Anonymous referee #1

RC: referee comment

AR: author response

AC: author's changes in manuscript

RC: p1: "Climate data record" is not discussed. It is also not clear until page 2 that the manuscript does indeed seek to develop a Essential Climate Variable (although it is unclear without reading Hollmann 2013 which of the 13 ECVs CC4CL will contribute to). A discussion of "essential climate variable", "CDR", and how this work fits in should be better discussed. It should also be discussed how it distinguishes itself from existing efforts in this regard (e.g., https://www.ncdc.noaa.gov/news/new-cloud-propertiesclimate-data-record).
AR: We will move p2 lines 97-106 to the very beginning of the introduction, and add some text to elaborate how our study fits into ESA CCI and the production of ECVs. We will now also mention CDR, but we do not elaborate on it, as it is not an essential part of this paper. We thus refer to Stengel et al., 2017, where this concept and resulting data have been published.
AC: "The European Space Agency has established the ESA Climate Change Initiative program (ESA CCI, 2015; Hollmann et al., 2013) in order to advance knowledge of the climate system through the generation of satellite based data records utilizing European and non-European assets. The CCI project's primary focus is the production of thirteen Essential Climate Variables (ECVs) covering ocean, atmosphere, and land geophysical variables. With these data records, CCI is aiming to fulfil highest climate requirements from the Global Climate Observing System (GCOS). This study presented here is part of the ESA CCI for clouds (ESA Cloud_cci), which has the objective to develop a state-of-the-art open-source community cloud retrieval algorithm which shall be capable of processing passive satellite imager data covering several decades. Both in part I and part II of this paper, we present the processing framework as developed within ESA Cloud_cci (CC4CL, part I), the detailed mechanisms of the optimal estimation retrieval (part II), and provide an initial assessment of the strengths and weaknesses of derived cloud parameters (part I). With CC4CL, several decades of passive imaging satellite data have been processed and are made available to the user. The resulting climate data records (CDR) are presented in Stengel et al., 2017."

AR: We will amend p2, line 106ff.
AC: p2, line 106ff: "In order to produce the cloud CDR presented here, we used satellite data from MODIS…"

RC: p1, l38: "shielding" is not a radiative transfer term - what is it in this context? Isn't it the same as "forcing"? Why are both terms used?
AR: We agree that shielding is superfluous, will remove.
AC: "Clouds considerably influence the global energy budget through direct radiative effects."

RC: p1, l38: "forcing": There is a difference between "radiative forcing" and "radiative effect"- which are the authors referring to? Probably the latter.
AR: Agreed, we are referring to radiative effect.
AC: See comment above.

RC: p1, l49: This sounds like the variables “propagate uncertainties” into the derived cloud properties, which would be incorrect.
AR: Will clarify.
AC: “Several secondary variables (state of surface and atmosphere, viewing geometry, sensor calibration and spectral response uncertainties) further complicate cloud retrievals, and insufficient knowledge on their state propagates uncertainties into the derived cloud properties.”

RC: p2,l14: While “auxiliary” instead of “ancillary” data have become almost interchangeable, the latter is more correct; “auxiliary” has the connotation of only being a replacement in case the “primary data” is not available (compare: auxiliary power, not ancillary). For satellite retrievals, ancillary expresses more accurately that data from other sources are ingested within the operational algorithm.
AR: Agreed, will replace auxiliary with ancillary throughout the text.

RC: p2,l25: “. . .not guaranteed to be radiatively consistent with. . .” It is unclear what that means (although the reviewer agrees with the statement). Please provide references. Also, does CC4CL perform “better” in terms of radiative consistency?
AR: The effect of COT/CER/CTH on the top-of-atmosphere (TOA) radiances differs between the different sensing bands as a function of atmospheric state. For example if you used just the 11 or 12 micron measurement to estimate CTH then you must assume something about the COT (usually that it is thick) and something about the CER (typically a climatological value). If the COT assumption is incorrect (e.g. cloud is not thick) so that more upward radiance is transmitted through the cloud than expected then the cloud top appears too warm and is located (incorrectly) lower in the atmosphere. On the other hand using an all channel fit will identify the cloud as optically thin (from the visible and near visible reflectance measurements) and will avoid this error.

We note that retrieving a specific cloud property from a specific channel is radiatively inconsistent (as example above) but it is theoretically possible to do a sequential optimal estimation retrieval. In this case one iterates through the channels improving the estimates of CTH/CER/COT with each step. The final result should be the same as an all channel optimal retrieval. This method is not adopted for our problem as it would be computationally less efficient.

See also the introduction in part II of this paper for a detailed definition a radiative consistency. We will not elaborate much on the issue here, as that already happened in part II.
AC: “However, the derived microphysical variables are not guaranteed to be radiatively consistent with independently derived cloud parameters, as most of the retrieval methods are separated into solar and thermal methods even though measurements in these spectral regions are not independent of parameters retrieved in the other.”

RC: p2,l39: “sees” into the cloud: A retrieval is not animate. Replace colloquial “see” with more appropriate wording.
AR: Will rephrase.
AC: “and beyond a penetration depth into the cloud corresponding to > 1 cumulative optical depth.”

RC: p2,l50: CONUS = contiguous US (conterminous is synonymous, but used much less frequently, also not by Sun et al., 2015).
AR: Will rephrase.
AC: “and contiguous United States”

RC: p2,l55-l57: This is an important statement: Cloud cover is not a good observable for trend detection because it depends on its definition (optical thickness threshold and/or reflectance threshold, sensor resolution) and instrument performance or calibration drifts. Even the CALIPSO-derived cloud information depends on which resolution is considered (because of sensitivity and SNR). A better

observable would be the optical thickness itself (or better still, the cloud radiative effect). Have the authors considered a different primary variable that is more amenable to trend detection than cloud cover? In fact, their approach of retrieving "pseudo CALIPSO optical thickness" seems to be going exactly in this direction - and in the reviewer's opinion, this would be the right way to proceed. But why then go a step backwards and convert ANNCOD into a binary cloud mask? Why isn't the retrieved ANNCOD not reported directly (in addition to the binary cloud mask outcome)?
AR: Here, we are referring to cloud cover as one of several other variables that were analysed for other retrieval frameworks in separate studies. In this study, we are only presenting a retrieval framework and an initial assessment of its data quality. We do not present trend data, but rather refer to other studies that assessed quality of other CDRs, including cloud cover but also CTH and CTP. The ANNCOD is a temporary retrieval product, from which we derive cloud cover. Cloud cover information is used to avoid processing cloud-free pixels and thus to reduce processing time. ANNCOD data are contained in ESA Cloud_cci L3U products (see Stengel et al., 2017).
AC:

RC: Related to the above [and also to material on p6]: Since CC4CL does keep cloud cover as primary variable, it should be explained whether the thresholds (table 2) vary (for example, with the specific sensor or orbit), or whether they are fixed once and for all, now that they have been optimized via the ANN technique. More importantly, do the weights as established during the ANN learning process vary? Are they a function of orbit, instrument, illumination, surface, topography. . .? Or else, are all of these dependencies incorporated in one single ANN? If so, how are commonly known problems with ANN (such as overfitting) avoided here? Using this cloud masking and thresholding technique, what is the (minimum) cutoff optical thickness, below which cloud are no longer detected? How do optical thickness detection thesholds vary with surface type and sun-sensor geometry?
AR: The thresholds in table 2 have been quantified through iterative optimisation rather than by the ANN technique (p 6, l 68). They are fixed for all sensors and orbits, and thus, as is shown in Table 2, only vary as a function of illumination and surface condition. There are no sensor specific thresholds, but we apply a simple viewing angle correction on the input satellite data. The ANN weights themselves have been trained with NOAA18 data, and we linearly adjusted input data for other satellites to better match NOAA18. The text already states which ancillary data have been used when training (p 6, l 38 - 42), including surface conditions, and also that several ANNs were produced (p 6, l 32 - 34). We are using 3 different ANNs (day, twilight, night), which reduced the overfitting problem mentioned by the reviewer. Also, overfitting was minimised by comparison with an independent test dataset while training.
We did not quantify a cutoff optical thickness as asked by the reviewer. Instead, our approach involved quantifying those threshold values for which the fit between CC4CL and CALIPSO cloud cover is best.
AC: p6, l 65 ff: "The thresholds themselves vary depending on illumination and surface conditions, namely land, sea, and snow/ice cover (Table 2), and were quantified through iterative optimization. They are fixed for all sensors and orbits."

RC: Related to the above [and also to material on p6]: The three elements of the ANN need to be described better. How well is the pseudo-CALIOP optical depth itself estimated with the ANN? Figure 2 illustrates the performance of the cloud mask after ANNCOD has be converted into a binary cloud mask. Since the ANN predicts ANNCOD and not the cloud mask itself, it should be the performance of the ANN with respect to ANNCOD that should be demonstrated here. In this context again: How is overfitting avoided?
AR: It was never the intention of creating a COD retrieval that can also be used to extract a cloud mask. We aim at creating a binary cloud mask. For a COD retrieval, we would have needed to train the full range of CALIPSO COD (approx. 0-15), but we cut off at a COD of 1 (and set all COD > 1 = 1). We assumed that CALIPSO COD values > 1 are clouds that will always be correctly detected by passive sensors. Considering that, we do not think that a comparison of ANNCOD with CALIPSO COD makes sense and thus should not be included here.
How can the non-linearities of radiative transfer be emulated with a single hidden layer?

AR: We agree that the use of at least one more layer could have improved the retrieval. However, in the CC4CL framework we used an IDL based library who does not provide more than 1 hidden layer.
What is the result for ANNCOD for the training data set as opposed to the test data set?
AR: The training dataset is only a small part of the collocation dataset. When training, the dataset was divided by 90/10 percent into a training dataset and a test dataset, i.e. we trained on 90 percent and tested simultaneously on the rest of the data (10 percent). So, the test dataset has only been tested while training. To avoid the overtraining, the training has been stopped when both RMSE (train/test) started to differ.
How is the correction for viewing angle done?
AR: We found that only a part of the whole viewing angle geometry was trained (0-35° out of up to 70° for AVHRR). We created an averaged ANNCOD with respect to each viewing angle and found a cosine-shaped dependency, which we corrected with an empirical cosine function.
$ANNCOD_{corrected}$ = ANNCOD - (1. / 12. * (1. / cos(satellite_zenith_angle * degree_to_radians) - 1.))
How many inputs does the input layer have; what are they?
AR: This depends on the illumination (day/twilight/night) and the availability of channels. We will add the input variables to the text.
AC: p6 l 37: "For the input layer, input variables are surface temperature, snow/ice cover, and the land/sea mask for all three cloud masks. Regarding sensor data, input channels are Ch1, Ch2, Ch5, Ch6, and Ch5-Ch6 for the day ANN, Ch4, Ch5, Ch6, Ch5-Ch4, and Ch5-Ch6 for the night ANN, and Ch5, Ch6, and Ch5-Ch6 for the twilight ANN."
What is the activation function? Are there bias perceptrons?
AR: Our activation function is the sigmoid function. We did include bias perceptrons.
What motivates the use of one single hidden layer, and why are there 50 neurons in it?
AR: Regarding the one hidden layer, see comment above. Will add text an number of neurons.
AC: p6 l37: "Through incremental testing, we found that 50 neurons was the value for which the trade-off between output quality and computing speed was optimal."
Is the network re-trained for every new satellite data set, or are the weights fixed?
AR: The weights are fixed, see p6 l 68 – p7 l 6. We tried to overcome the problem of having different shapes of the spectral response functions by applying linear regression coefficients (see Table 3). In a later version of the cloud mask, we applied a more sophisticated approach through using multispectral observations of IASI and SCIAMACHI. In the version presented here however, the linear regression is based on a one month triple collocation between AVHRR NOAA18, AATSR-ENVISAT, and MODIS AQUA.
How exactly were the threshold values from table 2 determined that are applied to ANNCOD to translate into cloud mask?
AR: We determined these thresholds through incremental application of a skill score analysis of the ANN cloud mask with CALIPSO for the whole collocation dataset (as a reminder, the training dataset is only a small part of the collocation data set). See previous comment, including a text change at p6, l 65 ff.
Finally, what is the quality of the thermodynamic phase retrieval, optical thickness and effective radius, depending on how close ANNCOD is to the cloud detection threshold?
AR: We did not quantify this relationship in detail. However, Figure 6 shows retrieval uncertainties of CTP, COT, and CER together with cloud mask uncertainty. The patterns do not appear to be clearly related. A quantitative analysis, e.g. calculating correlations between relative uncertainties, would certainly provide more detailed answers, but was out of this paper's scope.
AC: p18 l30: "It would also be worth investigating the relationship between the quality of retrieved variables (CTH, COT, CER, cloud phase) and cloud mask uncertainty."

RC: Essentially, the paper claims that a cloud retrieval is attempted if the optical thickness exceeds 0.4 over snow/ice during day light conditions. This would be a remarkable improvement over existing retrievals. MODIS usually does not detect clouds over snow covered areas in the Arctic unless they have an optical thickness significantly larger than 0.4 (around 7). CC4CL would be an improvement of an order of magnitude, and the question is whether the cloud retrievals would be of practical use,

especially when applying them to AVHRR instead of MODIS. The reviewer strongly believes that the only way to achieve detection thresholds on the order of 0.4 in optical thickness in snow/ice covered regions in the Arctic, one would need to use convolutional layers (i.e., use multi-pixel retrieval approaches).
AR: Please note that the ANNCOD is a *pseudo* optical depth. The threshold value, here 0.4 for daytime over snow/ice, is a relative value between 0 and 1. It does not provide any information on the absolute optical thickness value, but is rather a normalized optical thickness that attempts to fit CALIPSO measurements.

RC: p2,l78-80: "Consistency can be traded for continuity" needs clarification. Perhaps this can be done while elaborating on CDR (see comment above). This discussion will contribute to a better motivation of this study.
AR: Agreed, will clarify. See also reviewer #2 comments on the same issue.
AC: "Consistency in approach can be traded for continuity of results, and multi-platform algorithms could exploit additional data when newer sensors become available"

RC: p2,l90: "MODIS provides": is a partial repetition of material in the left column of the same page.
AR: Agreed, will delete that sentence, as it is redundant.

RC: p3,l18: "on other" > "over other"
AR: Will rephrase.
AC: "improvement over other established"

RC: p3,l20: Which "macrophysical" product do the authors have in mind here? What exactly does "radiative inconsistent" mean (supposedly, macroscopical products are inconsistent with microphysical products, but this is different from "radiatively inconsistent"; the reader is currently left to guess here). How exactly does the CC4CL approach ensure radiative consistency amongst all input satellite radiances (and all output products)? Indeed, other approaches have a cloud mask that may be independently derived from the microphysics products. Simply stating that CC4CL is "different" in this regard does not support the statement that it is more "consistent". More details are needed to add specificity.
AR: We are referring to macrophysical products such as CTT and CTH. Please see also our related answer on radiative inconsistency above, and the detailed description of the issue in part II of this paper.

RC: p3,l46: Quantify "very realistic", or just use "realistic"
AR: Will rephrase.
AC: "provides realistic estimates"

RC: p4,l35: Auxiliary > Ancillary
AR: Agreed.
AC: "Ancillary"

RC: p4,l38: Neural Network not yet defined at this point. May need the NN section prior to this statement.
AR: Will add a reference to the ANN section here. Most readers probably have at least a vague idea what a neural network is.
AC: "and as input to a neural network cloud mask (see Section 3.1.1)"

RC: p4,l73/l75: "optimal estimation", "cloud typing scheme". None of these have been described at this point in the manuscript. Sequence needs to be re-shuffled.
AR: Will add section references here. Again, these references should be sufficient, as the readers will have heard these terms before and do not require a detailed definition here to understand the following text.

AC: “The USGS data are used as a land sea mask within the optimal estimation retrieval (Section 3.3.3), as well as a land cover classificator within the cloud mask and the Pavolonis cloud typing scheme (Section 3.3.2).”

RC: p5,l1: “were” > “are”
AR: Cannot find “were” in that sentence.

RC: p5,l1: Reference and/or data source (link) needed for CALIPSO product
AR: Cannot find the CALIPSO product in that sentence. Maybe there is a linenumber mismatch?

RC: p5,l68-l72: multiple acronyms need to be introduced prior to first use.
AR: Most acronyms in that sentence were introduced in the first paragraph on page 2.
AC: “…Clouds from AVHRR Extended (CLAVR-X) (…) Global Cloud and Aerosol Dataset Produced by the Global Retrieval of ATSR Cloud Parameters and Evaluation (GRAPE) …”

RC: p5,l79: The outcomes of the study should be at least summarized here. Also, the use of “round robin” may not be ideal for an international readership as it is a cultural reference (British/American) that may not be commonly known. Consider paraphrasing the technique instead.
AR: Will rephrase.
AC: “was chosen from three competing algorithms in a “round-robin” (i.e. each algorithm is tested against all other algorithms) analysis. All algorithms have proven their maturity for deriving the considered cloud parameters (cloud cover, liquid and ice water path, cloud top height) from AVHRR and MODIS data (Stengel et al., 2015).”

RC: p5,l98: Do these channel numbers refer to the CC4CL IDs from table 1?
AR: Yes, will clarify.
AC: “The albedo of snow/ice covered pixels is set to globally constant values of 0.958 (Ch1, CC4CL ID as in Table 1), 0.868 (Ch2), 0.0364 (Ch3), and 0.0 (Ch4),”

RC: p6: Cloud detection: See multiple comments above (following p2,l55-l57 comment). Also: Are there any convolutional layers included in the approach? This would have allowed capitalizing on the context of a pixel.
AR: We did not add any convolutional layers in the ANN. See above comments regarding cloud detection and the ANN.

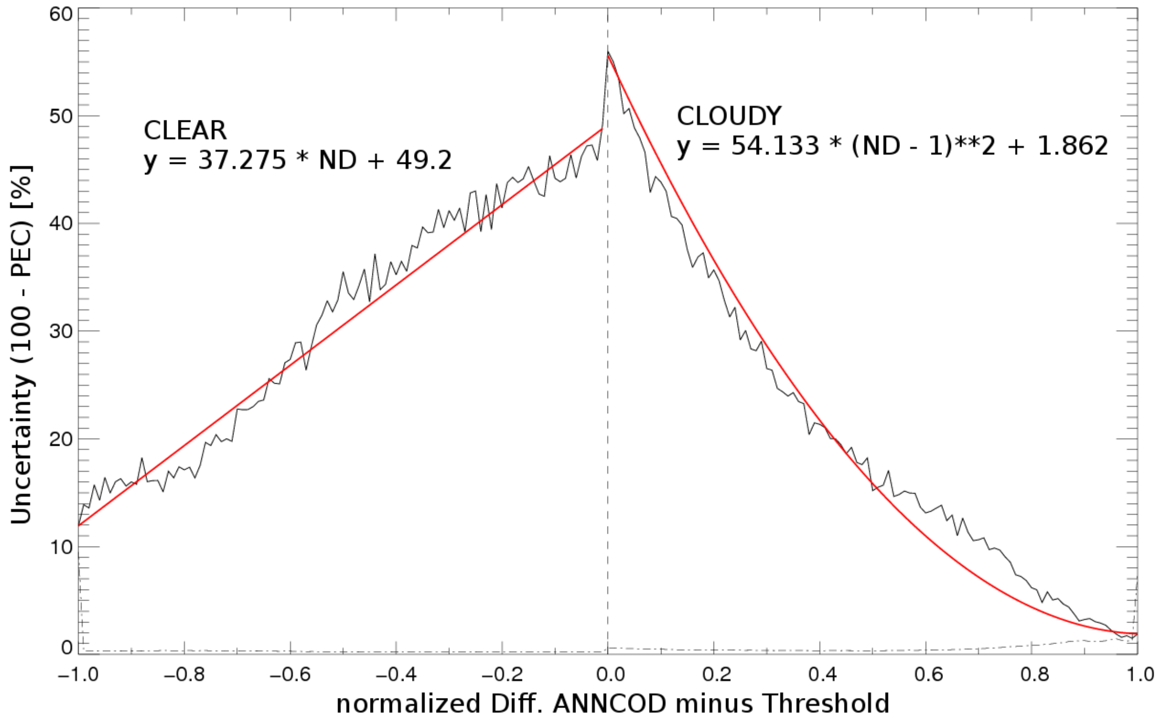
RC: p7,table 3: How was the regression done - based on radiance or irradiance, based on counts? Based on brightness temperature (for IR channels)? The offsets seem rather large; what is the explanation for significant offsets?
AR: The regression coefficients were calculated based on reflectance and brightness temperature data. The offsets might be a result of imperfect collocation, relative calibration differences, and mainly differences in spectral response functions. It is difficult to quantify the contribution of each, but spectral response probably explains most of the offset.

RC: p7,l49: VIIRS algorithm is used: What is the purpose of this statement? If it is kept, this needs to be elaborated (what does the VIIRS algorithm do differently). Also, there are various other algorithms that are improved over the heritage algorithms, which would probably all need to be mentioned here (or at least a subset thereof).
AR: This paragraph is a brief summary of the Pavalonis algorithm performance. Not surprisingly, it performs better if more spectral channels are used for cloud typing. The sentence emphasizes the generic limitation of using only AVHRR heritage channels, which does not only affect cloud detection or optimal estimation, but also cloud typing. We do not think that other algorithms using other data than the heritage algorithms need to be elaborated or mentioned here (however, we do so elsewhere). It is obvious that they perform better if using more channels, but that is not the point here.

RC: Figure 2: This is just one example where labels are too small, and are too pixelated. Generally improve the figure quality and enlarge labels. About the content: It is rather hard to interpret this figure. The x-axis is “normalized”. Does that mean that the difference of the ANNCOD-retrieved value and the threshold from table 2 is divided by the threshold value itself? Does “x=0” mean that the retrieved optical thickness equals the threshold per table 2? Does the “CLEAR” label refer to CALIPSO? For x=-0.2, we find an uncertainty of 40%. Does that mean that CC4CL misclassifies clear pixels as “cloudy” in 40% of cases?

AR: As mentioned in the text, the x-axis is normalized, i.e. the difference between ANNCOD and the threshold was divided by the threshold. Yes, x=0 means no difference between ANNCOD and the threshold. Again, please remember that this is a *pseudo* optical thickness. CLEAR means that the ANN cloud mask defined a pixel as cloud free. It shows that we need different equations to quantify uncertainty for clear and cloudy cases. The text also explains how uncertainty is calculated: 100 – PEC [%], with PEC = the ratio between all correctly classified pixels and the number of all pixels analysed. Also, if x=-0.2, CC4CL misclassifies cloudy pixel as 'clear' in 40% of the cases with respect to CALIPSO. The uncertainty defines the misclassification of CC4CL compared to CALIPSO for a certain combination of ANNCOD and the threshold used.

See below Figure 2 with larger labels and annotations. We also increased labels for Figures 3-7 (see responses to other reviewers).



RC: p8,l70: Why are largest uncertainties found for opaque clouds? Also, figure 10 does not show quantitative evidence for this statement - colors are harder to interpret than numbers on a graph. Can this somewhat counterintuitive statement be supported by a more succinct graph?

AR: We are referencing the wrong figures. Will correct.

AC: “COT uncertainties increase with COT magnitude, and largest uncertainties are found in cases of opaque cloud coverage (Figure 4 middle and Figure 6 topright).”

RC: p9,l3-5: :Validation is show for . . . rather than: Unclear. What is the difference between CTH and “its” retrieved value?

AR: CTH is a derived variable, i.e. derived from CTP, which is the retrieved value. Will clarify.

AC: “The validation is shown for comparisons of CTH (derived from CTP) rather than CTP (retrieved) to enable…”

RC: p9,l13: “TOA radiation is the *sum total* of emission and scattering throughout the atmospheric column” - please formulate this more accurately: What is a “sum total” of two processes? Also, the next paragraph more or less paraphrases Platnick’s vertical weighting function paper where this is formulated more accurately, and where the concept of a weighting function is well explained. Please cite that paper and use similar terminology here. As for multi-layer clouds, there is a fairly new paper by Wind, Platnick et al. (http://journals.ametsoc.org/doi/abs/10.1175/2010JAMC2364.1), but it is probably not applicable to this paper here because of the channel selection.
AR: Will clarify.
AC: “However, TOA radiation is the product of emission and scattering processes throughout the atmospheric column (Platnick, 2000).”

Platnick, S. (2000), Vertical photon transport in cloud remote sensing problems, J. Geophys. Res., 105(D18), 22919–22935, doi:10.1029/2000JD900333.

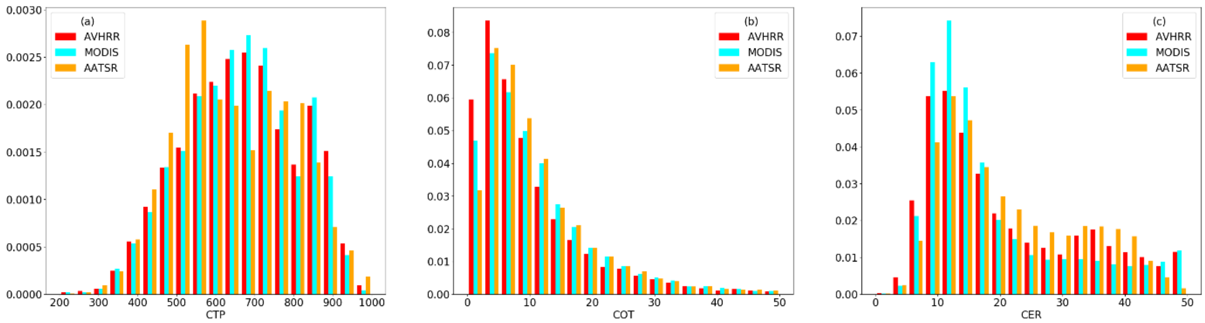
RC: p10,l1: How is the CTH adjustment done if the cloud base is not known? Where does cloud base (or cloud geometrical thickness) information come from?
AR: We approximate the observed temperature as emitted from one optical depth into the cloud. Assuming the cloud is vertically homogeneous with a constant lapse rate Γ, we can write the thickness-corrected CTT as,
$T_{cor} = BT(\lambda) + \Gamma / (\sigma N)$,
where BT is the observed brightness temperature, σ is the cloud particle cross-section, and N is the cloud particle number concentration. Using the observations at 11µm and 12µm provides two simultaneous equations in T_{cor} which can be solved, using σ values for a LUT.

RC: p10,l9: Does this statement about sectors refer to figure 9? Please match figures and text, otherwise figures become “orphans” that are not tied to the manuscript.
AR: Will add figure references.
AC: “CC4CL correctly classifies all pixels as cloud covered, with a few exceptions in sectors 3 and 4 (Figures 8 and 9).”

RC: p10,l14: Please define what is meant by “surface” in this case.
AR: Will clarify.
AC: “In the case of a (semi-)transparent cloud top layer, multiple surfaces (several cloud layers, Earth surface) contribute to the observed satellite data.”

RC: p10,l16: insert “a” before “single-layer”
AR: Agreed.
AC: “For a single layer, optically thick (COT > 1) cloud,…”

RC: Figure 7: please enlarge labels, as well as histograms; it is hard to compare the retrievals quantitatively otherwise. Also: It would really help if histograms were shown separately for snow-covered areas as opposed to dark surfaces. It is expected that retrieval quality would differ significantly depending on the surface conditions.
AR: We will enlarge labels and histograms. However, we do think it is sufficient to show histograms for all surfaces combined to make our point that there are differences between retrievals, which is also supported by the statistics.



RC: p12,l41: “performance of existing algorithms” What are the “existing algorithms” that CC4CL? Has the manuscript shown that these existing algorithms perform less well than CC4CL.
AR: Will remove the subordinate clause.
AC: “In general, the quantitative and qualitative agreement between CC4CL and CALIOP CTH is impressive.”

RC: p12,l88: “AVHRR” > “for AVHRR”
AR: Agreed.
AC: “and AATSR data than for AVHRR”

RC: p12,l89: Should “continually” be replaced with “consistently”? Unclear what this statement means. If it were “consistently” it would be more clear, but the word order should be fixed: “The CC4CL phase identification does not agree with any of the three CALIOP cloud flags consistently, which is reasonable given . . .”
AR: Agreed, we will rephrase as suggested.
AC: “The CC4CL phase identification does not agree with any of the three CALIOP cloud flags consistently, which is reasonable given . . .”

RC: p13,l19/20: “. . .insensitive to the specific instrument evaluated, such that the merged data set is sensible”. What does this statement mean? The paper does not actually present a *merged* data set, or was that the actual intent of the paper? It does evaluate collocated overpasses from different satellites, but these are not merged in the sense of a CDR. Please remove the statement about “merging” data sets unless this was the actual intent of the paper (in which case it would need to be modified considerably).
AR: We will remove the statement about “merging” datasets, which was once foreseen in the project but has not been done at the moment this paper was written.
AC: “In general, the retrieved values are insensitive to the specific instrument evaluated. Absolute…”

RC: p13,l31: “disagree nonetheless”: They disagree despite their channels are fairly close? Can this be re-phrased? The whole paragraph is a bit roundabout. There’s a 30-40% difference in reflectance, but “their” retrieval values are “much more similar”? Please make this statement more precise. “The difference to AVHRR and MODIS is largest for CER” - does this statement refer to AATSR again?
AR: Will rephrase, as these statements are definitively hard to understand. Yes, the last sentence refers to AATSR.
AC: “Also, even though spectral response differences are largest between MODIS and AVHRR (which results in a reflectance difference of up to 30–40 % (Trishchenko et al., 2002)), their retrieval values are much more similar. The difference between AATSR and both AVHRR and MODIS is largest for CER, so microphysical variables, which are derived from reflectance data only, appear to be most affected.”

RC: p13,l39: The t-test needs to be explained in much more detail. What is H0, what is mu1, what is mu2? Are we talking about the covariance between two data sets, which is assessed using the t-test approach? If so, are the data from the two different data sets (supposedly this is what “mu1” and

“mu2” refer to) re-gridded to one common grid before comparing them? The premise of this statement deserves at least one paragraph, if not half a page.
AR: This is a very basic t-test, using a well-defined symbology. It is a test for significance of the difference between the mean values of two populations (i.e. $\mu 1$ = mean of population 1, $\mu 2$ = mean of population 2). The data were indeed re-gridded to a common grid, which is all explained in section 2.3.
AC: “The differences between mean values ($\mu 1$ and $\mu 2$) are almost always significant (t-Test p-value < 0.1, H_0: $\mu 1 = \mu 2$).”

RC: p13,l45: “spatiotemporally collocated sensors”: The sensors are not collocated - is that the point of the statement? Or is this an explanation why the t-test “fails? What does “non-significant” t-test mean? Could the strictness of the comparison be relaxed by gridding the retrievals to a coarser common grid before making the inter-comparison?
AR: The sensors are not collocated, but the data are. And the collocation should minimize differences due to observation times and observation area. The significance level is now mentioned in the correction above, but can also be found in the caption of Table 6. As said above, the data were re-gridded.
AC: “…when driven with spatiotemporally collocated satellite data obtained from three different sensors.”

RC: p13,l57: “depending on the user’s application” - this needs to be clarified. For which applications can they be used interchangeably? Could a combined AVHRR and MODIS cloud data record constitute a CDR (would it meet the requirements)? As stated above, the manuscript does not actually “merge” data sets in this way, but more specificity would be helpful here.
AR: We added references to give examples. We do not think that the AVHRR and MODIS cloud data record should be seen as one continuous, consistent data record. Rather, AVHRR provides the opportunity of long-term data coverage back to 1982, providing data that are at least comparable to MODIS. That certainly excludes local analyses, but rather refers to continental to global applications.
AC: “depending on the user’s application, such as model validation, data assimilation applications, or climate studies in general (Liu et al., 2017, Yang et al., 2016).”

Liu, C., R. P. Allan, M. Mayer, P. Hyder, N. G. Loeb, C. D. Roberts, M. Valdivieso, J. M. Edwards, and P.-L. Vidale (2017), Evaluation of satellite and reanalysis-based global net surface energy flux and uncertainty estimates, J. Geophys. Res. Atmos., 122, 6250–6272, doi:10.1002/2017JD026616.

Yang, Qinghua, et al. "Brief communication: The challenge and benefit of using sea ice concentration satellite data products with uncertainty estimates in summer sea ice data assimilation." The Cryosphere, vol. 10, no. 2, 2016, p. 761.

RC: p13,l77: “we see that COT uncertainty scales with COT itself”: this is not shown in the manuscript. If it is, please refer to a figure or section.
AR: As mentioned above, it is shown in Figure 4 middle and Figure 6 topright.
AC: “we see that COT uncertainty scales with COT itself (Figure 4 middle and Figure 6 topright)”

RC: p13,l79-l88: Consider re-writing this section; simplify and use literature references; most of these observations have been documented before (large COT uncertainty as reflectance approaches asymptotic value; large uncertainties for bright surfaces).
AR: Will simplify and add references.
AC: “CC4CL COT values are at times unnaturally large, and the associated uncertainty reflects that. Also, it highlights under which conditions the optimal estimator converges to a solution with a relatively large divergence from the measurements, which here are associated with optically thick clouds or underlying snow/ice cover (see also Kahn et al., 2015, Wang et al., 2011). COT and CER

uncertainties are clearly largest, and reflect the limited information available with which to retrieve these values. For further possible explanations due to assumptions and limitations within the methodology applied, please see part II."

Kahn, B. H., M. M. Schreier, Q. Yue, E. J. Fetzer, F. W. Irion, S. Platnick, C. Wang, S. L. Nasiri, and T. S. L'Ecuyer (2015), Pixel-scale assessment and uncertainty analysis of AIRS and MODIS ice cloud optical thickness and effective radius, J. Geophys. Res. Atmos., 120, 11,669–11,689, doi:10.1002/2015JD023950.

Wang, C., P. Yang, B.A. Baum, S. Platnick, A.K. Heidinger, Y. Hu, and R.E. Holz, 2011: Retrieval of Ice Cloud Optical Thickness and Effective Particle Size Using a Fast Infrared Radiative Transfer Model. J. Appl. Meteor. Climatol., 50, 2283–2297, https://doi.org/10.1175/JAMC-D-11-067.1

RC: p15, l11: "otherwise are" > "otherwise they are"
AR: Will rephrase.
AC: "otherwise they are"

RC: p15,l15: "may it stem" does not work in English; consider "whether it stems from. . . or"
AR: Will rephrase.
AC: "whether it stems from a cloud or the Earth's surface"

RC: p15,figure 11: The table below the cross section is too small. Also, what happened at lat=61? Why do the active imagers pick up a cloud where CALIPSO does not?
AR: Unfortunately, the table itself cannot be increased due to space limitations and a bug in the Python library applied to produce the table. The colours show cloud phase. Cloud type numbers are not as important, and we could have removed them as for Figure 15. At latitude 61°, we see that there are broken cloud fields in the area, which might have appeared in the sensor's field of view but not in CALIPSO's.

RC: p16,l8: consider "a conscious decision was made to [deliberately] trade. . ."
AR: Will rephrase.
AC: "For ESA Cloud_cci, a conscious decision was made to trade spectral information for time series continuity."

RC: p16,l19: "on a first view" > "at first glance"
AR: Will rephrase.
AC: "At first glance, estimates of…"

RC: p18,l59: "synergic" > "synergistic"
AR: Will rephrase.
AC: "exploits synergistic capabilities of several EO missions"

RC: p18,l95: "accurate and precise": These two were not discussed separately. Where was this done? If not, please clarify this statement.
AR: We will remove precise, which stands for a low standard deviation of errors (not shown here). The results are accurate due to the relatively low bias.
AC: "optically thick cloud retrievals are very accurate when compared against CALIOP (bias < 240 m)"