Reviewer #1

The general impression from the response is that the authors are pretty much set on their approach. The use of ANNCOD in the background without quantifying its quality relative to CALIPSO is unfortunate, and it casts some doubt on the manuscript. I do think that a comparison of ANNCOD with CALIPSO makes sense, at least for COD<3. If this is what is used for the cloud mask, then it should be quality assessed. For bright regions, the authors' statement that passive sensors will always detect clouds for COD>1 (in their response) is incorrect. At some point, the authors comment that ANNCOD is a pseudo optical depth, which is "relative". But if that is the case, a formal definition should be provided what "pseudo" means. Following the current manuscript, ANNCOD seems to be unscaled, and trained with CALIPSO data, so it should reproduce CALIPSO data. However, the authors state that "ANNCOD does not provide any information on the absolute optical thickness value." At that point one has to wonder what it does do. If ANNCOD does not contain any information on COD, then the name "ANNCOD" is misleading, and the product should not be used as the basis for the cloud mask.

AR: We assume that CALIPSO correctly detects all clouds with COD > 1. Admittedly, this is not true for all passive sensors, in particular when only using the AVHRR heritage channel set. With that assumption in mind, we set all COD > 1 = 1 and thus do not account for the full spectrum of CALIPSO data. Moreover, the CALIPSO signal will be attenuated if thick clouds are present, and the lidar won't be able to penetrate to the surface if total column optical depth > 5. Thus, we think it is correct to refer to ANNCOD as a *pseudo* optical thickness. Although we trained the neural net with *true* CALIPSO COD data, our approach is no replacement of a Nakajima-King-style COD retrieval. ANNCOD is an unphysical quantity created by the neural net that is (a) observable and (b) correlated with cloud cover. An actual physical analogue is neither desired nor expected.

Otherwise, the added detail with regard to NN is helpful. However, questions remain how well it works with one single hidden layer. Just because the used IDL library does not provide more than one layer is not a good justification for limiting oneself to something that may not be adequate for the problem. Might this shallow NN be the reason that the authors don't show the validation against CALIPSO COD? In a similar vein, it is unfortunate that the authors do not want to show the histograms from figure 7 for snow bright surfaces (their response to the reviewer's comment to figure 7).

AR: We think that a single hidden layer is sufficient, and including one or more hidden layers would not produce significantly different results. See also:

K.M. Hornik, M. Stinchcombe, and H.White , „Multilayer feedforward networks are universal approximators“, Neural Networks, vol. 4, no. 5, 1989, pp. 359-366,

which shows that a neural network with

- one hidden layer
- a sufficient number of nodes
- nonlinear activation functions

is capable of approximating any real-valued continuous scalar function.

We did comparisons between ANNCOD and CALIPSO COD, and analysed them as a function of illumination condition and surface type, including bright snow surfaces. Snow/ice covered surfaces indeed show a lower degree of agreement between the two entities, in particular regarding correlation coefficient and standard deviation ratio (Figure 1). In these cases, correlations range between 0.6 and 0.7, but are clearly larger for snow/ice-free surfaces during day and night (~0.85).

Although our methodology could still be improved (such as almost any other approach), we are confident that it provides reliable results of very good quality. Problems remain for bright land surfaces and twilight illumination conditions, which however have been addressed in a new NN cloud mask version 3.0 (Figure 2). Please note that the paper shows results for version 2.0 only. We are not hiding any negative results and are happy to share any information with the reviewer. Please see also the Product Validation and Intercomparison Report (Stengel et al., 2018) for a more detailed validation of the ANNCOD-derived cloud mask against CALIOP.

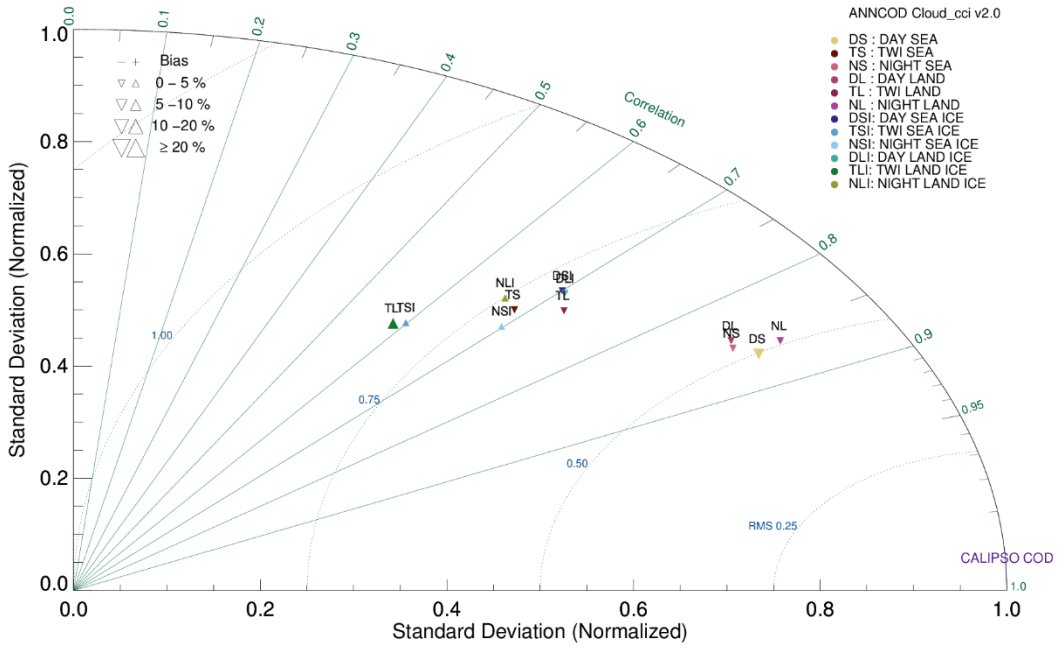

Figure 1: Taylor plot of ANNCOD (version 2.0) versus CALIPSO COD for various illumination conditions and land cover types. The agreement between ANNCOD and CALIPSO COD is perfect if x = 1.0.

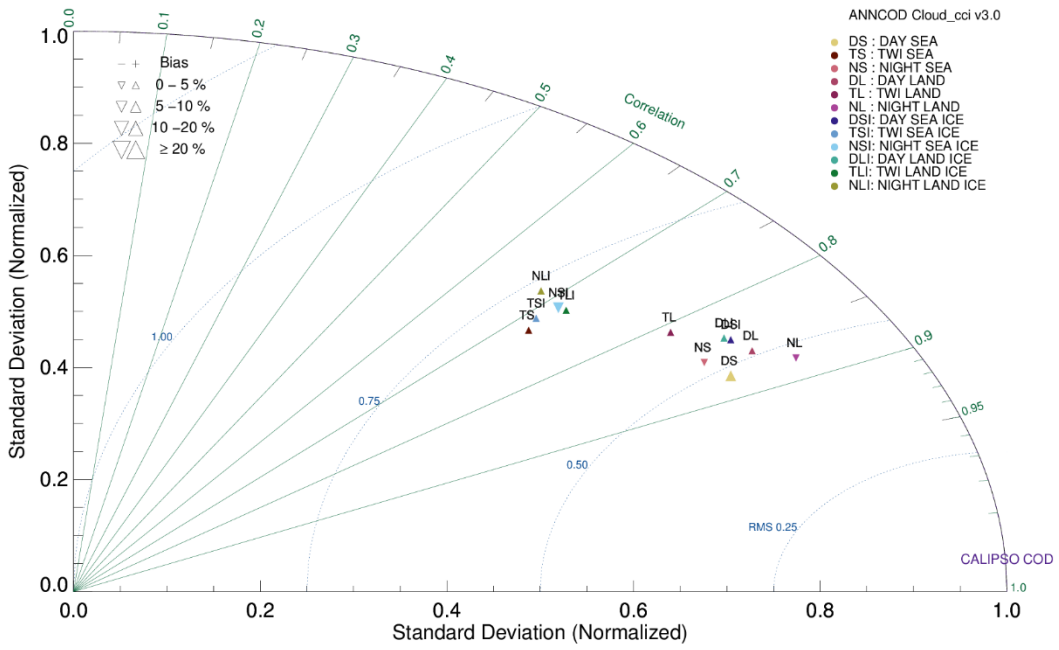

Figure 2: As above, but for version 3.0.

A few minor comments:

p2,l49: "insufficient knowledge...propagates uncertainties..." still doesn't sound quite right. It is not the insufficient knowledge that propagates uncertainties. Just a language issue.

AR: Agreed, will change text.

AC: “and insufficient quantification of their state propagates uncertainties…”

p2,l78: "Consistency in approach can be traded for continuity of results, and multi-platform algorithms could exploit additional data when newer sensors become available" is still unclear and needs elaboration.

AR: Agreed, we will elaborate that sentence.

AC: “There is a trade-off between using the information from a single sensor optimally and using the information from different sensors consistently. The former may provide the most scientifically accurate data but often results in sharp discontinuities as instruments are introduced. Towards the aim of producing a stable, self-consistent climate record, this paper focuses on evaluating data from a range of sources through a unified methodology.”

"For a single layer, optically thick (COT>1)" -- This is incorrect. Cloud with COD>1 are not optically

thick.

AR: We will rephrase that sentence.

AC: “For a single layer cloud with COT > 1,…”

Regarding the t-test (p13,l39): It is understood that readers will be familiar with a t-test, that was not the point of the comment.

AR: Our answer explains the purpose of the two-sample t-test, i.e. the null-hypothesis, and the meaning of the symbology. We are assessing whether the difference of the mean values of two independent sets of samples (drawn pairwise from AVHRR, MODIS, and AATSR data) is statistically significant. While we are aware of the statistical limitations of the t-test and its frequent misuse, this is a clear, simple, and frequently applied test to punctuate a simple point. As we do not aim to infer causation from these differences, and would prefer not to lecture the reader in an already long paper, we leave our text unchanged.

p13,l45: "spatiotemporally collocated sensors" The comment was not about the mathematical meaning of the intercomparison, but about the physical meaning. Why *are* the data different, even though they are collocated?"

AR: There are various reasons why the data are different despite collocation, such as sub-pixel cloudiness (the cloud varies on a scale smaller than that observed, so slight differences in pixel position produce statistically significant changes in value), cloud displacement between observation times, and differences in spectral response functions. Repeating the analysis on an even coarser grid would not improve any of these issues.

Reviewer #2

p 2, lines 34-35: I think the authors inserted the Marchant reference and description in the wrong sentence. That paper discusses the MODIS C6 optical property phase, not cloud detection/masking, and fits into the next sentence. Moreover, the next sentence discussing phase should include the Baum et al (2012) reference also since it is referring to the MODIS C6 IR phase which uses 8.5 and 11µm in addition to 7.3 and 12µm (the authors provide an incorrect description of this). There are two independent phase algorithms in MODIS C6 and both should be referenced.

AR: Agreed, we will add references as suggested.

AC: "For cloud masking, the retrieval frameworks apply various approaches such as Naïve Bayes (PATMOS-X), dynamic thresholding (CLARA-A1), or a battery of threshold test (MODIS C6). Finally, cloud phase or type is determined as a function of a combined convergence/cloud top temperature (CTT)-test (CLARA-A1), the Pavolonis et al. (2005) threshold algorithm (PATMOS-X), or a majority vote algorithm that combines four phase tests based on CTT, tri-spectral IR, 1.38 µm, and spectral CER data (Marchant et al., 2016, Baum et al., 2012)."

p 3, lines 6-7: "and beyond a penetration depth into the cloud corresponding to > 1 cumulative optical depth (Baum et al., 2012)." I'm not sure how this in itself describes an insensitivity to cloud top properties. The roughly 1 optical depth penetration depth is simply the reason why the MODIS C6 cloud top retrieval (and any other IR cloud top retrieval) loses sensitivity for optically thin clouds and why it cannot be considered the physical cloud top.

AR: Agreed, we will remove that sentence.

AC: "Still, the MODIS C6 cloud top retrieval loses sensitivity for optically thinner clouds (COT < 2, Menzel et al. (2010); Christensen et al. (2013))."

p 3, line 22: "However, the detection of optically thin ice clouds over warm, bright surfaces remains problematic (Marchant et al., 2016)." This should actually refer to the phase identification of optically thin ice clouds, which is what the Marchant paper discusses, not the detection, which is a cloud mask issue.

AR: Agreed, we will rephrase in order to refer to the correct product.

AC: "…, and the phase identification has been improved for liquid clouds. However, the phase identification of optically thin ice clouds over warm, bright surfaces remains problematic (Marchant et al., 2016)."

p 5, line 21: CALIOP is the lidar itself, CALIPSO is the satellite platform.

AR: Agreed, we will correct.

AC: "The Aqua satellite is a member of the "A-Train" constellation, which also includes the CALIPSO and CloudSat satellites."

p 6, line 8: “on-board” instead of “on-bard”

AR: Will correct, also removing repeated “is” in that sentence.

AC: “ATSR is equipped with on-board calibration capabilities, …”

p 16, lines 15-16: “In particular it would be worth investigating the impact of spectral response differences, which was outside the scope of this paper and the ESA Cloud_cci project.” Adding this statement is fine, but it is now at odds with previous statements (e.g., same page lines 1-2) that spectral response differences are unlikely the causes of retrieval differences, and with the immediately following statement.

AR: Agreed, we will tone down that sentence to be consistent with our previous and following statements that spectral response differences are not a main driver.

AC: “Although we found that spectral response differences are probably negligible, we still find it worth quantifying their impact, which was outside the…”