Interactive comment on "Evaluation of cloud fraction and its radiative effect simulated by IPCC AR4 global models against ARM surface observations" by Y. Qian et al.

Anonymous Referee #1

Received and published: 27 June 2011

General comments

This manuscript presents cloud fraction estimates from ARM surface observations and uses these to evaluate the monthly and annual cloud fraction, as well as the shortwave radiation at the surface, predicted by the AR4 global models. Doing this evaluation, for three point locations that have very different patterns of cloudiness, is worthy of publication and an important step towards making use of the wealth of observations available to learn about errors made in global models. There is a rich amount of results in the manuscript that could be very valuable for both the modeling and observation community. However, the manuscript misses to clearly outline important details of the methods used that may significantly impact the evaluation and conclusions reached. It also suffers from not being organized in a concise manner, and therefore fails to bring a clear message across. I recommend major revision based on the below comments.

1. Except for one paragraph in the last section, the paper does not describe how the cloud fraction profiles from the observations are constructed. The comparison of cloud fraction profiles from the models and the observations however is not at all trivial, and if the authors have thought carefully about the impact of vertical grid and temporal grid, they should discuss this explicitly earlier in the paper. In doing such a comparison, the observations should be mapped onto the same vertical grid as used by the models, at least. Otherwise it appears one cannot use Figures 9 - 14 to support statements in the text regarding whether models producing more, or less clouds, than observed at a given altitude. It is highly recommended to re-do these profiles as to allow a fairer comparison (which would greatly increase the value of the manuscript), or, the authors should very clearly highlight which pieces of information can in fact be derived from their comparison, given all the uncertainties discussed in the last section, and which are unique and new compared to previous model-satellite comparisons of cloudiness.

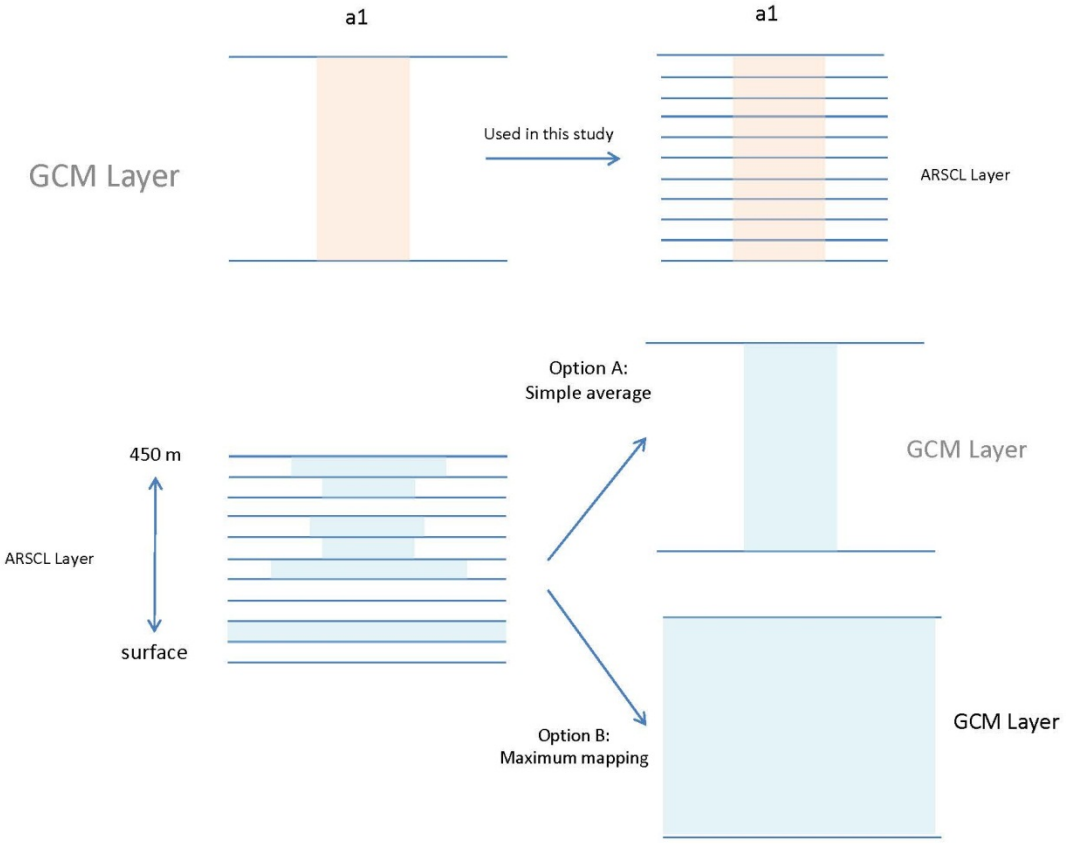
(1) Cloud fraction profiles

First of all, we need to clarify that we did vertical interpolation/mapping of the CF from each GCM into the ARSCL vertical grid in our original manuscript, so the comparisons presented in Figures 9-14 are based on finer ARSCL vertical grid. In the original manuscript, we discussed about uncertainty associated with the layer thickness of ARSCL data, but that was not meant for the CF that was not compared at the same vertical grid layer. We apologize for the confusing of the description in the original manuscript.

Since CF in GCMs is assumed vertically constant within each grid layer, in this study we evenly distribute the CF of each GCM layer (hundreds of meters thick) into the much finer ARSCL

layers (45 m thick), which is illustrated in the top panel of the attached figure below. There are two reasons why we mapped the CF from model vertical grids to ARSCL grids. First, if we map the CF from ARSCL to the model, we need to determine which GCM vertical grids should be used to represent all GCMs since the vertical resolutions of the GCMs are different. Otherwise, we would have to do a separate mapping for each of the GCMs. Moreover, mapping the CF from finer (ARSCL) to coarser (model) vertical grid may smooth out some meaningful vertical variability of CF.

Second, which is actually a more important one, is about how we vertically map the CF. The bottom panel of the figure below demonstrates two, out of many, different vertical mapping approaches. Assuming that the depth of one GCM layer is 450 m, spanning over 10 45-m ARSCL uniform layers, CF may vary in the 10 ARSCL layers. To derive the CF in that GCM layer, one can take the option A to make a simple average, i.e. (c1+c2+…+c10)/10, or option B to take the maximum CF (100% in this case) in the 10 ARSCL layers, which makes sense sometimes because a thick layer will be horizontally fully covered by cloud even if just a thin layer embedded within that thick layer is horizontally fully covered by cloud. Because the CF is assumed to be vertically constant within each GCM grid layer, option B apparently overestimates the mean CF for the GCM, especially for the purpose of calculating radiation transfer. We



believe that option A is more reasonable when mapping the observational ARSCL CF into GCM layers. The result of the option A (simple average) is similar to that using the approach currently applied in this study (i.e., evenly map the CF from coarse GCM to fine ARSCL vertical grid).

Therefore we believe that the comparison of model and observational CF currently presented is the "best" that we can do within the framework/scope of this study (i.e. we are not able to change the original vertical resolutions of either models or ARSCL observation).

We have added one new subsection (2.2.d) at the end of Section 2.2 rather than in the Conclusion and Discussion part to explain how the model CFs are vertically interpolated/mapped to the same vertical grids as observations in this study.

(2) Unique and new aspects compared to previous model-satellite comparisons of cloudiness

Due to their global nature, cloud properties and radiation budget in climate models have primarily been evaluated with satellite data from the space. However, in reality models need to produce correct radiative balances at both the top and bottom of the atmosphere. Studies by Wild and others cited in the manuscript have indicated that most models cannot accurately produce surface radiation balance, thus evaluation of model produced clouds and radiation from a surface viewpoint to complement the satellite viewpoint is clearly a useful concept. Both ground and satellite-based datasets have difficulty with different cloud types and thus a more comprehensive evaluation is also useful from that perspective.

We have modified the first paragraph of the introduction to include references to the fact that GCMs are more capable to represent the TOA than the surface radiation budget.

We have also added the following sentence to the introduction: *"A review by Wild (2008) indicates that the inter-model range in TOA SW flux is only 4% of its absolute value while the inter-model range in surface SW flux is 14% of its absolute value. This result is likely due to the relatively better availability of global satellite versus surface data and the adjustment of model cloud parameterizations to get agreement with global mean satellite observations. Thus, information on the relationships between clouds and surface fluxes is needed to further constrain the model parameterizations so that the correct radiation budget is obtained at both the top and bottom of the atmosphere. "*

2. The paper presents detailed background information on the prevalent cloudiness patterns (and the dominating dynamics) at each of the three sites (in section 5.1, 5.2, 5.3). It is recommended that this information is presented earlier and connected to figures that present the observations. Knowing the different cloud patterns, why can one expect one measurement of cloudiness being better than another, at a given site? If three different observations provide very different (or very similar) estimates of cloudiness, does that relate to the specific clouds being present or their variability, and based on that, what would one like to evaluate from the model? Presenting and discussing the differences in cloudiness at these three locations, and their impact on the accuracy of the measurements, could be used better to motivate the evaluation of say, a seasonal cycle, or a vertical profile of cloud fraction. For this it would be very good to add the results for SGP and NSA to Figures 2, 4 and 5 and bring out the (dis)similarities.

(1) Paper reorganization

The background information about cloudiness pattern over three sites scattered in sections 4.1, 4.2, 5.1, 5.2, 5.3 have been integrated and moved to Section 3 (3.1, 3.2 and 3.3) where the observations are first presented. We have also moved some background information or motivation to the introduction part to explain why it is useful to look at these 3 particular sites with ARM data. We also removed some sentences/phrases that seemed repetitive or caused confusion in the original version of manuscript. Now we have reorganized Section 3 into three subsections presenting the results for Manus, SGP and NSA, respectively. We believe that those changes have made the manuscript much clearer and more concise.

(2) Adding the results and discussions for SGP and NSA to Figures 2, 4 and 5

We have added the results for SGP and NSA to Figures 2, 4 and 5 and brought up the (dis)similarities for discussion.

It is highly desirable to see if the instruments can perform better or if GCMs can do a better job for a specific cloud type or at one of three sites. After comparing the observations and model performance at three sites, we didn't find a super site with a specific cloudiness feature that is best observed by instruments or simulated by the model. First, the inter-instrument differences are similar at both daily time scale (see new figures 2 and 4) and multi-year monthly time scale (figure 3) at three sites. Second, although the inter-model deviation and model-measurement differences vary across three sites, it is hard to compare the model performance for a specific cloud type among different sites since one type of cloud (e.g., cirrus) may exist at one site (e.g., Manus) but not at other sites. Another example, the GCMs perform better at SGP than at the other two sites in simulating the seasonal variation and probability distribution of monthly TCF. However, the models remarkably underpredict the TCF and the inter-model deviation and model-measurement difference for annual mean TCF are actually larger at SGP than at other two sites.

Nevertheless, we did find some common features among three sites and unique characteristics and added the new results in the revised version. Rather than presenting all new results and all changes we made to the manuscript, here (in this response) we only attach new Figures 2, 4 and 5 and cite a few paragraphs added in section 3.4 in discussing the similarities and dissimilarities in cloudiness pattern among three sites. Please see Sections 3.1, 3.2, 3.3, 3.4 and 4.1 in the revised version for more details about what we have changed. For your convenience, we attached a pdf file in which all changes we made to the original version of manuscript are tracked (This file is generated by Word by comparing the original and revised versions of manuscript so may not accurately reflect all changes in some cases).

"For all sites, the correlation coefficients are higher and RMSD are lower for the TSI/TSK comparisons than for the ARSCL/TSI comparisons. This is not surprising because ARSCL products were derived from the time-slice measurement with narrow lidar/radar FOV, but both TSI and TSK were from hemispheric observations……

The Manus and NSA sites both have large frequency of overcast cases and relatively few clear sky cases compared to SGP. At Manus, much of the overcast is likely due to ice anvil and cirrus associated with deep convective systems, while at NSA there is often extensive low-level cloudiness. At all sites, the ARSCL frequency is less than TSI when CF is small (< 0.3) and greater than TSI when CF is large (> 0.80). Also the difference between the TSI and ARSCL tends to be significant when CF is larger than 0.8

or smaller than 0.2. The compensating errors in lower and higher CF days result in small bias of TCF between ARSCL and TSI/TSK measurements as multi-year data is averaged."

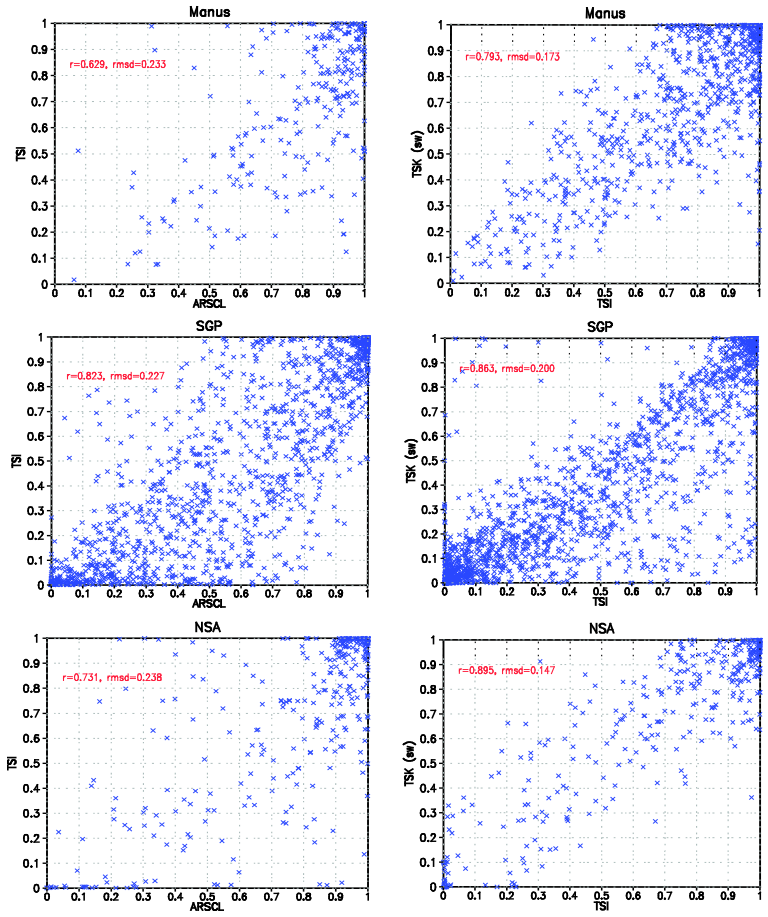
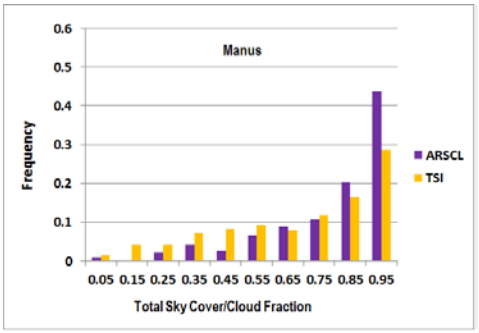
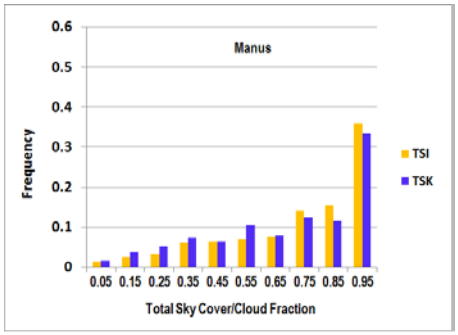


Figure 2





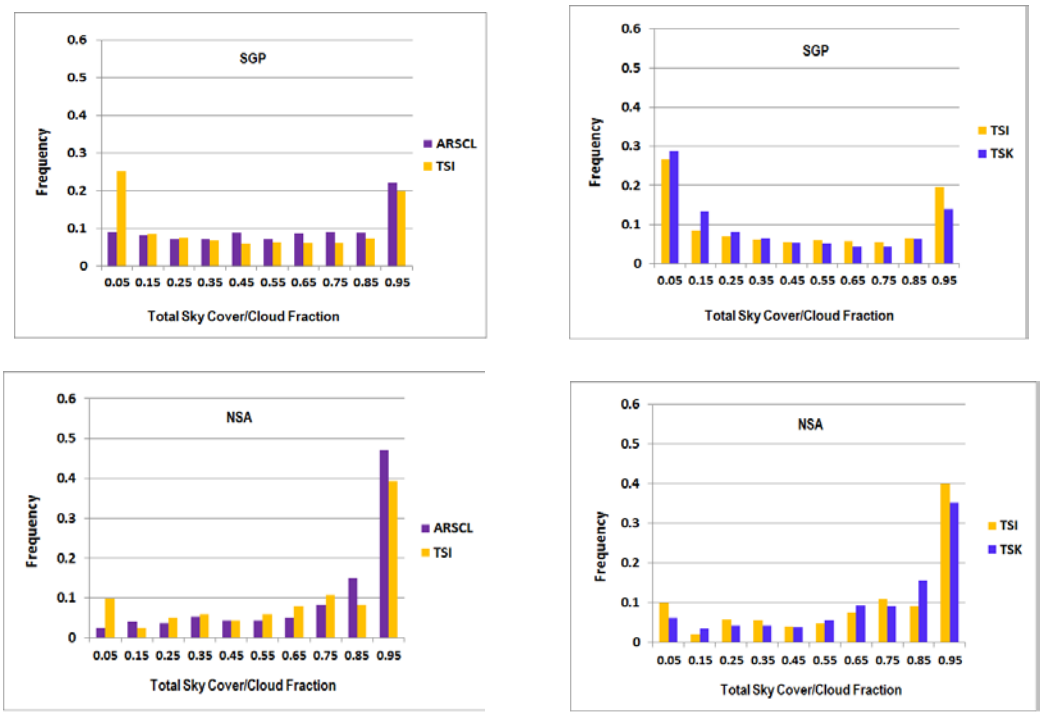
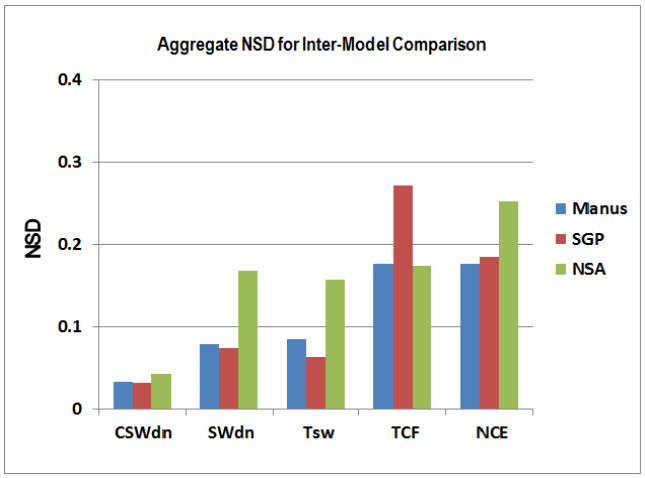

SGP
Frequency
0.6
0.5
0.4
0.3
0.2
0.1
0
0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
Total Sky Cover/Cloud Fraction
ARSCL
TSI
SGP
Frequency
Total Sky Cover/Cloud Fraction
TSI
TSK
NSA
Frequency
Total Sky Cover/Cloud Fraction
ARSCL
TSI
NSA
Frequency
Total Sky Cover/Cloud Fraction
TSI
TSK

Figure 4


Aggregate NSD for Inter-Model Comparison
NSD
0.4
0.3
0.2
0.1
0
CSWdn
SWdn
Tsw
TCF
NCE
Manus
SGP
NSA

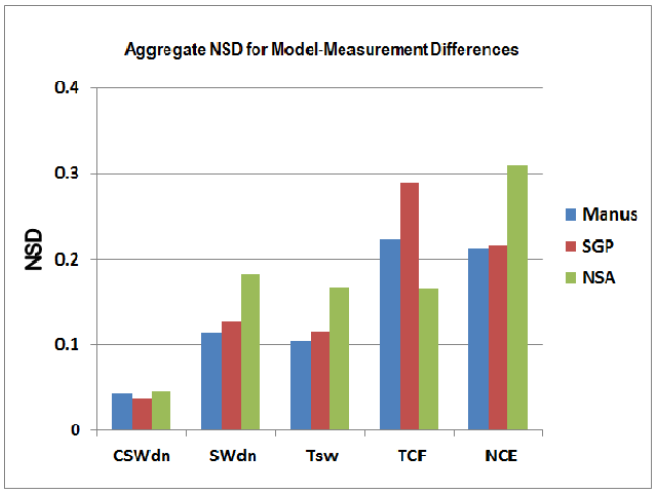


Figure 5

3. Different pieces of text, being either motivating comments, description of the methodology, background information or important results, are scattered throughout the document. However, relevant information on how cloud fraction profiles are constructed, or on why only a certain measurement is used to evaluate the models, should not be mentioned only once the reader gets to the conclusions. Also, motivating comments can be mentioned once in the introduction, but are distracting if they appear at the beginning or end of each paragraph in the remainder of the text. Including these pieces of text in their respective sections would make the manuscripts much more clear and concise.

Thanks for pointing this out. We have substantially revised the manuscript by comprehensively addressing the relevant comments of all three referees. Suggestions raised here have been taken into account. For examples, (1) we have moved most of the scattered motivation comments to the introduction and observational data section (Sections 3.1, 3.2, and 3.3). (2) We have reorganized Section 2.2 into four subsections, two of which are new subsections introducing the TSI data (2.2.b) and explaining why only a certain measurement is used to evaluate the models and how the vertical profiles of CF are constructed (2.2.d). (3) We have reorganized Section 3 into three subsections presenting the results for Manus, SGP and NSA, respectively. (4) We have moved some discussions on the uncertainty of measurements into the ARM data section (2.2). (5) We have revised Abstract and Conclusion/Discussion to more clearly highlight the uniqueness of this study and the key findings of this manuscript. (6) We have added a few paragraphs in Section 6 discussing the future work plan as reviewers suggested. (7) We also removed some sentences/phrases that seemed repetitive or caused confusion in the original manuscript.

We believe the presentation of has been much improved, as shown in the revised manuscripts with tracked changes. Please also see the response to general comment 2. In the attached pdf file with all changes tractable we can see substantial changes are made in reconstructing the paper.

Specific comments

1. P. 14936 line 16-20: What new light does this study shed on the findings in these cited papers (i.e., that total cloud fraction is better predicted in global models than the vertical profiles of cloud fraction)? Can you come back to this in the discussion?

One unique aspect of this study is that both inter-model deviation and model bias against observation are investigated in this study. Another unique aspect is that we use simultaneous measurements of CF and surface radiative fluxes to diagnose potential discrepancies among the GCMs in representing other cloud optical properties than TCF. The novel aspects and new findings of this study are summarized as follows. We have revised the conclusion and abstract to more clearly highlight the points.

1) This study for first time uses the long-term ground-based measurements to simultaneously evaluate the model produced clouds and radiative fluxes at the surface, including both inter-model divergence and model-measurement difference.

2) We provide a good sense on the uncertainty in different cloud fraction estimates as seen from the surface observations. The considerable differences in the various independent measurements of CF on a daily basis seem to be reduced significantly in the monthly (<10%) and annual (<5%) means, representing more statistically robust observations to evaluate climate models.

3) This study first archive the climatology of observed CF over three sites in different climate regimes and we find that the total sky imager (TSI) produces total cloud fraction (TCF) compared to a radar/lidar dataset in highly cloudy days (CF>0.8), but produces larger TCF value in less cloudy conditions (CF<0.3).

4) The model bias against the observation and the inter-model deviation (disparity) are much larger for total cloud fraction than that for the surface downward solar radiation and cloud transmissivity.

5) The climate models tend to generate larger bias against observations for those variables with larger inter-model deviation.

6) This study has also for first time provided quantitative and comprehensive details of comparison not only for total cloud amount but also the PDF, transmissivity and vertical profiles of cloud among GCMs and between model and observation over three different regions.

2. P.14938 line 7: Which definition of cloud fraction would hence lend itself more naturally to compare with the modeled cloud fraction? (hence later in the paper: why is the hemispheric cloud fraction TSK used in the comparison with the monthly total cloud fraction?)

While the hemispheric cloud fraction observed by TSK and TSI have more comparable definition to the model CF than the ARSCL does, each of the three measurements have their own advantages/disadvantages.

We use TSK in Figure 7 because 1) TSK has longer records than the other two observations. Here PDF is calculated based on monthly mean TCF since we only have monthly mean TCF from model side; Usually larger samples are required to calculate the PDF so we just use TSK that has the longest records; 2) the differences of PDF among three datasets are very small at

monthly mean level (much smaller than that at daily level as shown in Fig. 4). So we only choose one observation rather than three to compare with model results.

The reason we use TSK in figure 5 is that TSK TCF is accompanied by the surface radiation flux data. We can certainly use ARSCL TCF to calculate the model-measurement differences, but it will result in the NSD of TCF and NCE to be incomparable to the NSD of radiation flux because ARSCL TCF has different time period coverage than the radiation data. Also it will make less meaningful in calculating NCE if the TCF is not compatible with the surface radiation flux.

3. P.14939 line 10: To be consistent you removed some models from the pool that have missing radiation fluxes of vertical profiles, so why do the 11 models listed in Table 1 still have these missing variables? which ones are removed?

As shown in Table 1, there are 11 datasets for Total Cloud Amount but only 7 of them have Layer Cloud Fraction available. Four models (cnrm, mpi, gfdl and ukmo) have Layer CF data missing at the PCMDI website for unknown reasons. So we only have 7 GCMs presented when comparing the vertical profiles in Section 5. Only Total Cloud Amount and surface radiation flux data are needed for the analysis in Section 4, so we have included all 11 GCMs. This has been clarified in the revised manuscript.

4. P.14939 lines 20-25: Here would be a good point to mention how CF and total cloud cover in the global models are calculated, this aspect is critical to your study and should not just be referred to by giving the link to the documentation. What are the cloud overlap assumptions? What is the time period over which cloud fraction is averaged?

The time period over which CF is averaged is from January 1980 to December 1999, with starting and ending years slightly varied among the models (see Section 2.1).

It might not be realistic to provide detailed description of individual cloud schemes used in all GCMs due to the space constraints. The lengthiness of this paper has already exceeded a normal one. Detailed description of the cloud schemes will add significantly to the length (it will need 22 more paragraphs to describe the cloud scheme for all GCMs even if only two paragraphs are used to introduce the cloud scheme for each GCM). Instead, we summarize the CF parameterization schemes for all GCMs used in this study in Table 1 and gave the references where the schemes are described in much more details. Additionally, in Section 2.1 we added three paragraphs (attached below) briefly describing how CF and TCF are calculated and the overlap assumptions made in IPCC AR4 GCMs.

It is well known the representation of cloud is the most uncertain element in climate models and it is an extremely challenging task that may take several decades of effort just to improve the cloud scheme in a single GCM. This study aims to evaluate a group of GCMs rather than one specific GCM with a particular cloud scheme in simulating cloud and cloud-radiation scheme. For this reason, we are not able to go deeply to investigate the cause of biases of each model. On the other hand, only CF and surface radiation data are used in this study. Besides cloud overlap assumption, the cloud thickness and water content, height and shape of clouds, ice/liquid water

fraction, etc., all of which could affect the surface radiation, are not investigated in this study. Without comprehensively analyzing all those fields, we are not able to attribute the bias of surface radiation to cloud overlap assumption scheme or other factors. All of those works can be done in future study. Below is what we added.

"CF is a critical variable in climate models for determining the radiative flux through the atmosphere and at the surface. Depending on the complexity of the model, CF may also be used in many other physics parameterizations in the model such as cloud microphysics, aerosol wet removal and convective transport. In this study, we focus on the role of CF in radiation, where the area-averaged CF is used. As discussed in Brooks et al. (2005), although CF produced by most cloud schemes is volume-averaged, most GCMs assume that the cloudy area of a grid box fills the entire grid box in the vertical, thus essentially assuming area-averaged CF is the same as the volume-averaged CF. In GCMs, CF can be parameterized using statistic, diagnostic or prognostic approaches. Due to space constraints, we summarize the CF parameterization schemes for all GCMs used in this study in Table 1; for more details on each cloud scheme, including references, see http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php.

In GCMs, the vertical correlations between cloud layers have to be prescribed because cloud elements are usually smaller than a typical GCM grid cell and there is no general theory for how different cloud systems should overlap (Collins, 2001). Assumptions about vertical overlap of clouds can affect the exchange of energy between the atmosphere and other components in the model, influencing not only radiative heating rates but also atmospheric temperature and hydrological processes (Collins, 2001). In the IPCC AR4 models, the most common overlap assumptions are maximum/random (Geleyn and Hollingsworth 1979). One type of maximum/random assumption has maximum cloud overlap in each of three regions representing the lower, middle, and upper troposphere and random overlap between these regions (e.g., Chou et al. 1998). A second type of maximum/random overlap scheme has maximum overlap between clouds in adjacent levels and random overlap between groups of clouds separated by one or more clear layers (e.g., Zdunkowski et al. 1982). The latter form of maximum/random is the most consistent with a statistical analysis of observed cloud distributions (Tian and Curry 1989)."

Chou, M.-D., M. J. Suarez, C.-H. Ho, M. M.-H. Yan, and K.-T. Lee, 1998: Parameterization of cloud overlapping and shortwave single-scattering properties for use in general circulation and cloud ensemble models. J. Climate, 11, 202–214.
Collins, W. D., 2001: Parameterization of generalized cloud overlap for radiative calculations in general circulation models. J. Atmos. Sci., 58, 3224–3242.
Geleyn, J.-F., and A. Hollingsworth, 1979: An economical analytical method for the computation of the interaction between scattering and line absorption of radiation. Beitr. Phys. Atmos., 52, 1–16.
Tian, L., and J. A. Curry, 1989: Cloud overlap statistics. J. Geophys. Res., 94, 9925–9935.
Zdunkowski, W. G., W.-G. Panhans, R. M. Welch, and G. J. Korb, 1982: A radiation scheme for circulation and climate models. Contrib. Atmos. Phys., 55, 215–238.

5. Section 2 and 3: The limitations of each measurement, and what may or may not be usefully compared to the models, should be clearly discussed here (that is, before the conclusions). If not,

one gives the impression that after all, the comparison of cloud fraction from the models and the observations is not useful.

It would be very good to provide estimates of the contribution of rain particles to cloud fraction, as well as being clearer about the different sensitivities for detecting upper level cloudiness (for the ARSCL product) in a more quantitative matter if possible.

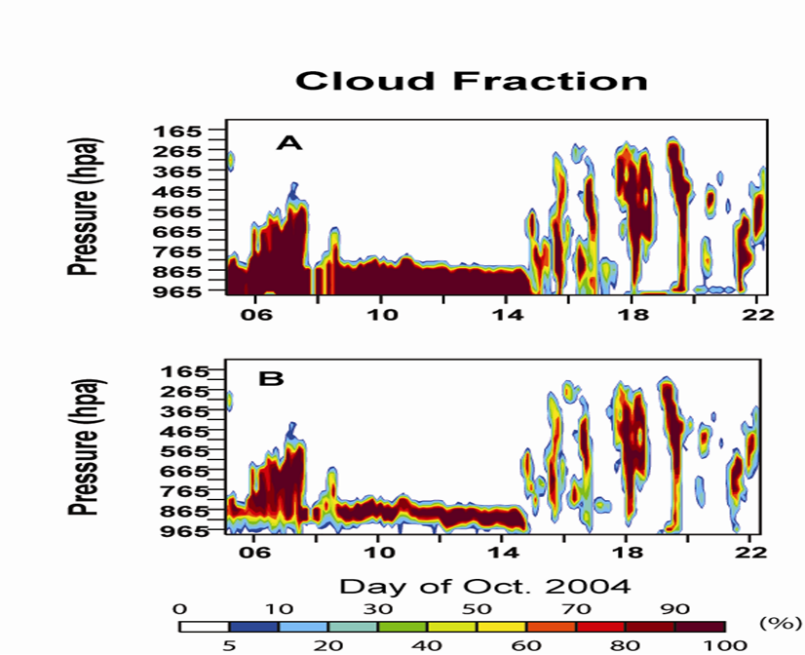
By giving the different products or instruments a subsection (with a separate header) this section as a whole would read much better. Some more detail on how cloud fraction is derived from the Total Sky Imager and how vertical profiles of cloud fraction are derived from the ARSCL product would be good.

We have reconstructed Sections 2 and 3. For example we have included three sub-sections with separate headers in Section 2.2 to introduce the three types of measurements respectively, including their limitations. We have also added a subsection 2.2.b giving more details on how CF is derived from the TSI and a subsection 2.2.d on how vertical profiles of CF are derived from the ARSCL measurements.

In the revised manuscript, we have also provided more details about the ARSCL measurements used in this study. As the reviewer pointed out, one issue with the ARSCL product is the contamination of ice precipitation near and below the MMCR radar detected cloud base. To minimize the potential consequence of this problem, the CMBE uses the ARM laser ceilometer and micropulse lidar (MPL) measurements, which are usually insensitive to ice precipitation (if the concentration of precipitation particles is not sufficiently large) or clutter, to determine the cloud base. As indicated in Clothiaux et al. [2000], the laser ceilometers and MPL can provide quite accurate cloud base measurements. Another issue with the ARSCL clouds is that cloud radar tends to underestimate the cloud top heights for high-altitude clouds because of detection limits and signal attenuation. The consequence of this problem has been mitigated with the use of ARM MPL in CMBE, which is more sensitive to small cloud particles. The ARSCL cloud statistics used in this study are calculated based on data during the period when both MMCR and MPL were in operation.

6. P.14941 lines 16 and 22-23: It is not clear how much precipitation has affected the results - can you give an estimate of how much of the rain influence is reduced with and without the ceilometer/lidar screening? For light rain events, such a screening might work well, because the ceilometer/lidar is still able to see cloud base and not rain, and this can be used to judge whether a cloud is overhead or not. For heavier rain, this becomes a more difficult task.

We did not quantitatively estimate how much precipitation has affected the results, but as shown in the attached figure, the ceilometers/lidar screening (top panel: without screening; bottom panel: after screening) effectively removes precipitating ice beneath cloud base for the cloud systems observed during the ARM Mixed-Phase Arctic Cloud Experiment, which was conducted during the period of Oct. 5-23, 2004 at the ARM barrow site. As we mentioned earlier, the ARM laser ceilometer and micropulse lidar measurements are usually insensitive to ice precipitation if the concentration of precipitation particles is not sufficiently large. But for very heavy rain, as pointed out by the reviewer, this may become more difficult. We have discussed this issue in the revised manuscript.



7. P.14942: The first paragraph (lines 1-17) might be better written along with paragraph 2.2 and can be more concise. What is the temporal resolution of the other instruments?

We have reorganized the Section 2.2 and first paragraph of Section 3 by having a subsection for each instrument so that the readers can more clearly compare the different datasets (See new Section 2.2).

TSK has 1-minute temporal resolution based on 1-second sampling. TSI does 30-second sampling.

8. P. 14943 and 14944: The authors use a 1 hour over which to derive cloud fraction (is that true for the profiles as well?) from the different measurements. As discussed in the text, one hour is short and leads to more frequent occurrences of cloud fractions being either zero (not shown) or one, for the narrow FOV ARSCL product (see also Boers et al. (2010) - JGR, Vol 115, D24116.). Does this period represent a spatial scale equivalent to the model grid box? As mentioned, as long as the one hour cloud fractions are averaged over long enough periods (a month, a year) they converge to other estimates of cloudiness, but the 1 hour cloud fractions reflect different spatial scales (as a result of different wind speeds), so it is not trivial that averaging 1 hour cloud fractions to a daily estimate, or a monthly estimate, would give the same results as when cloud fraction is determined over six hours, one day (a month). It would be good to think this through.

CMBE provides the CF data with 1-hour interval, which is probably for the convenience of user to look at features such as the diurnal cycle of cloudiness. One-hour time period probably does not represent a spatial scale equivalent to the model grid box. Assuming the wind speed is 4 m/s, cloud can move 3600*4=14.4 km within one hour, which is much smaller than the grid spacing

of current typical GCMs. That's why we averaged 1-hout CF to a daily estimate for Figures 1, 2 and 4, and to a multi-year monthly estimate for other Figures.

Would the CF for 1 day be the same if we average the CMBE 1 hour average to 1 day compared to if we average the actual data directly from raw resolution to one day? The answer is No. CMBE estimates the 1-hour CF first based on the number of data points within each hour of the day reducing the effect of missing data during the day, then calculate the daily mean based on the 24 hourly datasets. If the averaging is performed directly by averaging every point within the whole day regardless of missing data, then the average would be biased if missing data is not evenly distributed during that day if the diurnal variation of cloud fraction is substantial. A more apparent case is for solar radiation. If we have many missing solar radiation measurements near noon time, then we will underestimate the daily mean solar radiation if we just simply average all data points we have. More details see:

http://science.arm.gov/workinggroup/cpm/scm/best_estimate.html#CMBE_Stat_

9. P.14943 lines 5 - 10: Here the influence of wind speed on biases between the hemispheric and FOV CF estimates is discussed. First, can you plot the wind speed versus the biases (on a daily, 6 hourly, hourly basis)? It is very hard to see from Figure 1 that periods of low wind speeds correspond to larger biases, and it is not clear how much these biases reduce when moving from hourly to 6-hourly values, or 12 hourly (daytime) values. Second, periods with different wind speeds might correspond to periods with different cloudiness. May part of what you are seeing relate to the difficulty of measuring certain types of cloudiness that are predominantly present during periods of certain wind speeds?

We have added one more curve (i.e. ASCL-TSK difference) for Figure 1 to show the relationship between the wind speed and "bias". An anti-correlation can be found for some time periods (e.g., June 21 - July 7, 2006) but the correlation is not significant for other periods, which is probably related to the difficulty of measuring certain types of clouds that are predominantly present under certain meteorological conditions including wind speeds as the reviewer suggested.

Figure 1 shows the time series of daily (daytime) averaged total sky cover or cloud fraction and wind speed. Since both TSI and TSK are only available during daytime, it can be expected the daily (24-hour) or 6-hour mean will not generate big difference comparing with current daytime mean. As the reviewer pointed out, "one hour is short and leads to more frequent occurrences of cloud fractions being either zero or one for the narrow FOV ARSCL product (Boers et al., 2010)", we don't expect to see a meaningful relationship between wind speed and CF bias at 1-hour time scale. We have added more discussion in Section 3 and cited Boers et al. (2010).

Boers, R., M. J. de Haij, W. M. F. Wauben, H. K. Baltink, L. H. van Ulft, M. Savenije, and C. N. Long (2010), Optimized fractional cloudiness determination from five ground‐based remote sensing techniques, J. Geophys. Res., 115, D24116, doi:10.1029/2010JD014661.

10. P. 14943 lines 11-29: Why does the TSK have much larger day to day variability (see 1 - 11 May 2006 in Fig.1) than the TSI? Does a certain instrument perform better at one site, because of

the dominant cloudiness pattern? Is ARSCL TCF higher than TSI TCF over Manus because of frequent occurrence of high – level cloud?

The reason why the TSK has larger day-to-day variability for 1-11 May is probably that the clouds located right above are stable but clouds over remote area change more frequently. We also found the time periods with smaller day-to-day variability for TSK in other months/years.

All of the CF estimates are more likely to be biased (for potentially different reasons) in broken cloud situations where the TSI/TSK can be affected by cloud sides and the small sampling FOV of the ARSCL TCF may cause it to be unrepresentative and will perform best for overcast conditions. The ARSCL cloud fraction is slightly higher than the TSI and TSK over both Manus and SGP, but not over NSA, which may be due to the more frequent presence of thin high-level cloud at these sites.

11. P.14945: Would the discussion of daily value histograms fit better before the discussion of the monthly mean values (annual cycle)? Why does Figure 4 not show TSI versus ARSCL? Is TSK overestimating in partly cloudy conditions or times with low sun elevation? It would be valuable to show ARSCL, TSI and TSK in the same PDF and make the PDF for Manus, SGP and NSA.

Thanks for the suggestion. It's done. See new plots for TSI vs. ARSCL and new figures for SGP and NSA (Figure 4, see attached above).

We only chose the days in which both ARSCL and TSK or both TSI and TSK are available to calculate the frequency for ARSCL/TSK or TSI/TSK as shown in the original Figure 4. Because of different time periods with missing data in the three datasets, the numbers of sampling days used for TSK in ARSLC/TSK and TSI/TSK are different, so the calculated values of frequency for TSK in these two plots are slightly different. That's why we didn't combine the two plots into one figure.

In the new Figure 4, we compared the frequency of ARSCL and TSI, and TSI and TSK (the information of ARSCL/TSK can be inferred from ARSCL/TSI and TSI/TSK), and added two more plots for SGP and NSA. More discussions are added in Section 3.

12. P14947, line 21 - 22: Are there some missing references here, i.e., are you the first to show this?

We believe we are first to show this based on the results of a group of GCMs instead of an individual GCM.

13. P14948, line 1: Here you choose to do the comparison of the models' TCF with the TCF derived from the TSK observations: why? Is it just because you have longer records of TSK and because they are compatible with the surface radiation flux data? Why not ARSCL?

Yes, this is correct. The major reason is that TSK TCF matches with the surface radiation flux data. We can certainly use ARSCL TCF to calculate the model-measurement differences, but it will result in the NSD of TCF and NCE to be incomparable to the NSD of radiation flux because

ARSCL TCF has different time period coverage with radiation data. Also it will become less meaningful in calculating NCE if the TCF is not compatible with the surface radiation flux.

14. P14949: Errors in TCF and SW are due to errors in cloud overlap assumptions too, this should be discussed much earlier, when introducing how cloud fractions are determined for the global models.

Errors in TCF and SW could be due to errors in cloud overlap assumptions, also could be due to the cloud thickness and water content, height and shape of clouds, ice/liquid water fraction, etc., all of which could affect the surface radiation, but they are not investigated in this study. Without comprehensively analyzing all those fields, we are not able to attribute the bias of surface radiation to cloud overlap assumption only. All of those works can be done in future study.

15. P14951 line 3-4: This is an example of a sentence that has been frequently repeated and could be omitted.

Have removed this sentence according to the comment.

16. P14951 line 6: If you take out the cnf hires GCM then the Arctic site looks just as diverse (or even less diverse) as SGP and Manus.

We have slightly modified this sentence by removing the comparisons with other two sites.

17. P14952 line 4: In terms of the distribution, but not in their monthly mean (Fig 6b).

Corrected.

18. P14956 lines 16-17: This should have been mentioned much earlier when comparing the different estimates of TCF derived from the observations.

We have removed this sentence. Instead, we have added more discussion on how we tried to minimize the precipitation contamination problem in Section 2.2.c. Please see responses to specific comments 5 and 6.

19. P14956: line 26: If you would map cloud fractions onto the same vertical grid, what differences remain?

Done. See response to General Comment 1.

20. P14957 lines 20-21: This cannot be argued from the model-measurement difference (that is larger at higher altitudes) because the ARSCL may be less sensitive to seeing high cloud and you are using a different vertical grid.

We have removed “the model-measurement difference”.

21. P 14963 lines 18-29: Evaluating the models according to their cloud scheme, and cloud overlap assumptions, has not been systematically done in this paper. If this was the original goal, what part of it has been achieved? What can you conclude that was not known before?

Actually our original goal was not to systematically evaluate the cloud scheme and overlap assumption in the IPCC AR4 GCMs, although it would be a bonus if we can find a clear correspondence between model performance and specific cloud scheme. As discussed in the response to specific comment 4, it is well known the representation of cloud is the most uncertain element in climate models and it is an extremely challenging task that may take several decades of effort just to improve the cloud scheme in a single GCM. This study aims to evaluate a group of GCMs rather than one specific GCM with a particular cloud scheme in simulating cloud and cloud-radiation scheme. For this reason, we are not able to go deeply to investigate the cause of biases of each model. On the other hand, only CF and surface radiation data are used in this study. Besides cloud overlap assumption, the cloud thickness and water content, height and shape of clouds, ice/liquid water fraction, etc., all of which could affect the surface radiation, are not investigated in this study. Without comprehensively analyzing all those fields, we are not able to attribute the bias of surface radiation to cloud overlap assumption scheme or other factors. All those works can be done in future study.

The novel aspects and main conclusions obtained in this study but not known before are:

1) This study for first time uses the long-term ground-based measurements to simultaneously evaluate the model produced clouds and radiative fluxes at the surface, including both inter-model divergence and model-measurement difference.

2) We provide a good sense on the uncertainty in different cloud fraction estimates as seen from the surface observations. The considerable differences in the various independent measurements of CF on a daily basis seem to be reduced significantly in the monthly (<10%) and annual (<5%) means, representing more statistically robust observations to evaluate climate models.

3) This study first archive the climatology of observed CF over three sites in different climate regimes and we find that the total sky imager (TSI) produces total cloud fraction (TCF) compared to a radar/lidar dataset in highly cloudy days (CF>0.8), but produces larger TCF value in less cloudy conditions (CF<0.3).

4) The model bias against the observation and the inter-model deviation (disparity) are much larger for total cloud fraction than that for the surface downward solar radiation and cloud transmissivity.

5) The climate models tend to generate larger bias against observations for those variables with larger inter-model deviation.

6) This study has also for first time provided quantitative and comprehensive details of comparison not only for total cloud amount but also the PDF, transmissivity and vertical profiles of cloud among GCMs and between model and observation over three different regions.

Typo's/errors/graphics

1. P.14938 line 1: between themselves or among themselves?

Changed to “themselves“.

2. P.14945 line 15: seasonable should be seasonal

Changed.

3. P14950 line 12: is ’potentially’ at the right place here?

“Potentially” has been deleted.

4. P14957 lines 18: ’which is the height that the’ should read ’which is the height at which’ or ’which is the height where’.

Done. Changed to “which is the height where”.

5. P14959 line 11: reveals should be reveal

Done (at P14961, Line 11).

6. Graphics: Many figures show cloud fractions up to 1. 5. We know it cannot exceed 1, so why not make the y-axis go to 1. and include the labels outside of the main figure? For Figure 2, the x- and y-axis should have the same length (scale). For Figure 6: it is hard to distinguish the TSK (SW) line against the background grid, the same is true for the GCM’s mean.

All those Figures have been modified.