**Response to Reviewers**

We thank all of the reviewers for their insightful comments. We have carefully considered all the comments and revised our manuscript accordingly. Thank you for taking your time. We hope that we have addressed all the concerns on the manuscript.

Here we provide the point-by-point responses to the reviewers' comments. All reviewers' comments are in black, our responses are in blue, and corresponding changes in the revised manuscript are in red. We have also submitted a revised version of the manuscript and an alternative version that includes all the tracked changes.

We summarize our key messages as follows:

1. We include the practical accuracy of RSP in previous studies to make it easier to compare with the practical accuracy of HSRL-2.

2. We explain the purpose of the RSP MAPP algorithm is to develop a retrieval algorithm that balances both speed and flexibility. Both qualities are important for a successful field campaign. The result is that a simple data quality criterion is desirable during the post-processing stage.

3. We introduce MODIS as the third dataset to perform triple collocation analysis with the two retrievals from ACTIVATE. We show that the number of data points of the third dataset is important than its quality using a synthetic data approach.

4. We investigate the data quality of MODIS which is excellent most of the time. The local variation near the triple collocated data points is also small, suggesting a minimal influence of nearby aerosol gradients.

5. We at first use all available datasets to study possible reasons of disagreement between RSP and HSRL-2. We now add three case studies which include both subpar and high-quality RSP retrievals. Then it is possible to single out the most probable causes of poor retrievals.

**Reviewer 1**

- Line 143: version → versions

  Fixed.

  We updated L173 in the revised manuscript.

- Line 201: It is more correct to say North America. Both Canada and Mexico experience large biomass burning events. Smoke aerosols from those events regularly advect to the Western Atlantic Ocean.

  Agree.

  We updated L243 in the revised manuscript.

- Line 325ff: The discussion here is a little weak. Please check the relative height of the volcanic ash plume. Could there also be some influence from deep convection transporting smoke aerosols?

  One estimate of the plume top height is given in Kloss et al. (2021): 16–17 km from the Himawari-8 satellite. This altitude is above the lapse rate tropopause in the study region.

Liu et al. (2020) reported that climatologically midlatitude overshooting convection occurs more often over land than ocean, e.g., see its Figure 2. Most of these overshooting convections at the 30°–40° latitude bands (i.e., ACTIVATE study region) take place during MAM and JJA (Liu et al., 2020). If these overshooting events did have influence on the stratospheric AOD, the CALIPSO climatology should have shown maxima during these two seasons. However, the CALIPSO climatology shows the opposite pattern, i.e., maxima during SON and DJF but minima during MAM and JJA. Therefore, we conclude that overshooting convection plays only a minor role on stratospheric AOD budget.

We updated L399–413 in the revised manuscript.

**Reviewer 2**

Review of AMT-2023-214

Summary: During the 6-deployment ACTIVATE campaign over the Atlantic, the high-flying King Air supported the HSRL-2 (lidar) and RSP (polarimeter). While both can retrieve AOD, which is better? There are few ground-measurements, so retrieved AODs were triple-collocated with MODIS (satellite) -retrieved AODs to suggest that standard RSP retrievals are the outlier. Adding a stricter cost-function filter helps improve the RSP accuracy (reduce the offset compared to HSRL-2). Physical reasons for the differences between HSRL-2 and RSP includes cloud contamination, aerosols near surface, multiple aerosol layers, absorbing aerosols, non-spherical aerosols, and simplified retrieval assumptions. Neglecting stratospheric aerosol contributions appears to not be a major source of RSP bias (compared to HSRL-2). A conclusion was that "These results demonstrate the pathway for optimal aerosol retrievals by combining information from both lidar and polarimeter for future airborne and satellite missions"

Evaluation: This was one of those papers that seemed like a straightforward intercomparison between 2 products. On my first reading, I was willing to suggest only minor revisions, but then I read it again and gained major concerns. Not so much about what the authors did do (use MODIS as a proxy for sanity checking, and accept a higher level of QA filtering as cost function thresholds), but a lot of questions about what was not attempted or even suggested (examine the physical reasons for differences). A few examples:

- The authors mention *theoretical* $\pm 1\sigma$ accuracy of the MAPP algorithm (RSP) for AOD is $\pm 0.02$. They also mention *practical* accuracy of HSRL-2 as 0.01. Are there studies of the practical accuracy of RSP (compared to AERONET, AATS/4-STAR or other sunphotometer)? I don't know the RSP literature, but I know it was flown often and there must be such information available?

  The practical accuracy of RSP has been estimated in several field campaigns by comparing with sunphotometers but most of them took place over land. Knobelspiesse et al. (2011) found that the mean error of RSP relative to the Ames Airborne Tracking Sunphotometer (AATS-14) is 0.057 at 532 nm for a forest fire event during Arctic Research of the Composition of the Troposphere from Aircraft and Satellites (ARCTAS) field campaign. Di Noia et al. (2017) compared RSP aerosol retrievals against AERONET during the POlarimeter DEfinition EXperiment (PODEX) and Studies of Emissions and Atmospheric Composition, Clouds and Climate Coupling by Regional Surveys (SEAC$^4$RS) field campaigns in 2012 and

2013. The bias and RMSE of these retrievals at 500 nm are 0.02 and 0.04, respectively. Fu et al. (2020) compared measurements from four multi-angle polarimeters, including RSP, against AERONET during Aerosol Characterization from Polarimeter and Lidar (ACEPOL) field campaign over the western part of the United States in 2017. The standard deviations between RSP and AERONET are 0.035, 0.027, and 0.014 at 380 nm, 440 nm, and 675 nm, respectively.

We updated L97–107 in the revised manuscript.

- The RSP standard retrieval assumes a near-surface, non-absorbing coarse spherical sea salt. What are the sensitivities to assuming something else? Would that help improve the retrievals? Same questions about SSA of the fine modes. Would that reduce the cost function $\chi$ from 0.15 to 0.05, and thus lead to a more "successful" retrieval? I am not satisfied simply to change the cost function threshold.

  The RSP MAPP algorithm version 1.48, consisting of a vector radiative transfer model and a Mie code (Stamnes et al., 2018; Schlosser et al., 2022). The retrieval algorithm is geared for field campaigns which desire both speed and flexibility. Speed is important for field campaign because data are needed as soon as possible for preliminary data analysis and flight planning. To speed up computations, the Mie code assumes particles are spherical and homogeneous. Flexibility is also important because the traditional pre-computed Mie lookup table approach requires constant aircraft altitudes and aerosol locations, which limits the choice of flight patterns. To remain flexible, both the vector radiative transfer and Mie computations are done on the fly.

  While there have been advances on computing scattering properties with non-spherical particles, it is still a challenging task because aerosol properties such as size distribution, shape, and composition are still poorly constrained by observations. Because of these constraints, it is desirable to have a simple data quality criterion in the post-processing stage.

  We updated L76–82 in the revised manuscript.

- Since you are collocating with MODIS, what does MODIS suggest about cloud-cover/fraction and all the issues that might make collocation (and retrieval) difficult? Also, what about gradients of aerosol nearby the collocation?

  Thanks for raising this question. For MODIS 3-km aerosol product, each grid box consists of $6 \times 6 = 36$ pixels (i.e., each pixel is at 500 m resolution at the nadir). The MODIS aerosol over ocean algorithm is complicated (see, for example, Remer et al. (2005)). Generally, the algorithm checks the number of good pixels in each grid box using multiple criteria before doing the actual inversion. Therefore, the existence of cloudy pixels does not stop the retrieval as long as there are enough good pixels within the same grid box (Remer et al., 2013). Note also that the cloud pixels would not be included during the inversion.

  Since there can be a multitude of factors leading to a failed retrieval, it is much easier to evaluate the triple collocated MODIS data points using quality flag. There are four values from 0 to 3, with 3 being the best quality. Levy et al. (2013) recommend to use data with nonzero quality flags. Figure R1a shows that all triple collocated data points are at least 1 (red line), which is much higher than the MODIS data points within a 25-km radius of

RSP/HSRL-2 data points (blue line). Figure R1b shows that the cloud fraction of over 55% of triple-collocated data points is less than 0.1.

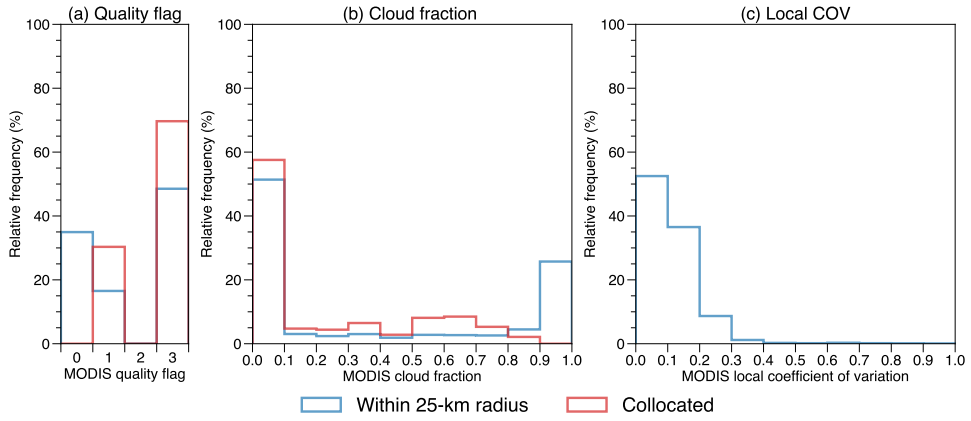


Figure R1: Histograms of the triple collocated MODIS AOD dataset: (a) Quality flag. (b) Cloud fraction. (c) Local coefficient of variation. Red color represents all the triple collocated MODIS data points. Blue color represents all MODIS data points within a 25-km radius of RSP/HSRL-2.

It is difficult to evaluate the influence of the spatial gradient of aerosols. We adapted the local coefficient of variation [LCOV$(r) = \sigma/\mu$] from (Anderson et al., 2003), which is the ratio of standard deviation and mean of all MODIS data points within a certain radius $r$ of collocated RSP/HSRL-2 data points. In this case, $r = 25$ km. Figure R1c shows that around 50% and 90% of LCOV are less than 0.1 and 0.2, respectively. LCOV of 0.1 means that for a normally distributed random variable, 68% of data points are within 10% of the mean (Anderson et al., 2003). Similarly, 68% of data points are within 20% of the mean when LCOV is 0.2. The mean HSRL-2 AOD is $\sim$0.1. Therefore, most of the local AOD variation is within 0.02, which is smaller than the expected error of the Dark Target algorithm. Interestingly, the current result is similar to the LCOV distributions obtained in Anderson et al. (2003) at similar spatial scale.

We updated L159–164 and L270–284 and added Figure R1 as Figure 4 in the revised manuscript.

- My biggest issue I think is the lack of attempting to look at the rest of the ACTIVATE data. There were 162 coordinated lower-level Falcon flights measuring in-situ properties. While not many, there are AERONET sites nearby (Bermuda, LaRC, Hampton, Wallops). What do they tell you about clouds? About aerosol layering? About the ocean surface? Etc?

  Thanks for your suggestion. We decided to combine with the next suggestion to have a full response.

- This is one of those studies where doing case studies (one collocation at a time) could be instructive. Pick one, learn something. Pick another interesting case, do it. Maybe one more. Then you can derive statistics and attempt to generalize.

Again, thanks for your suggestion.

We have checked all AERONET sites nearby. For those AERONET sites near NASA LaRC, King Air was at the initial climb or final descent stage. There are no RSP/HSRL-2 data for comparison because HSRL-2 is affected by the near range limit and RSP needs to be under minimal aircraft yaw and roll. For the AERONET site in Bermuda, only 2 days of data are available during June 2022 but there are no coincident King Air flights.

There are three new case studies for three research flights, including RF67, RF143, and RF170 (Figures R2–R4). Data from the low-flying Falcon are also used to provide near-surface information.

We updated L428–461 as a new subsection 4.5 and added Figures R2–R4 as Figures 12–14 in the revised manuscript .

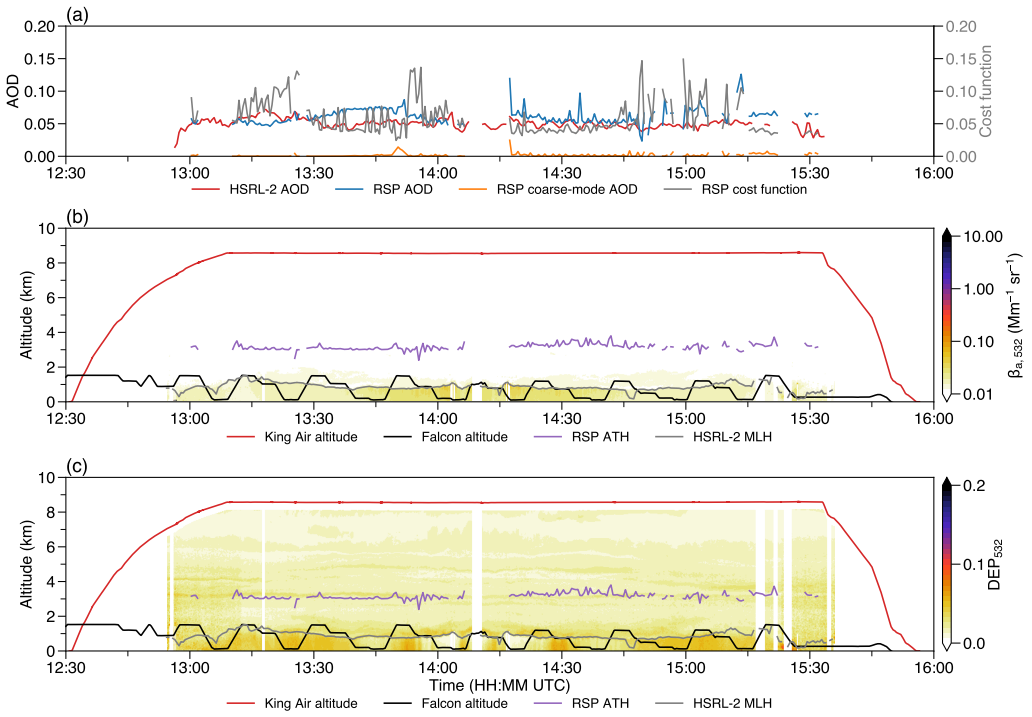


Figure R2: Time series of selected measurements during ACTIVATE RF67 on 19 May 2021. (a) HSRL-2 (red), RSP total (blue), and RSP coarse-mode (orange) AOD. RSP normalized cost function (gray). (b–c) King Air (red) and Falcon (black) altitude. RSP aerosol top height (purple). HSRL-2 mixing layer height (gray). All altitude values are in km. (b) HSRL-2 532 nm aerosol scattering coefficient $\beta_{a,532}$ in $Mm^{-1}$ $sr^{-1}$ (color). (c) HSRL-2 532 nm depolarization ratio (color).

- Finally, I am unconvinced by the conclusion statements regarding the need to combine information from lidar and polarimeter. Well, yes, I believe it, but right now I see the RSP retrieval being subject to poor assumptions, which I can tell are poor even before bringing the HSRL-2 retrievals to confirm. Pick one the cases, show us how the lidar information helps the polarimeter (and possibly vice-versa). Then I would accept that statement.

Thanks for your suggestion. The last question led us to add three case studies. The statistical analysis with case studies of both subpar and high-quality RSP retrievals shows that it is possible to single out the most probable causes of poor retrievals, such that RSP AOD retrievals can be optimally improved.

We updated L474–479 (Conclusion 4) in the revised manuscript .

Questions and comments related to text:

- Lines 78 then line 84: Introduces $\chi^2$ and then $\chi'$. Are these supposed to be the same terms?

  $\chi^2$ and $\chi'$ are not the same term. The RSP Microphysical Aerosol Properties from Polarimetery (MAPP) algorithm uses an optimal estimation approach to minimize cost function $\chi^2$ of the state vector $\mathbf{x}$ (Stamnes et al., 2018),

  $$\chi^2(\mathbf{x}) = \Phi(\mathbf{x})_{\text{data}} + \Phi(\mathbf{x})_{\text{prior}}, \tag{R1}$$

  where the terms on the rhs symbolically represent the data term and a priori term. The data term takes into account the difference between measurement $\mathbf{y}$ and forward model $\mathbf{f}$, and the measurement error covariance matrix $\mathbf{S}_\varepsilon$. The a priori term takes into account the state vector $\mathbf{x}$ and any prior knowledge on the state vector $\mathbf{x}_a$, and the error covariance matrix $\mathbf{S}_a$ of $\mathbf{x}_a$.

  The *normalized* cost function of the data term $\chi'$ is defined as

  $$\chi' = \frac{1}{m}\sqrt{\Phi(\mathbf{x})_{\text{data}}}, \tag{R2}$$

  where the square root of the data term is normalized by the total number of measurements $m$.

  We updated L83–96 in the revised manuscript.

- Line 104: Reference for the 10-s HSRL product?

  The output temporal resolution is usually field campaign dependent. The appropriate reference is the readme files from our ACTIVATE field data repository. For example, here is the readme file for 2022 HSRL-2 data. Note that higher temporal resolution files are available upon request.

- Line 176: Eq 10. Think subscript j should be replaced by subscript 1. Should there be a $\pm$sgn function too?

  Good catch on the subscript. We checked again equation 9 of McColl et al. (2014) and there is no sgn function.

  We updated L210 in the revised manuscript.

- Line 221: "introduce MODIS AOD as third dataset." As you clearly say, MODIS is not a ground-truth. Maybe this is a dumb question, but how "good" does this 3rd dataset need to be to be useful the RSP vs HSRL-2 sanity check? The authors of the MODIS dataset claim something like $\pm$(0.04 + 10%) for accuracy over ocean. Is this good enough?

Thanks for raising this important question. It will be helpful to study some synthetic datasets in order to generalize the results. The synthetic dataset approach has been used in some triple collocation studies (e.g. Su et al. (2014)).

Assume that we know the ground truth, which is given as a sinusoidal function, say, one full cycle with mean $= 0.1$, minimum $= 0$, maximum $= 0.2$. For example, in python, such an array with 5000 data points can be created using numpy library from the following code,

```
np.sin(np.linspace(0, 2*np.pi, num=5000))*0.1 + 0.1
```

Then we can add some white noise to the ground truth to generate synthetic datasets. The amount of noise is controlled by a signal-to-noise ratio (SNR), which is simply the ratio of ground truth variance and noise variance. To give some context using real world data, we can estimate the SNR of the datasets in this study by $b_i^2 \mathrm{Var}[\Theta]/\mathrm{Var}[\varepsilon_i]$ (Gruber et al., 2016; McColl et al., 2014). The SNRs of RSP, HSRL-2, and MODIS (2344 triplets in total) are 1.74, 5.98, and 2.79, respectively.

We now have the experiments 1–3 with three synthetic datasets A, B, and C, each having 5000 data points (Table R1). The only difference is the SNR of dataset C (1, 5, and 25). Note that all datasets are re-generated for each experiment, so, for example, dataset A in experiment 1 is not the same as dataset A in experiment 2.

Table R1: Summary of the synthetic experiments 1–3 with $n = 5000$ data points. Signal-to-noise ratio (SNR), error standard deviation $\sigma$, and error correlation coefficient $r$ with respect to the ground truth are provided for each synthetic dataset.

| Expt | $n$ | A | | | B | | | C | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SNR | $\sigma_{\varepsilon_A}$ | $r_A$ | SNR | $\sigma_{\varepsilon_B}$ | $r_B$ | SNR | $\sigma_{\varepsilon_C}$ | $r_C$ |
| 1 | 5000 | 5 | 0.0329 | 0.906 | 5 | 0.0311 | 0.915 | 1 | 0.0699 | 0.713 |
| 2 | 5000 | 5 | 0.0312 | 0.916 | 5 | 0.0314 | 0.915 | 5 | 0.0308 | 0.917 |
| 3 | 5000 | 5 | 0.0325 | 0.909 | 5 | 0.0315 | 0.913 | 25 | 0.0134 | 0.983 |

The result shows that all TC metrics of datasets A and B are not very *sensitive* to the amount of noise in dataset C (i.e., the quality of dataset C). We now repeat the same set of experiments but with only 500 data points (Table R2).

Table R2: Summary of the synthetic experiments 4–6 with $n = 500$ data points. Signal-to-noise ratio (SNR), error standard deviation $\sigma$, and error correlation coefficient $r$ with respect to the ground truth are provided for each synthetic dataset.

| Expt | $n$ | A | | | B | | | C | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SNR | $\sigma_{\varepsilon_A}$ | $r_A$ | SNR | $\sigma_{\varepsilon_B}$ | $r_B$ | SNR | $\sigma_{\varepsilon_C}$ | $r_C$ |
| 4 | 500 | 5 | 0.0344 | 0.891 | 5 | 0.0316 | 0.911 | 1 | 0.0735 | 0.671 |
| 5 | 500 | 5 | 0.0312 | 0.917 | 5 | 0.0319 | 0.909 | 5 | 0.0310 | 0.914 |
| 6 | 500 | 5 | 0.0311 | 0.918 | 5 | 0.0306 | 0.921 | 25 | 0.0175 | 0.972 |

All TC metrics of datasets A and B now have larger variations. This shows that the robustness of the TC metrics is *more sensitive* to the number of data points than the data quality. Therefore, in this study we try to maximize the number of data points, even it may introduce some small biases.

One powerful benefit of this synthetic data approach is that since both the ground truth and errors are *known*, we do not need to compute all the TC metrics via the equations given in the manuscript using synthetic datasets. That is to say, we can compute all these TC metrics directly from the ground truth and errors.

We added S1 in the supplement (in a separate file).

- Line 222: "we combine both datasets to maximize the number of files for collocation." What if Terra or Aqua are biased differently?

  Thanks for raising this interesting question. While Terra and Aqua are almost identical satellite sensors, it is not easy to directly compare between the two instruments because they have different overpass times. Terra and Aqua are often validated with AERONET measurements. For example, Remer et al. (2008) and Levy et al. (2013) concluded that the performance of two MODIS instruments is very similar. Another approach is adopted in Antuña Marrero et al. (2018): compare Aqua, Terra, and Aqua+Terra separately with measurement from one AERONET site in Cuba. Selected results from 7-year data are tabulated below (Table R3). MODIS AOD were obtained by the Dark Target algorithm. Generally, the differences in RMSD, MB, and $r$ between the three datasets are small. Therefore, it gives us confidence to combine Terra and Aqua data.

  Table R3: Summary statistics for different MODIS datasets versus AERONET in Antuña Marrero et al. (2018).

| | Terra vs AERONET | Aqua vs AERONET | Terra+Aqua vs AERONET |
|---|---|---|---|
| RMSD | 0.061 | 0.064 | 0.062 |
| MB | 0.007 | 0.017 | 0.010 |
| $r$ | 0.701 | 0.794 | 0.742 |

  We updated L266–268 in the revised manuscript.

- Line 225: Do you do any checking of the MODIS imagery to determine if there appears to be a rapid gradient nearby the 3-way collocation?

  We did check the MODIS imagery but it is difficult to quantify the influence of the spatial gradient of aerosols. We adapted the local coefficient of variation [LCOV$(r) = \sigma/\mu$] from (Anderson et al., 2003), which is the ratio of standard deviation and mean of all MODIS data points within a certain radius $r$ of collocated RSP/HSRL-2 data points. In this case, $r = 25$ km. Figure R1c shows that around 50% and 90% of LCOV are less than 0.1 and 0.2, respectively. LCOV of 0.1 means that for a normally distributed random variable, 68% of data points are within 10% of the mean (Anderson et al., 2003). Similarly, 68% of data points are within 20% of the mean when LCOV is 0.2. The mean HSRL-2 AOD is $\sim$0.1. Therefore, most of the local AOD variation is within 0.02, which is smaller than the expected error of the Dark Target algorithm.

  We updated L277–284 in the revised manuscript.

- Line 242–260: I find this super confusing in terms of its presentation. I can understand why you might think of restricting threshold for "fitting error/cost function" $\chi$, but don't see the logic of attempting to restrict by any of the retrieved fine or coarse AOD parameters.

Three filter candidates are sought, including RSP normalized cost function $\chi'$, coarse-mode AOD $\tau_c$, and fine-mode AOD $\tau_f$. $\tau_c$ and $\tau_f$ are chosen as they are part of retrieved parameters. For a successful retrieval, other than $\chi'$ being lower than 0.15, the algorithm also requires that $1 \times 10^{-5} \leq \tau_c \leq 0.3$ and $1 \times 10^{-5} \leq \tau_c \leq 0.7$. Because $\tau_c$ and $\tau_f$ are rarely larger than 0.2 and 0.5, respectively, lowering either of them may help eliminate outliers.

We updated L301–307 in the revised manuscript.

- Line 255–257: By reducing the cost function threshold to 0.05, it is at the expense of reducing available points by 50%. Is this acceptable?

  There is always some trade-off between quantity and quality for a specific threshold. This manuscript aims to evaluate the state-of-the-art instruments and their associated uncertainties. Therefore, it is acceptable to decrease available data points in order to reduce the uncertainties. Alternatively, if the goal is to characterize the long term aerosol climatology, it will be more desirable to keep more data points.

- Figs 5 and 6: Not authors fault, but really frustrating that the figures are shown >1 page after the text describing them.

  LaTeX tends to surprise us with the figure placement. We did try to make both figures smaller so that they can squeeze into one page. But the drawback is that the text in figures becomes too small.

- Line 272: Don't understand phrase "which indicates that the bias is possibly lagging behind the cloud influence."

  Thanks for pointing it out. Originally we speculated that extended period of cloud (i.e., higher cloud mask fraction; see also the next question on the definition of cloud mask fraction) may lead to higher AOD bias. The mean and median AOD biases do not change much for different cloud mask fraction. However, we accept that there is not enough evidence to prove that there is a delayed bias on the cloud influence.

  We updated L337–339 in the revised manuscript.

- Line 274: Don't understand what this means: "There is only one year of below-aircraft cloud mask data but it includes other properties such as dynamic cloud fraction which also considers the influence from glint." Which year? A summer and a winter? Please explain "dynamic cloud fraction" and the "influence from glint".

  Agree. We only have 1 year of cloud mask data that covers 2021 summer and winter deployments. To compute cloud fraction, we count the number of cloud pixels and divide it by the total number of pixels on a given camera imagery. However, sun glint on the imagery may skew the statistics (i.e., including or excluding glint pixels). In this study, we chose to use the dynamic cloud fraction which excludes glint pixels.

  We updated L333–336 in the revised manuscript.

- Line 276: I must be missing something here: Isn't it of course obvious that: "Higher cloud mask fraction is associated with higher below-aircraft cloud cover (Figure 7c)." Is there a difference between cloud cover fraction and cloud mask fraction?

We use the above-aircraft and below-aircraft cloud masks to detect the presence of cloud that leads to the contamination of aerosol retrievals. Cloud mask is a binary variable being either zero (absence) or one (presence). Because the temporal resolution of cloud mask is 1 Hz, we compute the fraction of cloud mask being one (i.e., cloud is present) during each 30-s window of a 3-km HSRL-2 data point.

It is not necessary that the presence of cloud is related to high cloud cover fraction, but higher cloud mask fraction is associated with higher below-aircraft cloud cover.

We updated L329–333 and L341–344 in the revised manuscript.

- Line 277: "This implies that there is not a linear relationship between the cloud cover and bias." I am wondering if there is a relationship between cloud cover (or mask) and the fitting error/cost function? Note that RSP-HSRL "bias" is plotted (Fig 7) in absolute terms, not in percentage terms, so maybe we don't care about small biases if the AOD is large.

  Thanks for your suggestion. Figure R5 below shows the RSP cost function as a function of above-aircraft and below-aircraft cloud parameters. While the range of cost function changes somewhat with all cloud parameters, the mean and median cost functions are pretty steady around 0.05, suggesting there is not a linear relationship between cost function and cloud cover fraction/cloud mask fraction.

  We updated L343–344 in the revised manuscript.

- Line 290: "Practically dusty mix is a mixture…" This makes no sense.

  Agree.

  We updated L357 in the revised manuscript.

- Line 292: This is backwards: "An SSA value from 0.90 to almost 1 represents a weakly to moderately absorbing aerosol" (0.90 is absorbing, approaching 1 is non-absorbing)

  Agree.

  We updated L359–360 in the revised manuscript.

- Line 305: Suggest insert particle to read "the shape of a sea salt particle depends".

  Agree.

  We updated L377–379 in the revised manuscript.

- Line 306: "reported that sea salt may cause high depolarization". Why is that? Because of shape? It becomes non-spherical?

  Generally, linear depolarization ratio is zero for spherical particles but markedly departs from zero for non-spherical particles (Mishchenko and Hovenier, 1995).

  As part of marine aerosols, sea salt is another potential source for contributing to high depolarization in the marine boundary layer due to its shape. Pure dry sea salt is non-spherical but its shape is sensitive to ambient relative humidity (RH). When RH increases, the sea salt particle remains solid until it reaches the deliquescence RH at which the solid particle absorbs water and forms aqueous solution. The aqueous-phase sea salt is nearly spherical and therefore has low depolarization. On the contrary, when RH decreases, the aqueous sea

salt solution can return to its crystalline form at the efflorescence RH. The deliquescence and efflorescence RH are usually different. Therefore, the shape of a sea salt depends not only on the magnitude but also the history of ambient relative humidity (Haarig et al., 2017). Ferrare et al. (2023) reported that sea salt may cause high depolarization at low RH (i.e., more likely to be non-spherical) within the boundary layer for around one third of flights during the first two years of ACTIVATE based on HSRL-2 and dropsonde measurements.

We updated L365–366 and L372–382 in the revised manuscript.

- Lines 320–325: For someone not as familiar with aircraft measurements, I am thinking a diagram might be helpful. What are each of these instruments observing above and below the aircraft, and what layering is "missing" because of the way these sensors must work?

  Thanks for the suggestion. See the following schematic (Figure R6).

  We updated L396–399 and added Figure R6 as Figure 10 in the revised manuscript.

- Lines 329–331: The chronology doesn't make sense: "The peak AOD in 2020 is very likely due to the stratospheric eruption at Raikoke (48° N, 153° E) in June 2019 (Kloss et al., 2021). Subsequently the stratospheric AOD increased up to 0.027 in October 2019 (not shown) and gradually decreased in the ensuing months." Nonetheless, 0.01-0.02 stratospheric AOD doesn't explain magnitude of 10 higher biases.

  Volcanic eruptions can have a prolonged influence on the stratospheric AOD. For example, (Kovilakam et al., 2020) compiled a dataset called the Global Space-based Stratospheric Aerosol Climatology (GloSSAC). Figure 15b of (Kovilakam et al., 2020) shows some of the most dramatic eruption in the last century, including El Chichón in 1982 (17° N), Pinatubo in 1991 (15° N), and Mount Hudson in 1991 (46° S), which soared the stratospheric AOD to well above 0.1 and lasted for more than 1 year. Since Raikoke is located in the northwest Pacific Ocean (between Japan and Russian Far East), it would take some time to build up the peak at the ACTIVATE study region.

  We agree that stratospheric AOD cannot explain the current bias.

  We updated L408–413 in the revised manuscript.

- Lines 359 (conclusion 4.): "It proves that the quality of AOD retrievals can be optimally improved by combining information from both lidar and polarimeter." I definitely believe it, but I don't see how this study proved that statement.

  We have added three case studies which serve as a complement to the original statistical analysis. The case studies include both subpar and high-quality RSP retrievals. It is then possible to single out the most probable causes of poor retrievals, such that the AOD retrievals can be optimally improved.

  We updated L474–479 (Conclusion 4) in the revised manuscript .

- Lines 360–365: To me this discussion is disjointed from the rest the paper. How does this study lead to desire to use 2-dimensional imaging? (which I don't even know what that is).

  The current HSRL-2 MLH retrieval algorithm detects the edge of aerosol gradient in each one-dimensional column. Instead, reading the whole two-dimensional flight curtain at once

and using two-dimensional wavelet transform will potentially speed up the computation and improve the detection accuracy.

We updated L484–486 in the revised manuscript.

**Reference**


Anderson, T. L., Charlson, R. J., Winker, D. M., Ogren, J. A., and Holmén, K.: Mesoscale variations of tropospheric aerosols, Journal of the Atmospheric Sciences, 60, 119–136, doi: 10.1175/1520-0469(2003)060⟨0119:mvota⟩2.0.co;2, 2003.

Antuña Marrero, J. C., Cachorro Revilla, V., García Parrado, F., de Frutos Baraja, A., Rodríguez Vega, A., Mateos, D., Estevan Arredondo, R., and Toledano, C.: Comparison of aerosol optical depth from satellite (MODIS), sun photometer and broadband pyrheliometer ground-based observations in Cuba, Atmospheric Measurement Techniques, 11, 2279–2293, doi: 10.5194/amt-11-2279-2018, 2018.

Di Noia, A., Hasekamp, O. P., Wu, L., van Diedenhoven, B., Cairns, B., and Yorks, J. E.: Combined neural network/Phillips–Tikhonov approach to aerosol retrievals over land from the NASA Research Scanning Polarimeter, Atmospheric Measurement Techniques, 10, 4235–4252, doi: 10.5194/amt-10-4235-2017, 2017.

Ferrare, R., Hair, J., Hostetler, C., Shingler, T., Burton, S. P., Fenn, M., Clayton, M., Scarino, A. J., Harper, D., Seaman, S., Cook, A., Crosbie, E., Winstead, E., Ziemba, L., Thornhill, L., Robinson, C., Moore, R., Vaughan, M., Sorooshian, A., Schlosser, J. S., Liu, H., Zhang, B., Diskin, G., DiGangi, J., Nowak, J., Choi, Y., Zuidema, P., and Chellappan, S.: Airborne HSRL-2 measurements of elevated aerosol depolarization associated with non-spherical sea salt, Frontiers in Remote Sensing, 4, 1–18, doi: 10.3389/frsen.2023.1143944, 2023.

Fu, G., Hasekamp, O., Rietjens, J., Smit, M., Di Noia, A., Cairns, B., Wasilewski, A., Diner, D., Seidel, F., Xu, F., Knobelspiesse, K., Gao, M., da Silva, A., Burton, S., Hostetler, C., Hair, J., and Ferrare, R.: Aerosol retrievals from different polarimeters during the ACEPOL campaign using a common retrieval algorithm, Atmospheric Measurement Techniques, 13, 553–573, doi: 10.5194/amt-13-553-2020, 2020.

Gruber, A., Su, C.-H., Zwieback, S., Crow, W., Dorigo, W., and Wagner, W.: Recent advances in (soil moisture) triple collocation analysis, International Journal of Applied Earth Observation and Geoinformation, 45, 200–211, doi: 10.1016/j.jag.2015.09.002, 2016.

Haarig, M., Ansmann, A., Gasteiger, J., Kandler, K., Althausen, D., Baars, H., Radenz, M., and Farrell, D. A.: Dry versus wet marine particle optical properties: RH dependence of depolarization ratio, backscatter, and extinction from multiwavelength lidar measurements during SALTRACE, Atmospheric Chemistry and Physics, 17, 14 199–14 217, doi: 10.5194/acp-17-14199-2017, 2017.

Kloss, C., Berthet, G., Sellitto, P., Ploeger, F., Taha, G., Tidiga, M., Eremenko, M., Bossolasco, A., Jégou, F., Renard, J.-B., and Legras, B.: Stratospheric aerosol layer perturbation caused by the 2019 Raikoke and Ulawun eruptions and their radiative forcing, Atmospheric Chemistry and Physics, 21, 535–560, doi: 10.5194/acp-21-535-2021, 2021.

Knobelspiesse, K., Cairns, B., Ottaviani, M., Ferrare, R., Hair, J., Hostetler, C., Obland, M., Rogers, R., Redemann, J., Shinozuka, Y., Clarke, A., Freitag, S., Howell, S., Kapustin, V., and McNaughton, C.: Combined retrievals of boreal forest fire aerosol properties with a polarimeter and lidar, Atmospheric Chemistry and Physics, 11, 7045–7067, doi: 10.5194/

acp-11-7045-2011, 2011.

Kovilakam, M., Thomason, L. W., Ernest, N., Rieger, L., Bourassa, A., and Millán, L.: The Global Space-based Stratospheric Aerosol Climatology (version 2.0): 1979–2018, Earth System Science Data, 12, 2607–2634, doi: 10.5194/essd-12-2607-2020, 2020.

Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., and Hsu, N. C.: The Collection 6 MODIS aerosol products over land and ocean, Atmospheric Measurement Techniques, 6, 2989–3034, doi: 10.5194/amt-6-2989-2013, 2013.

Liu, N., Liu, C., and Hayden, L.: Climatology and Detection of Overshooting Convection From 4 Years of GPM Precipitation Radar and Passive Microwave Observations, Journal of Geophysical Research: Atmospheres, 125, doi: 10.1029/2019jd032003, 2020.

McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., and Stoffelen, A.: Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target, Geophysical Research Letters, 41, 6229–6236, doi: 10.1002/2014gl061322, 2014.

Mishchenko, M. I. and Hovenier, J. W.: Depolarization of light backscattered by randomly oriented nonspherical particles, Optics Letters, 20, 1356, doi: 10.1364/ol.20.001356, 1995.

Remer, L. A., Kaufman, Y. J., Tanré, D., Mattoo, S., Chu, D. A., Martins, J. V., Li, R.-R., Ichoku, C., Levy, R. C., Kleidman, R. G., Eck, T. F., Vermote, E., and Holben, B. N.: The MODIS aerosol algorithm, products, and validation, Journal of the Atmospheric Sciences, 62, 947–973, doi: 10.1175/jas3385.1, 2005.

Remer, L. A., Kleidman, R. G., Levy, R. C., Kaufman, Y. J., Tanré, D., Mattoo, S., Martins, J. V., Ichoku, C., Koren, I., Yu, H., and Holben, B. N.: Global aerosol climatology from the MODIS satellite sensors, Journal of Geophysical Research: Atmospheres, 113, doi: 10.1029/2007jd009661, 2008.

Remer, L. A., Mattoo, S., Levy, R. C., and Munchak, L. A.: MODIS 3 km aerosol product: Algorithm and global perspective, Atmospheric Measurement Techniques, 6, 1829–1844, doi: 10.5194/amt-6-1829-2013, 2013.

Schlosser, J. S., Stamnes, S., Burton, S. P., Cairns, B., Crosbie, E., Van Diedenhoven, B., Diskin, G., Dmitrovic, S., Ferrare, R., Hair, J. W., Hostetler, C. A., Hu, Y., Liu, X., Moore, R. H., Shingler, T., Shook, M. A., Thornhill, K. L., Winstead, E., Ziemba, L., and Sorooshian, A.: Polarimeter + lidar–derived aerosol particle number concentration, Frontiers in Remote Sensing, 3, 1–13, doi: 10.3389/frsen.2022.885332, 2022.

Stamnes, S., Hostetler, C., Ferrare, R., Burton, S., Liu, X., Hair, J., Hu, Y., Wasilewski, A., Martin, W., van Diedenhoven, B., Chowdhary, J., Cetinić, I., Berg, L. K., Stamnes, K., and Cairns, B.: Simultaneous polarimeter retrievals of microphysical aerosol and ocean color parameters from the "MAPP" algorithm with comparison to high-spectral-resolution lidar aerosol and ocean products, Applied Optics, 57, 2394, doi: 10.1364/ao.57.002394, 2018.

Su, C.-H., Ryu, D., Crow, W. T., and Western, A. W.: Beyond triple collocation: Applications to soil moisture monitoring, Journal of Geophysical Research: Atmospheres, 119, 6419–6439, doi: 10.1002/2013jd021043, 2014.

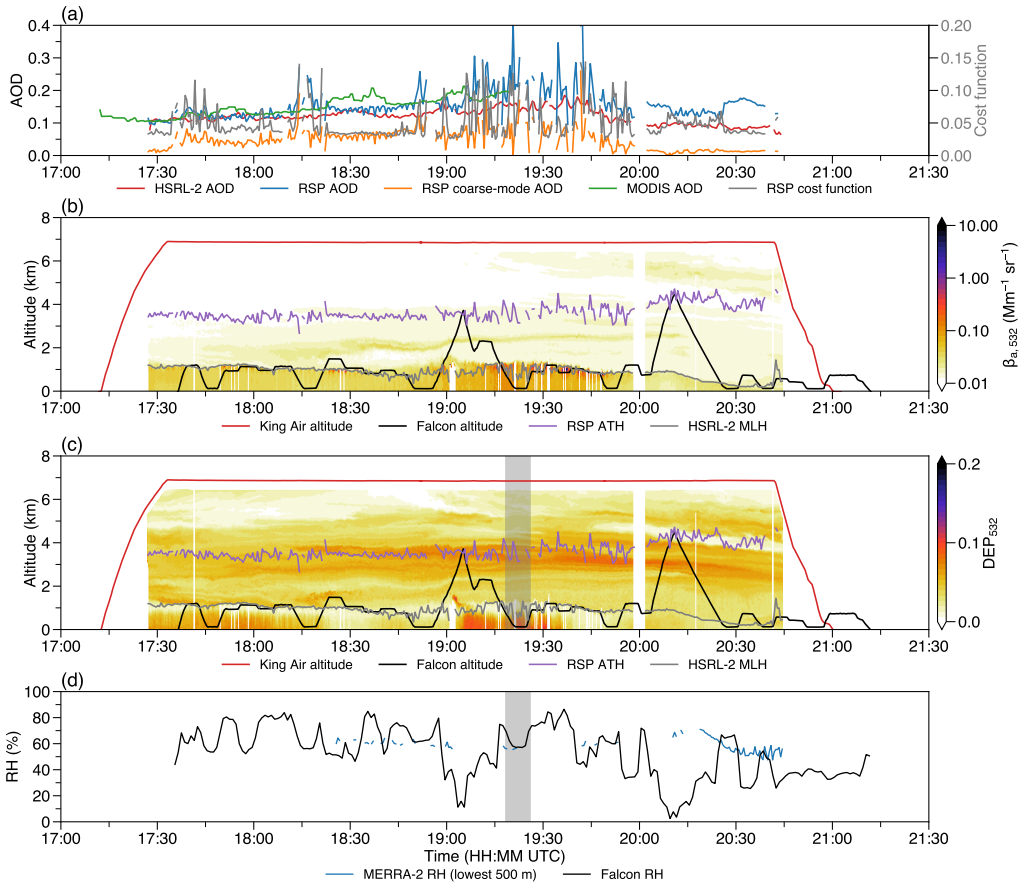


Figure R3: Time series of selected measurements during ACTIVATE RF143 on 22 March 2022. (a) HSRL-2 (red), RSP total (blue), RSP coarse-mode (orange), and MODIS (green) AOD. RSP normalized cost function (gray). (b–c) King Air (red) and Falcon (black) altitude. RSP aerosol top height (purple). HSRL-2 mixing layer height (gray). All altitudes are in km. (b) HSRL-2 532 nm aerosol scattering coefficient $\beta_{a,532}$ in $\mathrm{Mm}^{-1}$ $\mathrm{sr}^{-1}$ (color). (c) HSRL-2 532 nm depolarization ratio (color). (d) Lowest 500 m MERRA-2 (blue) and Falcon (black) relative humidity in %.

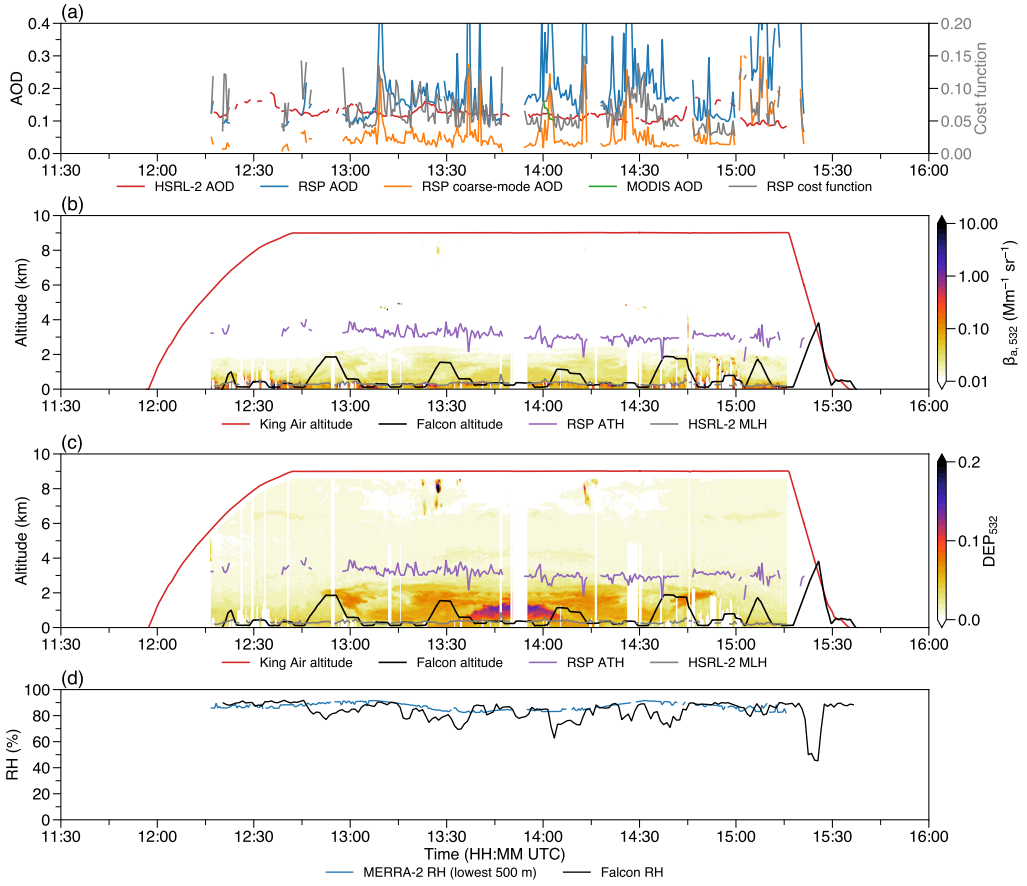


Figure R4: Time series of selected measurements during ACTIVATE RF170 on 10 June 2022. (a) HSRL-2 (red), RSP total (blue), RSP coarse-mode (orange), and MODIS (green) AOD. RSP normalized cost function (gray). (b–c) King Air (red) and Falcon (black) altitude. RSP aerosol top height (purple). HSRL-2 mixing layer height (gray). All altitudes are in km. (b) HSRL-2 532 nm aerosol scattering coefficient $\beta_{a,532}$ in $\mathrm{Mm}^{-1}$ $\mathrm{sr}^{-1}$ (color). (c) HSRL-2 532 nm depolarization ratio (color). (d) Lowest 500 m MERRA-2 (blue) and Falcon (black) relative humidity in %.

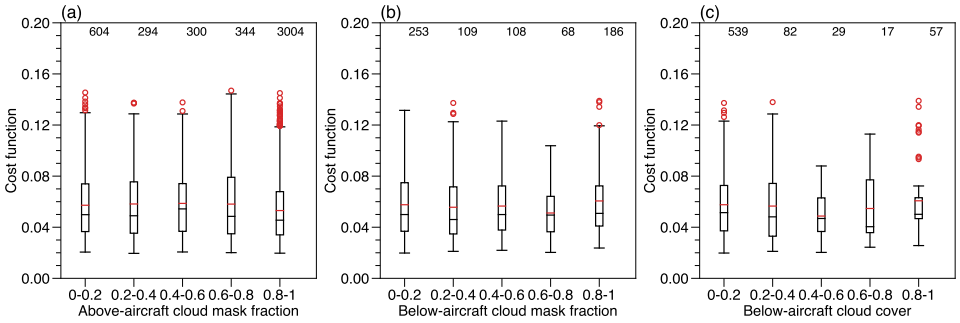


Figure R5: Cloud impact on the RSP cost function. (a) Grouped by above-aircraft cloud mask fraction. (b) Grouped by below-aircraft cloud mask fraction. (c) Grouped by below-aircraft cloud mask fraction. Numbers at the top indicate the number of each data group. The box and whiskers indicates the variability of each group (box, first and third quartiles; black line in box, median; red line in box, mean; whiskers, a distance of 1.5×IQR beyond the first and third quartiles; red circles, outliers).

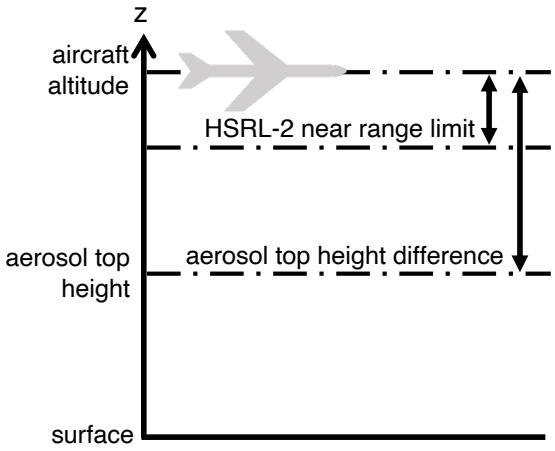


Figure R6: Schematic of altitudes associated with the high-flying aircraft King Air and its onboard instruments RSP and HSRL-2.