The authors thank the reviewer for some insightful comments and suggesting areas that require further discussion. We have attempted to address all the reviewer's comments below (in blue). When the text required significant revisions we pointed to the section of the updated manuscript where the revisions were made. We have included some of the reviewer's suggestions in section 6 (Conclusions and future work).

The present study describes the validation of ammonia (NH3) data retrieved from the AIRS and CrIS satellite instruments using the MUSES algorithm. The study focuses on comparing NH3 profiles and columns derived from aircraft measurements obtained during the DISCOVER-AQ campaign in California and Colorado, specifically in source regions of ammonia. Additionally, it includes a comparison with three years of surface NH3 measurements from a monitoring network in Idaho.

The manuscript is well written, properly structured, and aligns well with the scopes of AMT. It contributes to the extensive validation efforts of various NH3 products derived from satellite measurements, which is crucial considering the growing utilization of NH3 satellite data. There is an evident need for such validation. Overall, the comparisons between satellite and in-situ data are well executed and yield interesting results.

However, I have some general concerns regarding the significant uncertainties inherent in the comparisons involving satellite-derived mixing ratios and in-situ measurements at specific altitudes/pressure levels, particularly at the surface. This concern arises due to the absence of vertical information that can be obtained from trace gas retrievals like NH3. Additionally, I believe the manuscript lacks thorough discussion and investigation into the reasons behind the remaining biases/uncertainties. Addressing these aspects would provide valuable material for the paper's conclusion.

Therefore, I recommend publication once the following major comments listed below are addressed.

Major comments

A major conclusion of the study is that a portion of the biases between satellite and in-situ measurements can be attributed to smoothing errors and unaccounted error sources, which can be substantial. This is evident in Figure 5, where the standard deviation of the biases remains large even after applying the AVKs. However, I believe the study could go deeper into understanding the sources of these biases and uncertainties, and additional tests within this framework would be beneficial. Specifically:

- The manuscript should discuss the uncertainties and errors associated with temperature and H2O profiles, as they often have significant impacts on trace gas retrievals from nadir-viewing satellite observations, particularly for NH3 that mainly resides in the boundary layer. It would be helpful to explore whether the remaining biases between satellite and in-situ measurements are dependent on errors in these two variables.

Additionally, what is the influence of uncertainties in temperature and H2O profiles on the overall NH3 uncertainty budget?

This is certainly an important topic, and prompted by the reviewer's question on why systematic errors were not included we reviewed the text and realized this had been stated incorrectly and has now been now fixed. The MUSES NH3 retrieval step includes the estimated errors in NH3 due to water vapor and temperature errors in the systematic error. We now mention the range of systematic errors in the text. We also state the estimated errors in Figure 6 are sum of the measurement error and the systematic error, and that since the measurement error is usually small, these estimated errors are nearly equivalent to the errors due to water vapor and temperature errors. Given the good agreement between the estimated errors and the actual uncertainties in Colorado below the MLH, we believe these uncertainties have captured the error sources. In California it is obvious from Figure 6 that the retrieval errors are underestimated, either because the systematic errors are incorrect or there are missing error sources. We have also added the following to section 6.

This study has not attempted to untangle the impact of the errors in the retrieved water vapor from those of temperature on the NH_3 errors.. Such an analysis is an important and ongoing task, as global maps of CrIS NH3 have revealed artificial hotspots of NH3 over tropical oceans, where humidity is high. There is a weak water vapor line in the spectral region used in the NH3 retrievals, which is possibly leading to these artifacts.

- Information about the DOFS associated with the AIRS and CrIS NH3 retrievals would be interesting to discuss. Did you exclude observations based on their DOFS before conducting the validation? Do the biases between satellite and in-situ measurements decrease when filtering out observations with low DOFS? Exploring this aspect would provide valuable insights.

  The small number of aircraft profiles led us to exclude only truly poor retrievals. We have modified the text where we discuss the QC as follows:

Retrievals were checked for quality by ensuring that for all retrievals the root mean square error (RMSE) of the residuals was less than 5.0. The MUSES cloud optical depth (COD) values were also evaluated but since the maximum COD for the retrieved profiles was 0.25 no retrievals were rejected due to large COD. Four CrIS profiles over Colorado were rejected due to very high estimated uncertainties. The DOFs ranged were between 0.8 and 1.1, except for two CrIS profiles over California, four AIRS profiles in California and six AIRS profiles in Colorado, for which the DOFS were smaller (0.2 to 0.7). Given the small number of profiles in each dataset, we did not exclude any profiles based on the DOFs.

- The choice of the a priori profile is crucial. Are the selected a priori profiles representative enough for the conditions in California and Colorado during the in-situ measurements? How would the biases between satellite and in-situ measurements change if you adopted an a priori profile that peaks closer to the surface? Would such a choice help reduce the biases? This is an aspect worth investigating and discussing.

  The authors are well aware of the impact of the a priori on the retrieved profiles. In the column comparisons in section 4.1 and 4.2 we now discuss how the choice of a background or moderate a priori (see new Figure S1 in the supplement) leads to a very different vertical distribution than the enhanced prior, which does peak at the surface. Comparing the a priori profile shapes with the visualization of the aircraft profiles shown in Figure 2 suggests that over California the enhanced a priori profile reflects the shape of the measured profiles, albeit with a shorter tail and more gradual rise. Over Colorado an enhanced prior with simply a steeper rise would have been ideal. These modified a priori profiles might have led to better agreement with the aircraft data, though please note the new discussion about the detection limits of the PTR-MS over California (7.0 ppbv) and Colorado (3.0 ppbv). However, using modified a priori profiles would have defeated the purpose of this paper, which is to evaluate the MUSES retrievals as they are now and to provide users who have utilized these retrievals an estimate of the current uncertainties. In the next phase of the algorithm development we will experiment with modified retrieval shapes and the addition of a super-enhanced prior with a long tail. We have taken the reviewer's suggestion and added this issue to the discussion of future work in section 6.

- I have concerns when comparing trace gas mixing ratios retrieved from ground-based or spaceborne observations at a specific altitude/pressure level directly with in-situ measurements taken at the same level. It's important to recognize that, with the optimal estimation, the value retrieved at this level alone lacks meaning as it heavily relies on information obtained from other levels. This is particularly relevant for trace gases like NH3, where only a single piece of information (total column) can be obtained. It becomes even more complex when comparing surface in-situ measurements with near-surface mixing ratios derived from retrieved profiles, considering the decreased sensitivity of satellite IR sounders in the lowermost tropospheric layers. Although the sum of each row of the AVKs shows some sensitivity to the near surface, the influence of the a priori profile remains substantial in these layers. Furthermore, the shapes of the AVKs indicate a tendency for the retrievals to overcompensate in the free troposphere (Figure 4), suggesting that part of the information used to estimate near-surface values originates from higher levels. Additionally, I assume that the constraint matrix restricts the variability in these layers, incorporating inter-layer correlations through the extra-diagonal elements, in order to prevent abnormal oscillations in the retrieved profile and maintain it within a reasonable range relative to the a priori. Considering these factors, it is crucial to thoroughly investigate and discuss the extent to which these retrieval characteristics impact the comparisons between satellite and in-situ data before drawing conclusions.

> The authors agree with every point the reviewer has made here. Yes, the constraint matrix was designed to restrict variability in the free troposphere. We have added the text below in the section describing the SRAK, and added further comments in the column section on the impact of the a priori choice.
>
> *The sum of the rows of the averaging kernels (SRAK) (Figure 4, top two panels), which provides an estimate of the retrieved information at each level originating from the measurement rather than from the a priori, shows for both AIRS and CrIS that while the information from the radiance data peaks just below 700 hPa, it also significantly contributes to the retrieved surface values. This is driven by the structure of the covariance matrix (S_a). As noted in the introduction of the DISCOVER-AQ section, the DOFS for AIRS and CRIS NH3 ranged mostly between 0.8 and 1.0, signifying the retrieval provides only one piece of information, basically a column amount. By building off-diagonal correlations in a priori covariance matrix between the surface level and a few levels above, this information is vertically distributed in such a way that it restricts unphysical oscillations in the retrieved profile and deviations a priori profile shape. Each of the three a priori profiles is associated with a different covariance matrix. The enhanced a priori retrieval tends to load the profiles at the surface, while the moderate and background profiles push NH3 to the free troposphere*

In Guo et al. (2021), it was demonstrated that notable differences exist between the ascent and descent aircraft profiles of NH3 measurements obtained during the DISCOVER-AQ campaign. These profiles are utilized for validating the AIRS and CrIS NH3 observations in the current study. The observed disparities arise due to the slow response time of the PTR-MS instruments, leading to a sampling lag when the aircraft moves from the boundary layer to the free troposphere, and vice versa. Since the majority of NH3 is concentrated in the boundary layer, this response lag results in an overestimation of NH3 in the free troposphere during upward spirals and an underestimation of NH3 in the boundary layer during downward spirals. These biases between ascent and descent profiles have been identified as significant and, when combined with the 35% uncertainties associated with NH3 measurements, can considerably impact the comparisons with satellite measurements. However, this point is not discussed or accounted for in the present study.

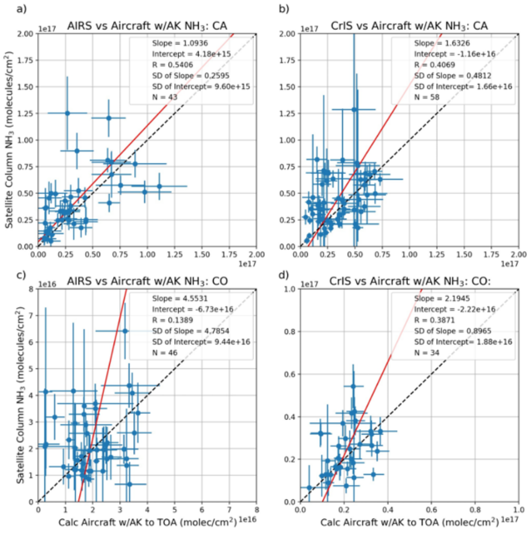
The authors agree that Guo et al. (2021) demonstrate this effect very clearly. We have added the following text that we hope addresses the reviewer's concerns:

The CrIS retrieval layers are fairly coarse and therefore the median value of the PTR-MS is derived from a set of hundreds of measurements spanning the layer, and from both up and down flight paths, thus possibly reducing to some degree the biases from entering and leaving the boundary layer discussed above.

However, since the initial draft of this paper was submitted to AMT we learned we had assumed the wrong detection limits (based on Müller et al. 2014) for NH3 during these campaigns, and the true limits were much higher: 7.0 ppbv in California, and 3.0 ppbv in Colorado. Examining the PTR-MS data we found that basically almost all the observations above the MLH were below these limits and effectively noise.

In the comparison of NH3 columns between satellite and in-situ measurements (Figure 7), it is evident that both AIRS and CrIS tend to overestimate NH3 columns in California and Colorado, as indicated by the slopes ranging from 1.6 to 2.3. However, the application of AVKs has not been performed on the aircraft profiles in this comparison. Since satellite sounders exhibit reduced sensitivity and obtain less information in the near-surface layers, where NH3 is predominantly abundant, it is expected that the AVKs would diminish the influence of these layers when computing NH3 total columns from the smoothed aircraft profiles. Consequently, the aircraft NH3 columns may be lower, potentially accentuating the overestimation of NH3 by AIRS and CrIS. This aspect should be discussed.

The authors have invested considerable effort trying to understand the total column results. We did try calculating total aircraft columns by integrating the extended profiles with the AK applied (profiles shown in the new Figure 5), but found the results difficult to understand (see plot to the left). In fact we tried a number of different approaches with and without applying the AK and in the end opted for a simple experiment. Since the aircraft data above the MLH is unreliable, we continued to use the aircraft data integrated to the MLH, and did the same for the AIRS/CrIS data (new Figure 8). We feel that comparing the total and partial columns, and considering the effect of the a priori choice and the aircraft uncertainties, along with the fact mentioned by the reviewer that July and August are fire season in Colorado, has provided some useful insights. Please see the text in the column section for our conclusions.



On the other hand, in section 5.2, it is revealed that CrIS NH3 exhibits an overall low bias when compared to surface in-situ measurements (slope of 0.65). This finding contradicts the previously noted large overestimation observed for CrIS NH3 columns (as mentioned in my previous comment). This discrepancy raises additional concerns regarding the reliability of quantitative comparisons between surface in-situ and satellite data. It suggests that such comparisons at the surface level are particularly uncertain and prone to biases.

Please keep in mind that Figure 7 (and the new Figure 8) compares columns, while Figure 11 compares surface values. In the section on columns the authors now discuss why the AIRS/CrIS total columns are biased high with respect to aircraft columns, which only include reliable observations up to the MLH. Columns can be high while the surface values are low because the retrieved vertical distribution does not match the true vertical distribution, as the reviewer has also noted. So yes, the surface level retrievals are prone to biases, but a large fraction of in situ data is surface data, so it is worthwhile doing the comparisons. The authors also feel that the time series in Figure 10 show little bias in the warm season; in cold weather not only are emissions lower, and therefore NH3 concentrations are closer to the AIRS/CrIS detection limit (0.5 to 1.0 ppbv), but lower temperatures lead to lower thermal contrast and weaker signals..

The study identifies sampling differences between satellite data and in-situ measurements as another significant source of uncertainties and biases, particularly following the application of satellite AVKs. To gain further insights into this aspect, it would be beneficial to conduct tests on the co-location criteria in terms of both spatial and temporal alignment. For instance, reducing the co-location time to 30 minutes could improve the representativeness of the satellite measurements with respect to the in-situ data. Conversely, extending the co-location time would provide a larger statistical dataset for the comparisons, enabling a more robust analysis of uncertainties and biases.
The authors considered this request, but given that GUO2021 had already shown that a one hour and 15 km window was the most appropriate one for Colorado, we decided to adopt that window for Colorado, and for consistency sake, also for California.

Minor comments / typos

- Lines 226-228 and 399-401: I find it unclear whether the retrievals are conducted over cloudy scenes as well. If retrievals are performed over cloudy areas, it could pose a concern since the presence of clouds can impact the baseline temperature of the spectra and the thermal contrast, consequently affecting the retrieved column. These effects should be taken into consideration, as they have the potential to introduce biases and uncertainties in the NH3 retrieval process.

  As stated in the text, MUSES retrieves cloud optical depth in a previous step, and this optical depth is included in the radiative transfer calculations for the subsequent steps, including NH3. The maximum cloud OD for the set of DISCOVER-AQ profiles was 0.25, and in general the cloud OD was less than 0.1. In the Magic Valley there were more cloudy days; in the correlation plot we only included pixels with COD less than 1.0, while in the time series we allowed retrievals with COD up to 2.0, in order to highlight the seasonality.

- Lines 263-264: Why are the systematic errors not evaluated for CrIS?
  See response to first major comment.

- Lines 288-290: I believe truncating AVK matrices in such a manner may not be appropriate. It implies that the influence of the upper layers on the remaining layers is entirely disregarded, consequently affecting the smoothing of the aircraft profiles. A possible workaround could involve complementing the aircraft profiles at the top with additional information, such as model profiles scaled to background NH3 values, and applying the complete AVK matrices. Although this approach would introduce an impact from the profile at the top, it seems more reasonable to me than truncating the AVK matrices entirely.

We no longer truncate the AVK. We have extended the aircraft profiles to the TOA by blending the in corresponding AIRS/CrIS a prior profile above the aircraft. The new versions of Figure 5 and Table 1 reflect this change, but note that the change is small.

- Line 292: Müller
  Corrected.

- Lines 350-351: Why are the observations from CrIS/JPSS-1 not used here? They could have filled the gap of CrIS/SNPP in 2019. And it would have been interesting to check the consistency between these two CrIS instruments for NH3.
  The authors considered the reviewer's suggestion, but after some debate concluded that adding the CRIS/JPSS-1 data would have required some substantial processing and would not have significantly changed the results of the Magic Valley analysis. Comparing SNPP and JPSS NH3 is a task for another paper, which would do so globally.

- Lines 365-361: Could you be more specific on the filters you applied to the retrieved data?
- *Retrievals were checked for quality by ensuring that for all retrievals the root mean square error (RMSE) of the residuals was less than 5.0. The MUSES cloud optical depth (COD) values were also evaluated but since the maximum COD for the retrieved profiles was 0.25 no retrievals were rejected due to large COD. Four CrIS profiles over Colorado were rejected due to very high estimated uncertainties. The DOFs ranged were between 0.8 and 1.1, except for two CrIS profiles over California, four AIRS profiles in California and six AIRS profiles in Colorado, for which the DOFS were smaller (0.2 to 0.7). Given the small number of profiles in each dataset, we did not exclude any profiles based on the DOFs.*

- Figure 2: It might be useful to superimpose the mean and standard deviation of the profiles on these plots, as it was done for Figure 3.
  This was a very useful suggestion: the mean and standard deviation have been added to Figure 2

- Line 437: "*It has been argued*". As is, it sounds a bit weird. Do you refer to a specific study?
  We apologize; this statement was made to us at several conferences, but never in a paper. We have removed the sentence.

- Line 477: What is the typical detection limit of NH3 for these satellite instruments?
  These limits have been added to the text:
  The MUSES total columns are compared against the integrated aircraft columns, using orthogonal linear regression (Figure 7a and 7b); the intercept has been allowed to vary, as both AIRS and CrIS have detection limits, (~1.0 ppbv, for thermal contrast above 5K), as does IASI (3.0e15 molecules/cm2, for thermal contrast above 5K).

- Lines 482-485: I don’t understand what is meant here.
  This section has been extensively rewritten. Please refer back to the new text.

- Line 495: “some poor quality CrIS retrievals”. It is surprizing, since CrIS has much lower instrumental noise compared with AIRS.
  Four CrIS retrievals over Colorado had very large estimated errors, due to large systematic errors (from the water vapor and temperature retrieval steps). We have modified the text to make this clearer.

- Line 535 and line 545: There is likely a full stop punctuation missing at the end of each of these two lines.
  Corrected, thank you.

- Line 559: Could the winter values be underestimated because of the general weaker thermal contrast (measurements closer to the detection limit)?
  Yes, thermal contrast can certainly play a big role. We have rewritten this section as: *At every site, CrIS clearly captures the seasonal cycle, though winter values are usually underestimated: this can be attributed to weak radiative signals due to low temperatures and low thermal contrast. The CrIS level of detectability is normally cited as ~1.0 ppbv (Shephard and Cady-Pereira, 2015), but at low thermal contrast this level increases significantly.*

- Line 559: *“(possibly,”*? There might be a typo here.
  We apologize but could not find the typo referred to here.

- Lines 587-588: “*at the high end of the values reported in the literature*” Please provide references.
  We added a reference to van Damme et al., 2015b, and pointed back to the introduction.

- Figure 10: The shapes/limits of the subplots are not consistent between NH3 data vs. number of dairies. Please consider using the same projection. Also, the values of the NH3 colour bars are hidden.
  Figure 10 is now Figure 11. Color bars are now visible. We have tried to make the projections as close as possible.

- Line 631: delete the full stop punctuation after “measurements”
  The text in this section has been changed significantly and this sentence is no longer present.