Response to Quentin-Libois

This paper is a follow-up of a previous study (Feng and Huang, 2018) that investigated the feasibility to retrieve stratospheric water vapor profiles above thick convective clouds using infrared hyperspectral observations. The present study explores in details the performances of various retrieval strategies (based on the optimal estimation method) and in particular the added value of using ancillary observations (ice cloud properties from the DARDAR synergistic product and atmospheric profiles in the vicinity of the scene) to retrieve water vapor, temperature, and ice water content profiles. Including vertical information on the distribution of ice near the top of the cloud greatly improves the retrievals compared to the slab-cloud hypothesis where an optically thick cloud is located at a single level in the retrieval algorithm, mostly when the top of the cloud exhibits low ice water contents. Using additional atmospheric information also improves the retrieval compared to using climatological a priori values.

The various retrieval strategies are generally well presented and detailed. The paper is well written and the figures appropriate. The manuscript, rather technical, perfectly fits for AMT journal. However the abstract lacks information and the introduction is probably too short to highlight the originality of the work compared to previous studies. A particular critical point is the use of the term "OSSE" to describe something that is probably more a standard retrieval, although tested on atmospheric fields originating from an NWP model. The multiplicity and relative complexity of the different cases described makes it sometimes difficult to follow the optimal estimation framework, but local rephrasing can certainly improve the readability. The results should probably put in a broader context to help the reader figure out whether the improvements for the various retrievals mentioned are significant, and how they compare to existing methods. This paper deserves being published in AMT, but the authors are encouraged to make their best to make the reading as easy as possible. Various suggestions are made below towards this.

Thank you, Quentin Libois, for your detailed and thoughtful comments. Following your comment, we have greatly expanded the abstract and introduction, added necessary description about the forward model, and expanded the discussion on cloud-induced uncertainties.

Specific comments

1) The abstract is not an abstract so far. An abstract should indeed be a very short summary (very similar to the conclusion actually) of the work. It could be a bit longer, emphasizing for instance that generally sounders reject cloudy pixels (at least in NWP models). It should highlight quantitative results. An abstract should be self-sufficient and provides details : what instruments are used, how much are the performances improved etc. ? The abstract needs rewriting.

Following this comment, we re-wrote the abstract.

2) The term "OSSE" is used, in particular in the title, to define the type of experiment performed in this study. Although NWP model outputs are used to test the retrieval algorithms, the present study is not an OSSE, in the sense that no forecasts are performed with the NWP models. In general, an OSSE implies to compare two NWP forecasts, one of those including the assimilation of observations from a supplementary instrument compared to the other. The gain in forecast skill between both forecasts can then be attributed to the additional use of this new instrument. See for instance Arnold and Dey (1986).

We agree that the term "OSSE" is typically used to describe the weather OSSE that assesses the added value from satellite observations to operational observation systems (e.g., NWP forecast).

However, many studies (such as Feldman et al., 2013, Smith et al., 2020, Crow et al., 2001) also used this term to describe simulation experiments that derive observational signals from numerical weather forecast model or climate model to evaluate the information content carried by the new instruments, without the assimilation of the observations.

Following your suggestion, we use 'simulation experiment' instead.

3) Some descriptions of the retrieval algorithms are not very clear, in particular when it comes to the

details of the optimal estimation method, with the definition of observation vectors, state vectors and covariance errors. In particular Case 5 experiment is not sufficiently clear. It is essential that the combination of the text and equations allows the reader to reproduce the algorithm. The same holds for the definition of DFS relative to each state variable for instance. Likewise, the way atmospheric profiles are sometimes modified for sensitivity studies is not always clear (see examples below).

To avoid confusion, we now refer to Cases 1-4 as 'retrieval strategies', and avoid referring Case 5 as a retrieval. Necessary clarifications on atmospheric profiles for sensitivity studies are added.

4) The study deliberately focuses on high thick clouds to probe UTLS atmospheric properties. However the title seems to be more general and does not exclude other types of clouds, including for instance low clouds or thin ice clouds. To make the paper more general it'd be useful to extend the study to other types of clouds. On the contrary, if the focus is only on high thick clouds it has to be more clearly stated. More generally the evaluation of retrieval algorithms on a single atmospheric simulation is probably hard to support, although this does not question the algorithm itself.

A focus on high, thick clouds is now stated in the abstract, as well as the introduction. We do not expand it to cloudy-sky cases in general, considering: 1) a general application using a similar single-footprint OE framework has been discussed in Irian et al., 2018 and DeSouza-Machado et al., 2018, and 2) the way our radiative transfer model is configured is to neglect the contribution from lower atmosphere.

Technical corrections

l.4: The combined use of OSSE and retrieval is a bit misleading. Although both approaches rely on the optimal estimation, retrieval aims at retrieving optimal parameters, while OSSE generally results in weather forecasts that can be compared to forecasts without the novel instrumentation included.
See response 2.

[l.5: "Detecting" is unclear or at least too qualitative, and different than retrieving (quantitative)

l.12: *and* ice particles]
Abstract has been re-wrote.

[l.17: what do you include in "atmospheric composition"?
l.22: "severe" is qualitative, what does it mean in terms of cloud fraction or optical thickness for instance?
No clear-sky portion in nine (3x3) FOVs (Overcast cloud cover)
l.22: more details should be given on cloud-clearing, and why clouds generally make it difficult to retrieve atmospheric composition. More generally demonstrate more convincingly that what you're proposing has never been done before.
Done
l.24: "large sampling footprint" is unclear. Do you mean vertical resolution? Give some numbers to illustrate that resolution is too coarse, and detail what's the typical vertical resolution that you need to investigate water vapor injection from convective towers
Horizontal resolution. Done.
]
We re-wrote this paragraph to describe the current in-situ and satellite observations in UTLS water vapor. See L19-43.

l.26: clarify "single footprint". What quantity do you refer to for "retrieval"? If it is water vapor, make it clear because so far it's mixed with "atmospheric composition"
The cloud-clearing algorithm uses adjacent nine footprints while the references given in this paragraph uses only one footprint.

l.29: how can the reader know whether these DFS values are much or not?

We re-phrased this sentence. However, whether DFS values are much depends on readers' study of interest, which is the purpose of giving DFS values.

l.34: maybe just say the "actual complexity of clouds"
We re-phrased this sentence

l.37: rephrase this. Suggestion: "Moreover cloud IR emission, that depends on cloud microphysics (or something similar to explain why it's critical to know the details of the clouds), largely contributes to satellite...
We re-phrased this sentence.

l.38: what does this lapse rate look like?
We re-phrased this sentence. Deep convective clouds are typically associated with a cold cloud top, due to radiative and dynamical cooling. It is however warmer within the cloud as a result of latent heat release and radiative heating. Therefore, temperature decreases dramatically near the cloud top.

l.46: precise "*in situ* observations" and clarify with respect to the observations used in Feng and Huang (2018)
To answer your question, the observation used in Feng and Huang 2018 is from Anderson et al., 2012, which is also cited here. Using the in-situ observations carried by aircraft campaigns some limitations in validations of the retrieval. The 'collocation' is never exact. For safety/technical concerns, aircraft samples around a convective system, instead of flying above them. There will always be a certain time/space mismatch.

l.47: more details about OSSE should be provided in the introduction. It's not clear why this cannot be named "sensitivity study based on synthetic observations from NWP model outputs". OSSE generally refers to NWP to estimate the potential impact of a new type of observations on the quality of the forecast. It implies to run a NWP model with data assimilation.
See response 2.

l.56: why studying a tropical cyclone rather than a convective tower?

One purpose of developing the retrieval approach is to detect stratosphere humidification. Tropical cyclone events can create horizontally extensive cloud cover with overshooting tops, which the retrieval method described in the manuscript can be effectively applied and then contrast the overshooting tops with the 'background' to show up the effect of overshooting. For mid-latitude storms,

We stated the focus on high, thick clouds. Compared to a single convective tower in the mid-latitude, a single tropical cyclone event can produce large high, thick cloud cover for several hundreds of kilometers. This is also relevant to your later question, 'L80', so that a single event produces relatively large variability in temperature, water vapor, and cloud fields.

This event particular event is selected because of the availability of collocated aircraft campaigns (ATTERX) and AIRS overpass, which can help further validation of the retrieval algorithm. However, we chose not to mention it in the paper because such validation studies are not performed here.

l.60: again this really looks like a retrieval rather than an OSSE.
See response 2.

l.70: how do the 13 hPa relate to the altitude targeted in the paper?
29.1 km.

l.71: should any temporal spinup or exclusion of outer points of the domain be mentioned here?
Done.

l.76: can you precise whether you handle partial cloud fractions or only cloud fractions of 0 and 1 at each level
The NWP simulates partial cloud fraction and we consider cloud fraction when generating Figure 1(a). However, we find it is not necessary to consider cloud fraction as in this time instances (Figure 1 (a)), most deep convective clouds have a cloud fraction of 1. See 140, the definition of overshooting deep convective clouds and Figure 1(a).

l.79: are these liquid clouds accounted for in the radiative transfer simulations? The information should be put here

No. We add clarification in the text.

l.81: ~~the~~ generates
Done.

l.80: is this snapshot used in Figure 1 only, or it's the only one used in the present study? What's the limitation of building and evaluating a retrieval algorithm on a single atmospheric situation?

It is the only one used in the present study. For evaluating high, thick clouds only, we think using this snapshot is sufficient, as it provides enough samples and generates different combinations of temperature, water vapors profiles. By estimating the standard deviation for every 10 minutes simulation throughout 24 hours, we find this snapshot is representative of the variability of this event.

As tropical cyclone systems are unique in generating an outflow that affects a large spatial domain, other overshooting convections in mid-latitude storms can be different. We will incorporate mid-latitudes storms to generate a similar simulation study for the application of the method on convective storms other than tropical cyclones and evaluating more detaily on the averaging kernel and sampling footprints.

l.81: what do you mean by "suitable"?
See previous response. Descriptions added.

l.83: it is surprising to look at brightness temperature to see the cloud distribution. Looking at condensed water (liquid + ice) paths or cloud fraction would be more standard. Unless you're interested in the vertical distribution here, in which case this should be made clear.

I agree that the water path is typically used to indicate cloud distribution. However, the purpose of showing Figure 1 is to illustrate a hypothetical infrared satellite image that is very commonly used for infering the distribution of clouds, especially overshooting and deep convective clouds, from satellite observations.

l.85: overshooting should be rigorously defined from a cloud perspective

In the original manuscript, we do look into precipitation, cloud fraction, and the vertical extent to identify overshooting from a cloud perspective. It was explicitly stated.

Not all OT-DCCs pass the BT criterion, because the simulated cloud fields near the edge of DCCs can have very low IWC, even though they meet all the criteria described above. The BT criterion is designed to minimize the false positive rate. We chose not to justify the BT criterion here because it is quantitatively examined by the cited previous studies and by an accompanying paper.

l.85: explain the physics behind this brightness temperature difference (BTD), accounting for temperature profile in the UTLS and water vapor and cloud optical properties at these channels. What kind of BTD are expected?
Done.

l.87: "this" criterion is not clear because it was loosely defined.
Done.

Figure 1: remove "of upwelling radiances".
Done.

l.103: clarify what are the pressure levels of MODTRAN, and explain how interpolation from GEM to MODTRAN is performed
Done. Section 2.2 has been expanded. Although we do not perform interpolation in this simulation experiment. Layers from 13.5 hPa to TOA (0.1hPa) are added to radiative transfer model.

l.104: "cloud information" is unclear. Clearly state "optical properties". Clarify how liquid clouds are treated when present in GEM.
Section 2.2 has been expanded.

l.106: what input is taken from GEM for clouds? Ice water content, liquid water content, number concentrations, effective radius? How are clouds specified in MODTRAN? Via optical thickness, single scattering properties? Please clarify how GEM information (vertical profiles) is converted into MOTRAN-required information

IWC at every vertical level is inputted. Particle size distribution, effective radius, roughness, habit are used to build a look-up table for the extinction coefficients, single-scattering albedo, and asymmetry factors, defined over ice water content.

l.109: ok for liquid clouds, but the information probably comes too late (see above). Can you then give the range of optical thicknesses obtained for ice clouds to support this choice
Optical thickness of ice is greater than 100 in a typical DCC sample. Explanation added to Section 2.2

l.111: reference for AIRS should appear earlier on
Done.

l.113: cm^{-1} should not be italic. Holds elsewhere, including figures captions.
Done.

l.113: what's the spectral resolution of AIRS? Is the resolution of MODTRAN sufficient to simulate AIRS radiances?
Yes. The Full Width at Half Maximum of the AIRS spectral response function is around 0.4 to 1 cm^-1. Berk et al., 2017 (Figure 8) validated the line-by-line algorithm of MODTRAN 6 with AIRS spectral response function. It is more convenient to use MODTRAN for calculating cloudy radiances as it includes DISORT. It is also faster than the traditional LBLRTM.

l.114: remove "reference level"
Done.

l.115: does random mean "normal distribution of noise?
Yes.

l.115: would you have any reference to support the uncorrelation between spectral channels? In particular, any spectral shift is expected to alter similarly all channels.
Correlation between spectral channels can be a concern; however, we do not find references/data sources that quantitatively document how spectral channels are correlated. Therefore, synthetic noise is assumed to be uncorrelated and the Se error covariance matrix is assumed to be diagonal.

l.119: is optical thickness conserved when effective radius is varied?
No. Optical thickness is not conserved. See response to 'l.106'. The radiative transfer model is driven by IWC at every level and ice optical properties over IWC. The ice optical properties over IWC are strongly dependent on effective radius.

l.121: this sentence is unclear
Rephrased.

l.123: DARDAR product should have been fully defined at the first occurrence
Done. See introduction.

l.124: clarify whether this DARDAR selection has been done by the authors or in the cited paper
Done.

l.129: does cloud top correspond to the topmost layer of DARDAR? In the legend of Figure 2 it corresponds to 100 hPa. Make sure that this is consistent
Rephrased.

l.132: how is roughness quantified? If solid columns are chose, is the sensitivity to habit investigated later on, as suggested in the introduction?

1) roughness has a negligible impact on the spectral range of interest, as it mainly affects the scattering

angle.

2) Crystal habit indeed has some impacts and can be important. However, we lack the habit fraction dataset, which makes it difficult to assess the impact. In the updated manuscript, we chose to include the radiance uncertainties due to the neglection of habits other than solid columns in the forward model error.

l.134: does this average conserve the whole mass of ice? Is the average performed over a single profile (to make it vertically homogeneous) or across all profiles to have comparable profiles?

mass of ice is always converged because IWC and effective radius is both reported by DARDAR-Cloud product and we use them independently. Average is performed across all profiles.

l.135: clarify whether Re is a profile here

After the revision, Re is not a profile in the retrieval process but it is a profile for generating the synthetic observation. We now add the radiance uncertainties caused by neglecting the vertical variation in Re to the forward model error.

l.138: all this paragraph is unclear, it's not clear how the profiles are built to compute the mean and standard deviation of brightness temperatures. In particular because the vertical extent of various clouds is different.

After revision, we clarify the t,q profiles used, and also add more description of how effective radius affects the radiance spectral in the radiative transfer calculations. Also see previous response that the description of radiative transfer calculation has been expanded.

l.144: what are the ranges of IWC and effective radius resulting in the observed variability?

See revised Figure 2.

l.147: is there a good reason to compute the spectra over the far-infrared, if not used further? This kind of information could be kept for discussion maybe.

The far-infrared spectra is kept for readers' interests. We kept this spectral range in the plot. To avoid confusion, the spectral range used in the retrieval is marked in Figure 4.

Note also that for FORUM you can now cite Palchetti et al. (2020).
Done.

l.147: cm^{-1} should not be italic
Done. For other units as well.

l.154: why not to conserve IWP in this slab assumption? Unit should be $g\ m^{-3}$. Is this IWC commonly observed? What about the value of Re in this slab?

1) Conserving IWP in the slab is a good idea. Conserving the IWP in such a slab can have IWC around 15 g/m^3 (mean value in the DARDAR-cloud dataset for overshooting DCCs). This value is much higher than reality. We choose 1.5 g/m^3 to match the upper limit we found around 14 km for DCCs. The choice of IWC actually does not matter much in our study of interest, because any value greater than 0.2 g/m^3 for 500 m can lead to saturation (optical depth of 10).
2) Yes.
3) Re is 34 micros

l.157: not clear what these residuals mean. I'd say that they highlight the uncertainties due to the slab-cloud assumption.
Done.

l.161: could this tilted bias be mostly related to the penetration depth of radiation, or equivalently to the absorption coefficient of ice (actual emission comes from different effective heights depending on wavenumber)

Yes, exactly. The dependence of extinction coefficients on wavenumbers determines the penetration depth of radiation at different wavenumbers.

l.161: is this statement about RMSE and bias general? In the case you'd just have a spectrally flat bias, bias and RMSE would probably be the same, and would both vanish with debiasing.

It is not flat, because altering IWC and Re affects the optical depth and hence thermal emission, which affects the overall brightness temperature.

Figure 3: probably $R(Re_0, IWC)$ for the second line
Done.

l.170: F/x should probably be derivatives
Done.

l.176: clarify whether K is computed at x_0, or whether it's computed iteratively and the final one is that on the best estimate
Done. It is iteratively computed. Although updating K (or not), except at the beginning two time steps, does not make much difference to the retrieved values.

l.178: isn't Eqn. 5.16 of Rodgers (2000) more appropriate?
Done.

l.183: again, some convincing justification of this uncorrelation would be appreciated
Done.

Figure 4: *at* a spectral resolution
Done.

l.187-188: probably redundant
Done.

l.202: what about the *a priori* on IWC and Re?
IWC uses the same *a priori* dataset as temperature and water vapor. Re is based on the DARDAR-Cloud Datasets. There are no correlation between the two in the covariance matrix.

l.203: it would be interesting to see how the final estimates of Re and IWC differ from the DARDAR *a priori*
Results for Re are added to Table 2 and IWC profiles are added to Figure. We did not show how final estimates of IWC compared to y_{iwc} because it is hard to visualize the difference.

l.205: I don't understand the tentative justification for using log(IWC) in the state vector
Mainly to avoid getting negative values and also because DARDAR outputs their posteriori uncertainty in log scale. Actually, we also run retrieval tests using IWC without log scale by adding a positive constraints on IWC. The final estimates and DFS do not differ much from what you see in this manuscript.

l.210: not clear why you perturb the profiles in this way. Could you simply propagate the observation error as it is done for the radiances? In this case the retrieval might be optimistic, but the associated error would reflect the IWC measurement uncertainty. It depends whether you're looking at retrieved values or final uncertainties.

There are DFS in IWC and Re so we would like to update them. In practice, using it as the first guess for iwc or as 'y_{iwc}' gives similar results (if not identical). We choose to formulate the retrieval this way because we want to use the climatology (a priori) in the x0 and Sa and use observations only in y, to distinguish the source of error.

l.212: microns
Done.

l.229: what's the rationale for this choice of synthetic observation? Does it correspond to a particular satellite? What if cloud conditions have completely changed? How to compute the observation vector for this? Is unity matrix chosen?
This choice is arbitrary.

l.235: why not using y_{atm} as the *a priori* instead, and increasing *a priori* error? This holds for DARDAR observations as well

See previous response and the Turner and Blumberg (2018). In this study, because we do not have information in lower-altitudes from infrared radiances, this framework is especially helpful. The state vector in lower-altitudes sticks to the y_atm, (i.e., reanalysis). We find that use 'y_{atm}' as the observation constraint, rather than *a priori*, greatly improves the convergence when using real observations. There might be other technics to improve the convergence, but I find this framework best describes the source of uncertainty.

l.237: what variables are you focusing on to assess the performance? All of them or just water vapor?
See Table 2.

l.242: this experiment is surprising, you just expect to converge to DARDAR and atmospheric observations don't you?
Yes. In the revision, we avoid referring to case 5 as a retrieval. It is the posterior estimation of the state vector without infrared radiances. We conduct it so that readers can statistically compare it to Case 4 and infer the improvement from infrared radiances.

l.244: this short paragraph is surprising. What's the use? Why does it appear now and not earlier?
For instance, information content has not been defined so far
Deleted.

l.247: water vapor "perturbations" is too vague. Where? At which temporal frequency?
Rephrased it.

l.249: why focusing on slab-cloud assumption only here?
Rephrased it.

l.257: I guess t_{cold} should have been defined in the previous paragraph as it is used in Fig. 4.
Done.

l.257: CO_2 not italic

Done.

l.264: shouldn't the index 'i+1' be removed?
Done.

l.266: not clear whether the cases remain the same if you now only keep radiance observations. Do you remove other observations only when computing A, but do the calculation at the retrieval point estimated with all observations? To be clarified
Clarified in Section 3. Yes, I only remove other observation only when computing A, because I want to show the DFS from radiance observation alone.

l.269: "As expected" is not obvious because Fig. 4 suggests that the slab-cloud assumption does not really impact the information content
Yes, you are right. We rephrased it.

l.275: maybe say "highlights the strong sensitivity"
Done.

l.277: reference to Fig. 4b is not obvious. Do you mean reference to slab-cloud error?
Yes. Clarified.

l.278: here the DFS for IWC with only radiances is considered. Does it prove that there would still be an added value when considering direct observations of IWC through DARDAR?
Yes, exactly.

l.279: can you elaborate on this high sensitivity to low IWC on a physical basis?
Done.

l.281: that DFS for Re is surprisingly low, even though Figs. 3a and 3b suggest it should be less than DFS for IWC
Thank you for pointing it out. You are right. There was a mistake in calculating kernels. Corrected in the revision.

l.288: you probably don't need this recap if the cases have been clearly defined before (and ideally defined when needed, that is not too early)
Deleted.

l.292: "there are some DFS values" is awkwardly
Rephrased.

l.301: the patterns are not obvious. Do you refer to differences between the center of the cyclone and the points in the South-East corner of the domain, or within the cyclone? In general, Figure 6 is not as informative as the others in the way it is used in the text.
Rephrased.

l.305: check syntax
Done.

l.305: where does this 13.5% come from?
exp(-2), the transmittance through a layer with optical depth of 2.

Figure 7: why not showing truth and retrieval for IWC? Not clear what 7c and 7f show: true or retrieved profiles?

Clarified in Figure 7. Because truth, retrieval, and y_iwc are visually similar. It is hard to tell which is which. We added them to Figure 7.

l.311: "broad"

Done.

l.311: moistening is not so clear in Fig. 7b for Case 1
Yes, because of the smearing effect as explained. You can still see over a higher mixing ratio over a 40 hPa vertical range.

l.323: "better resemblance of Truth" is awkward.
Done.
 Improvement from Case 1 to 3 is not obvious in Fig. 6 for water vapor profile
Yes, still because of smearing effect. Including y_atm increases DFS, leading to narrower vertical averaging kernel, and therefore a narrower peak in Case 4.

Figure 5: RMSE missing in title of 5f. Not clear why y_{IWC} is biased. Maybe clarify how it is built, as I was expected some random white noise around the truth.
Thank you for pointing it out. We found that the number generated by a Matlab function was not really random. We have overcome this issue.

l.335: This last sentence is not clear
Rephrased.

l.338: what is the uncertainty range? Clarifying differences between using observations to refine the a priori or to complete radiance observations would help.
We work on the uncertainty range more on an accompanying paper. Please check Figure A2, https://doi.org/10.5194/acp-2021-154

l.340-341: that's not easy to follow. Consider reexplaining how y_{atm} error is set and whether various errors are tested
Done.

l.343: "vertical" variability?
Done.

 l.354: "synergistic". Also check "synergetic" earlier on
Done.

l.372: "best resemble" is awkward
Re-phrased.

l.377: consider removing the citation to an apparently non-existing paper
Removed.

Table 1: in caption, 4 or 5 cases? What does 20 refer to in the DFS? Why are the DFS values similar across cases? Why are DFS values shown here and not in Table 2?

Table 2: in caption, 4 or 5 cases?

1) **Clarified. We now state four 'retrieval' cases, because case 5 is not a retrieval. 2) '20' refers to the number of model layers used to calculate DFS in t and q. 3) DFS is similar because uncertainties caused by the slab-cloud assumption is small. 4) Because it is hard to merge Table 1 and 2 because the limit of page size.**