We thank all the reviewers for carefully reading our manuscript and for the detailed feedback aimed at helping us to further improve the manuscript. Below we address the raised concerns in a point by point fashion. Changes are highlighted in the attached revised version.

Anonymous Referee 1

This paper presents a comparison of the SEALDH-II hygrometer with the German PTB water vapor standards. The essential aspect of SEALDH-II is that it is calibration free and this validation effort closes the traceability chain with the German water vapor standard. While I have only minor comments regarding the comparison with the German standard itself, this paper raises significant questions regarding its position within the water vapor observation community. The questions, which I outline below, need to be addressed before this paper can be published. Therefore I would evaluate this paper as accept after major revisions.

General comments:
I have only minor comments regarding the technical work itself and go into detail these below. However, the larger concern is the novelty and importance as expressed in this paper. The authors claim that SEALDH-II is a novel hygrometer, which is a calibration-free, tunable diode laser spectrometer that bridges the gap between metrological water vapor standard and field deployed hygrometers. However, last year the authors published work (Buchholz et al., 2016) on the novel Hygrometer for Airborne Investigations (HAI), which they developed in cooperation with the Research Center Jülich. The claims made in that paper read very similar than the claims made about SEALDH-II. Both are claimed to be calibration free, with some level of metrological traceability. However, the HAI paper by the same lead author is not even referenced here, which is quite odd. It is not clear what the connection is between these two instruments and which is more novel than the other. The authors should clarify the connection between these two instruments, before claiming that SEALDH-II is a novel hygrometer.

==> Thank you very much for that comment.
The instruments HAI and SEALDH-II are only similar on the first glance: The used technique (dTDLAS with a calibration-free evaluation approach) is similar, both instruments can be deployed on airborne platforms, and both are hygrometers. To grasp the differences between the instruments, we should first take look at the most important features of HAI: HAI is a multi-channel and multi-phase instrument developed for the German research aircraft HALO. The term "multi-channel" means here, that HAI measures with two different laser systems at 1.4 and 2.6 μm at two different measurement locations. The first measurement location is the "open-path" cell, which is installed on the fuselage of HALO, i.e. outside of the airplane. The second location is a "close-path" cell inside of the cabin aircraft. Additionally, HAI comprises two separate close-path cells, one for each wavelength.
The open-path cell only analyzes pure gas-phase water. By connecting the closed path cells to a forward-facing trace gas inlet, in order to sample "total water" (i.e. not only gas phase but also ice particles and droplets), the entire configuration of open and closed path cells allows "multi-phase" water measurements. In flight sections through clouds, liquid and/or frozen water particles are sampled via the inlet and then evaporated in heated sampling lines. Thus, the closed path cells determine the so called "total water". Both instruments follow on slightly different path of the concept to ensure a permanent, highly defined, multi parameter instrument control for each individual data point during the application. Goal of this is to minimize uncertainty about the validity and correctness of each measurement point. To keep the SEALDH paper concise, we maybe didn't elaborate clearly enough on this connection between the instruments.

Indeed, the novelties of HAI and SEALDH-II are in different fields.
HAI is more complex, more versatile. Its particular novelty is the multi-phase, multi-wavelength approach and its field application and validation on one of the most modern airborne carriers. HAI's unique setup allows for the first time a highly redundant overall configuration which is enabling different versions of in-flight validations e.g. of sampling errors etc. Thus, HAI is clearly the most complex and versatile, calibration-free, airborne, multi-phase hygrometer.
HAI was until now not compared and validated at a primary standard and hence does not (yet) achieve SEALDH-IIs metrological correctness/reliability/trust level. All HAI papers are more focused on field related science questions: For example, we wrote a paper about how to validate the dTDLAS relevant pressure in the open-path cell [1] which allowed us to be sure about the accuracy of HALO's gas pressure readings. However, this is not achieved for the open path gas temperature so far.

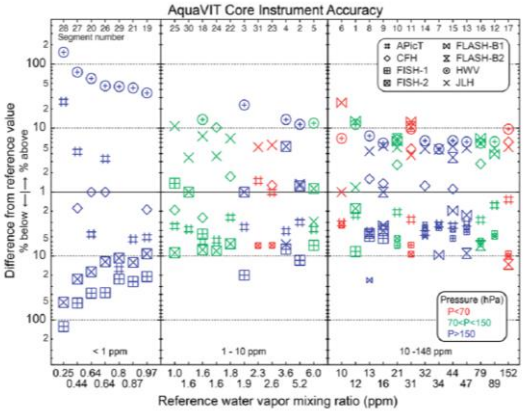
SEALDH-II on the other hand, compared to HAI, is significantly more advanced from a metrological point of view, i.e. with respect to data control, data treatment, metrological linkage, reliability level. In particular SEALDH-II is much more stringently validated at the highest primary metrological water standard. The novelty of SEALDH is thus clearly focused on the **metrological validation**.
Therefore, the full quality control from SEALDH-II still needs to be further developed, transferred and implemented in HAI. The present SEALDH manuscript deals exactly with these advantages of SEALDH II and with the results of the more stringent metrological validation experiment realized with SEALDH II. The novelty of SEALDH against HAI therefore is clearly given by the experimental rigorousness of the metrological validation. A further important novelty is that the metrological validation is realized over the full concentration and pressure range of the most relevant atmospheric regions of the UTLS and below. We added that some explanations to the manuscript, thank you for that hint!
We tried to make these differences clearer in the manuscript.

The authors also claim that this effort is the first metrologically validated humidity transfer standard. This may or may not be completely true.
==> We might have formulated our point not well enough, which could have led to some misconception by the reviewer. In essence, every metrological transfer standard which is used e.g. to compare different national primary standards with each other, or to transfer the representation for the metrological humidity scale to industry is a "metrologically validated humidity transfer standard. However in contrast to the reviewers statement, we wrote more detailed: *"With this validation, SEALDH-II is the first metrologically validated humidity transfer standard* ***which links several scientific airborne and laboratory measurement campaigns to the international metrological water vapor*** *scale."* We believe that this statement is true and justified, given the present manuscript and our previous papers on SEALDH. If the reviewer knows any publication about a full metrological validation of airborne laser hygrometers which contrasts or amends our statement, we would be thankful to include those in our paper and to adapt our statement.

During the AquaVIT -II campaign this community made a dedicated effort for a metrological validation of a number of instruments. The authors collected all observations, but never released the metrological reference observations. However, they presented this work at several conferences. That work may actually be the first metrological validation of several transfer standards.

==> It would have been very insightful if they had done exactly that. The peer reviewed paper and the White-paper [2], [3] of AquaVIT-1 explain in detail the deviations (an exemplary work!) but they are NOT linked back to any metrological source or device. The reason is obvious: At that time, the only available "metrological linked" devices were dew point mirror hygrometers (DPMH) which can be

traced back via a comparison with a primary generator. DPMH are often seen as the simple way of getting "metrological links". The DPMH used during Aquavit-1 was - to our knowledge - "only" traced back at 1 bar under static conditions. However, most of the AV-1 campaign was done at reduced pressures where the traceability was not given. What's more, it is questionable whether the traceability can be transferred from an extractive device (the DPMH) to an 80 m³ large vessel (AIDA) which was operated in a quasi-static mode, but which – in particular at the low humidity levels and gas pressures – was far from reaching a static equilibrium as it showed significant drifts over time. It is fundamentally wrong to believe, that an instrument which is used outside its validated range (pressure, temperature, concentration,…) and operation conditions for which is it built (time response, gas matrix, impurities,…) can be used as a "metrological" standard even if it was once metrologically calibrated/validated. Doint mirror hygrometers only show correct values when they are in equilibrium (ice layer, humidity, mirror temperature). At most of the AquaVIT conditions, it also has to be considered that even short (1-2 m) sampling pipes to DPMHs required equilibration times at the low humidity levels (e.g. ppmv range) of many hours up to a day in metrological institutions. Especially instruments with a low flow (e.g. DPMHs) have great problems to reach a stable equilibrium and are prone to long term drifts. Hence, the best reference instrument for AIDA would have been a sampling-free instrument like APICT, which was – however – due to many reasons not chosen as reference. We assume that the authors of the AQUAVIT papers were aware of that and therefore both papers do NOT claim any metrological linkage. Instead, all instruments were compared to the mean value of the core instruments. AquaVIT, thus answered mainly the question about the "spread between the instruments" Due to the lack of a suitable reference, the AQUAVIT comparisons are not suitable to give answers about the level of absolute accuracy. The argument which is sometimes made, that the mean value of AquaVIT is like the "true H_2O value" is as wrong as to claim that one instrument was more accurate than another.



Referring to AquaVIT2, it is a fact that no final results and certainly no full papers have been published; neither from the simultaneous comparison exercise inside AIDA (called AV2a), which by the design of the experiment would very hardly ensure tractability, nor from the sequential metrologically motivated comparison with a PTB transfer standard generator (called AV2b). It also needs to be noted that the sequential comparison was only applied to a quite small sub group of the AV2 participants and hereby mostly to the group's internal standards and to a much lesser extend to very few instruments used in AIDA. The AV2b comparison thus had a much reduced participation compared to the AV2a part. Even if the AV2b data once will be published, they will for sure be not comparable in rigor (setup, number of data points, concentration levels, and pressure levels) and in quality (control over system, time for validation) with the data presented here. As a consequence, AV2b is of much more indirect nature, which needs up to three calibration levels/steps to link the instrument of the relevant group to the primary standard at PTB. Second, the time planned for each individual comparison in AV2b was quite short to match the total time available for AV2a. Thus, for each of the instruments/group standards only very

few data points at 1000 hPa could be measured, with a quite short measurement time per data point. The pressure dependence could not be studied in detail.
To summarize, there are no papers and certainly no peer-reviewed full papers from AV-2b or 2a. We did work on this data and published some posters on preliminary results.

One of the drivers for the AquaVIT campaigns was the disagreement between some aircraft and balloon borne observations. Water vapor observations of less than 5 ppmv were the most important range of this disagreement. Given the uncertainty of SEALDHII, this instrument would not contribute to this concentration range. Despite the un-questioned quality of the observations presented here, this raises the question of the importance of their results, especially given the frequent reference to AquaVIT. The authors should clearly point out, whether they have achieved a metrological validation of a field deployed instrument that can measure stratospheric water vapor (100 hPa, 5 ppmv) with an uncertainty of less than 5-10%.
==> Thank you for that comment; nevertheless we are also a bit puzzled by it.
SEALDH is NOT an instrument designed especially for UT/LS application, nor is it one of the common "standard" even commercially available laser hygrometers like the ones from WVSS, Picarro, LosGatos, etc. Instead, SEALDH is designed as a metrological transfer standard which is still suitable for field applications and ensures full traceability to the primary water scale. Thus, the application range is wider than just UTLS and the requirements are also much more rigid and demanding. It is clearly stated in the abstract that we claim a "primary validation", that covers the relevant mid to upper tropospheric and lower stratospheric (MT/UT/LS) concentration and pressure range. This was done via "15 different H_2O concentration levels between 5 and 1200 ppmv". At each concentration level, we studied the pressure dependence at 6 different gas pressures between 65 and 950 hPa." The introduction states (as well as the referred paper in section 2 [4]) that the instrument is built for a concentration range from 3 - 40000 ppmv; 0 -1000 hPa with a calculated uncertainty of 4.3% ± 3 ppmv. The lower detection limit is defined at that point where measurement value equals uncertainty (as explained in the paper: "The calculated mixture fraction offset uncertainty of ±3 ppmv defines the lower detection limit."). The measurements in figure 8 show how conservative these estimations for the calculated uncertainties were since the "real" deviations are only in the 0 - 20% range (with the lowest level at 35%). SEALDHs science and metrology targets are thus much wider and of different perspective than those of the AQUAVIT 1 or 2.
The reasons why we explained the AquaVIT study in such a detail are the following: Firstly, it is the most extensive comparison exercise which was carried out on such a level (representative for the community, externally reviewed, blind submission, clear statements, exemplary work, etc.) that allows a reliable relative performance statement about the state of the art of airborne hygrometry. Secondly, as you can see in the graph above, AquaVIT covered the range up to 150 ppmv. Most instruments deployed during AquaVIT had a smaller concentration range compared to SEALDH-II because they were specifically developed for stratospheric measurements. With reference to the reviewer's question, we have to comment that one should not compare apples with oranges. SEALDH-II is intended as a transfer standard for the troposphere AND lower stratosphere and thus designed as a wide range hygrometer. This asks for pragmatic compromises which have consequences. If we had in mind to make a comparison **only** for the stratospheric range, we could easily replace the 1 m long extractive cell with a fiber-coupled multipath cell having 10x - 50x more path length. This would drastically improve SEALDH-II's sensitivity AND offset stability in the low concentration range. As the sensitivity and offset largely scales with the cell path length, this would lead for SEALDH-II to a calculated uncertainty (see [4]) of 4.3% ± 0.3ppmv (@ 10 m) and of 4.3% ± 0.06ppmv (@50 m) respectively, which seems quite well suited for a stratospheric application.

It also has to be kept in mind that we stated conservative calculated uncertainties; the "real" experimental deviations are usually smaller, as described and shown in this paper. In our experience – also as participant in AV 1 and 2 - we made the experience, that metrological humidity traceability options in the atmospheric sciences are more or less ignored, despite the availability of a fully validated metrological reference scale and infrastructure. It seems highly likely for us that an improved traceability would lead to better absolute accuracy of airborne hygrometry. SEALDH was designed to enable this link between metrology and field sciences. With the series of SEALDH papers and in particular with the one here under review, we intend to present the data that demonstrate this capability in a metrological sense and not just present another UTLS hygrometer.
So the argument of the reviewer that only improvements in stratospheric sensing would be required or relevant is a bit short sighted and ignores to a large extent our focus towards implementing general traceability for airborne hygrometry over the full range from LT to LS. Atmospheric science studies the entire atmospheric water cycle and not just its subsections in the stratosphere and as such, it would definitively be beneficiary to establish traceability with a wide range instrument such as SEALDH in order to improve the comparability between all atmospheric sub-compartments and not just between UTLS focused instruments.
We added a few lines to explain why we compare to AquaVIT.

Specific comments:
Line 18: I believe that the term 'bridges this gap by implementing an entirely new concept' is overselling their result. While their work is important, TDLAS technology is not new and has been around for quite a while. The authors own work on HAI show that this is not an 'entirely new concept'. Furthermore, given the relatively large metrological uncertainties at true stratospheric water vapor concentrations, I don't see, where a gap is being bridged.
==> We deleted the word "entirely" which was colloquial English.
Furthermore, the reviewer should be reminded that "The "new concept"" is not TDLAS. The new concept is validated traceability employed via a special, first principles TDLAS approach based on high-accuracy line data! It is a very common misconception that researchers who only "use" TDLAS think that all instrumental approaches have more or less identical properties – in particular with respect to the requirement for calibration. The statement to be "calibration-free" is also very often misused by lab focused TDLAS-groups without airborne experience as well as by commercial TDLAS instrument manufacturers. This is actually the reason why we term our approach "dTDLAS", in contrast to alternative evaluation approaches. Thus, overselling happens at other points of the TDLAS community, certainly not at our point. TDLAS in itself only stands for "Laser absorption spectroscopy with a tunable diode laser". "TDLAS" says nothing about the modulation approach or the data extraction and evaluation approach, so one has to be precise here.
Furthermore, SEALDH-II fortifies TDLAS with an unprecedented concept for a "complete" internal quality control mechanism, which yields a large and almost complete set of environmental/boundary parameters, the so called "housekeeping data". This ensures that every single (!) captured raw scan and hence every reported H_2O value of the instrument has an assigned multi-parameter derived quality statement. This in combination with our first-principles calibration-free approach and our own high accuracy spectral data truly separates our instrument from any other, and allows us to achieve trustworthy and reliable measurements. It also ensures that a metrological validation (like the one in this work) in a protected laboratory environment can be transferred to harsh field environment (see [4] where the compensation of external effects are demonstrated with an environmental simulation chamber). It would be

helpful if the reviewer states any papers were a comparable or even better dTDLAS concept has been realized and was validated with a national primary standard, with our rigorousness.
Until then, we recommend not to generalize over an entire instrument class. But if one wants to do so, it is unavoidable to acknowledge, that instruments, which have been developed and published so far, do NOT have a data quality control concept at the level of SEALDH-II. This is often caused by the lack of a full physical model between the measurement value and the response of the instrument – the same reason why these instruments eventually have to be calibrated. The instrumental SEALDH paper [4] describes several important quality control scenarios which can be detected during SEALDH-II operation just because of the rich house-keeping data sensing any change in boundary conditions of the instrument and the measurement zone.
==> We revised the sentence.

Line 25, 'first metrologically validated': Aren't the AquaVIT-II and to some extent even the AquaVIT-I measurements metrologically validated?
==> As discussed above: AquaVIT was a major step towards assessing the quality of airborne hygrometers which we wrote and honored in the paper as well. Though, AquaVIT was not linked to the SI.
The Aquavit group could not / or didn't want to select a single reference instrument to serve as the main reference point, as there was obviously no knowledge about a traceability path of any of the participating instruments. As one of the participants in Aquavit 1 (and AV2), I remember well that there was no acceptance of the traceability concept and no discussion about a traceable reference instrument. We could not reach the conclusion that a single instrument was qualified as a reference type "gold standard". Hence, the best compromise the group could find was to take an average over the most mature "core" instruments. Without the concept of a unique primary reference and without any acceptance for traceability, it can't be stated that Aquavit1 is traceable, not even partially.
In Aquavit2, the only possible traceability path was planned via an extra, sequential, comparison of instruments (called AV2b) with a mobile traceable transfer standard from PTB which was transported to KIT. However, only a small number of instruments participated in the traceable comparison and no final data evaluation has been published yet. Therefore, the outcome of the traceable side-comparison called Aquavit2b is still open. Due to the large size of the AIDA vessel, it is also difficult to directly transfer the traceability to the AIDA experiment, which definitively limits the metrological impact of the Aquavit1 and 2 comparisons. Furthermore as mentioned above, AV2b was suboptimal with respect to the number of data points per instrument, the lack of comparison data at reduced pressure, and the short measurement time per data points which made it difficult to reach a stable equilibrium. Furthermore, mainly reference generators and not instruments were tested in AV2b, which results in a fairly long traceability chain over 3 and more levels, which significantly increases the uncertainty.
Note: It is unclear what the reviewer means with "to some extent"? Traceability is a binary property - so either "Yes, the instrument/data is traceable" or "No, it's not".

Lines 38 and 57: The tropical tropopause is highly relevant for atmospheric water vapor and may show values of less than 1 ppmv. This lower limit is a common value for some regions and seasons.
==> That's correct – even though to my knowledge averages lows are more in the range of 3-4 ppm and values below 1ppm are actually not very common. Nevertheless, we added the word "typically" to keep the statement general.

Line 46: The target accuracy for field weather stations is certainly a lot lower than 15%. Field weather stations report relative humidity and 2%-5% accuracy (in RH) are more common requirements.
==> That's was indeed confusing; we meant relative deviation and had more standard instrumentation in our mind rather than sophisticated instrumentation which is rarely used.
Note: Maybe we were talking about different things. We meant "relative accuracy", i.e. an absolute uncertainty of 5% rH at 50% rH or even 30% rH corresponds to a relative target uncertainty of 10% or even 16.5 %. So maybe we are not that far away from each other.
We revised the sentence.

Line 50: WVSS-II instruments are another variant of TDL instruments. They are commercially available instruments but probably not standardized.
==> WVSS is wavelength modulation spectroscopy based sensor which definitively has to be calibrated.
We revised the sentence

Lin 55: Currently, in situ observations of water vapor are done at least up to 10 hPa and at gas temperatures of less than -90 deg C.
==> Changed

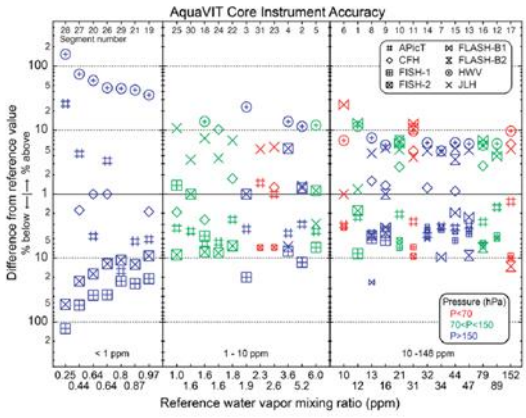
Line 79: : : : 'does not facilitate a clear accuracy assessment'. This study is still highly valuable and able to characterize the status of in situ observations during that campaign. The fact that large differences, seen in earlier campaigns, were not repeated there is of great value, even though there is no direct metrological connection.
==> We agree with your sentence – it's not in contradiction to ours. If you look at the (Rollins et al., 2014) paper, you cannot draw a clear picture of the reason or cause for deviations: Were these deviations "real" e.g. because of different gas samples? Or: Was it caused by effects which modify the gas sample? Or by external driving forces like an increase of the instrument temperature? Environmental parasitic humidity or external pressure change? Calibration issues? Or one of the many other "airborne typical issues"?
Total comparisons like that are indeed valuable but they make it quite difficult and often even impossible to assess the internal causes of the deviation. They show more a "comparative state of the art", but make individual instrument assessments and assignments of systematic errors to certain problem or cause difficult. Hence, they do not show where improvements have to be done in the future. AquaVIT instead tried to distinguish between different effects (e.g. only the stationary, flat segments of the comparison where AIDA was as constant as possible). Deviations are only calculated within these "quasi-stationary" areas. This was done in order to suppress/minimize dynamic effects. The evaluation and comparison of the dynamic sections in AQUAVIT could still be done; to our knowledge, there is no intention to do so.

Line 87-88: Differences of a factor of two stimulated AquaVIT. Disagreements of 10% were largely considered within the individual instrument uncertainties.
==> Let's take a look at the AquaVIT paper: Uncertainties in the range above 5 ppmv: APicT <5%; CFH 4%; FISH-1 6%; FISH-2 6%; Flash 10%; HWV 5%; JLH 10%.
The results of AquaVIT (see plot above) show that two instruments deviate by up to 20% from each other. Keep in mind that the reference value is, as discussed above, was just a mean value – the "true" value is therefore still unknown. Even if we consider the mean value could be the "true" value, we would have the problem that the uncertainty statements of the AQUAVIT participants are insufficient. An uncertain-

ty is NOT the typical deviation of an instrument [5], [6]. One also could rephrase the statement and state: "AquaVIT shows that the at least some instruments underestimate their uncertainty". However, I think this statement would be short-sighted since in the atmospheric communities "uncertainty", "error", "deviation", "accuracy", and "reproducibility" are often used interchangeably which causes this dilemma.
Each Aquavit participate had to state the performance or reliability or trustworthiness range of its instrument by giving an "offset term" and a relative "error". Those numbers were not analyzed nor had to be justified or even mathematically deduced and were more or less estimates by the instrument PIs. As there was no "true" reference, i.e. no validated absolute value, the error statements of the AV instrument could only be coarsely checked. Thus again, a metrological evaluation of the uncertainty by referencing an instrument to a metrological standard and to derive a metrological uncertainty is a completely different almost complementary approach.

Line 90f: The goal of AquaVIT was to evaluate instruments under controlled conditions, not to rigorously evaluate each instrument's uncertainties. No gold standard was included since no recognized standard was available for this setup.
==> Yes! We fully agree !
This is exactly the reason why our primary SEALDH validation is an indeed novel and highly relevant for the field! Contrary to a reviewer statement made earlier.

Line 128: Systematic differences of 20% and more were seen during AquaVIT at the lowest mixing ratios, i.e. below 3 ppmv. The authors should point out that SEALDH-II would not help addressing this concentration range.
==> I guess this question is referring to your third comment and hopefully already answered there. One more comment: The "span of up to 20%" describes the following graph. Keep in mind that these instruments are usually flown by themselves. If instrument A and instrument B report two measurement values, they could deviate 20% from each other (in the lower range even more).



Line 131: I doubt that this is the 'first comparison' with a metrological standard. Water vapor has been measured for a long time and a lot of validation efforts have happened, not all published. The AquaVIT-II activities, in which the authors have played an important role, is just one example.
==> The reviewer needs to be precise here: We do NOT claim the "first comparison with a metrological standard" in general, that would not be true. Instead, we present in this paper the "first comparison of an airborne hygrometer (SEALDH-II) with a metrological standard for the atmospheric relevant gas pressure (65 – 950 hPa) and H2O concentration range (5 – 1200 ppmv)". We believe that a detailed statement like this is true. In addition, we think it can be stated here, that AquaVIT-I was by no means

linked to the SI as discussed above and thus has nothing comparable to offer. For AquaVIT-II we don't understand the *diffuse, undocumented* "doubts" casted by the reviewer. We certainly don't know of any peer-reviewed final publication originating from Aquavit-2 which could challenge our claims, neither A) from the simultaneous comparison exercise in AIDA, which by the design of the experiment would be very hard to make traceable (but as said there is no full paper on AV-2), nor B) from the sequential comparison with a PTB transfer standard generator. The sequential comparison was furthermore applied only to a quite small sub group of the AV-2 participants and hereby mostly to the group's internal standards (not the instruments used in AIDA for AV-2a). This comparison part, termed AV-2b, did not have the wide participation like the AIDA part. In addition, it also did not cover the large validation range (in pressure and concentration nor did it invest sufficient amount of time as our work presented here (23 days! Permanent operation)). Even if the AV-2b data once will be published, they will for sure be not comparable in rigorousness and quality with the data presented here. One reason for this is the more indirect nature of the comparison, which needs up to three calibration levels/steps to link the instrument of the relevant groups to the primary standard at PTB. PTB just recently modified its primary standard to provide different gas pressure levels; we can for sure say that no such a comparison was linked to PTB (Germany) in the past. So again, there are no papers and certainly no peer-reviewed full papers from AV-2b. To our knowledge, there still needs to be a final data discussion meeting amongst all participants to be announced in order to decide about the further evaluation and use of the AV2 and the AV2b data.
We are of course very interested in learning which comparisons the reviewer has in mind, and would be grateful if the relevant citations could be forwarded to us. This includes publications of comparisons as well as any other metrological validations of airborne laser hygrometers, we would be very delighted to read, assess, discuss, and if suited add them.
Until then, we think our above statement is valid as well as justified and it's on the reviewer to provide references and detailed information if he is aware of comparable earlier work. To our best knowledge our data are unique in quality and in impact.

Line, 158, 165, 345-349: The lower limit of 3 ppmv is a significant limitation, since the <10 ppmv range is essential for stratospheric observations. At 5 ppmv an uncertainty of 3 ppmv makes the measurement effectively useless for stratospheric research. This should be discussed in greater detail.
==> We looked through the paper to make sure at any point that SEALDH-II is not presented as a stratospheric instrument as we also explained at other comments. That has never been our intention. We added some words.

Lines 207ff: There are other calibration free instruments. HAI, published by the authors is one of them. Some of the frostpoint hygrometers, which are being used on aircraft and balloons may be considered calibration free in the same sense. Other TDL instruments are equally considered calibration free under the definition of the authors.
==> We indeed understand the point which the reviewer wants to make, but - as explained above - we don't "oversell" and we have given in the paper presented here detailed technical explanations what we mean with "calibration-free" and even more important we have given experimental justification why we can claim this properties. It is certainly quite decisive not just to CLAIM that the instrument is calibration-free (which (too) many groups do) but also to proof it and to VALIDATE this property as we do here via a side by side comparison with the world's highest accuracy water scale and one of the few primary water standards. To our knowledge, none of the airborne TDLAS instruments has ever done this. For organizational reasons and HALO mission deployments, HAI was never available long enough in order to experimentally validate this property at the primary standard. Even though there were HAI papers coming out recently, none of them was focusing on the validation of the absolute accuracy, as

this exercise was done with the SEALDH instrument which wasn't in such long and frequent airborne missions. Again, we ask the reviewer to state papers from other air- or balloon-borne TDLAS instruments which proof a primary metrological validation exercise. We don't know any. Essentially there are many groups which claim that property but no one so far challenged, validated, and quantified experimentally the degree of accuracy which can be achieved.
Referring to a frost point instrument, we can make the following statement. Calibration-free would mean the following: The instrument is assembled and e.g. the temperature/pressure sensors (if not first principle methods) are calibrated. Then the "target" gas is guided through the frost point hygrometer and it has to report an absolute value of e.g. "500 ppmv". In an ideal world, this might be possible (using the Sonntags equation [7], ideal gas law, etc.); however, in a real world it is not. This also explains that even the most sophisticated frost/dew point hygrometer have to be calibrated regularly. Reasons for that are e.g. technical problems such as: The temperature sensor has to measure the surface (!) temperature of the mirror; the constant airflow above the mirror transfers heat away. I agreed, all this could be modelled and corrected in principle – however nobody (to our knowledge) has every published such a work. Operational problems such as: Any kind of hydrophobic substance, any dirty, micro scratches on the mirror, etc. change the ice layer behavior locally. Therefore, the "frost point" can be shifted slightly which is corrected during calibration. Physical problems: The enhancement factors (which describe the dew point shift if the gas is not ideal (Sontag equation is not accurate in this case)) cannot be calculated due to the lack of a full physical model.

Line 216, 'accuracy': JCGM (2008) recommends not using this term in a quantitative sense. The authors should explain what they refer to here.
==> Line 216 contains: "… philosophy leads to measurements which are very reliable with respect to accuracy, precision" and the comment is probably referred to the line: "measurement accuracy" defined as "closeness of agreement between a measured quantity and a true quantity value of a measurand. NOTE: The concept 'measurement accuracy' is not a quantity and is not given a numerical quantity value. A measurement is said to be more accurate when it offers a smaller measurement error"
As mentioned above, we believe that in non-metrological, application and field science oriented journals like AMT, we have to find a bridging language which can convey and explain the general idea and results of the scientific work. For this paper here, we decided to stay as often as possible within the terminology of the AMT community and to find a compromise between the common understood language and the rigorous metrological terminology. We strongly believe that our paper would not profit, probably be less understandable, and thus have less impact if we add not commonly used expressions such as trueness and closeness. Therefore, we assume that every reader will understand the usage of "accuracy". A metrologically much more rigorous discussion of the implications of our work will be subject to upcoming papers in common journals of the metrological community like "Metrologia" or the like.

Lines 220ff: The authors point out later in the manuscript, that calibration in the strict sense improves the measurements only, if the ambient conditions can be replicated during the calibration. They should elaborate on this topic and consolidate the various paragraphs throughout the manuscript.
==> Long-term stability (long = relatively longer than the drift of the instrument) of the instrument is a requirement for any calibration. We wrote: "(…) a calibration can only improve the accuracy for the relatively short time between two calibration-cycles by adding all uncertainty contributions linked to the calibration itself to the system."
This sentence referrers to every calibration process. The replication of the ambient condition is necessary if the impact of the ambient conditions on the calibrated device is unknown. As in the comment for Line

216, we will give a more rigorous discussion in planned papers for the metrological community. In the interest of keeping this manuscript in an acceptable length to content ratio we can't discuss general topics like this in the given journal and paper.

Lines 264-267: Delete. These sentences contribute nothing and could be deleted.
==> We don't understand why these lines should be deleted! These lines contain: "In this paper, we present data from a permanent 23 day long (550 operation hours) operation in automatic mode. Despite a very rigorous and extensive monitoring of SEALDH-II's internal status, no malfunctions of SEALDH-II could be detected. One reason for this are the extensive internal control and error handling mechanisms introduced in SEALDH-II, which are mentioned above and described elsewhere (Buchholz et al., 2016)."
This comment might link to the question above "why SEALDH-II has a new concept" which we hopefully have been clarified. Typical TDLAS instruments such as e.g. the WVSS-II (mentioned above) just give a value with a minimalistic "set" of operational data. A definite assessment whether the measurement process went well or had issues inflicting the quality is not possible. SEALDH-II embodies a new, holistic, embedded concept. Therefore, this question should now be answered and the content of the sentence should be clear: "No error" with an instrument such as SEALDH-II means: "no error" and not "maybe / probably a correct measurement"

Lines 301f, 'One has to compare : : :' No, this comparison does not have to be done. The purpose of AquaVIT was very different and a metrological standard was not available at that time. This statement should be deleted.
==> We have already revised this sentence according to the comments above. It seems that the reviewer presumes that we do not appreciate the tremendous contribution which AquaVIT brought to the hygrometer community. We wrote (line 89) "AquaVIT was a unique first step to document and improve the accuracy of airborne measurements in order to make them more comparable". If the reviewer would like to add an even more appreciative sentence, we are happy to include it!

Lines 306fff (section 4.1): Isn't the point of controlled static setups to minimize the impact of dynamic effects on the uncertainty estimation? Fundamentally the uncertainty of the SEALDH-II cannot be better than that of the THG. Therefore, the authors should quantify the impact of the THG dynamic effects on their static uncertainty estimation of SEALDH-II, if that is possible.
==> This statement is correct. The figure 2 shows exactly that: The uncertainty stated there (line 643) is "The uncertainties of every individual calibration point are stated as green numbers below every single measurement point". Figure 5 left allow double checking of this statement. The max peak-to-peak deviation is approx. 4 ppmv at 1250 ppmv equals 0.3%. The Figure 5 right is an extreme case where the dew point mirror data are oscillating. A section like that is not preferable/acceptable for a validation.
We added that for clarification.
As a site note: A few comments above, the reviewer stated "2%-5% accuracy (in RH) are more common requirements". Figure 5 right shows that a measurement which has a small dynamic change can easily oscillate a frost point mirror hygrometer in the order of 2%. Therefore, a reliable accuracy of 2% is not "easily" achievable with frost point mirror hygrometers which are often seen as the most accurate hygrometers.

Line 314: What is an 'indirect, inertia, thermal adjustment process'? The authors should find a better term for what is meant here.
==> Revised

Line 342: PHG should be Primary Humidity Generator.

==> Revised, thank you

Lines 365ff, 'It is important : : :': What does this sentence mean? Any uncertainty estimate always implies that the true uncertainty could be smaller. It could also be at the estimate. Lines 376 through 378 are somewhat contradictory. The authors place great value that the measurements presented here are the first metrological validation of SEALDH-II. How would non-metrological validations done previously provide contribute?

==> Response to the first question: We have already answered this question to some extent above: The words "uncertainty", "error", "deviation", "accuracy", and "reproducibility" are often used interchangeably in different communities: E.g. if an instrument has a typical "deviation" of 2%, the uncertainty is sometimes just defined as 2%. From a metrological point of view, such an "uncertainty definition" has to be seen critical. An uncertainty budget/consideration should include all significant factors to understand their individual contributions. An aggregated value, such as a deviation, does not facilitate further statements about the "reliability" of this so defined uncertainty. The more the deviation measurements includes other parameters such as instrument temperature, external/internal pressure, power supply fluctuations, vibrations etc. the more reliably a deviation statement can be transformed into an uncertainty - even with a black-box-like system. SEALDH-II is designed as a fully white box system to avoid in the first place these kinds of "deviation=> uncertainty" discussions.

==> Response to the second question: We described AquaVIT as the largest, most representative, non-metrological laboratory comparison exercise. From our point of view, AquaVIT is unique in terms of quality level and the conclusions which could be drawn from it. Therefore, we did not describe other airborne or laboratory campaigns since a detailed analysis would lengthen this paper unnecessarily and would not add value since the overall statements remains. If the reviewer is interested to read more about other water vapor related comparisons: here are some entry points [8]–[15]

Do the authors imply that non-metrological validations are equally useful or even better suited to address the uncertainty issue at low pressures and low mixing ratios? As shown in this manuscript, the uncertainty of SEALDH-II at true stratospheric values (low pressure and low mixing ratios) is too large to be scientifically relevant.

==> First part of the question: A non-metrological validation has by definition no direct link the SI units. Therefore, it is well suited for comparing instruments relatively to each other but as soon as the absolute value is of strong interest, a metrological link should be established, to provide comparability between different instruments. This is a general recommendation and does explicitly not depend on any pressure and/or mixing ratio range.

==> Second part: SEALDH-II is – as mentioned frequently – not designed as a dedicated stratospheric instrument. It could be adopted for LS conditions but so far it isn't ; we clarified this in this rebuttal several times as well as in the paper. SEALDH-II could, as explained above, set up with a longer optical path length to serve as a standard for stratospheric values. But currently, SEALDH-II is set up for tropospheric concentrations up to 40000 ppmv and with its uncertainty of 4.3% ± 3 ppmv not well suited for single digits ppmv measurements. The measurement principle, however, could be easily adapted to LS conditions.

Lines 428f: Why the authors would want to suppress this systematic pressure dependence? In instrument comparisons and atmospheric measurements the systematic biases are often the determining factors.

==> We don't suppress. If we did, we would only show the "corrected" i.e. calibrated data. As we have a clear physical explanation for the pressure dependence, which is an inherent deficit of the VOIGT line

shape, we certainly know that by expansion of our physical measurement model – i.e. inclusion of a higher order line shape profile – we could "remove" this pressure deviation at the cost of higher computational efforts and an enlargement of the number of necessary fit parameters. Recent developments in higher order line shape models (such Speed-dependent-Voigt (SDV) [16] or Hartmann-Tran-Profile (HTP) [17]) are most suited for such an improvement, which we have planned for future updated versions of our fitting model.

Line 667f: What do the authors want to say here? The sentence as is doesn't make sense.
==> Maybe we talk about different lines: "The different dynamic characteristics of SEALDH-II (fast response time) and THG (quite slow response) lead in a direct comparison to artificial noise." We added here a sentence.

Figures 6-8: The abscissa should be shown as the Log of P. This makes it easier to relate the altitude and emphasizes the lower pressures, where water vapor is more challenging.
==> Thank you for that comment. SEALDH-II is a wide range (3 - 40000 ppmv) hygrometer. Therefore, the entire pressure range is equally important. This comment is probably inspired by the fact that the reviewer assumed SEALDH-II might be a purely stratospheric instrument. We clarified that (see multiple comments above)

The authors use the term 'calibration-free' excessively and should reduce it to the necessary amount. The term is defined in a dedicated section and does not need to be repeated subsequently. The authors do not seem to be completely familiar with the water vapor observation community. Stratospheric water vapor is also observed on large and small balloons reaching all the way into the middle stratosphere. These measurements use a variety of techniques, none of which are referenced, but should be referenced. Water vapor is also measured using remote sensing (Raman and DIAL lidar), which are technologies comparable to their own. In particular DIAL measurements are considered calibration free and traceable measurements.
==> Thank you for this comment. SEALDH-II is not a stratospheric hygrometer; therefore, we didn't focus on this single atmospheric region. The target of this paper is a "Pressure dependent absolute validation from 5 – 1200 ppmv at a metrological humidity generator" and not to write a review about different water measurement techniques even up to the middle stratosphere. As an entry point for such a review for different measurement methods: [18]
We would be interested to learn more about the "default" traceability of DIAL measurements. The technique certainly requires traceable absorption coefficients, which are very difficult to get, due to the lack of traceable spectral data. We would be delighted to receive some references about this specific topic.

Lines 61-64: These statements are much too broad and even incorrect. The vast majority of water vapor observations has been quite sufficient for validation studies of models. The limiting factor in model validation is usually the availability and coverage of these observations, not their quality. The authors should change this statement.
==> The statement was: "These and other impacts complicate reliable, accurate, long-term stable H_2O measurements" followed by a brief outline why water vapor measurements remain a quite difficult in-situ measurement in the field, even if they are nearly always needed in atmospheric science. Up to now, the lack of sufficient accuracy may have limited important scientific interpretations (Krämer et al., 2009; Peter et al., 2006; Scherer et al., 63 2008; Sherwood et al., 2014)."

==> We added the comment about availability and coverage – thank you for that.

We would like to note that we can't follow the reviewer's opinion that "quality" is not a limiting factor in atmospheric science. This is definitively too general. I remind e.g. of a paper about observations of atmospheric super saturations of 300% in the UTLS, well beyond the homogenous nucleation threshold! It is not unlikely that these were caused by significant offsets of non-traceable instruments, which then led to this extreme numbers. Other examples could open discussions on sampling effects in gas lines to extractive water sensors. Here, our 4-fold redundant HAI sensor could show in the future a quantification of such effects. We are convinced that data quality and quantification of disturbances is actually often a topic in atmospheric H_2O detection.

Technical comments:
Line 52: standard (singular)
==> thank you
Line 60: delete 'a quite'
==> thank you
Line 60: measurements (plural)
==> thank you
Line 73: Better: The latter is particularly important for investigations in heterogeneous regions in the lower troposphere as well as for investigations in clouds.
==> thank you
Line 101: : : : inside the aircraft: : :
==> thank you
Line 281: Replace ')(' with ', '
==> thank you
Line 359: Delete '(primary standard = calibration-free)', which is a meaningless repetition here. Also delete 'calibration-free' in the same line, which is again a repetition.
==> We revised to emphasis that two calibration-free and one calibrated device are compared with each other.
Line 155, 389, 413: What is the meaning of 'holistic' in this paper? Better to delete this term.
==> We introduces this term for a situation where a "state of the art detection principle" is embedded in a sophisticatedly controlled/supervised environment. I guess the comment above "why is SEALDH-II a new concept" and this comment are linked. We emphasized that better in the text.

[1] B. Buchholz, A. Afchine, and V. Ebert, "Rapid, optical measurement of the atmospheric pressure on a fast research aircraft using open-path TDLAS," *Atmospheric Measurement Techniques*, vol. 7, pp. 3653–3666, (2014), doi:10.5194/amt-7-3653-2014.

[2] D. W. Fahey, H. Saathoff, C. Schiller, V. Ebert, T. Peter, N. Amarouche, L. M. Avallone, R. Bauer, L. E. Christensen, G. Durry, C. Dyroff, R. Herman, S. Hunsmann, S. Khaykin, P. Mackrodt, J. B. Smith, N. Spelten, R. F. Troy, S. Wagner, and F. G. Wienhold, "The AquaVIT-1 intercomparison of atmospheric water vapor measurement techniques," *Atmospheric Measurement Techniques*, vol. 7, pp. 3159–3251, (2014), doi:10.5194/amtd-7-3159-2014.

[3] D. Fahey and R. Gao, "Summary of the AquaVIT Water Vapor Intercomparison: Static Experiments," *source: https://aquavit.icg.kfa-juelich.de/WhitePaper/AquaVITWhitePaper_Final_23Oct2009_6MB.pdf (last accessed: May 2014)*, (2009).

[4] B. Buchholz, S. Kallweit, and V. Ebert, "SEALDH-II—An Autonomous, Holistically Controlled, First Principles TDLAS Hygrometer for Field and Airborne Applications: Design–Setup–Accuracy/Stability Stress Test," *Sensors*, vol. 17, no. 1, p. 68, (2016), doi:10.3390/s17010068.

[5] Joint Committee for Guides in Metrology (JCGM), "Evaluation of measurement data - An introduction to the 'Guide to the expression of uncertainty in measurement' and related documents," *BIPM: Bureau International des Poids et Mesures, www.bipm.org*, (2009).

[6] JCGM 2008, "JCGM 200 : 2008 International vocabulary of metrology — Basic and general concepts and associated terms (VIM) Vocabulaire international de métrologie — Concepts fondamentaux et généraux et termes associés (VIM)," *International Organization for Standardization*, vol. 3, no. Vim, p. 104, (2008), doi:10.1016/0263-2241(85)90006-5.

[7] D. Sonntag, "Important new Values of the Physical Constants of 1968, Vapour Pressure Formulations based on the ITS-90, and Psychrometer Formulae," *Meteorologische Zeitschrift*, vol. 40, no. 5, pp. 340–344, (1990).

[8] M. Heinonen, "A comparison of humidity standards at seven European national standards laboratories," *Metrologia*, vol. 39, no. 3, pp. 303–308, (2002), doi:10.1088/0026-1394/39/3/7.

[9] J. M. Livingston, B. Schmid, p. b. Russell, J. Redemann, J. R. Podolske, and G. S. Diskin, "Comparison of Water Vapor Measurements by Airborne Sun Photometer and Diode Laser Hygrometer on the NASA DC-8," *Journal of Atmospheric and Oceanic Technology*, vol. 25 (10), pp. 1733–1743, (2008).

[10] S. J. Abel, R. J. Cotton, P. a. Barrett, and a. K. Vance, "A comparison of ice water content measurement techniques on the FAAM BAe-146 aircraft," *Atmospheric Measurement Techniques Discussions*, vol. 7, no. 5, pp. 4815–4857, (2014), doi:10.5194/amtd-7-4815-2014.

[11] M. Helten, H. G. J. Smit, and D. Kley, "In-flight comparison of MOZAIC and POLINAT water vapor measurements," *Journal of geophysical research*, vol. 104 (D21), pp. 26087–26096, (1999).

[12] S. J. Abel, R. J. Cotton, P. A. Barrett, and A. K. Vance, "A comparison of ice water content measurement techniques on the FAAM BAe-146 aircraft," *Atmospheric Measurement Techniques*, vol. 7, no. 9, pp. 3007–3022, (2014), doi:10.5194/amt-7-3007-2014.

[13] R. A. Ferrare, S. H. Melfi, D. N. Whiteman, K. D. Evans, F. J. Schmidlin, and D. O. Starr, "A comparison of water vapor measurements made by Raman lidar and radiosondes," *Journal of Atmospheric & Oceanic Technology*, vol. 17, pp. 1177–1195, (1995), doi:10.1175/1520-0426(1995)012<1177:ACOWVM>2.0.CO;2.

[14] A. K. Vance, S. J. Abel, R. J. Cotton, and A. M. Woolley, "Performance of WVSS-II hygrometers on the FAAM research aircraft," *Atmospheric Measurement Techniques*, vol. 8, no. 3, pp. 1617–1625, (2015), doi:10.5194/amt-8-1617-2015.

[15] H. Vömel, V. Yushkov, S. Khaykin, L. Korshunov, E. Kyrö, and R. Kivi, "Intercomparisons of Stratospheric Water Vapor Sensors: FLASH-B and NOAA/CMDL Frost-Point Hygrometer," *Journal of Atmospheric and Oceanic Technology*, vol. 24, no. 6, pp. 941–952, (2007), doi:10.1175/JTECH2007.1.

[16] C. D. Boone, K. A. Walker, and P. F. Bernath, "Speed-dependent Voigt profile for water vapor in infrared remote sensing applications," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 105, no. 3, pp. 525–532, (2007), doi:10.1016/j.jqsrt.2006.11.015.

[17] J. Tennyson, P. F. Bernath, A. Campargue, A. G. Cs??sz??r, L. Daumont, R. R. Gamache, J. T. Hodges, D. Lisak, O. V. Naumenko, L. S. Rothman, H. Tran, J. M. Hartmann, N. F. Zobov, J. Buldyreva, C. D. Boone, M. D. De Vizia, L. Gianfrani, R. McPheat, D. Weidmann, J. Murray, N. H. Ngo, and O. L. Polyansky, "Recommended isolated-line profile for representing high-resolution spectroscopic transitions (IUPAC technical report)," *Pure and Applied Chemistry*, vol. 86, no. 12, pp. 1931–1943, (2014), doi:10.1515/pac-2014-0208.

[18] ManfredWendisch and J.-L. Brenguier, *Airborne Measurements for Environmental Research: Methods*

and Instruments. WILEY-VCH Verlag GmbH & Co. KGaA, (2013).