Reply to Referee #1

Comment 1:
Optimal estimation without *a priori* information is arguably not optimal estimation. The use of a prior for regularisation is what distinguishes the method from other manners of regularisation, such as smoothness constraints (e.g. Pounder et al. 2012, 10.1175/jamc-d-10-05007.1). I appreciate that the influence of the prior on the final solution is a focus of much critique of the OEM, so the removal of the prior is 'desirable' in one sense but it's an unfair representation of the information available. Why not change the grid but retain the prior, which should then have a minimal impact on the points you currently retain?

The use of priori information has two purposes. One (a) is to make the solution of the inverse problem compliant with the Bayes theorem. The other (b) is to make an unsolvable because ill-posed inverse problem solvable. Our resampling offers an alternative to the use of prior information with respect to (b). We do not claim that our solution provides a maximum a posteriori probability, thus the Bayes theorem is of no concern for us. Instead we provide a maximum likelihood solution which is a fair representation of all measurement information available. The Bayesian data user can easily transform the result into the Bayesian world by calculating the optimal average between the assumed prior and the maximum likelihood result. Conversely, if a non-Bayesian wants to remove the contamination by prior information to obtain a maximum likelihood result, much more complication arises and much more auxiliary data were needed. So technically the data user is helped much more with an a-priori-free representation of the data.

Furthermore, it is not true that the prior has minimal impact on the retained gridpoints. Depending on the assumed a priori variances and on the difference between the maximum likelihood profile and the a priori profile, the impact can be significant. And there are numerous reasons why contamination of the retrieval by prior information is undesirable: (1) Time series of profiles with varying averaging kernels are hard to analyze. (2) Data containing climatological a priori must not be averaged because the the a priori contents in the results to be averaged are not independent. (3) The choice of an adequate prior has its own problems.
For these reasons, we prefer to entirely remove the prior information from the data.

Comment 2:
As you point out on P2L28, it would have been preferable for you to use a different prior. I am not surprised that a single, globally representative profile is a biased prior. Other sources of information are available so why not use them? The reanalyses and forecasts of GFS and ECMWF seem like excellent candidates. If you wish a more climatological prior, there are statistical analyses generating a set of representative

atmospheres, such as §3 of
https://www.ecmwf.int/sites/default/files/elibrary/2008/11040-generation-rttov-regression-coefficients-iasi-and-airs-using-new-profile-training-set-and-new.pdf
(which is, admittedly, a tad obscure). Or use your record of radiosonde launches to make something more locally representative. A globally representative profile makes sense for a moving lidar, but for the permanent installations you discuss, a more specific prior seems appropriate.

Answer:
The reviewer suggests many options to construct a prior. Any of these options will lead to a different result. Prior information is a concept of subjective (as opposed to frequentist) statistics. But the data user does not want to know which prior information we believe in most but what we have measured. Once a stable maximum likelihood solution is available, the user can easily transform it to any Bayesian solution based on her preferred prior.

We agree that the CIRA and US Standard model are not the most accurate models. However, it is necessary to use a consistent *a priori* throughout a climatology or trend analysis study to avoid inducing trends or bias into the results. To make it clear to the reader that this is a matter of choice, we can add a sentence to the conclusion stating that if one decides to use an *a priori* closer to the retrieval the effect may be smaller.

The lidar averaging kernels are so sharply peaked (that is they are at the resolution of the retrieval grid) for most of the profile, that in fact an *isothermal a priori* temperature profile could be used. Hence, for Rayleigh-scatter temperature using ECMWF would have little benefit (in addition to the fact it does not use measurements above 80 km altitude). For water vapour, the retrievals were designed with operational use in mind which requires a minimal number of dependencies in the code as possible, and preferably no need for internet. Additionally, it requires consistency over an entire data set. Therefore, we chose the US standard model over other reanalysis models like ECMWF which would require constant updates.

Comment 3: In my opinion, the primary use of removing the prior from optimal retrievals is for averaging multiple retrievals together. Combining retrievals that contain a prior gives that assumption undue influence on the average and biases Level 3 data. Had you considered this?

Answer:

We agree with you and this point is mentioned on P21L12 of the original manuscript.

Comment 4:
Though you never clearly state so, I can see that adapting the retrieval grid to the information content provides a better representation of what the measurement actually told us. The paper would benefit from an explanation of why you prefer this statistical regridding approach to defining a single inhomogeneous, vertical grid that has decreasing resolution with height. I suspect it's because of dry layers, such as Fig. 5, that 'waste' state vector elements on heights where there is minimal information content available, but explicitly explaining that somewhere before §6 would be helpful. It would also be worth considering how using uneven grids makes it difficult to average retrievals and can confuse inexperienced users.

We agree that a few sentences stating why we consider this useful would be good to add to the paper. One reason we do not choose a single inhomogeneous grid is that if you choose any grid from the beginning, it is not guaranteed that the *a priori*-free retrieval would be stable.
Additionally, as water vapour is highly variable, the spacing would have to change from night to night (or day to day) and would be almost impossible to automate, which is the ultimate goal for our retrievals. Any ad-hoc choice of the retrieval grid might be too dense in one region, which would give rise to instabilities in the retrieval, and might unnecessarily sacrifice resolution at other altitudes.
If the retrievals are conducted by hand and the researcher can gauge an appropriate spacing based on the traditional water vapour calculation, it would be possible but not practical, and would introduce possible biases into the results. An information-based approach is impartial, easily automated, and based on the conditions of the atmosphere. A consistent variable retrieval grid for the temperature retrievals might be more feasible, but could possibly become more problematic on nights with strong gravity waves and could result in unwanted smoothing.

It is true that working with uneven grids is difficult for averaging multiple retrievals due to the variation in vertical resolutions for each retrieval. We have discussed in the conclusion that we do not recommend the reader do this. A grid which is optimal for one atmospheric state will in most cases be close to optimal for a similar atmospheric state. Then the error bars might be a bit different, but the averaging kernels will still be unity and the vertical resolution will be the same, which makes it easier for the user to work with the data over long measurement periods.

Comment 5:

In Figs. 5 and 9 please explain the meaning of the % unit. Is this fractional uncertainty (e.g. uncertainty ÷ value), fraction of the uncertainty (e.g. statistical uncertainty ÷ total uncertainty), or a conversion of the unit of mixing ratio? The presence of a point > 100 % implies the first option, but it then makes little sense why the uncertainty tends to zero with height, as the magnitude of uncertainty should tend towards the value given in the *a priori* covariance matrix. I agree with your arguments at the bottom of P9, but if your prior is controlling the retrieval and uncertainty is tending to zero, I suspect that your *a priori* covariance is far too small at these heights.

This is fractional uncertainty where the absolute uncertainty has been divided by the value. The uncertainty tends to zero with height because the covariance matrix decreases with height. The variance of the water vapour is chosen as 100%, therefore, it decreases with height as the water vapour concentration decreases. The uncertainties are a function of the covariance matrix, therefore, they also decrease with height.

Comment 6:

Throughout the experiments, you argue that the highest retrievals aren't viable. Could you please clearly state the conditions whereby you define a retrieval to be invalid? It appears to be based on the magnitude of uncertainty but it's unclear (a) if you are considering total uncertainty or the statistical component thereof, (b) where your threshold lies, as values from 60–100 % are mentioned, and (c) if the % there means the unit of mixing ratio or a fractional uncertainty.

We agree that this should be clearer throughout the paper. We have used the total fractional uncertainty of 60% as the threshold for where we consider the retrieval to no longer be meaningful and will add a sentence clarifying this to the paper. However, to be clear this is merely our preference; if a researcher thinks a greater or less uncertainty is a better choice that is up to them.

Comment 7:

Figs. 6 and 10 are a somewhat unfair comparison. Considering the full resolution, does the coarse grid retrieval have a lower RMS summed over the entire valid profile? Also, how do you interpolate the radiosonde profile onto the grid of each lidar retrieval? In theory, a fair comparison would use the averaging kernels of each retrieval to make a weighted average of the radiosonde for each profile and compare to that (e.g. Rogers and Connor 2003, 10.1029/2002JD002299).

Originally, we interpolated both the fine grid retrieval and the radiosonde measurements onto the coarse grid. However, we did not state this in the paper which would lead readers to reach your same conclusion.

We agree that when comparing measurements from two instruments, and averaging kernels are available, it is best to use the averaging kernels to weight the other instrument's measurements (in this case the radiosondes). However, doing so in the paper does not illustrate how using a ML retrieval on a coarse grid changes the results compared to the fine grid. Therefore, we would prefer to keep the comparison figures in the paper the same.

However, an analysis comparing the results using the averaging kernels as weighting functions is worth doing, therefore, we have conducted the analysis you suggested by using the fine grid OEM averaging kernels to weight the radiosonde measurements and compared them to the ML coarse grid results. We did not weight the radiosonde measurements to compare to the ML coarse grid because the ML averaging kernel values are unity. We have conducted this analysis for the two case studies as well as for the 8 additional nights and 5 additional days we have added to the paper. However, we will not be adding this discussion to the paper as we do not believe it highlights the differences between the two methods**.**

Figures 6 and 10 have been recalculated here using the fine grid averaging kernels to weight to radiosonde. However, these figures will not be added to the paper and are here to illustrate how using the averaging kernels also decreases the differences between the radiosonde and the lidar measurements.

Weighting the radiosonde with the fine grid averaging kernels decreased the differences in the dry layers, therefore, it is not possible to conclude that the coarse grid will improve the comparisons of the lidar and the radiosonde in these regions and we will remove these statements from the paper as they appear to have been coincidental. Therefore, we see that there is little difference between the coarse and fine grid retrievals, except that the coarse grid retrieval reaches higher altitudes than the fine grid retrieval if we consider the last point on the fine grid as the 0.9 measurement response cutoff height. When comparing the other 5 days we saw similar behavior and no difference within their respective uncertainties between the fine grid and coarse grid retrievals when compared with the radiosonde.

Figure 6: The fractional percent difference between the radiosonde measurements and the lidar measurements. The 1-sigma uncertainties for percent difference are shown as

shaded regions. The fine grid results are shown in blue and the coarse grid results in red. We can now see that there is no significant difference between the two measurements within their respective uncertainties.

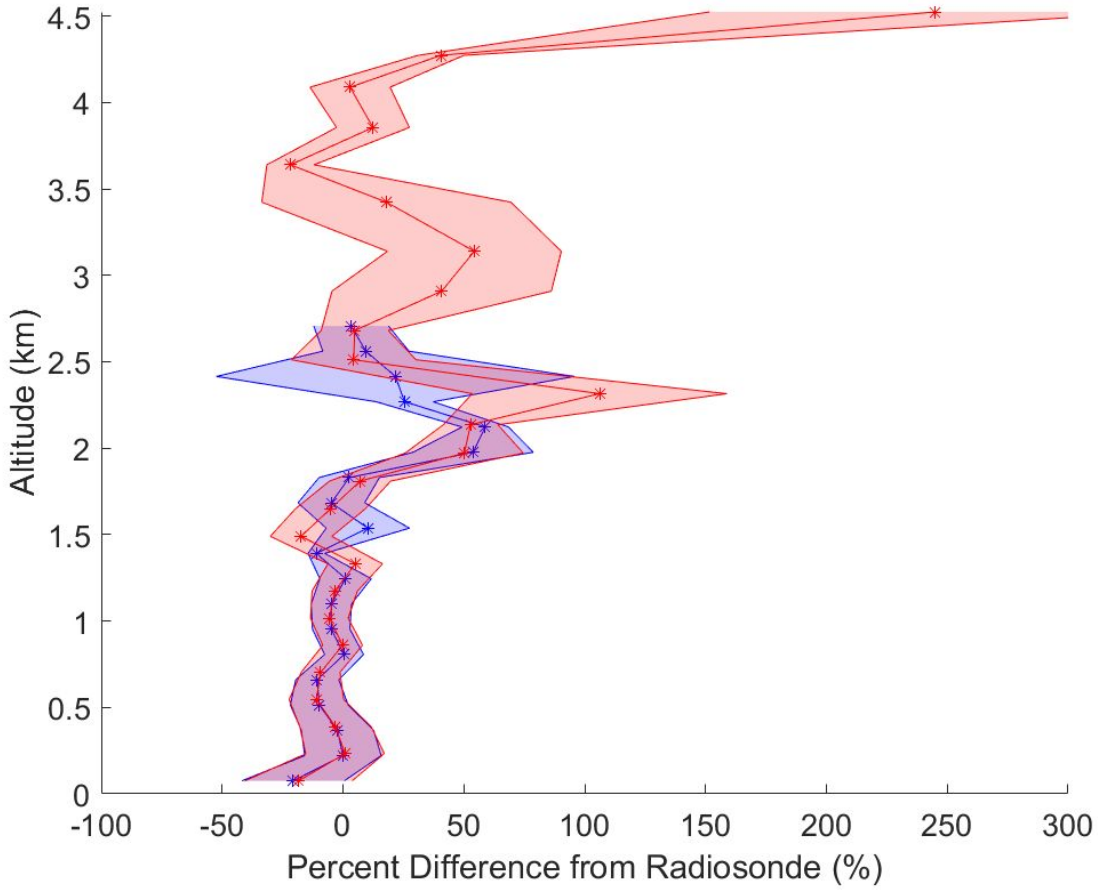
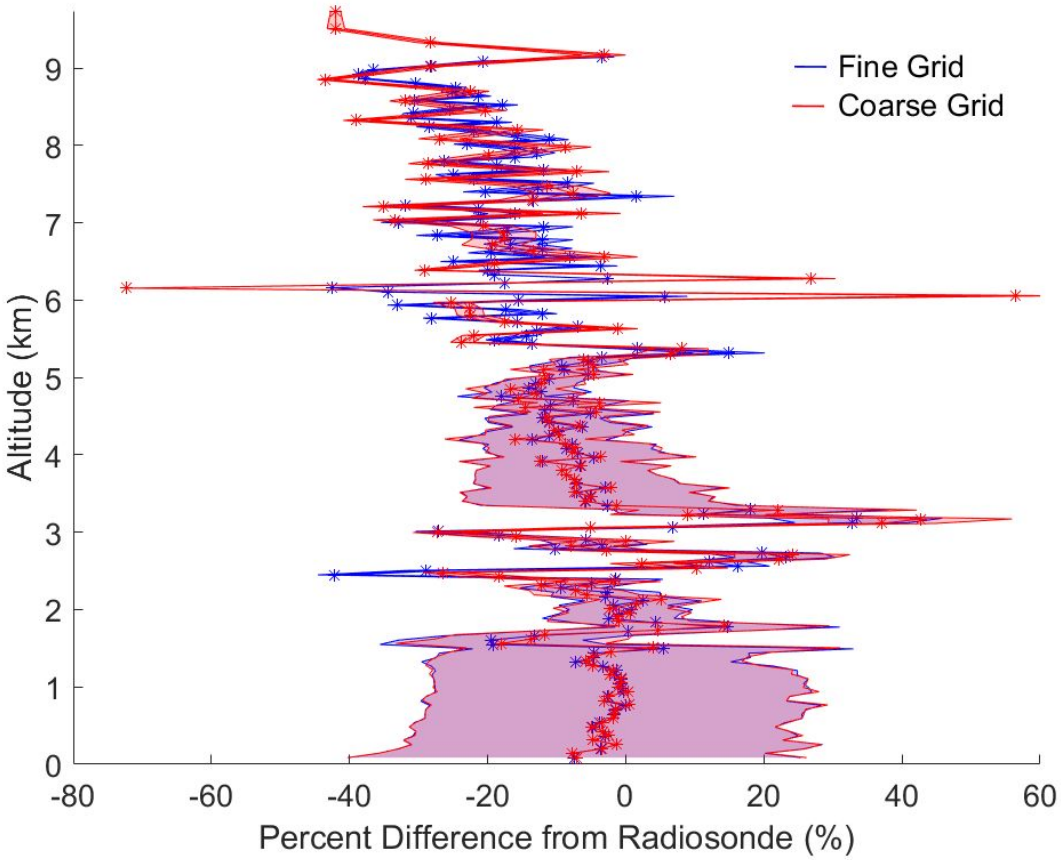


Figure 10: The fractional percent difference between the nighttime fine and coarse grid OEM retrievals with respect to the GRUAN radiosonde on 2013-04-24 00UT. The fine grid OEM results are shown in the blue, and the coarse grid results are shown in red. As with the daytime results, we now see that there is no significant difference between the two methods.



Similarly to the daytime results, the percent difference between the fine and coarse grid retrievals with respect to the radiosonde showed no difference between the two methods within their respective uncertainties (Figure 10). The uncertainties for the nighttime percent differences are more variable than the daytime percent difference uncertainties due to the fact that we used a GRUAN RS92 radiosonde on this night which calculates the uncertainties of the radiosonde as a function of altitude. This was not possible for the RS92 radiosonde used for the daytime results, therefore, uncertainties were assumed constant using the values cited in Dirksen et al. 2014.

Comment 8:

P9L8 The red line first drops below 0.9 at about 2.2 km. I think what you meant to say is the red line *last* drops below 0.9 at 2.7 km.

Yes, that is correct. We will clarify this in the paper.

Comment 9:

P19L24 I strongly disagree with your implication that retrievals using a prior are invalid. The erroneous belief that prior information overwhelms statistical retrievals is too widely held and doesn't need assistance. Your retrievals were likely 'invalid' because of an overly general prior (possibly with unreasonably small uncertainties) on data too noisy to fully constrain the problem at that point.

Unfortunately this seems to be a misunderstanding due to poor wording in the paper. We are not trying to say that OEM retrievals using *a priori* profiles are invalid. That sentence was meant to imply that by using an information-centered method the entire ML coarse grid profile may be considered useful (depending on the uncertainties), whereas at some point in the OEM fine grid profile the *a priori* takes over the retrieval and the measurements do not contribute to the retrieval anymore. Jalali et al. 2018 has studied the point where this occurs and showed that the point where the measurement response last equals 0.9 is a good estimate of this turning point in the fine grid retrieval. We would like to refer you to the signal-to-noise ratio document which suggests that the 0.9 cutoff height is also probably appropriate for the water vapour retrievals as well. The *a priori* covariances were chosen from literature using the NDACC LWG guidelines given by Leblanc et al 2016 when available, therefore we do not believe that they are too constrained and accurately represent the uncertainty of the chosen *a priori* values (Sica and Haefele, 2016).

Comment 10:
P19L26 I also disagree that your paper shows that this removal method 'improves the validity of the retrieval.' You showed one example where the agreement qualitatively improves. On the one hand, you could have potentially achieved the same result through the use of a better prior. On the other hand, it may be that all possible priors are wrong, in which case you need to add another section to this paper.

The word choice in that sentence was extremely unfortunate on our part as it seems to have implied the wrong thing. "Validity" was not the correct word to use in this case. A better wording would have been "the *a priori* removal technique improves the OEM retrievals with respect to the radiosonde and generally increases the cutoff height of the retrieval" which was the original intent of the sentence. The original sentence will be changed accordingly.

It is true that we may have retrieved better or different results by using a better prior, unfortunately, we don't typically have better prior knowledge. For the Rayleigh-scatter measurements it is unusual to have near coincident temperature profiles, and many satellite instruments such as MLS do not get high enough in altitude. We could have

used different *a priori* profiles for the tropospheric water vapour measurements, however, as these retrievals have been designed for operational use, we found it more appropriate to use a smooth and constant *a priori* profile, as was discussed in the answer to your second comment.

It would be preferable to make a statistical analysis of the change in RMS vs. radiosonde across many days of observations, as it is possible that beneficial results shown in the three examples are coincidental. No one ever provides such an analysis during revisions, but that won't stop me from asking.

We agree that is an important study but beyond the scope of this paper, whose main purpose is to demonstrate the effect of the *a priori* profile on the temperature and water vapour retrievals, and show for application of OEM to high-resolution active sounding measurements that, if required, the impact of the *a priori* profile on the retrieval can be minimized at the greatest heights.

Some more minor comments:

P2L14 It took a few minutes to work out that this paragraph was referencing two different figures, one of which isn't in this article. Since this article won't be published until 2019, it may be easier for future readers to distinguish, but it may just be easier to reproduce that figure here.

We apologise that this wasn't clear to the reader. We feel that the sentence on P2L14 is clear in the text, however, we don't make a clear transition to discussing Figure 1 of this paper, therefore, we will clarify the sentence on P2L22 where the figure is first introduced. We don't think we can copy the other figure into this paper and we would prefer not to add in more figures to the paper as we will already be adding at least 3 more to show the additional retrievals we have added to the study.

P2L22 The difference may be smaller than random uncertainties, but is it smaller than the *a priori* uncertainty?
Yes it is smaller. The *a priori* uncertainty (CIRA-86 and US standard) is considered 35 K (Sica and Haefele, 2015).

Eq.6 You omitted the prior term. The equation should read,
$\hat{x} = x_a + A(x - x_a) + G\varepsilon = (I - A)x_a + Ax + G\varepsilon$.

The second form follows Rodgers (3.12), which I find a better illustration of the impact of the averaging kernel. If the averaging kernel is a unit matrix, the first term is clearly zero so the prior has no influence. However, your mileage may vary.

Thank you for noting the error, we have added the prior term to the equation.

P8L2 It isn't 'necessary' to remove the regularization term; it's merely desirable. Perhaps word this sentence, 'We then remove the regularization term in Eq. 4 by choosing an arbitrarily large *a priori* uncertainty.'

We agree that it is not necessary, and can reword the sentence as you requested.

P9L3 I can see why you want Fig. 5 after 3 and 4. Perhaps remove the reference here so the reader doesn't worry they've missed something from the out-of-order mentions?

Yes, we can remove the reference so that it is not confusing.

P9L20 It may be better to say that these retrievals aren't 'useful' or 'meaningful', as 'viable' implies that something about the retrieval process failed. The retrieval still works, it's just that your observations are swamped by noise such that the retrieval isn't telling you anything useful.

We agree that the word viable maybe isn't the best and can change it to "useful" or "meaningful" instead.

P9L22 Why keep telling us only the statistical component of the uncertainty? The other components are important too!

We agree that the systematic uncertainties also are important but usually the statistical uncertainty is the largest one and already above the aforementioned threshold, therefore there was no need mention them. We also discuss all of the uncertainties in the following paragraph.

Fig.4 The caption implies that the reason the points are rejected is their large span. The text implies they are rejected because of their large uncertainty. Please clarify.

Apologies that this was not clear. The points are not included due to their large uncertainties. Technically it is possible to retrieve water vapour values at these

resolutions, therefore they cannot be excluded due to the resolutions. We will clarify this in the paper by changing the sentence to:
“The last two points have vertical resolutions of several hundred meters, but are not considered meaningful points as they have total uncertainties of larger than 60%.”

Fig.6 The fine and coarse grid appear to have the same vertical coordinates (and number of points, which must be wrong).

Yes, you are correct that both the fine grid and the coarse grid cannot have the same grid points. However, the figure is not wrong, because we had interpolated both the fine grid and the radiosonde onto the coarse grid. This has now been changed and we have left the fine grid as it is and now not interpolated it to the coarse grid to better illustrate how the ML coarse grid retrieval changes the results with respect to the fine grid.

P11L12 You mention a ‘traditional method’ which is not shown in Fig. 9.

Apologies, “traditional method” should be removed as it is not shown in the Figure 9. We will remove it.

P12L12 You didn’t show invalid retrievals in Fig. 6 so why show them here?

We can remove them and cut the figure off at 9.7 km so that it is clear that we are not comparing the two methods at altitudes above where we cut the retrievals.

Fig.13 Are you sure ‘systematic’ is the right term here? Optimal estimation models all uncertainties as Gaussian, which would be referred to as random errors. I know the various retrieval parameters are systematic in that they affect all points in the profile equally, but the definition of a systematic error is more like a bias and something that optimal estimation requires care in dealing with.

From “systematic” we mean uncertainties in the retrieval due to the uncertainties of forward model parameters, that is uncertainties whose contribution to the retrieval can not be reduced by averaging like random uncertainties. OEM assumes that the uncertainties given for the covariance matrices are Gaussian in nature, however, the uncertainty budget that is calculated is not random.

P20L14 I would find it more honest to say ‘by 2 km and 600 m in the examples shown.’

This will be changed accordingly and expanded upon to incorporate the results from adding the extra retrievals.

P23 DOIs are an efficient means of locating papers and their inclusion in references makes life easier for readers.

That is true, and we will add the relevant DOIs.

P23L2 The page number of Boersma 2004 is D04311.

Thank you, we will add it in.

Strange though it may seem, I also have several comments on the review from J. Kgaran.

I see no reason to show the reader the value of the cost function through each iteration. The precise path by which a retrieval achieves convergence is of mini- mal interest to anyone, unless the cost surface is very complex (and the author's previous papers provide no reason to suspect that). I also know of no operational retrieval scheme which retains such information. If you mean to compare the costs from the two schemes, that is complicated by the removal of the *a priori* from one retrieval, such that the costs are not directly comparable. (It occurs from their later comments that Kgaran may believe that the change in grid is done after a single iteration of the retrieval. My understanding was that two full sets of iteration are completed. If the authors *are* breaking into the iteration scheme, I agree with Kgaran.)

There is no breaking in the retrieval iteration scheme and we agree with referee 1. Please refer to our comment to Kgaran.

P20L12 I agree that this statement is overly general.

We agree that this statement was overly general. However, now that we have added more nights to the paper, we believe we can make a more appropriate conclusion. We will not discuss improvements with respect to the dry layers because similar results can be achieved simply by weighting the radiosonde using the fine grid averaging kernels. Instead we will discuss the overall increases in maximum retrieval altitudes by using the removal technique.

I hadn't noticed that the Au threshold changed from 0.8 to 0.9 between papers. This change should be made explicit.

We showed in Jalali et. al. 2018 that the 0.8 value for Au in not a good choice and it is mentioned in the conclusion. There was no change between the two papers.

Fig.6 Error bars would indeed be helpful. Perhaps rendered as shading rather than whiskers (as in Fig. 10 of Sica and Haefele, 2015)?

We have added error bars as shading to Figures 6 and 10, as shown in the answer to your previous comment.

The new figures which will be added to the paper are shown here:

Figure 6: The fractional percent difference between the radiosonde measurements and the lidar measurements. The 1-sigma uncertainties for percent difference are shown as shaded regions. The fine grid results are shown in blue and the coarse grid results in red.

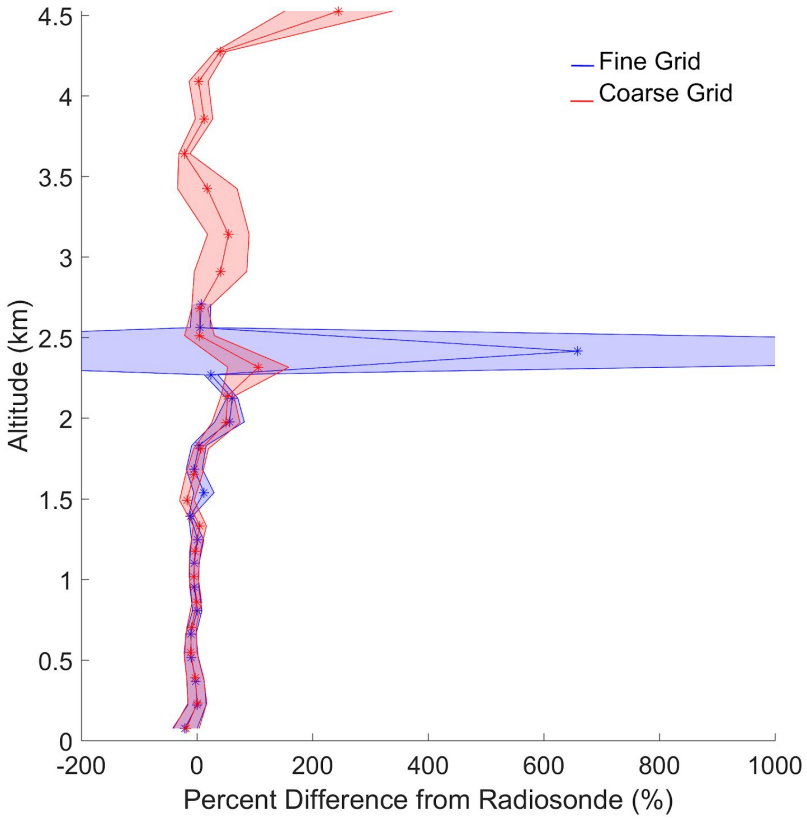
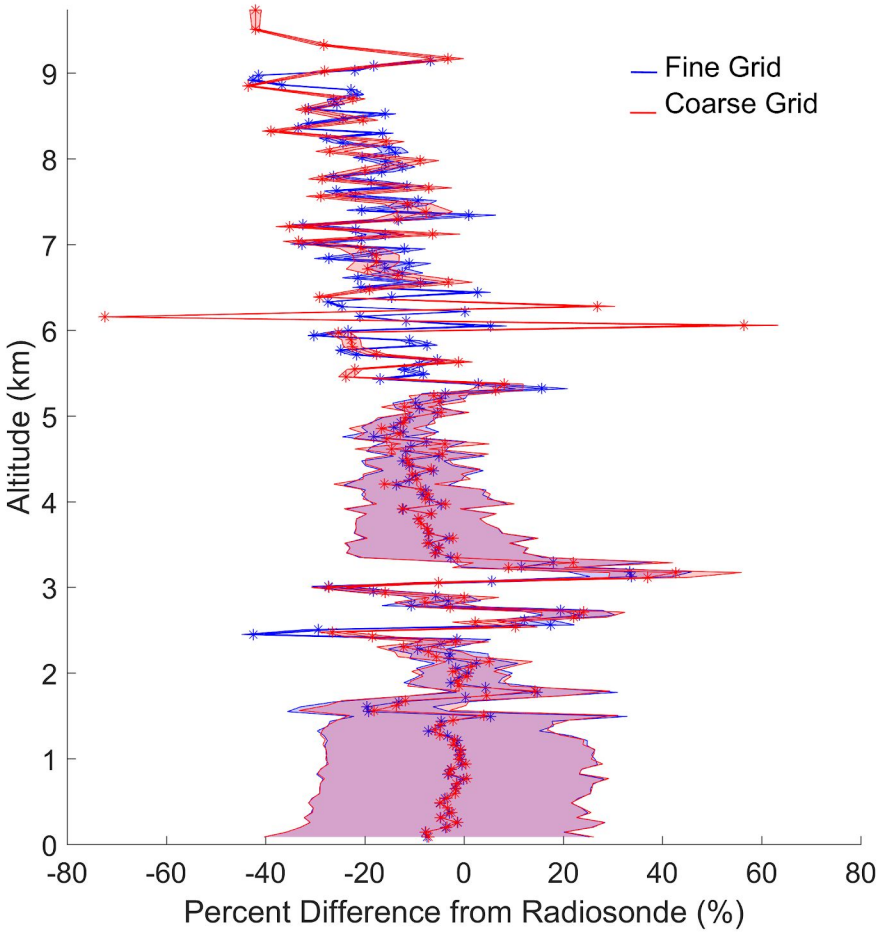


Figure 10: The fractional percent difference between the nighttime fine and coarse grid OEM retrievals with respect to the GRUAN radiosonde on 2013-04-24 00UT. The fine grid OEM results are shown in the blue, and the coarse grid results are shown in red. As with the daytime results, we now see that there is no significant difference between the two methods.



P13L10 Yeah, it's the retrieval that has sensitivity (not the kernel) and, because you used the prior to make the retrieval grid, there will by a small, but non-zero, contribution to the eventual product so you should probably tone those statements down a bit.

Maybe the sentence should be changed from
Fig. 11b shows that at the coarse grid points, the averaging kernel is completely sensitive to the measurements and therefore there is no *a priori* contribution.
TO

Fig. 11b shows that at the coarse grid points, according to the averaging kernel, the temperature retrieval is completely sensitive to the measurements and therefore there is no *a priori* contribution.

I don't agree that the large averaging kernel value is 'artificial'. It's not exact (in that a simultaneous observation would give subtly different values) but it's a fair guess of the information content. In essence, the technique is averaging observations until the SNR is large enough to permit a retrieval. These authors reduce the number of state vector elements to better match the quantity of information available. The idea is discussed (obtusely) in §5.8.3 of Rodgers (2000).

Thank you for your response. We think we have clarified this issue also inside our response to Kgaran which we repeat here:

"It is sometimes possible to see artificial features in the water vapour profile above 12 km at night, or above 8 km during the day due to high noise. However, it has been established that lidar averaging kernels are on the order of unity, and much larger than passive instrument averaging kernels, and having a measurement response of 1 is well known. It is certainly possible for the maximum value of an averaging kernel to be smaller than 1 and still have a trace of 1 as the the information is spread across multiple altitudes and is corroborated by the increase in vertical resolution."

I'm sympathetic with Kgaran that removing the *a priori*, by definition, cannot add information to the retrieval and therefore cannot increase the height of available information. I think that's being overly exact but isn't conceptually wrong.

We think there has been some confusion as to what we mean by increasing final retrieval altitude. It is correct that it is not possible to add information to the retrieval and therefore the entire retrieval cannot go beyond the last point on the fine grid retrieval (which we see it does not in the coarse grid averaging kernels). In this case, we mean that the coarse grid increases the altitude at which we consider the retrieval to be meaningful. We have also clarified this in our response to Kgaran.

P20L21 The uncertainty increases because they've decreased the information available. The resolution increases because it's a fairer representation of the information available. Their previous retrieval was unduly confident as the authors appear to believe there is no unbiased prior. Perhaps by adding a measure of the RMS vs. radiosondes, the authors can more clearly explain the merit of their adjusted data?

This is a good point, but not something we can answer in presenting this method. An individual investigator would have to decide what statistical uncertainty of the lidar measurements were useful relative to any geophysical variations they want to measure. Whether the prior is biased or not is accounted for in the averaging kernels; when they are very close to unity the prior makes little contribution to the retrieval. When they aren't close to unity, this is when the removal technique should take over.

P9L22 I suspect this is the same point I made in the PDF, whereby in English 'increased resolution' means 'more fine detail' but the authors have used it to mean 'greater distance represented by each pixel.' That should be fixed throughout as it's quite confusing.

Yes, thank you for explaining this and for catching the error as well. We have changed the language so that it agrees everywhere.

P19L12 Isn't the dry layer the area where the measured mixing ratio is substantially lower than the *a priori* profile?

Thank you for pointing out that we did not define "dry layers" in our text. While we had originally defined it as less than 1 g/kg of water vapour, but a better definition is using relative humidity. Therefore, we can define a dry layer has one having less than 40% relative humidity, or the point at which humans consider the air to be dry. The dry layer in Figure 5 extends from 1.5 km to 4 km where the radiosonde measured relative humidities that steadily decreased from 25% at 1.5 km to 10% at 4 km. Above 4 km the relative humidity sharply increased to above 50%.

We have defined a dry layer as a region where the water content is below 25% relative humidity.

Regarding supplementary comments:

Thank you for the edits in the supplementary pdf you provided. We have made the suggested changes.