We are grateful to the two reviewers' constructive comments. We've revised the manuscript accordingly. The two major revisions are: (1) move section 4.4 to Appendix C. (2) add the error analysis in section 4.4.

Comments from Reviewer#1

The core topic of the manuscript is retrieval of ice water path (IWP) from MHS-type sensors. The only (at least well-known) dataset of this type is MSPSS, but it has been shown to have a large bias and, accordingly, the manuscript treats a very important subject. It can here be mentioned that a similar dataset is also in the process of review. It is denoted as SPARE-ICE, see www.sat.ltu.se. This fact does not make the work of Gong and Wu less important. The opposite, it is a sign on that improvement of these retrievals is urgently needed, as well as that their basic methodology (building an empirical forward model by collocations with CloudSat) is a good idea.

We highly appreciate the reviewer's recognition of the value of our work. We also thank the reviewer for pointing out the effort on SPARE-ICE, which we didn't pay attention to until now. The relevant paper (Holl et al., 2013, under review) has been added in the citation here.

Although the major product (IWP) is the same, we want to point out some fundamental differences between SPARE-ICE and our product:

(1) SPARE-ICE retrieves IWP from combining IR, VIS and MW channels, while our method only uses MW channels. They both have their advantage and caveats. For SPARE-ICE, AVHRR channels see the Earth's surface. Therefore, the retrieval uncertainty tends to increase over complicated surface conditions (e.g., ice, snow, etc.), where both the re-analysis and the most sophisticated radiative transfer models cannot handle well. *Holl et al., 2013* also pointed out this deficiency. Our algorithm thrives all types of surface condition by using 183 GHz channels that barely penetrate to the surface (note that 150 GHz channel is only used over ocean and thick cloud over land). However, our algorithm hasn't been tested over high latitudes, so the validity needs to be further evaluated. Also, the sensitivity of our algorithm is lower than SPARE-ICE due to the different subset of channels we specified.

(2) SPARE-ICE doesn't use 150 GHz channel while we used it over ocean and thick cloud scenes over land. This channel helps on the retrieval of IWP from very thick/precipitating clouds that 183 GHz channels and othe IR channels are already saturated.

(3) SPARE-ICE uses neural network as the "retrieval model" while we used a pure "forward model". The method for cloudy-scene screening is hence also completely different.

(4) Our algorithm is relatively simpler than the method used to produce SPARE-ICE product, and therefore our method is easier to be extended to other MW measurements (e.g., NPP ATMS, GPM) to produce real-time operational IWP product.

Since these two algorithms are completely independent from each other, it will be a good effort to compare these two datasets in the future.

Hence, the scientific importance of the manuscript is accordingly high, but there are shortcomings in both the presentation and efforts performed that require consideration. The main issues are:

1. No error analysis is presented. My opinion is that a retrieval is of little value until an useful error estimate is given. I understand that exact values not can be given for these retrievals, but an estimate must still be provided. The errors in the CloudSat retrievals propagate into this dataset, but there are also additional error sources. One such source is collocation mismatches. Another is that the CloudSat data just cover a part of the MHS footprint and the effect of so-called beam filling is not totally covered by the collocation approach. It would also be useful to have an estimate of the fraction of false cloud defections (more below).
   Some of the error analyses are blended in the main text. For example, the CloudSat error, the additional error that is associated with mis-match and inhomogeneity within one MHS footprint (i.e., beam filling effect). We thank the reviewer for pointing out the lack of analysis of the retrieval error. Now the error analysis is presented in a separate section, including the retrieval error analysis.

2. The scope is unclear. Will the dataset be made public available? Can it be extended to higher latitudes? It is not clearly defined how IWP is here defined. What is the exact lower limit? And with respect to readers with a model background, is IWP here including both "cloud ice" and "precipitating ice" (snow)?
   Yes, this dataset is publicly available under request. As a matter of fact, we believe that the relative simple algorithm which is described step by step in the main text makes it easy to follow and to build the code by the reader from scratch.

   So far the algorithm is only valid to 30 N/S. This has been described and validated in the main text. We have some preliminary results for higher latitudes, which is a slight alternation of the presented one. The entire algorithm will be made publicly available in the near future, which should be valid up to 70 N/S.

   Although technically speaking, 183 GHz channels barely have a weight down to the surface, meaning it can only detect partial column IWP. However, our method using CloudSat to “extrapolate” the capability of 183 GHz channels to the total column IWP. This point is now emphasized in PXXLXX.

   Satellite cannot tell you whether the ice is “cloud ice” or “precipitating ice” (snow or graupel). We compared our retrieved PDF with a WRF model run (3-km resolution with cumulus parameterization being turned off), and the comparison results suggest that the IWP retrieved from our algorithm is mostly from snow, as the detection threshold is relatively large in our algorithm. For more details, please see: http://atmospheres.gsfc.nasa.gov/climate/science/slides.php?sciid=38

3. The comparison between the two radiative transfer models provides no useful information and should be removed. In fact, as it is presented, it rather gives an incorrect view of the performance of forward model (more below).
This entire section about problems of radiative transfer models is not meaningless or irrelevant. Rather, it points out the common issue of the state-of-the-art forward models at high-frequency MW channels (frequency > 89 GHz) in simulating ice cloud/frozen precipitation. As ice scattering dominates the high-frequency MW channels, they are more and more popular to be implemented in current and future missions to detect ice clouds/snow precipitation (e.g., Aura MLS, GPM GMI, NPP ATMS, and MetOp series). Therefore, the lack of capability of simulating ice cloud/precip by forward models, especially those are operationally used, is detrimental to use these observations.

   The two forward models discussed in this paper (CRTM and CRM) are the center-pieces of the NCEP/NOAA assimilation system and delivery of Aura MLS retrieval products, respectively. Many other works have touched base on the difficulties in assimilating MLS data to deliver the operational forecast (see references added in the revised manuscript). Therefore, it is of great meaning to point out the issues of these two models, and to give out a quick and easy way to solve the problem by swapping the portion of those channels with our model.

   Having said that, we realize and acknowledge the reviewer's opinion that this section is not tightly associated with the main content. Rather, it can be treated as an "implement" of our model. Therefore, we move this section to the appendix in the revised manuscript.

4. Essential information is lacking in several places. On the other hand, quite a lot of irrelevant information is found. Examples are given below. The manuscript could be made considerably shorter.
In summary, the main changes I suggest is to add an error analysis, while removing the radiative transfer comparison.

   Suggestion acknowledged. Now the error analysis is included as a separate sub-section, and the radiative transfer comparison has been moved to the appendix.

   Detailed comments:

   P8188L9: Unclear sentence. In addition, I would suggest to use "empirical relationship" instead of "forward model". Yes, a forward model can be empirical, but "forward model" gives the impression of something more advanced than the very simple relationships derived here.

   Now the wording is changed to "empirical forward model" here and elsewhere.

   P8188L15: I found this comment highly misleading. There exist several forward models that can treat all relevant aspects of the simulation problem. The issue is instead big uncertainties in the input to the forward model, mainly what single scattering properties to apply. This distinction must be made clear, here, in the abstract and in other places.

The word “operational” is added in front of the “current radiative transfer models”. We agree with the reviewer that the main problem comes from the lack of knowledge of the “input cloud”. Unfortunately, this is always the case for the real situation. In other words, one can never have perfect information of a real cloud. On the other hand, the “operational RTMs” use coarse vertical resolution, two-stream or other assumptions to speed up the level-by-level radiative transfer calculation, so part of the large uncertainty also arises from the RTM itself.

We now emphasize in the appendix that we focus on discussing the possible improvement for “operational RTMs”。

P8188L20: In LaTeX terminology \citep should be used for Stephens et al. This error is repeated at other places.

Suggestion adopted.

P8190L13: This is not generally true, only valid for relatively low Tcir, as made clear by Eq 2. Or if true, only a small part of the possible range of Tcir is actually used, and the IWP-range of the retrievals is low, which is a drawback.

We tend to not agree with the reviewer at this point. The Tcir ranges in these MW channels are indeed very large, especially for 157 and 190 GHz channels. As also shown in Fig. 3 for all the collocated cases at the tropics, Tcir is far from reaching its saturation value. That means cloud IWP can hardly reach a 10 kg/m2 value in reality, as also shown in the CloudSat PDF (Fig. 6). Therefore, a linear relationship is in general true for cloud IWP and Tcir at these microwave channels.

P8190L18: It is reversed, a RTM or forward model relates radiative properties to radiance.

Suggestion adopted.

P8190L19: Is there any empirically RTM?

Now “RTM” is changed to “models”.

P8190L20: This comment is irrelevant.

It’s not irrelevant. RTMs always perform better under “clear-sky” situation than under “cloudy-sky” condition. It’s dangerous to trust a RTM under “cloudy-sky”, and that’s the value of empirical model here.

P8190L22: Please, distinguish between forward model input and simplifications made for efficiency reasons. And that there exist forward models without any such simplifications.

The current language shows no intention to say that every RTM uses simplifications.

P8192L24: Another missing \citep.

Suggestion adopted.

P8193L5: More common is to say that this is valid in the Rayleigh limit. Anyhow, the expression "Mie scattering" is vague, used differently in different contexts.

Doesn't ice particle scattering fall into "Mie scattering" range at these frequencies? That's my understanding.

P8194P5: Several comments here are not very relevant as those instruments not used.

These comments are listed for the purpose of not using CH#3.

P8195L14: Why making an interpolation of CloudSat's IWC. Both IWP and pIWP can be calculated with the original data.

It's simply because it is easy to compare the cloud top height (ht) and categorize clouds into different groups according to the cloud top height, as it's the dominant factor to alter the Tcir-IWP relationship in the tropics.

P8195L16: This comment indicates that retrievals are tested for different pIWP, but this is never done. Or do I miss something? Anyhow, please clearly define the IWP product that you output. A comment here is that setting a lower altitude limit for the IWP is not ideal. A more common approach is to specify the lower limit as a temperature. This as the main IWP retrieval problem for CloudSat is to discriminate between liquid and ice particles. The limit between liquid and ice has rather a temperature threshold, than an altitude one.

You are absolutely correct. In the tropics, there's basically no ice below the freezing level (~ 5km) in CloudSat data due to the factor that the retrieval is purely temperature based. On the other hand, MW channels cannot penetrate down to the surface in the thick ice cloud case. Therefore, we tested to start IWC integration from the surface, and from 5 km, and the Tcir-IWP relationships are almost the same, which makes us confident that the MHS channels are indeed measuring a column-wised IWP even the cloud is thick, as long as CloudSat is treated as the "truth". This is generally true in the tropics, as ice below the freezing level is snow/graupel.

P8195L27: CloudSat can be "truth" for creating the empirical relationships, but its errors must still be considered when making an error assessment.

Yes. This point is now mentioned in section 4.4, where the error is defined as CloudSat error + retrieval error from our algorithm + mismatch error.

P8195L28: Comment not totally logical. Yes, microwaves penetrate deeper and should be better for IWP retrievals, but you can apply CloudSat collocations also for IR/VIS (as done in SPARE-ICE).

Analogous way can surely be applied to IR/VIS channels. However, as they don't have the capability to see through a thick ice cloud, I really doubt if a large IWP value can ever

be retrieved from IR/VIS channel only as the information is “invisible” to IR/VIS channels. In my understanding, SPARE-ICE integrates IR, VIS and MW channels (correct me if I’m wrong).

Eq1: Explain the meaning of CIR and CCR. This makes it easier to remember what e.g. Tccr stands for.

Thanks for the suggestion. It’s now added in the text.

P8196L15: "sate" should be "state".

Suggestion adopted.

P8196L20: How is "significant" defined here? Or why not just give a value of the maximum difference?

Statistically significant. Only the PDFs of Tccr are compared, and student t-test was applied to test whether the PDFs are significantly deviate from one another assuming they obey normal distribution. Now the wording is altered to “no statistically significant difference of Tccr distribution…”

P8197L8: What is a "clear-sky surface"?

A slash is missing here: clear-sky/surface. Thanks for pointing it out.

P8197L28: Specify if 5K is one standard deviation or something else.

One standard deviation.

P8198L6: Should be 157 GHz here.

Thanks! It’s a typo.

P8198L20-29: Several unclear points here (but anyhow removed if the forward model comparison is removed.

This paragraph is summarized into one sentence now, and the content is partially moved to the appendix.

P8199L18-22: Irrelevant comments.

The sentence is removed.

P8199L25: Repetition of information.

It’s never been stated before that NOAA-18 and CloudSat have the closest Local Solar Time at the equator.

P8200L1: 60 000 samples, does that mean 60000 MHS footprints? Or a lower value of

footprints, where each footprint is connected to several CloudSat profiles? That is, it would be helpful to describe the complete collocation approach here. Are single CloudSat profile used, or averages? If averages, are all these CloudSat profiles inside 10 km? Or suffice that one CloudSat profile is inside 10 km?

Yes, MHS footprint. It's now clarified in the text. The collocation criteria are first applied, and then we average all CloudSat profiles that fall into the same MHS footprint to get the CloudSat IWP and ht. Averaging process is described later in the text.

P8200L1: The latitude range is only given inside parenthesis, despite it is crucial information. Part of the actual scope of the work.

The latitude range is emphasized several times in the main text and in the figure captions.

P8200L25: A repetition of lambda^4 relationship! Apparently, the Rayleigh limit is discussed here. In this limit, the particle extinction is proportional to radius^6,, while the particle mass (IWP is a mass measure) is proportional to radius^3. Thus, the first part of the sentence is incorrect.

Please see first paragraph on PP51 of Wu and Jiang (2004). I think this ATBD is downloadable from the internet, but if not, please ask me for a copy. Thanks.

P8200L21: Why this threshold for defining ht? It is probably reasonable, but a motivation is needed. In fact, a small sensitivity analysis of the threshold would be highly interesting. And a general comment: In abstract and conclusions, it should be made clear that ht is not the "radiative" top altitude, but a more instrument specific altitude, more related to mass.

The threshold is the highest level that IWC >=10 g/m3. Wu and Jiang (2004) did a bunch of sensitivity tests to sort out the dominant and secondary factors on altering the Tcir-IWP relationship, and they found out that ht (lapse rate) was the primary factor in the tropics (mid-latitude). These experiments were all conducted with the MLS RTM. Now for the first time that we demonstrate that these are true from purely observational evidences (we have also tested mid-latitude for the impact from the lapse rate).

Now the definition of ht is included in the 5th paragraph of section 3.1.

P8201L10: Incomplete sentence.

"which is" is added to lead the subordinate clause.

P8201L24: Does "linear regression" here mean assuming a direct linear relationship between IWP and Tcir?

Yes.

P8202L8: Remaining 52%, at intermediate levels or outside?

Most are at intermediate levels.

P8202L26: The comment here indicates that CRM only can treat limb sounding geometry. If true, why comparing CRTM with a model with such a strong limitation? If false, this part must be made clearer.

No, CRM can calculate radiance from nadir to limb geometry as long as you specify the profile along the LOS.

P8203L12: I can not find any error bar for "without height separation".

That's no necessary as the spread of the PDF can be clearly seen from Fig. 3.

Eq5: Don't see the point in citing Livesey et al. here, as they did not add anything to what is found in Rodgers (2000). Write out the part that is here denoted as Sx (Sx is here just recombination of the other matrices, I got confused as Sx is frequently used for the matrix here denoted as Sa). Seems unnecessary to be needed to download Livesey et al. to understand the equation. Write out that this is just the Gauss-Newton version of OEM.

We basically follow Livesey et al. [2006] for the retrieval procedure, so we tend to keep the current symbols.

I don't quite understand your question: Sx is the covariance of the input variable, while Sa is the covariance matrix of the apriori.

P8205L2: First of all, I don't follow this. Anyhow, my interpretation of the text is that x(q) is always equal to a. If correct why include this term in Eq5 at all, as it is then always zero?

You are correct. Eqn. (5) is the general format, and Sa does matter as it is a component of Sx (see Eqn. 7).

P8205L13: This is a common misunderstanding about Sa. This matrix describes only the a priori uncertainty, it does not control the step length, nor sets a limit on the final solution (all solutions has probability>0 and can occur, especially if the input radiances are corrupt in some way).

I don't quite understand your statement: when shrinking Sa to [1, 1], it does slow down the convergence speed. Also, what do you mean by "corrupted radiance"? The radiance was pre-selected to remove those outliers.

Sec 4.1: It is indirectly described that the IWP PDF depends on horizontal averaging, would be good to express this more clearly. Is the same averaging of CloudSat used here as applied in the collocation process? It seems not, why change?

Yes, the same averaging, except that for the collocation process, whatever CloudSat footprints that fall into the same MHS FOV criteria are averaged together, so the total CloudSat footprint number ranges from 1 to 18 or so for the averaging. To compare with

the CloudSat PDF, we cluster every 15 CloudSat footprints along the orbit to get an average value.

P8207L9: <10%, refers to change in individual average IWP or in PDF values.

In PDF values. It's clarified in the text now.

P8208L25: This aliasing is not clear to me. Please explain. One significant aliasing effects should be diurnal variations. I would say that the "spottiness" is mainly due a non-sufficient sampling of non-Gaussian statistics.

For polar orbiting satellite, there is always a westward drift of the orbit from the previous one. If the instrument across-track swath width is too narrow to partially cover each other between two consecutive orbits, the westward moving clouds may coincide with the gap, or with every orbit, leading to a falsely low or high "cloud occurring frequency". This occurs frequently in the tropics due to the dominant easterly wind.

Diurnal variation is not likely to be the cause, as the equator local passing time is always the same everywhere for CloudSat.

P8209L14: Surface emissivity probably a more important issue.

Agreed. Added.

Sec 4.2: My guess is that a large part of the difference to CloudSat is due to "false detections", at least over oceans. See eg. higher values east of Australia for MHS. This section should discuss this possibility. And somewhere an estimate of false detection rate is needed.

Sec 4:3: Why a special study for +-(25-30) deg? This is a relatively small "extrapolation" from the -25to+25deg used for setting up the retrievals. The basic question is if the methodology can be extended to give global coverage (are there sufficient collocations)?

As we briefly discussed in the discussion section, that temperature lapse rate should become the primary factor of altering the Tcir-IWP relationship. Where does the separation begin? Where does the current algorithm lose its validity? This is not a simple "extrapolation" problem.

The number of collocation samples increase with latitude, so they are sufficient for us to explore the mid-latitude situation, as we are currently working on.

P8211L1: Did Wu and Jiang study this for limb or downward looking observations? If limb, the comment is not relevant.

See my answer to question P8202L26.

Sec 4.4: A meaningful comparison between the models requires likely a complete study in itself. The differences are significant and many tests are needed to understand the

source the discrepancy. Anyhow, many aspects are here unclear, such as what simplifications do the models use, are both models actually using the same micro-physical assumptions and can some difference originate in different parameterisation of water vapour absorption? If the motivation is to plan for model development, why not compare to a validated model where the simplifications are kept as low as possible? Such a model is probably slow, but this amount of testing can anyhow easily be performed.

As suggested by the reviewer, now the model comparison part is moved to the appendix. The main motivation is not to dig out the under-lying reason of the discrepancy of RTMs, but to point out that main-stream operation RTMs have strong biases at these channels, and one easy and fast solution is to swap the look-up-table inside the RTMs with the relationship computed here. This is especially efficient for CRTM, while for MLS RTM, we just suggest that uncertainties increase with ice cloud thickness during the retrieval, which was not realized before.

Sec 5: The conclusions should be revised, considering comments above.

Revised accordingly.

P8215L14: This is just the "zenith angle", not the solar one. Same issue in text of Fig A1.

Corrected.

P8215L19: Took me a long time to figure out that the value 2.1 comes from the figure.

P8216L13: The term "limb darkening" is probably not known to everybody. What is Tside?

Tside is T from off-nadir view. This subscript is described in Appendix A.

P8216L15: I would guess that surface emissivity assumed is the main cause to the bias.

Suggestion adopted.