Review of « An observing system simulation experiment (OSSE)-based assessment of the retrieval of above-cloud temperature and water vapor using infrared hyper-spectrometers », by Jing Feng et al.

General comments

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.

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.

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).

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).

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.

<u>Technical corrections</u>

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.

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

l.12: *and* ice particles

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?

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.

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

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"

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

l.34: maybe just say the "actual complexity of clouds"

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...

l.38: what does this lapse rate look like?

l.46: precise "*in situ* observations" and clarify with respect to the observations used in Feng and Huang (2018) 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.

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

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

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

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

l.76: can you precise whether you handle partial cloud fractions or only cloud fractions of 0 and 1 at each level

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

l.81: ~~the~~ generates

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?

l.81: what do you mean by "suitable"?

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

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

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?

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

Figure 1: remove "of upwelling radiances". Inte~~r~~grated

l.103: clarify what are the pressure levels of MODTRAN, and explain how interpolation from GEM to MODTRAN is performed

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

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 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

l.111: reference for AIRS should appear earlier on

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

l.113: what's the spectral resolution of AIRS? Is the resolution of MODTRAN sufficient to simulate AIRS radiances?

l.114: remove "reference level"

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

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

l.119: is optical thickness conserved when effective radius is varied?

l.121: this sentence is unclear

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

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

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

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

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?

l.135: clarify whether Re is a profile here

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.

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

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

l.147: cm^{-1} should not be italic

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?

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

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)

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.

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

l.170: F/x should probably be derivatives

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

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

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

Figure 4: *at* a spectral resolution

l.187-188: probably redundant

l.202: what about the *a priori* on IWC and Re?

l.203: it would be interesting to see how the final estimates of Re and IWC differ from the DARDAR *a priori*

l.205: I don't understand the tentative justification for using log(IWC) in the state vector

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.

l.212: microns

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?

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

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

l.242: this experiment is surprising, you just expect to converge to DARDAR and atmospheric observations don't you?

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

l.247: water vapor “perturbations” is too vague. Where? At which temporal frequency?

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

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

l.257: CO_2 not italic

l.264: shouldn’t the index ‘i+1’ be removed?

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

l.269: “As expected” is not obvious because Fig. 4 suggests that the slab-cloud assumption does not really impact the information content

l.275: maybe say “highlights the strong sensitivity”

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

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?

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

l.281: that DFS for Re is surprisingly low, even though Figs. 3a and 3b suggest it should be less than DFS for IWC

l.284: maybe gather all information relative to retrieval performance here (see l.237)

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)

l.292: “there are some DFS values” is awkwardly

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.

l.305: check syntax

l.305: where does this 13.5% come from?

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

l.311: “broad”

l.311: moistening is not so clear in Fig. 7b for Case 1

l.323: “better resemblance of Truth” is awkward. Improvement from Case 1 to 3 is not obvious in Fig. 6 for water vapor profile

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.

l.335: This last sentence is not clear

l.338: what is the uncertainty range? Clarifying differences between using observations to refine the *a priori* or to complete radiance observations would help.

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

l.343: “vertical” variability?

l.354: “synergistic”. Also check “synergetic” earlier on

l.372: “best resemble” is awkward

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

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?

<u>References</u>

Arnold Jr, C. P., & Dey, C. H. (1986). Observing-systems simulation experiments: Past, present, and future. *Bulletin of the American Meteorological Society*, *67*(6), 687-695.

Palchetti, L., Brindley, H., Bantges, R., Buehler, S. A., Camy-Peyret, C., Carli, B., ... & Serio, C. (2020). unique far-infrared satellite observations to better understand how Earth radiates energy to space. *Bulletin of the American Meteorological Society*, *1*(aop), 1-52.