Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC2
https://doi.org/10.5194/amt-2021-229-RC2, 2021

Comment on amt-2021-229

Anonymous Referee #2

Referee comment on "Differential absorption lidar measurements of water vapor by the High Altitude Lidar Observatory (HALO): retrieval framework and first results" by Brian J. Carroll et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-229-RC2, 2021

September 7, 2021

Summary:

The submitted manuscript analyzes the performance of the NASA HALO lidar system in its water vapor and HSRL configuration during 5 Aeolus calval flights from 2019. The lidar system is described in some detail with the processing steps described in more detail than the lidar hardware. Data comparisons are presented where possible including range resolved comparisons to dropsondes, an in-situ diode laser hygrometer, AIRS and IASI, with passive column measurements also compared to AIRS and IASI as well. WV data is presented covering 3 orders of magnitude, which is impressive from any sensor, but especially from a DIAL sensor that must be mounted on an aircraft. Detail related to this new advanced system is certainly relevant within the scope of AMT.

Overall, the analysis methodology presented is reasonable, albeit with flawed ancillary data, and the quality of the work is high. The explanations given are mostly clear, though I have suggestions below for areas of my personal confusion. The caveats the authors give that are used to qualify and describe the extent of the statements given are very refreshing and in my personal opinion rather brave, i.e. “we understand the limits of our analysis are here and they are not what we would have hoped”.

However, as a reviewer, I see three overarching issues that can not be ignored, upon which my below comments will expand. First and foremost, the quantity and quality of the ancillary data used for validation does not give me confidence as a reviewer that this paper provides a true validation of the HALO instrument in the presented configuration. This is a major weakness that I believe disqualifies the presented data set from being used for validation without substantial additional information. Second, the retrieval

framework is presented with hardly any detail related to specific error sources and magnitudes; this makes it very difficult to evaluate this new system within the context of previously described lidar systems, or indeed other non-lidar sensors. Basically, I find it very difficult to judge if the results observed are "The right answer for the right reasons". Third, it seems the authors are trying break up a large amount of analysis and data over a few manuscripts (an instrument paper, a validation, and a description of a flight campaign, i.e. Bedka et al. 2021). While this is reasonable in principle, the execution leaves some areas of importance unexplored and leaves the interested reader to have to dig through multiple resources containing somewhat redundant information looking for details. This seems to be rather like threading a needle of detail. There is significant overlap of conclusions, comparisons, and information with Bedka et al. 2021 in particular. This seems to be necessitated by publishing a campaign description before the validation before the instrument paper. This significant overlap of information is detrimental to the impact of this paper by itself. Said more concisely: this paper is not, nor does it seem to be intended to be, a definitive resource. Furthermore, it promises more work to be presented in the future related precisely to the scope of the presented manuscript, which is not really admissible evidence in the context of assessing the overall merit of the presented manuscript.

I would suggest that major revisions are required before publication of this work. In truth, perhaps "major additions" are what is needed, in my opinion, as the data and methodologies presented seem perfectly reasonable. It seems that the authors have done about the most they can given the highly limiting constraints of the chosen ancillary data set. HALO data is analyzed with an incomplete error description applied to a flawed data set that causes too many caveats and externalities to be ignored. I have broken my comments into major and minor comments as well as suggestions which should be understood as very minor comments.

Major Comments:

- If I may bluntly summarize the data set used for validation, it consists of a set of untrustworthy dropsondes, a single vertical profile from the DC 8, in situ comparisons of the DIAL system's first range bin, and satellite data that is coarsely range resolved in the region of interest. Taking none of the below comments into account, I question very strongly whether that is sufficient for a validation. It seems the authors had hoped the ancillary data set would be more conclusive, but were left with an impossible task of salvaging poor data. I applaud the author's honestly with lines 631-635: "Although substantial conclusions were drawn, the opportunities for validation of HALO WV during the Aeolus cal/val campaign were not ideal due to the focus and brevity of the campaign. Further validation of HALO will be performed during future campaigns to quantify any potential systematic or other sources of error beyond the statistical uncertainty that is currently reported, though no such errors are evident to date." In my opinion, this is disqualifying of the draft in its current form without the future data, which is not yet available. I do not think the promise of further validation efforts is admissible in this case as the work is not done nor presented. These future efforts are precisely within the scope of this manuscript. It is effectively assumed here than no further issues are present. To the contrary, should it be assumed that there is an error in the system that is as yet undetected, having a single validation in the publication record that did not well cover the system's deployment potential is both confusing and

risks being contradicted.

- I am surprised at the lack of treatment of retrieval errors. I would argue that analysis of all known systematic and random errors is very much in scope of a paper claiming validation of an instrument. This is especially true where the retrieval methodology contains a significant amount of flexibility for human intervention and interpretation. This flexibility is emphasized in this manuscript. Only statistical errors are presented in my reading of this manuscript. The comment from Lines 201-202 about future error discussion given by Nehrir et al. is not sufficient, especially as error bars and error data are extensively presented as in Figures 3, 4, 6, 11, and 12. Relevant for data from, for example Figure 4, where single digit uncertainty is reported on water vapor quantities less than 0.03 or 0.04 g/kg; this is both potentially a demonstration of a disruptive technology and also potentially misrepresenting the underlying data. I would expect multiple single elements of a full error budget to be in excess of that amount at that level of precision/accuracy. I agree with the authors that much of the theoretical groundwork for general DIAL based errors has been laid and does not need to be rehashed. However, the errors presented resulting from this retrieval framework specific to HALO have missed an enormous amount of detail that is critical for anyone seeking to use this data in the future or anyone trying to evaluate HALO data in the context of non-lidar sensors. Some examples are:
  - How do you balance averaging, thereby reducing statistical error, with increasing the range of observation, thereby increasing your error in representativeness? Line 331-332 suggest there is no consideration to uncertainties from data smearing. See for example Hayman et al. 2020 on this topic.
  - Presumably HSRL data come with an error estimate. That error is not described here at all as far as I can tell. How do those errors propagate into your Rayleigh-Doppler corrections?
  - Angstrom exponents are used to convert HSRL data to the 935 nm wavelength. Is that an exact number or range and does the uncertainty in that exponent cause further uncertainties?
  - Are there errors in calculating the Noise Scale Factor or is that assumed to be known/retrieved exactly?
  - What are your assumed spectroscopic errors based on the Hitran 2016 database?
- It seems like this instrument really requires monotonically, or nearly monotonically, increasing water vapor content with observed range. In principle with a canonical water vapor distribution, this is reasonable. However with a non-standard distribution, this utility might be affected. If for example a moist layer sits on top of an extremely dry layer, the sensitive DIAL pair λ_1 and λ_2 would be extinguished before its data was useful. It is in my opinion therefore crucial to examine cases where water vapor profiles do not follow a canonical increasing trend with range from the instrument. This comment comes with a few questions:
  - Are extremely dry layers between moist layers expected to be observed well?
  - Does your processing method of assembling a profile based on decreasing moisture absorption with increasing range from the instrument fundamentally limit your ability to discover unusual atmospheric situations that do not follow this canonical form?
  - Taking as an example Figure 4, how sensitive to the extremely dry layer should your least sensitive DIAL wavelength pair be? Should this data simply be flagged out or trusted, or simply have massive error associated with it?
  - Is the dry layer observed by the dropsonde in Figure 6 panel d real? It appears that there is a lack of evidence to refute it other than a general misbehavior of the sondes (though that sonde comparison otherwise looks pretty good if you do a mental shift of the data in range).
- Section 2 feels out of place and repetitive. In my opinion, it is difficult to understand what is going on and why it is being presented before the HALO system and capabilities are described. Furthermore, I believe too many comparison-specific details are presented in section 2, causing repetition. I would suggest that the following might improve the readability of this manuscript.

- A more logical location in my opinion for the current section 2 is between the current sections 4 and 5.
- Data specific to the nature of comparisons (lines 116-120 for the Dropsondes, lines 132-141 for the DHL, lines 156-164 for AIRS, and lines 176-182 for IASI) should probably be moved to Section 5 to avoid repetition.

- Lines 235-238: Regarding the use of sequential wavelengths as DIAL wavelength pairs, I am not sure the reasoning presented completely follows. It is clear that you are using the most closely spaced pairs optimizing your assumptions vis-à-vis scattering cross sections and extinction. However, my guess is it places unnecessarily limits the analysis you can perform. Does an increase of wavelength separation of 0.2 nm quantitatively impact your error budget in any significant way? Furthermore, at 200 m/s flight speed and 1000 Hz laser rep rate, you are talking about 20 cm between shots. It shouldn't matter if you take online or offline data first or second, so the laser shots should not be further than 40 cm apart in any combination. If you do data analysis at the 2 Hz rate of DAQ reporting or if you do analysis with the multi-second averaging you report, the time between shots should be negligible. I do not understand how it matters at all what order you take data in. However, I do see a couple of significant reasons to use a single laser wavelength as the offline laser including minimizing Rayleigh-Doppler broadening and maximizing available signal. This seems to be suggested for example in Figure 1 of Nehrir et al. 2017, albeit with a different offline wavelength. I believe this analysis needs to be expanded here to at least address the following:
  - What effect, if any, does picking a different wavelength for the offline channel have on the results reported? Said differently, should you not get the same answer aside from statistical noise with any available choice of offline wavelength given the same online wavelength? Here I mean specifically is data from λ_1 and λ_2 substantially different than would result from using λ_1 and λ_3 or λ_1 and λ_4 and is λ_2 and λ_3 substantially different than λ_2 and λ_4?
  - Is the difference in measured water vapor from different DIAL wavelength pairs from question 5a indicative of or help quantify any error?
  - Does the choice of using sequential descending wavelength sensitivities vs. using a single offline wavelength have any negative consequences?
  - Does the size of the Rayleigh-Doppler correction increase substantially for λ_2 or λ_3 as offline channels vs. λ_4? λ_4, by sitting on the side of a line and not on top, would seem to me to be the least sensitive to Rayleigh-Doppler broadening.
- It is unclear to me how you handle data dropouts from the HSRL (for example roughly 3 UTC on Figure 4). If you need that data to perform spectroscopic corrections, how do you handle that lack of data? Does this lack of data reflect in increasing error quantities? I would think some mention of the way this is handled is necessary.
- What order of magnitude are the applied Doppler corrections? In particular with your extremely dry data set, I would think it should be very sensitive to Rayleigh-Doppler broadening as both wavelengths are parked on the peak of an absorption feature. In the manuscript the fact that this correction is applied is given as a matter of fact but the magnitude is likely important, i.e., as an example with made up magnitude, if it is a 2% correction it is no worse than statistical noise, 20% would be worth knowing, and 200% would possibly mean it should be flagged as bad and ignored. The work of Späth et al. 2020 might be a useful reference here.
- Section 4.3 seems incomplete in a number of respects. You claim in the abstract that this IPDA-type technique is presented. However IPDA is not novel per se as is referenced to the work of Barton-Grimley et al. 2021 among others. There is also no data presented from the test flights in this section other than comparisons to dropsondes, which have been noted to be unreliable, and to satellite sensors, which have been noted to have poor measurements near the ground. Finally, without data presented, there are a number of claims of empirically determined processing steps that are supported by vague statements that don't seem supportable.
  - Lines 398-400: Optimal signal strength is claimed with no data presented over urban or rural land or ice. Presumably they are different but what then is optimal. Does it

change flight to flight or hour to hour?
    - Lines 402-403: It is unclear what you mean by improvement here as you have not defined a reference measurement or error as a comparison method.
    - Lines 404-405: Empirical methods to reduce outliers seems rather heavy handed as no data is presented to prove an outlier is not valid. By what criteria are outliers identified?
    - Lines 474-476: Here a claim of validation of the surface result is achieved with sondes that have potentially several hundred meter altitude offsets to a return calculated with the last approximately 100 meters of HALO data?
- Lines 452-454: This section suggests to me as a reader that detailed analysis of the issues experiences with moisture from the dropsondes can be found in Bedka et al. 2021. However, that manuscript says simply: "During the Aeolus campaign, a new RH (relative humidity) sensor, deployed for the first time within the sonde, was found to have lag in response and did not have adequate sensitivity to vertical WV gradients. An initial view of this is provided by Fig. 14a above 5 km altitude, which will be further discussed in Sect. 4. Due to this response lag, sonde WV profiles will not be discussed in detail in this paper." Bedka et al.'s Fig. 14a seems equivalent to Fig. 6f in the current manuscript. I see no detailed analysis of the errors with the dropsondes nor one suggesting how to understand the limits of interpretation that can be afforded these drop sondes. This manuscript describes no possible correction for dropsonde data. Furthermore, simply playing devil's advocate to the stated dropsonde performance, if I assumed that the dropsondes actually have no error, a plausible explanation of the differences could also be timing lag in HALO's measurements. Finally, there also appears to be a range shift of the comparisons from HALO to the satellite sensors (Figure 12). How do you sort out this discrepancy in a rigorous manner?

Minor Comments:

- Line 38: I don't see "PBL" previously defined before use here.
- Line 68: "Deutsches Zentrumvfür" should be "Deutsches Zentrum für"
- Line 99: You have already defined "LaRC" in line 67.
- Line 104: As a minor follow on question to Major Question 6: Given that you need HSRL data to correct your WV absorption data for the different transmit and received spectra (and possibly your methane spectra as well?) is the WV/Methane combination truly possible? I assume this sentence is meant to say that HALO can be configured and fielded in this manner, but can you practically afford a lack of HSRL data for spectroscopic correction? Does a WV/Methane configuration alter the steps described in Figure 2?
- Line 154: I don't see "ITCZ" (I assume it is Intertropical Convergence Zone) defined.
- Lines 191-195: There are a lot of assumptions needed for DIAL, but I think that I might include the following 2:
    - You are assuming there are no interfering absorption species.
    - You are also assuming single scattering otherwise your range interval might be longer than expected.
- Lines 198-199: Why break out uncertainties in absorption cross section? How is this different than a systematic error?
- Lines 256-266: The discussion about resolution is mostly clear. However, I can find nowhere where the pulse width is described. I assume you are oversampling, which is fine, but the pulse will serve to smooth the features you see. It also seems (though I don't know for sure) like it should also add range correlations to your data that affect your NSF. It is arguably unclear to say you have 1.25 meter timing resolution and 15

meter detector bandwidth, when some portion of your range resolution and smoothing might originate from a larger laser pulse. Additionally, this impacts the 45 meter standoff distance you need from clouds and the ground (line 298). I would suggest adding the pulse width to Table 1.

- Figure 1: Listing the altitudes (0 and 12 km) used for your Hitran reconstruction is less helpful than the temperature and pressure in my opinion. Did I miss a reference to a standard atmosphere model where the temperature and pressure are known and linked to height?
- Table 1: What is the receiver field of view of the 1064 nm channel?
- Figure 2: The top gray box implies to me that you do everything highlighted in yellow for both online and offline for 3 sets of DIAL wavelength pairs. This would be 6 sets of calculations. I assume this is a misunderstanding and you really only calculate look up tables once. Is that true? If so, I suggest modifying this figure to suggest preprocessing in yellow is done for each wavelength then each wavelength is used.
- Figure 2: How do you apply Doppler corrections to a spliced profile? Your DIAL wavelength pairs should be differently sensitive to this effect. Should the corrections be applied to each wavelength pair before splicing?
- Lines 443-447: I find this explanation a bit confusing. Are you using a different Angstrom exponent for aerosols and molecules? Presumably you are using 4 for the molecular channel and something else for the aerosol. Is that scene dependent for aerosol type? What range of values are you using?
- Figure 9: Do both of these instruments measure WV in terms of g/m^3? If so, they both require conversion to include the mass of air. If so, would it be more reasonable to compare as g/m^3?
- Figure 9: As Figure 9 results from 3 separate measurements spliced together, it seems like it would be reasonable to break out the source of the measurement by wavelength pair. Because this comparison is so close to the plane, I would think this data is heavily dominated by the DIAL wavelength pair λ_1/λ_2. That said, that wavelength pair is not really what is doing the bulk of the data measurements in the 10^{-1} to 10^{1} range in your standard operating concept.
- Figure 9: Following on to the above comment, are you worried about the offset of 0.03 g/kg? That sounds tiny at first over all 3 orders of magnitude, but it is something like 30% bias on your driest measurements. Do you expect nearly stratospheric data measurements to increase by this quantity given the 400 or so meter difference in range?

Suggestions:

- Line 13: Suggest changing "…uses four wavelengths at 935…" to "…near 935…"
- Lines 69-70 and 101-102: There is repetition here where you say in effect that HALO is the successor or LASE. Suggest removing one.
- Line 146: Should "…generally fell with a 5%..." be "...within a 5%..."?
- Line 167: Is NWP ever used again? If not, I would suggest removing this definition.
- Line 200: I would suggest using "e.g." before your reference list here.
- Line 215 (Eq. 3): Do you need some sort of reference here to scale the optical depth to account for laser power output and receiver optical path and sensitivity issues? If you take data from Figure 5 and mix up high and low sensitivity channels, you would get different answers.
- Line 217-218: Suggest removing the optimal optical depth being 1.1. Your analysis right below more accurately accounts for systematic issues and more completely describes your target OD.

- Lines 242-243 and Figure 1b: Why are the count profiles shown from the low sensitivity setting? It seems like an unnecessary departure from your convention of only using high sensitivity data. I would suggest keeping all your figures constant and just using the High/High data here.
- Line 408: Is “IPDA” ever used again? If not, I would suggest removing this definition.

Suggested references:

- Hayman et al: "Optimization of linear signal processing in photon counting lidar using Poisson thinning," Opt. Lett. 45, 5213-5216 (2020)
- Späth et al: "Minimization of the Rayleigh-Doppler error of differential absorption lidar by frequency tuning: a simulation study," Opt. Express 28, 30324-30339 (2020)