 2020; 12(24):4096. https://doi.org/10.3390/rs12244096). While the tau retrieval agreement is quite good later in the paper, did the authors assess the relative radiometry between RSP and MODIS/GOES? It's possible that the good agreement is fortuitous and may be masking larger heterogeneity effects.

RSP is calibrated in the GSFC calibration facility pre- and post- mission, or the ARC calibration facility. In both cases the radiance calibration is traceable to the NIST irradiance standard and uses an integrating sphere as described in https://www.nist.gov/sites/default/files/documents/calibrations/sp250-20.pdf. We have found that RSP is radiometrically stable to within 2% over a period of 5 years, and RSP is thermally controlled to the same room temperature (20°C) at which it is calibrated. Moreover, the agreement between the GLAMR/SIRCUS detector based calibration and the older lamp/irradiance based calibration was generally within 2%, and for the key 865 nm band was within 1%. Figure 8 in McCorkel et al. (2016) shows that Landsat8 OLI and RSP agreed to within about 2% for window channels. Assessing radiometry between MODIS/GOES and RSP is not easy in the North Atlantic because RSP has a smaller footprint than MODIS/GOES and cloud variability would make it very difficult to assess differences at a meaningful level of fidelity. However, Figure 10 of Vermonte et al. (2016) suggests that OLI and MODIS agree well radiometrically at least for the red and NIR bands.

McCorkel, J., Cairns, B., and Wasilewski, A.: Imager-to-radiometer in-flight cross calibration: RSP radiometric comparison with airborne and satellite sensors, Atmos. Meas. Tech., 9, 955–962, https://doi.org/10.5194/amt-9-955-2016, 2016

Vermote, E., Justice, C., Claverie, M., and Franch, B.: Pre- liminary analysis of the performance of the Landsat 8/OLI land surface reflectance product, Remote Sens. Environ., https://doi.org/10.1016/j.rse.2016.04.008, 2016.

- Page 6, lines 177-178: See my comment above on the vertical weighting functions of polarimetry versus total reflectance. I guess for 3.7/3.9μm, the difference in weighting with respect to polarimetry is reduced compared to, say, 1.6μm, but this is a little hand-wavy and there may still be differences.

We agree with the reviewer. Our previous statement was inaccurate. We revise sentence to read: "…the satellite-RSP consistency in the sense that RSP $r_e$ is mostly sensitive to the cloud top (τ~1), comparable to GOES and MODIS (τ~ 2, Platnick; 2000)." Additional information is also provided in our response to reviewer's comment concerning lines 115-117.

- Page 6, lines 204-205 and Fig 4: I suggest adding error bars to this plot similar to those in Fig. 5. For the MODIS vs CDP plot, can you stratify these results by the MODIS 250m heterogeneity index (Liang et al. (2009), again similar to what is done in Fig. 5)? Also, what about sensitivity to the width (effective variance) of the observed droplet size distribution? The satellite retrievals are making an assumption on veff (later on defined as 0.1) – how do these results stratify as a function of divergence of that veff assumption from the observations? Veff can be calculated from the observed DSDs, so I suggest doing that analysis.

Figure was updated following the recommendation of the reviewer. Regarding the inhomogeneity index estimated from satellites, a disadvantage of such calculation is that GOES-13 imager and MODIS pixel resolution is dissimilar. This implies that heterogeneity indices are satellite dependent, which is not ideal. In contrast, RSP is advantageous for deriving a heterogeneity metric as the same index, at the same resolution (which is much higher than MODIS) can be applied to any satellite sensors, becoming a more absolute way of quantifying inhomogeneity. The inherent assumption is that the RSP sampling is statistically representative of the wider area viewed by the satellite sensors.

Concerning veff, it is a challenging analysis due to several reasons. First, the range of variability observed at the cloud top (with most of the samples with veff<0.1, Figure 12) was narrow and the number of matched GOES-RSP samples was insufficient for a robust statistical calculation (see our response to the comment below). In addition, based on ongoing work, the dependencies are highly non-linear and vary with viewing geometry, solar zenith angle, and particle size. The issue is complicated enough to be addressed in a standalone work. While we agree with the reviewer about the scientific value of pursuing a more comprehensive analysis, this is left for future work.

- Page 7, lines 213-215: Using RSP to define the heterogeneity index only provides information in one direction, i.e., along the flight track. Both satellite imagers have

footprints much larger than the width of the RSP footprint, so across-track heterogeneity may be missed. Using the MODIS 250m heterogeneity, as I suggest above, would be helpful. Also, following my previous comment, what is the veff retrieved by RSP for these comparisons? Are the RSP veff generally consistent with CDP, at least where the two can be reasonably compared? I see the RSP veff are shown later in Fig. 12, but there is no stratification of CER retrieval differences as a function of veff deviation similar to what was done for VZA and scattering angle, or even heterogeneity. Veff sensitivity should be a no-brainer to add here.

While the fine resolution of the 250-m MODIS channel is well-suited for inhomogeneity calculations, the coarser resolution of GOES-13 (1 km) would yield inhomogeneities indices that are not comparable with its MODIS counterpart. Instead, as previously discussed, we decided to use RSP as it detects cloud features at much higher spatial resolution, and the calculation can be matched and applied to both GOES-13 and MODIS.

Concerning veff, comparisons between and RSP and CDP droplet size distributions during NAAMES in Alexandrov et al. (2012) suggest that RSP and CDP veff are generally consistent. The scatterplot below depicts $r_e$ differences between GOES-13 and RSP as a function of the effective variance from the RSP. It is unclear from Figure S1 GOES biases are related to veff. However, we need to carry out a more comprehensive analysis to test the hypothesis that cloud retrievals could be sensitive to veff near the rainbow.

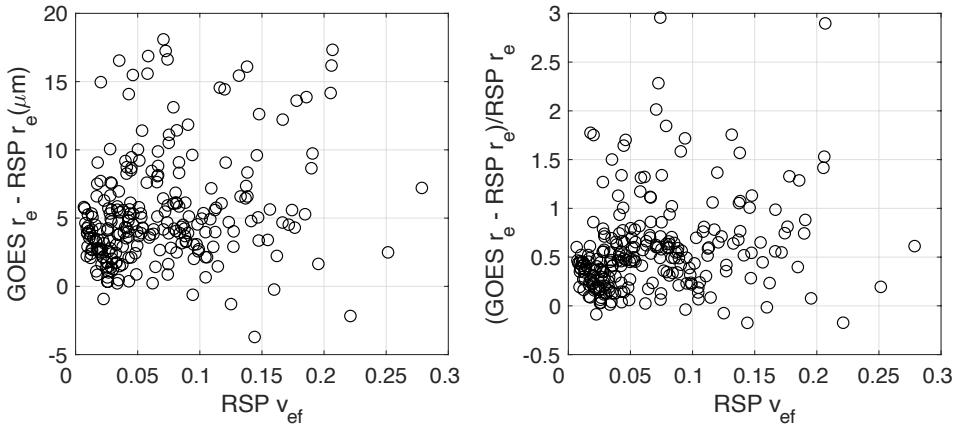


Figure S1: Dependence of GOES-RSP $r_e$ differences relative to RSP effective variance. Absolute differences (left panel), and differences relative to RSP $r_e$.

- Page 7, lines 230-243: Perhaps the MODIS vs GOES re retrieval differences are tied to the rather large central wavelength difference (3.75 vs 3.9µm) and may point to a different forward model issue? Specifically, the liquid index of refraction assumed in the calculation of the cloud single scattering properties – see Platnick et al (2020) for a discussion of re sensitivities to refractive index and temperature (Platnick S, Meyer K, Amarasinghe N, Wind G, Hubanks PA, Holz RE. Sensitivity of Multispectral Imager Liquid Water Cloud Microphysical Retrievals to the Index of Refraction. Remote Sensing. 2020; 12(24):4165. https://doi.org/10.3390/rs12244165). Note that paper shows MODIS 3.75µm vs VIIRS 3.7µm re differences on the order of those shown here, though I admit the impacts of heterogeneity are difficult to disentangle. Can the authors at least comment on the implications of this on their MODIS vs GOES results?

We use the liquid index of refraction from Hale and Querry (1973) for water at 25˚C. Regarding differences in the forward model between MODIS and GOES-13, rather than deriving lookup tables based on GOES-13 central wavelengths, we compute lookup tables using weighted-average optical properties based on the spectral response function. This method should minimize the channel differences between GOES and MODIS.

We have added the following paragraph to address the reviewer's comment:

" In addition to pixel resolution and viewing geometry differences, the dissimilar spectral response between MODIS and GOES-13 imager could yield retrieval discrepancies if the sensor differences are not properly accounted for in the algorithm, especially considering the spectrally wider GOES-13 channel. To circumvent this problem, rather than deriving optical properties for the central wavelength, we derive solar reflectances (lookup tables) for GOES-13 using weighted-average optical properties based on the instrument's spectral response function. An aspect more difficult to address is the retrieval dependence on the index of refraction dataset. Platnick et al. (2021) found that retrieval differences that arise from the choice of refractive index dataset could explain $r_e$ differences between MODIS and the Visible Infrared Imaging Radiometer suite (VIIRS) on Suomi NPP over ocean of about 1 μm for the 3.7-μm band. While the use of a specific refractive index dataset needs to be scrutinized, we note that pixel resolution and viewing zenith angle (Figs. 9 and 11a) well could explain most of the 2 μm bias of $r_e$ GOES-13 relative to MODIS."

- Page 8, lines 255-258: This may be more challenging, but I think you can at least plot the MODIS scattering angle distributions within each GOES scattering angle bin (perhaps as an accompanying box plot). That should indicate scattering angle sampling differences. You should also plot Terra and Aqua MODIS separately, since the scattering angle sampling may be quite different.

We appreciate the reviewer's suggestion. We followed the reviewer's recommendation and separated Terra from Aqua in Figure 10.

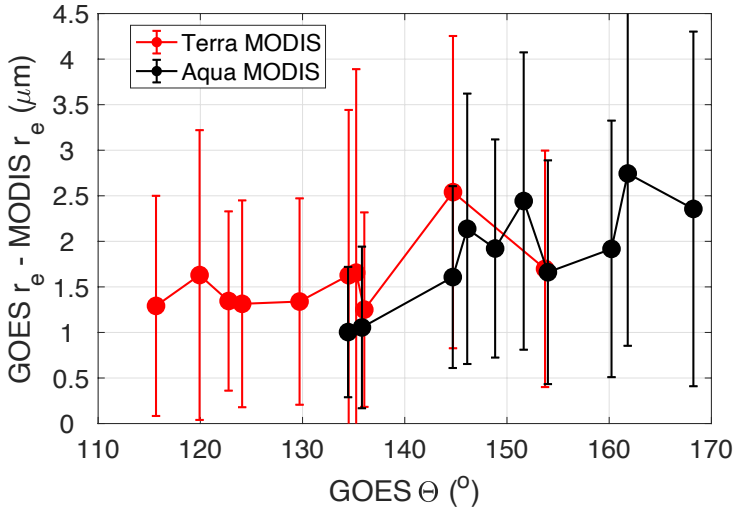


Figure 10: $r_e$ differences between GOES-13 and Terra MODIS (red) and Aqua MODIS (black) binned in deciles of GOES-13 scattering angle ( Θ). Error bars represent the root mean square difference for each bin.

We also analyzed MODIS scattering angle, but did not include this analysis in the paper because it is not possible to disentangle the scattering angle effect from the viewing zenith angle (Figure S2). On average, MODIS scattering angles are typically less than 140˚, with a small fraction of samples with angles >140˚ and a wide range of viewing zenith angle. To reflect this, we wrote in Section 3.3 the following explanation: "A similar analysis applied to MODIS Θ is more challenging because the range of MODIS Θ variability is narrower than GOES, and VZA and Θ cannot be fully disentangled."

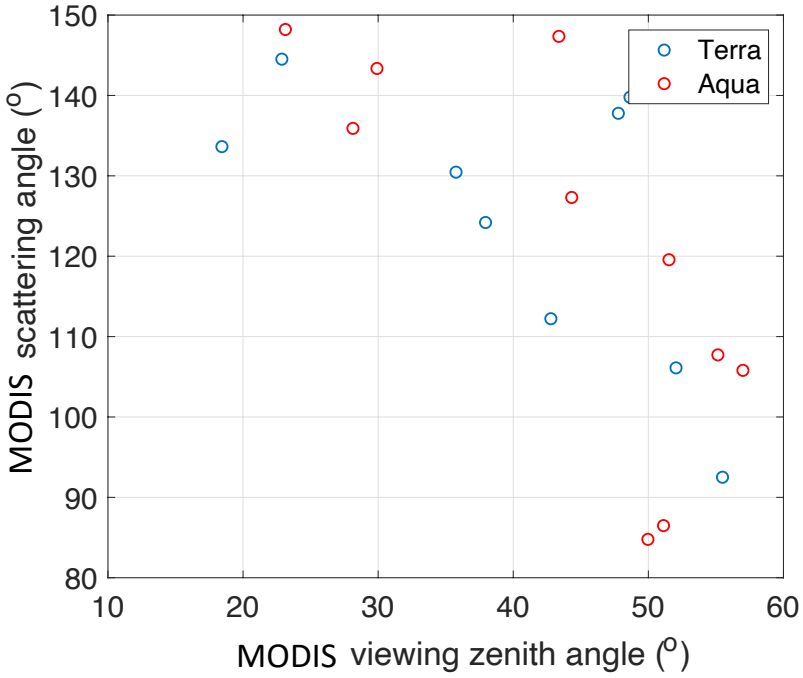


Figure S2: MODIS viewing zenith angle and scattering angle for Terra (blue) and Aqua (red) binned as a function of GOES-16 satellite scattering angle depicted in Figure 12.

- Page 8, line 264-267: While the precipitation likely isn't aliasing into the satellite retrievals, how do the DSDs observed by CDP itself change between precipitating and non-precipitating clouds? If it's significant, it's possible that there may be a correlation with re differences given the assumed veff may deviate more/less from reality.

We compared the CDP effective variance against the precipitation liquid water path from the CIP probe (Figure S3). Interestingly, the amount of near cloud-top precipitation and effective variance are uncorrelated.

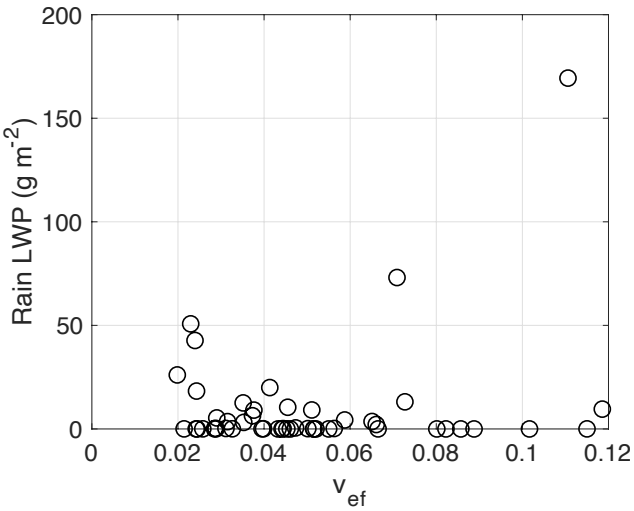


Figure S3: Relationship between CDP (cloud mode) effective variance and liquid water path derived from the CIP probe.

- Page 9, line 278-279: Besides spatial resolution differences between GOES-13 and 16, what about scattering angle differences? This was pointed to as a key player in the MODIS vs GOES-13 differences, and GOES-13 and 16 weren't viewing from the same orbital location.

Scattering angles differences between GOES-13 and GOES-16 are modest, less than 0.5˚ (viewing geometries are nearly identical).

- Page 10, lines 323-325: I don't think investigating veff impacts needs to wait for future work, nor does it require using veff as an additional input to the satellite retrievals (i.e., using various veff in the forward models). As I suggested above, you can simply look at re retrieval differences as a function of RSP veff (or CDP veff). You already have these data from RSP, and can calculate veff quite easily from CDP, so the hypothesis at least can be partially tested here. I suggest the authors do this analysis.

See our previous response to the veff comment.