We would like to thank the Referee for the thoughtful and critical comments that helped us to enhance the clarity of our manuscript. We addressed each comment individually and made revisions in response to their suggestions, as detailed below. Our replies to the Referee comments are highlighted in blue. Modifications to the manuscript text are indicated in *italic*.

Referee #4 (anonymous)

The authors present a strong, detailed, extensive study on a performance evaluation and optimization of their balloon-borne MIR-TDLAS-hygrometer ALBATROSS, which is definitively worth publishing.

Thank you for your positive feedback

Introduction: Due to the significant importance of the topic and the long lasting efforts of the airborne hygrometer community dating back way into the 1980s and 1990s I think the introduction should be revised to include essential representative and important work in airborne hygrometers, e.g., as FPH are mentioned as golden standard the review paper by Hall https://doi.org/10.5194/amt-9-4295-2016 should be relevant and mentioned. Further, the FISH hygrometer by C Schiller, reviewed by M Krämer in https://doi.org/10.5194/acp-15-8521-2015, is one of the most extensively used airborne hygrometers and one of the key reference instruments in AQUAVIT, and should be taken into account and mentioned.

Also there are plenty of airborne TDLAS hygrometers instruments e.g. by Sargent https://doi.org/10.1063/1.4815828, dating back in the 90s by Durry https://opg.optica.org/ao/abstract.cfm?uri=ao-38-36-7342 or Scott and Herman https://opg.optica.org/ao/abstract.cfm?uri=ao-38-21-4609. Particular relevant should be high flying ballon-borne open-path direct TDLAS hygrometers previously used for UT/LS sounding: *CHILD* by Gurlit et al https://opg.optica.org/ao/abstract.cfm?uri=ao-44-1-91 should be cited here. There is also work on new open-path hygrometers based on cMPC e.g. by Witt et al https://www.mdpi.com/2076-3417/11/11/5189 which I think should also be mentioned.

Particular stratospheric H2O vapor accessed via balloon platforms is strongly influenced by photochemical conversion of CH4 to H2O, asking for the need to simultaneously monitor traces of H2O and CH4, which is covered by some balloon sensors e.g. https://opg.optica.org/ao/abstract.cfm?uri=ao-44-1-91. A fact which should be considered and mentioned too.

We are well aware that there is a large amount of published work available about water vapor measurements in the upper atmosphere. In fact, the amount of research is so large that it would be valuable to write an up-to-date full review paper. However, this was not our aim, and therefore, we consciously limited the focus of this paper to "techniques demonstrated for lightweight balloon platforms" (page 2, line 15). Except for FPH by Hall et al. (2016), which is cited in the paper (page 2, line 12), none of the instruments mentioned above falls in this category. Although some of them were deployed on large atmospheric research balloons, their weight category is very different from our instrument, e.g: SDLA by Durry and Megie (1999), 20 kg; HWV by Sargent et al. (2013), 65 kg; ALIAS-II by Scott et al. (1999), 36 kg; CHILD by Gurlit et al. (2005), 20 kg; FISH by Schiller et al. (2015), 30 kg, while ALBATROSS (this work) is merely 3.5 kg.

Furthermore, we note that some of the suggested literature is already cited in our paper: besides Hall et al. (2016) (page 2, line 12), the FISH hygrometer is acknowledged through the review paper by Krämer et al. (2009) (page 2, line 5), and the most recent work by Durry and colleagues (Durry et al., 2008) about picoSDLA is also cited (page 2, line 17). Nevertheless, we underlined the large amount of available research in the field by adding the following paragraph to the introduction:

"Since the pioneering work by Brewer and Dobson (1951), a large amount of scientific research has been done on the water vapor distribution and variability in the upper atmosphere, based on a wide range of platforms and analytical techniques (e.g., Scott et al., 1999; Rosenlof et al., 2001; Gurlit et al., 2005; Sargent et al., 2013; Buchholz et al., 2014; Meyer et al., 2015)."

The dominant topic of the paper is hygrometer validation. The community has realized in the past a few fundamentally different type of validations: A) Field comparisons (Problem: lack of repeatability and lack of boundary parameter control) ; B) "Lab-like" parallel comparisons, e.g. AQUAVIT, (Problems: maintaining identical or at least comparable measurement and sampling conditions for all instruments and implementation of metrological references), and C) Rigid single instrument validations, preferentially to a SI-traceable water vapor source (Problem: large total effort, lack of H2O sources suitable for atmospherically relevant conditions, i.e. accurate definition of trace H2O levels - variable low gas pressures - low air temperatures). These differences should also be part of the introduction in order to avoid comparing "apples with oranges". In addition to AQUAVIT other validations of the above categories should cited / taken into account/ analyzed, e.g:

Buchholz + Smit > field comparison https://doi.org/10.1007/s00340-012-5143-1;
Filges + Gerbig > field comparison https://doi.org/10.5194/amt-11-5279-2018;
Buchholz + Ebert > metrological standard https://doi.org/10.5194/amt-11-459-2018 ;
Buchholz > metrological primary standard https://doi.org/10.1007/s00340-014-5775-4

For AQUAVIT (which was the largest parallel hygrometer comparison under variable p-T-H2O conditions), the authors should not highlight the (insufficient) performance of very young - not matured – instruments ("exceeding 100%") and give their underperformance the same weight like the very mature CORE hygrometers, which have been used and improved over decades. The performance of the non-calibrated, absolute, open-path TDLAS "APicT" in AQUAVIT certainly also relates well to the paper here, and could be mentioned in the paper. Some of the main findings of AQUAVIT were indeed the still quite large total discrepancies (-+ 10% relative) between the very mature "core" hygrometers. Also it took a complicated decision making process to define a "comparison reference" i.e. a suitable metrological H2O source or metrologically validated reference instrument which is compatible with the special (low temperature) boundary conditions of the AIDA chamber and their huge size.

We fully agree that the intercomparison of the measurement techniques is a very challenging issue. Even more, we have gained extensive knowledge and experience in this topic as we recently participated in the last AquaVIT-4 intercomparison campaign that took place at the AIDA chamber in April 2022. We are preparing a separate manuscript dealing with all the aspects of such intercomparisons, and thus we focus our current paper on the assessment of our spectrometer.

We believe that a sufficient level of detail is given to the reader regarding the different methodologies employed by the intercomparison papers, which are cited. In particular, concerning the discussion of the AquaVIT-1 intercomparison results (Fahey et al., 2014), we note that both aspects underlined by the referee ("core" instruments within ±10 %, other instruments exceeding ±100 %) are in fact already mentioned in the manuscript (see page 2, lines 9-11).

Experimental:

The authors target a SI-tracebale validation of ALBATROSS, where ALBATROSS is claimed to be a gas sampling-free, and calibration-free open-path Mid IR spectrometer.

The open-path approach promises to avoid H2O adsorption problems. However, open path also causes a very complicated tradeoff in system design, due to the complete lack of "sample control" during field-use, so that gas pressure, gas temperature, residence time, sample homogeneity must be measured, evaluated or assumed. Additionally p and T are often not measured within the optical sample volume but outside of it, leading to further heterogeneity errors in measured p and T wrt to the gas sample.

In my understanding a validation of the open-path ALBATROSS was not tackled or described in the paper. Instead a closed-path version of ALBATROSS was used, which I think, is a big change with respect to the initial claim. Of course, the validation of the closed-path version is highly important and demanding, but closed-path studies are certainly not fully sufficient to validate the open-path version and certainly not under UT/LS field conditions. The title of the paper is therefore misleading and should be considered to be changed (e.g. > *"validation of a closed-path Albatross"*).

We share many thoughts of the referee, and we are well aware of the challenge given by the fact that the "perfect" validation does not exist (see options A – C above, described by the referee). Therefore, we have adapted the title to clearly state that the validation was not done ON a balloon-borne spectrometer but FOR a balloon-borne spectrometer, i.e.:

"SI-traceable validation of a laser spectrometer for balloon-borne measurements of water vapor in the upper atmosphere"

Nevertheless, we would like to note that it is common practice to determine spectroscopic parameters in a close-path system and apply them to open-path measurements. In fact, this is the scientific ground for all remote (including satellite) measurements, and there is no need to fundamentally question this concept.

Our laboratory experiments in a closed-path, but otherwise identical, configuration are likely the best, or even only, approach to characterize the spectroscopic performance. We see no reason to doubt its validity for the open-path configuration, and the paper makes it clear that the experiments are done in closed-path configuration (proof given by the comments of the referee).

The other technical aspects mentioned by the Referee, i.e. gas pressure, gas temperature, residence time, and sample homogeneity apply to all field instruments, regardless of the measurement technique. In fact, it is hard to think of a setup that is more representative and resilient than ALBATROSS under slow flight conditions and in the UTLS. Nevertheless, this has been (Graf et al. 2021) and will be (AquaVIT-4, manuscript in preparation; further in-flight comparisons) discussed in more detail. This is discussed in the conclusions, making it clear that this is not the last, and likely not the final assessment. Adding a very extensive discussion here would add more noise than weight to this paper, which we consciously kept in a focused manner.

In order to deduce the performance of the open-path version careful consideration and evaluation of p, T sensor location and calibration is needed, which is however not given in the paper. Witt et al in https://www.mdpi.com/1424-8220/23/9/4345 recently evaluated a comparable open-path C-MPC under dynamic situations and found considerable systematic deviations caused by spatial gas temperature inhomogeneity and by the un-even statistical spatial weighting caused by the special C-MPC beam pattern. This findings are probably of high relevance for open-path in-field-use as well as for high-accuracy validations in closed-path cMPCs as presented in this manuscript by the authors.

We agree that p and T measurement in an open path configuration during flight is a topic under debate and up to know there is no 'golden'-solution for solving this issue. This aspect is an inherent problem to all measurement techniques, independently of the platform used. Thus, efforts as described by Witt et al. are very useful studies and deliver important insights in optimizing p and T measurements (especially for sampling aboard airplanes), applying the necessary corrections or estimating the uncertainties.

In our paper, we explicitly communicate that the assessments are done under well-controlled laboratory conditions where these parameters are kept constant and homogeneous.

For the sake of completeness, we note that the paper by Witt et al. (2021) is dealing with the evaluation of gas temperature and concentration inhomogeneities in dynamic tube flows. While doubtlessly an important investigation, the targeted application there is rather different from the conditions of a balloon flight, not to mention the low-flow and constant temperature conditions used in our laboratory assessment. Even in the atmosphere during flight, the temperature and the water vapor concentration are locally much more homogeneous than the conditions investigated in the Witt *et al* paper, i.e. "strongly heterogeneous T fields generic for industrial process application, e.g., in pipe flows" and "temperature range from 293 to 473 K at 1 atm of pressure."

Interestingly, Witt *et al*. found, despite the harsh conditions, that "for the case of a strong thermal boundary layer with a delta-T of 180 K (…) would lead (…) to a relative deviation of −5.3 % between the “true” and the calculated concentration." Considering our well-stabilized setup, with $\Delta T < 0.1$ K, the impact of such effects on the accuracy is negligible.

The authors aim on calibration-free first-principles evaluation of the hygrometer signals, which is indeed a very powerful capability for field use (see e.g the airborne HAI Hygrometer). In a cal-free mode, however, the TDLAS-instrument integrates any H2O spectral absorption over the full light path i.e. anywhere between the laser chip and the detector chip. Any “parasitic” = unwanted water along the absorption path “outside of the absorption cell” will lead to systematic, potentially drifting offsets and needs to be carefully evaluated and removed. Particularly complicated are situations where the gas pressure also is heterogenous along the path (e.g in sealed laser or detector housing). This problem is carefully described in Buchholz 2014 https://iopscience.iop.org/article/ 10.1088/0957-0233/25/7/075501. How this is solved / or avoided in the present study must also be described, in particular if ALBATROSS is claimed to be cal-free. It is unlikely that this problem is completely absent in the ALBATROSS design. Parasitic water vapour offsets can of course be removed to first order by calibration, but not in a cal-free TDLAS hygrometer.

This is a very valuable and important remark. The "parasitic" absorptions along the free space path are indeed challenging. Our solution is to keep the distance between laser-MPC and MPC-detector as short as physically possible. In our case, this amounts to 2.7 cm. This short path is then enclosed by a flexible tubing that is purged with dry N_2 such that the absorption contribution from the residual water drops below the detection limit of the instrument. The technical details of this solution will be discussed in our next paper, describing the flight measurements, where the influence of this artifact can have a substantial impact, mainly due to the similar pressure conditions within and outside the SC-MPC.

We added the following clarifying text to Section 2.1 of the revised manuscript:

"*Furthermore, the free-space path between the key optical elements, i.e. laser-MPC-detector (kept by design as short as physically possible, in our case, 2.7 cm) was enclosed by a flexible PEEK-tubing that is purged with dry N_2 and maintained slightly above atmospheric pressure, to avoid any "parasitic" H_2O absorption from these external path sections.*"

With respect to this topic it should also be analyzed where the zero air blank values (e.g. 1.46 ppm in Fig 2 compared to 0.59 ppm in fig 9) comes from, how stable they are and e.g how much of this is caused by parasitic water in the spectrometer itself.

The non-zero H_2O content (~1.5 ppm) observed in the laboratory while measuring dry synthetic air originates exclusively from the memory effects of the setup, mainly tubing surfaces and the large surface-to-volume ratio of the closed SC-MPC (see also the replies to Referee #1).

As mentioned above, the free-space OPL of 2.7 cm was efficiently purged and maintained slightly above atmospheric pressure to avoid "parasitic" H_2O absorption from these external path sections. The large pressure difference between inside and outside SC-MPC would also allow for an easy separation of the two contributions during the spectral fitting. Because of this, "parasitic" water in the spectrometer itself can be definitely excluded. Nevertheless, we acknowledge that the above description of the purged design was missing, and the Referee's critique was therefore well motivated.

In the spectral evaluation a "spectrum normalization" via a division through an "empty cell spectrum" is used. As the "empty cell" still had non-stable "zero" water levels of 1,5 ppm, the spectrum normalization actually also introduces an effective offset (and to a certain extend a parasitic water vapor) correction. This approach and the offset correction cannot be used in the open-path configuration. The alternative approach "polynomial baseline reconstruction" does not provide offset correction so that the parasitic contributions should be effective. The authors should add data on this if possible or discuss this effects and their quantitative influence on the absolute accuracy of both ALBATROSS versions.

The empty-cell spectrum normalization procedure applied to the laboratory validation data is correctly described by the referee. However, this offset correction is not required in the open-path configuration (i.e., under flight conditions), where there is no sampling line, the gas flow is much larger, and the surfaces of the SC-MPC are drastically reduced, as the lids are removed and the gas-surface interaction is limited to the narrow inner circumference of the cell. This aspect was already discussed in a reply to Referee #1, and a clarifying sentence was added to Section 2.4 of the revised manuscript.

Used preparative water vapor references:

Primary Permeation source: It should be better clarified which components in the entire setup (fig 1) define the "SI-traceable permeation source". Is this the permeator only, or the permeator and MFC1 and 2, or even more components? This needs to be clarified as only this subsystem provides the property of being SI-traceable. As I see it now, the "permeation source subsystem" is embedded into a larger gas mixing system containing further MFCs plus an additional pressure controller(s?), pressure sensors, gas and cell body temperature measurement. These all should be shown on fig 1. and better explained in the text. For the entire validation to be "SI traceable" all relevant measurement data need to be SI traceable. Traceable calibration data and accuracy and expanded uncertainties should be provided for all measurement parameters (p, T, flow etc) required for the TDLAS evaluation procedure, which is not the case. Figure 1 lacks also an excess flow outlet before MFC3.

In Figure 1, the SI-traceable part of the magnetic suspension balance (MSB) is indicated by the grey shaded area. This entire unit represents the core of the metrological-grade solution used at METAS and it was validated in several intercomparison studies.

The overall uncertainty of the spectroscopic retrieval was already given in Section 3.1 of our manuscript.

The absolute accuracy and stability of the reference H2O concentration and the gas handling system will influence the TDLAS validation and e.g. depends on accuracy and stability of the H2O blind value, which needs to be determined and should be given in the text.

As we used a solely spectral retrieval (i.e. purely deduced from molecular and environmental parameters), none of the spectroscopically determined values were calibrated or linked to the SI-traceable values. These latter data were used for comparison purpose only. Therefore, each value represents basically a blind value. The absolute accuracy and stability of the generated reference H_2O concentration is already discussed in Sect. 2.2 in our manuscript.

Due to the lack of traceability information for the used validation setup I can’t see that the “entire validation setup” is SI traceable. Due to this deficit, the paper claims and the title should better changed to, e.g *“Validation of a closed-path balloon-borne spectrometer with a permeation-based SI traceable H2O-source”.*

As explained above, the source is traceable, and so are all other elements upstream of the spectrometer. The corresponding uncertainties are also given in the text. The referee is well aware (see discussion of types of validation) that there is no such thing as an artificial, traceable stratospheric chamber in which an open-path instrument can be flown on a balloon. Within this fundamental constraint, the high-level and traceable experiments described in the manuscript are adequate. The main limitations are discussed in the text (with additional notes following the valuable comments of the referee), and further validation steps (e.g. flights and chamber measurements) are mentioned in the conclusions. Overall, we believe, that this gives sufficient and balanced information to the reader.

Secondary water standard: The bottled H2O mixture generated is analyzed (if I got it right) only by the closed-path ALBATROSS. Hence the assigned bottle value of 181 +- 0.06 ppm “collects” all uncertainties (and all systematic errors) from the closed-path Albatross validation using the primary permeation source. The +-0.06 ppm (=3,3 E-4 relative!) can thus only be “precision”. Here the accuracy and the uncertainty of the bottle assignment should be added and discussed, which then needs to be taken into account for the “expected H2O amount fraction” in fig 9, and for the evaluation of the uncertainty of the linearity relation. Looking at the fitting function in fig 9 the differential linearity and the 1,008 slope seem certainly excellent. However, the large offset of 590 ppb (which is very close to 2% ! at 30 ppm and would extrapolate to 15% at tropopause concentrations of 4 ppm) definitively needs further explanations by the authors. For me this indicates an accuracy problem of closed-path ALBATROSS or/and this secondary standard setup. Also for both values (m and b) uncertainties should be provided.

Other points to be considered are the likely dependance of the H2O amount fraction form the bottle pressure, as well as sampling influences by the sampling line including the pressure reducer, more information on the sampling system and the adsorption minimization would be helpful.

The systematic overestimation of the measured H_2O amount fractions in the linearity assessment was discussed in the context of the replies to Referee #2. This artifact was unfortunately due to a mistake in calculating the expected H_2O amount fraction levels. Applying the correct value, the systematic overestimation is removed, and all measurements lie within ±0.6 % of their expected values. Accordingly, the linear fit results are also improved. Figure 9, Table 4 and the text of Section 4.3 were updated to the revised values and results. We thank the referee for this comment and apologize for the mistake.

Taking all this into account, bottle-based secondary standards might be useable as a high concentration H_2O source, but to be useful to a broader community they certainly need more evaluation work.

The secondary water standard does not fulfill SI-traceability, and it is subject to well-known stability issues. Its sole purpose is to assess whether ALBATROSS is capable to measure significantly higher water vapor amounts than can be generated by the permeation method (see also the replies to Referee #2). We have added a sentence to Section 2.3 of the revised manuscript to further clarify this aspect:

"It should be noted that this custom-made secondary reference gas does not fulfill SI-traceability, and it is subject to well-known long-term stability issues. Its sole purpose is to assess whether ALBATROSS is capable to measure significantly higher water vapor amounts than can be generated by the permeation method."

Spectroscopic retrieval:

The spectroscopic retrieval section is quite extensive and specialized for publication in AMT.

In my view the full fitting model is not sufficiently described: It is not clear how many and which water lines (or other interfering species) are fitted, or e.g how large the pressure dependent influence from neighboring lines is and how and if it is compensated. Which H2O isotopic composition is assumed? An H2O stick spectrum showing the fitted as well as the ignored lines would be helpful here. Also the description of the physical model behind the spectral evaluation and in particular a complete set of input parameters and their total uncertainties is not given. A total uncertainty evaluation of a cal-free system seems therefore not possible.

The employed quadratic speed-dependent Voigt profile (qSDVP) is well established and recommended by IUPAC. The spectral line (1662.809 cm^{-1}) is given in the manuscript. At the relevant pressure (< 250 mbar) and concentration (<35 ppm) there is no spectral interference from neighboring water lines and, thus, the fit considers only the isolated line. The closest and most relevant line would be from $H_2{}^{18}O$ (181) at 1662.3353 cm^{-1} with an absorption amplitude of 3×10^{-4} that has no impact on our retrieval.

The line profile parameters used in our study are listed in Table 2 and illustrated in Figure 5.

The selected absorption line belongs to the main water isotope (161). The isotopic composition of our water standard is not known, but it can be assumed that of typical tap water, i.e. ~-60 ‰ δ^2H and -10 ‰ δ^{18}O. Considering the low abundance of the heavy isotopologues (0.27 %), the total estimated uncertainty would be about ± 0.02 %. This additional uncertainty has been introduced in Section 2.2 of the revised manuscript, and the reference levels generated by the dynamic-gravimetric method were scaled by a factor 99.73 %.

"*As the isotopic composition of our water standard is not known, we estimate an additional uncertainty of about ± 0.02 % on the total H_2O amount fraction, by assuming that the liquid water standard has a signature of typical tap water, i.e., −60 ‰ δ^2H and −10 ‰ δ^{18}O. Considering the low abundance of the heavy isotopologues and a natural distribution, the total amount fraction contains about 99.73 % of the light water isotopologue ($H_2{}^{16}O$). Since ALBATROSS measures only this water isotopologue species and not the total H_2O amount fraction, the reference values generated by the dynamic-gravimetric method were scaled by this factor for the accuracy assessment.*"

Consequently, the results of the accuracy assessment (i.e., Figure 8, Table 3 and the text of Section 4.2) were updated in the revised version manuscript, after rescaling the reference levels generated by the dynamic-gravimetric method by a factor 99.73 %. The main conclusions of the accuracy assessment are unaffected by this change. We thank the referee for this valuable remark.

If the cal-free evaluation is the goal, then all spectral parameters plus all auxiliary measurements needed (= p, T, L ....) must be stated with their (expanded) accuracies/uncertainties. Here I would expect an uncertainty table for all input parameters, as well as more information on p-, T-sensors their location and traceable calibration, which is not given.

See our reply above. The uncertainties of the *p* and T measurements are already given in the manuscript (see Section 2.3).

Also the uncertainty influence of the fitting process itself as well as e.g. the uncertainties of the linearization of the spectral axis/laser tuning should be discussed.

We agree. This aspect and especially the impact of the uncertainty of the laser tuning is now discussed in more detail in Section 3.4 of the revised manuscript (see also our replies to Referees #2).

Particularly in gas spectrometers the real gas temperature in a weakly thermally conducting low pressure gas can cause problems. Concerning the HMP110 used here: this T sensor is specified by the manufacturer with an accuracy of 0.4K (not 0,2K) for the extended T range (needed for UTLS use). Also comments by the authors are recommended if/how they deal with the systematic T-offsets /uncertainties (and the effect in the TDLAS evaluation) caused by invasive air temperature measurements, i.e. evaluation of PT100's self-heating and thermal gas to sensor transfer problems (particularly at low pressure). EURAMET project 1459 "Air Temperature Metrology – ATM" could be considered here.

Again, the overall uncertainty of the spectroscopic retrieval has been specified in our manuscript (see reply above). The largest source of error is the permeation source (1.5 %), while the absolute uncertainties on the measured p (0.12 %) and T (0.06 %) play a secondary role. Concerning the HMP110, we consider only the conditions that are representative for our assessments, i.e. room temperature, where the value of 0.2 K applies. For the flights, we use other calibrated temperature sensors, which are much smaller than the HMP110. For the sake of clarity: in-flight uncertainty may be different, e.g. because of such small adaptations and because of other effects, such as contamination by the balloon wake (Graf et al., 2021). This is quite fundamental to any validation approach as the reviewer rightly pointed out at the very beginning.

The magnitude of the "temperature problem" also strongly depends on the spectral line selection: H2O line identification and lower state energy of the fitted lines therefore should be given in the paper.

The selected line has a lower state energy of 79.5 cm^{-1}, which makes it largely insensitive to the "temperature problem". For example, a temperature change of 1 K would give < 0.13 % change in the observed absorption amplitude. Since we take this effect into account in our calculations, and because the ambient temperature is maintained constant, the impact of this term can safely be considered negligible. The following sentence was added to Section 2.1 of the revised manuscript:

"This absorption line corresponds to the transition 221←212 with a lower state energy of 79.5 cm^{-1}, which makes it largely insensitive to temperature. For example, a temperature change of 1 K would give < 0.13 % change in the observed absorption amplitude. Since we take this effect into account in our calculations, and because the ambient temperature is maintained constant, the impact of this term can safely be considered negligible."

The influence of individual spectral data uncertainties can be quite large and often strongly limits the achievable total uncertainty of cal-free spectrometer realization. As I understand the authors paper, they are taking fixed line strength S(296K) and broadening G_0 for the spectral evaluation from HITRAN, further they need T dependance of broadening and S (which also comes from HITRAN with their uncertainties) and the line pressure shift Do (again HITRAN + uncertainty) and then finally the “new” qSDVP braodening parameter G_1 (which also needs to get an uncertainty from the parametrization). With typical HITRAN uncertainties for S 1-10 % (depending on the line selection) Voigt broadening (another U= 2-5 %) , broadening T coefficient (5 -20% and more), plus p, L, T, fit process, tuning and spectral axis uncertainties it certainly takes further explanations how the closed-path ALBATROSS reaches 1.5% total uncertainty in cal-free mode. The best short-term accuracy can certainly be achieved by a hygrometer calibration to a very good reference and not via a spectroscopic cal-free approach, due to the large amount of spectral input parameters with fairly large uncertainties.

The estimated uncertainty of the line strength for the selected transition is specified in the HITRAN database with code 7, corresponding to a relative uncertainty ≥ 1 % and < 2 %. This represents the largest uncertainty of the spectroscopically retrieved values. This information was now added to Section 2.1 of the revised manuscript.

We demonstrate an agreement between the SI-traceable sample and the spectroscopically retrieved (measured) values within the uncertainty (1.5 %) of the sample when using the qSDVP profile with the parameters given in Table 2. As discussed below, please note that the qSDVP broadening coefficients are determined by letting the MSF minimize the fit residuals for all the pressure levels at once, i.e. this is the only criteria to determine the line-profile parameter values which are thus independent of the absolute value of the water vapor source.

Nevertheless, for the sake of clarity, we decided to not use the term "calibration-free" in our manuscript, but simply state the fact that the measurements show that ALBATROSS achieves an accuracy better than ±1.5 % with respect to the SI-traceable reference at all investigated pressures and H_2O amount fractions. In this context, it may be valuable to recall that accuracy is also understood as the closeness of agreement between measured quantity values that are being attributed to the measurand (see e.g. VIM 3).

The transfer of the closed-path validation presented in the paper to the open-path balloon-version, depends particularly strong on the accuracy of the spectral data i.e. H2O line selection or the temperature coefficients of the broadening. What measurements this requires and how this should be described is shown e.g in Pogany et al for traceable H2O strength https://doi.org/10.1016 /j.jqsrt.2015.06.023 and in Nwaboh https://doi.org/10.3390/app11125341 for traceable determination of H2O broadening incl T dependence for TDLAS.

The only parameters not covered by our study are the uncertainties of the temperature dependency of the line strength and the temperature coefficient of the broadening parameter. This limitation is clearly stated in our paper. As mentioned above, we have now added supporting information about this in Section 2.1 of the revised manuscript.

Although, this does not apply to our measurements performed at room temperature, temperature dependency of the line strength and the temperature coefficient of the broadening parameter might have an impact on the atmospheric water vapor measurements in the UTLS. However, their effect is still expected to be of minor compared to the uncertainty (~1 %). of the line strength.

Line shape study:

The authors compare the applicability of two line shape models: Voigt (VP) and quadratic speed dependent Voigt (qSDVP) and then optimize the qSDVP approach. Their VP evaluation is not very extensive and based on a single fixed set of parameters taken from HITRAN: The to be expected pressure dependent line shape deficits are not taken care of. It should be noted that Buchholz AMT 2018 had proposed to correct this parametrizable, perfectly long-term stable, systematic deviation caused by the Voigt profile deficits e.g via a look-up table approach. This correction approach allows faster fitting and avoids too many fit parameters, which have caused in his spectrometers noise-like fitting instabilities.

This paper is not primarily a study of line-shape models, and more extensive studies on VP have been published. Nevertheless, we give a very thorough insight that will be appreciated by the community (see e.g. the three previous reviewers), e.g. a) we clearly show the VPs difficulty to properly describe the pressure effects (see Fig. 8), b) we show that with "fine-tuning" of G_0 and lines strength an excellent agreement can be obtained, but the pressure-bias would still dominate the uncertainty (see Fig.4 and consider the case $G_0 = 0.0954$), and c) we indicate the fit residuals and their dependency on pressure and H_2O concentration on Fig. 6a,b.

The look-up table approach, proposed by Buchholz et al., although computationally efficient, relies on normalization factors, which have no physical meaning and may strongly depend on instrumental parameters. We avoid such an approach and rely on physically motivated and well-studied quantities instead.

For the qSDVP evaluation the authors use a restricted qSDVP parameter set, allowing only one additional “broadening parameter”, and hence should be better called “simplified qSDVP” to be precise. As I understand the paper, the authors use the Albatross-permeation standard-comparison via an iterative parametrization to “determine” the “optimal” qSDVP broadening parameters for their setup (while checking S). The parametrization of the width parameter of the simplified qSVDP follows two goals A) to match the spectrometer response function (Albatross H2O concentration) and the reference concentration (permeation source plus mixing system) and B) to minimize the fit residual i.e. to remove systematic deviations in the line shape fitting (optimization of QF). However, there is no uncertainty provided for the outcome of this process, leaving it open, how accurate this qSDVP parameter really is. Literature comparisons of this spectral parameter are also not given, making it somehow an instrumental parameter.

As the goal of the parametrization was to improve the “correlation” between the permeator and the TDLAS system, it is “no surprise” that the parametrized qSDVP evaluation yields pretty much the “input data set”, while the QF optimization improves the apparent system precision by minimizing the fit residual. However, the problem I see is, that the reference permeator information was used twice: First for the determination of the spectral information and then the “trained” qSDVP-TDLAS was compared to its previous reference in the learning situation. And the result of the second step is not really surprising, it’s a pretty good match.

Obviously, there is a fundamental misunderstanding of our approach. We do not use the permeator values in our spectral retrieval at any stage. The line profile parameters are optimized by the fitting routine using the Levenberg-Marquardt least squares algorithm. The only criteria for selection of the line profile parameter values is the fit residual from the MSF, i.e. when the highest QF is obtained. The "real" H_2O concentration is not used anywhere in the fitting process as a constrain or "training" set. The model is based on first principles using exclusively molecular parameters and environmental/physical (p, T, OPL) values to describe the observed absorption signal.

In our opinion, it is staggering to find such an outstanding match between the purely spectroscopic values and the "real" H_2O concentration values. This indicates the high quality of the molecular parameters listed in the HITRAN database and their well confined uncertainty, at least for the H_2O transition selected for our study. Furthermore, it also demonstrates that beyond-Voigt line-profile models have the capability to accurately describe the observed shape of the absorption line under varying pressure conditions. We added the following paragraph to our manuscript to strengthen this point:

"*It is important to realize that the line profile parameters are solely determined by the QF. The MSF algorithm is not aware of the target (or "true") value of the H_2O concentration, it simply tries to minimize the sum of the squares of the residuals, i.e. the difference between the observed and the fitted value provided by the model. Here, the model is based on first principles using the molecular parameters and the physical quantities (p, T, OPL). The generated SI-traceable H_2O concentrations are used for comparison purposes only. There is no calibration involved.*"

Furthermore, we extended our discussion to include the important finding:

"*This excellent agreement reflects the high quality of the molecular parameters listed in the HITRAN database and their well confined uncertainties, at least for the H_2O transition selected for our study. Furthermore, it also demonstrates that beyond-Voigt line-profile models have the capability to accurately describe the observed shape of the absorption line under varying pressure conditions opening the path to a highly accurate quantification of the observed data.*"

For me this approach seems essentially like a more elegant way to calibrate the spectrometer response function by using the reference H2O concentration. The uncertainties of this process are not sufficiently discussed. Also the high correlation caused by that approach is not studied or taken care of. An elaboration of this problem would require a further comparison with a third independent preparative or analytical H2O system, which has not been shown in the paper.

Therefore I think that the authors cannot claim a demonstration of a calibration-free hygrometer. Not in closed-path configuration and even less so with an open-path cell.

See our reply above. Our spectral evaluation is not correctly understood and, therefore, the conclusion is not appropriate. We do not apply any calibration, but simply compare the spectroscopically derived values with the SI-traceable ones. The main uncertainty is given by the reference material. In principle, the spectrometer would be able to achieve accuracies that are of similar level than its precision.

As we explained above, the close cell configuration is actually more challenging as it involves surface and memory effects that are not present in the open cell case. The physics of light-matter interaction is independent of the MPC configuration. We are confident that our assessment is similarly valid for both cases, but we acknowledge that during flight, additional factors may have an impact on the accuracy. This is, however, related to the field conditions and not to the assessment and validation steps as presented in our study.

Essentially they have developed a novel (?) calibration procedure instead. In contrast to a classical calibration they are not aiming on a direct correction of the instrument function but realized a “physics-informed approach” to remove line shape deficits. This is also valuable(!) but it remains a calibration process. Also I think further work is needed to investigate the accuracy and (longer term) stability of this parametrization / parametric calibration, and how often it needs to be repeated. But I think that the claims derived from the data and the results should be carefully and conservatively revised. What I see is a “Use of a SI-traceable permeation source for the characterization/calibration of a closed-path Mid-IR QCL TDLAS hygrometer suitable for balloon-borne, extractive UTLS-hygrometry”

Once again, we use well established and recommended (see IUPAC) line shape models to best capture the observed line profile. The line parameters are obtained by applying standard mathematical methods (least squares fitting) to optimize the residuals between the observed and calculated line profile. At this point, we would like to cite the remark of Referee #3: "This work also illustrates that the Voigt line profile is largely inadequate for an accurate description of collisionally broadened molecular absorption lines. This is actually a well-known fact since the time that high-resolution spectra are being obtained by using narrow laser spectral sources (…). Nonetheless, it is very instructive to see the impact of the choice of spectral line shape on the retrieval of molecular mixing ratios, and to see that linear and accurate results may be obtained by using an advanced line shape model - with parameters determined by a multiline fit at several pressures, as shown in this work."

It is unfortunate, that apparently we were not able to clearly communicate our spectral retrieval. We take the opportunity to improve this aspect in our revised manuscript correspondingly (see also our replies to the other referees). As the values retrieved by ALBATROSS are exclusively relying on first principles, molecular parameters and environmental/physical (p, T, OPL) values, their comparison to the SI-traceable values is the best available way to show their quality. Furthermore, we are preparing another manuscript where we show intercomparison results from balloon flights using the CFH as reference. This will demonstrate the capabilities of ALBATROSS also in the open-path configuration under flight conditions.

Nevertheless, we made a slight change to the title of our manuscript to accentuate even more that the assessment was not done ON a balloon-borne spectrometer but FOR a balloon-borne spectrometer:

"SI-traceable validation of a laser spectrometer for balloon-borne water vapor measurements in the upper atmosphere"