Response to Review

Referee Comment

Effectiveness of Cirrus Detection with MODIS Cloud Mask Data

by Nguyen Huu et al.

Referee #1

In my opinion, this paper is not yet suitable for publication. A stated objective, the data analysis methods, and the conclusions seem to be flawed. In my opinion, the authors have not adequately addressed the reviewers' concerns, and in revision, have managed to add additional confusion that further reduces the quality of the manuscript. The results described in the manuscript indicate that the statistical comparisons between MODIS and CALIOP are not very good, yet it is claimed without providing evidence that MODIS provides a reliable cirrus mask when compared to CALIPSO. A significant shortcoming is that the classification schemes for the two sensors are not well described, and the data filtering and matching procedure may be inadequate for conducting a fair comparison between the two sensors during daytime and nighttime, particularly with the inclusion of very thin stratospheric clouds only detected by CALIPSO. The presentation of the results is confusing and not well explained. I am concerned that the results presented here may misrepresent the accuracy and utility of the operational MODIS cloud products. If further consideration is to be given for the publication of this manuscript in ACP, I highly recommend that it be sent to someone from the operational MODIS cloud team for their opinion as they would be able to better interpret the MODIS results presented here.

Major concerns

1. As stated in the abstract, the objective of this paper is "to determine if a MODIS product exists that detects cirrus with the same accuracy as CALIOP". This objective seems off base since several publications have shown that the CALIOP active sensor is more sensitive to cirrus than the MODIS passive sensor. If MODIS is less sensitive, then obviously, it will be less capable of detecting some cirrus. Therefore, the authors should revise the objective stated in the abstract. Something like that stated in the introduction on line 73 would be more reasonable, i.e. "Our objective is to determine how well the MODIS products can be used to identify cirrus clouds compared to CALIPSO." Another objective stated on lines 74-75 does not seem to be addressed in the paper (that I could find), i.e. "we aim to assess whether MODIS cloud detection tests used to generate MYD35 operational data can be re-used for a time-effective masking of cirrus." Beside the fact that the meaning of 'time-effective' is unclear, there is no evidence presented in the manuscript that the temporal consistency of the MODIS products was evaluated. Therefore, this objective should either be removed or supported with data.

The objective in the abstract has been revised to reflect the focus of the study better.

Regarding the term "time-effective," our intention was to emphasize the practical approach of assessing how much can be extracted from the existing MODIS cloud detection tests (e.g., those used in the MYD35 product) without developing a new cirrus detection algorithm from scratch. The focus is on evaluating the potential of current operational algorithms with minimal additional processing. We have revised the text to clarify this point.

2. Unfortunately, the authors did not adequately clarify and defend their definition of 'cirrus' as used in this study, nor how consistent that definition is for the two data products being compared, as requested by the reviewer(s).

We have now addressed this point by adding a detailed explanation of the physical definition of cirrus clouds and clarifying how cirrus are detected and defined within both CALIPSO and MODIS data products.

The evaluation of the 6 spectral tests for cirrus may be of modest interest to algorithm developers as they provide performance metrics against CALIOP in a relative sense. However, these tests are not meant to stand alone for cirrus detection. From a practical standpoint, the ATC test which combines the results of all six tests could be a more useful gauge as to how well the MODIS data product can be used to identify cirrus overall, provided that the population of data being tested is evaluated in context with the total population of cloudy pixels. Unfortunately, from what I can tell, this isn't done. It isn't clear if the ATC collection of six tests encompass all possible cirrus pixels determined by the mask or if there are other information contained in the MODIS data products that could lead to a different population. The reader should not have to guess at this. This is important because if there are other cloudy pixels as determined by the mask for which there are other indicators (that these pixels may be cirrus (e.g. cloud phase and height), then the statistical comparisons don't have much meaning and could even be misleading regarding the accuracy and utility of using the MODIS data products to discern cirrus. Is there a population of cloudy pixels for which it is unknown whether these could be cirrus or not which? Compared to the cirrus screening used here, would the population be the same if all cloudy pixels were included as determined by the mask that are also determined to be ice phase, either anywhere in the vertical column or above some height level? Are such other tests not possible due to failure rates in the cloud optical property and/or height algorithms?

The parameter called ROP (rate of observations performed) is not meaningful to me as it is defined for a specific test to be the fraction of observations evaluated in the test to the total observations. The problem is that it isn't explained what population the total observations represents? Is it meant to be all cloudy pixels, or all cloudy pixels evaluated with the six tests, or something else?

If concerns relate to the fact that our analysis focuses on six specific tests out of a broader suite of MODIS Cloud Mask tests and that we may not have fully addressed the possibility that other unexamined tests could also contribute to cirrus detection, we would like to clarify that the selection of these six tests was intentional and grounded in prior literature, which highlights their particular physical relevance and sensitivity to cirrus detection, especially in identifying high, optically thin ice clouds.

Additionally, of the calculated statistics were based on all available observations (pixels), without excluding any based on their classification as cirrus clouds, other cloud types, or clear sky, according to the data sources used. In other words, every pixel within the dataset was considered, regardless of whether it was categorized as cirrus, another cloud type, or clear.

Furthermore, to account for other relevant parameters, we also presented ISCCP tests incorporating factors such as optical depth and cloud top pressure.

I hope that the answer explains the issue addressed in the questions raised.

Furthermore, regarding the CALIPSO data, it isn't clear why the optically thinnest clouds that are impossible for MODIS to detect, particularly those in the stratosphere, are included in these comparisons. Stratospheric ice clouds are important for atmospheric chemistry, but MODIS is not a suitable sensor to study stratospheric clouds that it cannot detect. The CALIPSO products themselves are also not consistent between day and night due to inconsistencies in the lidar sensitivity. So why are the optically thinnest clouds included, particularly those in the stratosphere? The authors need to discuss this and justify the rationale for including stratospheric clouds detected by CALIPSO. At the very least, the statistical comparisons should be conducted, or stratified, using data with and without the thinnest CALIOP clouds. It doesn't seem that this has been done (more on ths regarding figure 10 below). This would provide perhaps a fairer comparison and more informative performance metrics, but certainly a more informative comparison across daytime and nighttime where the CALIPSO sensitivities are much different.

Clarified in the manuscript. We considered all Cirrus clouds detected by CALIPSO, regardless of the COT. Clouds above the tropopause, namely the polar stratospheric clouds (PSCs), were NOT included. They state a separate feature type category in the CALIPSO data. Hence, we were able to filter them out as one of the first steps during data reduction.

We agree that COT for PSCs is low (<0.3; Noel et al. 2008, doi: 10.1029/2007JD008616), and comparable to optically thinnest Cirrus. The value coincident with the cloud detection limit of MODIS (~0.3-0.4; Holz et al. 2008, doi: 10.1029/2008JD009837). The chance of MODIS data being 'contaminated' by PSC is, therefore, extremely low, if any.

Additionally, PSCs are relatively rare phenomena, limited to polar regions and summer conditions. Based on that, we conclude the PSCs had no impact on our results. For the same reason, Fig.10 and the corresponding discussion distinguish no special case for PSCs but only stratify data for various COT ranges of Cirrus.

3. In the discussion section, it is stated that this study proved that MODIS ready-to-use cloud mask product can be used for producing a reliable cirrus mask, however, it is totally unclear how this conclusion is arrived at. By what metrics levels is the mask deemed to be reliable and how are those levels of 'reliability' determined? The 'goodness of fit' parameters shown in table 3 for example are not particularly impressive, especially for nighttime as pointed out in the manuscript. Whatever potential the paper has up to this point really becomes confusing and seems to fall apart near the end when figures 8-10 are introduced.

Clarified in the manuscript . The 'reliability' term only referred to daytime conditions, and the conclusion was supported by numbers: overall accuracy of Cirrus detection at 73% (kappa 0.5), probability of detection > 80%, and false alarm rate of 35%. Indeed, the night-time performance is significantly poorer, and cannot be deemed reliable (although overall accuracy is of 60%, the kappa coef. of 0.2 indicates a random agreement, rather than an actual effectiveness of the Cirrus detection).

Regarding figure 8: This shows a remarkable inconsistency (factor of 4 difference) between the daytime and nighttime cirrus coverage as determined from MODIS that I can't understand. Is this day/night difference representative of the difference in high cloudiness as determined from MODIS in other studies or is this a result of the cirrus screening procedure adopted in this study? In other words, are the operational MODIS products really this inconsistent with respect to the ability to identify high clouds consistently during daytime and nighttime?

The figure reports Cirrus frequency day and night based on the ATC approach developed in this study. The inconsistency is true and results from the very low Cirrus detection skill of the ATC approach.

MODIS thermal infrared-only tests for high clouds in MODIS operational cloud mask product are insufficient to detect Cirrus night-time (as compared to CALIPSO) effectively. The most notable is the lack of a unique MODIS 1.38 μm channel, a 'cirrus band', introduced specifically to detect high ice clouds (Gao and Kaufman 1995, doi: 10.1175/1520-0469(1995)052<4231:SOTMCF>2.0.CO;2).

Additionally, at night, the reference sensor (CALIOP lidar) detects more thin cirrus clouds due to the absence of solar background noise and increased nocturnal convective activity, which enhances cirrus formation. These combined factors explain why MODIS shows significantly reduced cirrus detection rates at night compared to daytime.

Regarding figure 9: This figure shows a comparison between the MODIS and CALIOP cirrus cloud cover for daytime and nighttime. First, it isn't clear what the individual points represent as this is not explained in the text. Are these annual regional means? Second, the daytime comparison is awful (MODIS considerably oveestimates cirrus cover compared to CALIOP), while the nighttime comparison is much better, which seems to contradict the discussion regarding the goodness of fit analyses that imply much more significant issues at night than during daytime. It's acknowledged on line 466 that "MODIS will inevitably miss a significant portion of cirrus clouds due to its lower sensitivity. This comparison offers valuable insights into the practical efficiency of the MODIS instrument." Yet, there is no attempt to explain the large daytime overestimates (false alarms) in cirrus cover from MODIS. It's impossible to know whether these are MODIS errors or the result of something related to the obscure definitions and confusing analysis methods undertaken in this study.

To clarify, each point in Figure 9 represents the mean annual cirrus cloud amount within a 5-deg grid box.

Regarding the second point, we did address this by noting in the manuscript: "Although the majority of fit metrics show improved performance during the day, the high number of false alarms ultimately results in the nighttime fit being more accurate when cirrus coverage is examined in the subsequent analysis."

The seeming discrepancy between the single observation-based metrics (e.g., POD, FAR, kappa) and the aggregated cirrus cloud cover comparison can be explained by the difference in scale between these analyses. While the kappa coefficient indicates that MODIS achieves better pixel-level agreement with CALIOP during daytime (kappa = 0.46) than at night (kappa = 0.19), the scatterplots of aggregated cirrus cover reveal a better linear relationship at night. This is likely due to MODIS generating more false alarms during the day (FAR = 34.86%) compared to night (FAR = 6.90%), leading to an overestimation of cirrus cover when aggregated. At night, MODIS is more conservative in cloud detection (lower POD). However, the lower false alarm rate

results in better agreement in total cirrus cover with CALIOP, despite the weaker pixel-level correspondence.

Our analysis shows that a significant portion of the daytime false alarms (approximately 28 out of the total 35% FAR; Table 3) can be attributed to the so-called "inherited" detections in the MODIS ATC procedure. These detections are primarily linked to the 1.38 µm cirrus test, which is commonly regarded as the best spectral test for identifying high-level clouds. However, while this test delivers a high POD for cirrus detection, it is also known to generate a substantial number of false alarms, especially during daytime when sun-glint and surface reflection can influence the signal.

This behaviour is indeed reflected in both our pixel-level analysis (POD/FAR metrics) and the aggregated cloud cover comparison, where daytime MODIS tends to overestimate cirrus cover relative to CALIOP. Importantly, since the 1.38 µm channel is not used in MODIS nighttime retrievals, this overestimation pattern largely disappears at night, which is consistent with the improved FAR and the more accurate agreement with CALIOP during nighttime conditions.

We added this clarification to the manuscript to improve the understanding of the limitations related to the MODIS cirrus detection approach.

Regarding Fig 10: This figure shows the detection accuracies as a function of COT. The authors don't clarify which sensor the COT is from. One might assume this is from MODIS since the CALIPSO signal saturates near a value of 4 (any values beyond that, if they exist, would have no meaning). However, the MODIS cloud mask misses many of the thinnest clouds that CALIPSO can detect (as shown in this study!) and the MODIS cloud optical property algorithm has a somewhat high failure rate for the thinnest of clouds that are detected. Therefore, if the results in fig 10 are with respect to MODIS COT, it isn't clear how representative the values are at the low end of COT when matched with the MODIS pixel populations used to compute the statistical comparisons agains CALIPSO. Did this population all have corresponding successful COT retrievals? Or, is CALIPSO COT used in Fig 10? We don't know! Also, it's stated that "The most noticeable changes occur at COT values close to 10, though these may be influenced by the sample size, as the occurrence of cirrus clouds with a COT near 10 is limited or may represent a misclassification by CALIOP." It's difficult to know if this is an interesting finding or not since there is little discussion or attempt to explain it. I would like to know how the higher COT's could be associated with CALIOP misclassifications? How can that be? It's impossible to understand without a better explanation of the classification schemes adopted here for the two sensors.

Clarified in the manuscript. COT values for the analysis were based on CALIPSO data.

Regarding the noticeable changes at COT values close to 10, this refers to a small number of cases where optically thicker layers might have been classified as cirrus in CALIOP data due to limitations in classification(i.e. cirrus-like top of a strong cumulonimbus cloud).

Additionally, higher COT values may be associated with uncertainties stemming from CALIOP's limitations (Winker et al., 2024). Specifically, at such high optical depths, lidar signal attenuation often prevents accurate detection of lower cloud layers, leading to overestimating COT. As noted in the manuscript, even small uncertainties in the assumed lidar ratio can significantly affect the accuracy of optical depth retrievals. Additionally, there are potential issues with cloud phase classification, particularly in the presence of horizontally oriented ice crystals, which may lead

to misclassification of thin layers as optically thicker clouds. We will expand the discussion in the manuscript to address this aspect.

Minor concerns:

The title doesn't make sense to me. Cirrus detection is done with a cloud mask algorithm rather than cloud mask data. The data are the result of applying the algorithm. I suggest that you consider modifying the title. Here are two suggestions:
Comparison of Operational MODIS Cirrus Cloud Detections with CALIPSO data
Evaluation of the Operational MODIS Cloud Mask for Detecting Cirrus Clouds

We adopted the second suggested title: *"Evaluation of the Operational MODIS Cloud Mask for Detecting Cirrus Clouds"* in the revised version of the manuscript.

Line 8-9: I suggest rephrasing to the following: "Our objective was to determine how well the operational cloud mask from the MODIS Science Team can be used to infer the presence of cirrus clouds relative to data products derived from the highly sensitive CALIOP instrument by the CALIPSO Science Team."

Corrected, as suggested.

Line 28-31: Suggest the following: "Globally, it's been estimated that cirrus clouds have a net warming effect of 35.5 Wm-2 (Campbell et al., 2016; Kärcher, 2018;Lolli et al., 2017; Oreopoulos et al., 2017) in part because they trap and reduce outgoing longwave radiation more efficiently than they reflect solar radiation back to space."
The following are probably not the most appropriate original citations for these phenomena but do provide examples. Therefore, it is appropriate to add citations for the original findings or cite in the following way:
Line 61: (e.g. Kortaba and Nguyen Huu…)
Line 67: (e.g. Heidinger and Pavolonis…)

Corrected, as suggested.

Section 3 and later: Consider using references to the 5-degree areas as 'regions' rather than 'pixels'
We have added a clarification in the text regarding using the term "pixel" in our study. In this context, we refer to a "pixel" as a 5-degree grid cell representing a spatial unit of analysis. We hope this clarification addresses your concern.

Line 85: change 'in the range of' to 'at least'
Corrected, as suggested.

Line 107: clarify what a 'middle threshold' is or remove
Corrected, as suggested.

Line 163: change 'other' to 'passive'

Corrected, as suggested.

Line 296: It isn't clear what P.P. means. Please clarify.

Corrected, as suggested.

Line 297: Please briefly describe finding from Kortoba and Nguyen-Huu with regards to what you mean by 'sampling frequency' and how it affects the estimate of cirrus cloud fraction. Are you referring to occasional missing time periods in the CALIPSO record? If so, maybe it is more clear to say "can vary significantly due to occasional gaps in data availability due to instrument or spacecraft issues."

In this context, reference to "sampling frequency" was not intended to imply gaps in the CALIPSO record due to instrument or spacecraft issues. Rather, we referred to the limitations described in Kotarba and Nguyen-Huu (2022), who examined the spatial and temporal mismatches between the CALIPSO lidar observations and ground-based SYNOP cirrus reports. Specifically, they demonstrated that the narrow footprint and orbital characteristics of the CALIOP sensor result in relatively infrequent co-locations with SYNOP observations, leading to a very low match rate (0.022% of SYNOP reports paired with CALIPSO overpasses).

This sparse sampling directly impacts cirrus cloud fraction estimates. Since CALIPSO samples only a narrow swath along its track, many cirrus events visible to surface observers within a broader hemispheric view are missed, potentially leading to bias in satellite-derived cirrus occurrence statistics.

Line 300: should read '…detected at nighttime are 2-3 times higher those detected during daytime'
Corrected, as suggested.

Line 315-320: This argument does not make sense to me. CALIOP is more sensitive at night, which means it should detect more thin clouds (higher cloud cover) which would lower the average COT, relative to daytime. It seems more likely to me that your analysis that indicates higher nighttime COT from CALIOP is either due to a real diurnal change in the nighttime cirrus COT, a retrieval algorithm artifact, an artifact of your screening method, or some combination of all of these.

We have incorporated the necessary adjustments for greater precision.

Line 411: What do you mean by 'reliable'? By what measure? For what applications are these measures deemed to be reliable and how is that determined? These questions should be answered if you are going to make such a definitive and broad statement. I suggest that you back off a bit and simply focus on summarizing the statistical findings.

We replaced the term "reliable" with "accurate" to reflect the context better. Our goal is to create a cirrus cloud mask (Ci) that can be used to analyze long-term trends based on MODIS data. This clarification will help to focus on the measurements' accuracy rather than making broad statements about reliability.

Line 444-445: This statement seems overstated also and is more likely an assumption. What evidence have you shown that supports the contention that the detection accuracies you find are high enough to accurately monitor climate quality long-term changes in cirrus clouds?

While this study is based on one year of data, the number of observations was substantial. Our results indicate that the ATC test provides a relatively high probability of detection during daytime and acceptable agreement with CALIOP, but exhibits limitations at night and for optically thin cirrus. We now emphasize that MODIS, due to its extensive temporal coverage and spatial resolution, has the potential to contribute to cirrus climatologies.

We acknowledge that our statement regarding climate-quality monitoring may have been overstated. In long-term studies, the most critical factor is not necessarily the absolute accuracy of detecting individual cirrus clouds in each observation but rather the systematic stability of the detection process over time, including the consistency of potential biases.

MODIS, with its continuous and global observations, provides a unique dataset for trend analysis. While we recognize that MODIS has limited sensitivity to optically thin cirrus compared to CALIOP, the key aspect is that this detection threshold has remained stable over time due to the instrument's and algorithm's consistency. Therefore, even with a lower absolute detection rate, MODIS remains valuable for assessing spatial and temporal variability in cirrus cloudiness.

We also acknowledge that in the context of detecting subtle trends (e.g., changes of 0.5% per year), the lower kappa values and false alarm rates could introduce uncertainties or mask weak signals. However, the long-term stability of the MODIS instrument and its cloud detection algorithms mitigates this concern to an extent, as any systematic bias would likely affect the full-time series uniformly.

Nevertheless, we agree that further multi-year validation and intercomparison studies would be beneficial to strengthen the evidence for reliably using MODIS data for monitoring long-term climate-quality changes in cirrus clouds. In such applications, the key factor is whether the bias remains stable over time.

We have clarified this aspect in the revised manuscript.

Response to Review

Referee #2

General comments:

The Introduction is now more clear, more focused, and much improved. The simple example added to the discussion of bootstrap sampling is helpful and will make it clear to the community why bootstrapping was used to compute the performance metrics. Overall, the manuscript is much improved but a few additional changes are necessary to be ready for publication.

Specific comments:

Line 168 – the criteria for classification as Category 6 (pressure at cloud top less than 440 mb and nonopaque) should be mentioned here so the reader understands how this class is selected.
Corrected, as suggested.

Lines 174-175 – When the CAD algorithm gives a CAD score near zero, the algorithm finds the probability of aerosol and the probability of cloud are nearly equal. This is most often because the detected 'layer' did not match the characteristics of either an aerosol or a cloud, often because the detection algorithm triggered on a noise spike or other signal artifact and not on an actual aerosol or cloud layer.
Thank you for the clarification!

Line 288 – I found the use of "pixel" here to be confusing. I think this refers to a 5-degree lat-lon grid cell.

Corrected, as suggested.

Lines 301-305. Sassen (2009) used an earlier version of the cloud product and also used different criteria for selecting and screening cloud layer data. Both of these likely contributed to differences when compared with the later results. Higher cirrus occurrence at night is primarily due to better sensitivity due to a lack of solar background (see Winker et al. 2024). The true diurnal difference in cirrus occurrence is complicated, as convective clouds have different diurnal cycles depending on geographic region. The day-night difference in background noise likely produces an artificial diurnal difference which outweighs the true diurnal differences.

Thank you for the clarification and your help in interpreting the results. We greatly appreciate your input.

Lines 317-319 – Yes, lidar systems are more sensitive to optically thicker clouds, but they also have much greater sensitivity at night due to a lack of solar background and higher signal-to-noise ratio. Whether higher frequency of cirrus detection at night is (partly) due to increased

nighttime optical depth is open to debate.
Corrected, as suggested.

Line 325 – I am still confused by “parameters that precluded the use of the test”, used in line 325 and the caption of Table 2. What does ‘parameters’ refer to? To me, a parameter is something like reflectance or radiance. The authors provided a clear response to my previous comment on this: By the phrase "precluded the use of the test," we meant that the specific indicators in question reach values that, in our judgment, make it impossible to use these tests directly for identifying Cirrus cloud masks. Given that the tests indicated by numbers in bold do not help in identifying cirrus for the cloud mask, do the bolded numbers in Table 3 give us a threshold value of the metric (ROP, POD, FAR, etc) where the test is not useful below (or above) that threshold? A little more explanation is necessary.
In fact, it's not about a specific threshold above or below which the test is excluded. Instead, if a particular test metric, such as ROP, significantly deviates in a negative direction compared to the others, it is considered "precluded." This means that when a metric performs considerably worse than others, the test is deemed ineffective for cirrus cloud identification, regardless of a specific threshold value.

Line 471 – Figure 10 shows results as a function of cloud optical depth, up to an optical depth of 10. Winker et al. (2024) points out that CALIOP retrievals of cirrus with optical depth become very uncertain when the optical depth is greater than 2 or 3. Optical depth uncertainty can grow to much larger than 100%. The authors should consider whether this large uncertainty at large optical depths might impact the results shown. Technical corrections: There are several instances of ‘p.p.’ which I think should be ‘%’
Thank you for your helpful comment. The findings from the referenced article were useful in refining the paragraph.
"p.p." (percentage points) refers to the difference between two values expressed as percentages, while "%" (percent) indicates a value as a part of a whole. For example, if a value increases from 10% to 15%, we say it increased by 5 percentage points (p.p.), not 5%. In short, **p.p.** measures the absolute change between percentage values, while **%** represents a value as a fraction of 100.

Line 181 – polar orbits with 16-day revisit cycle

Corrected, as suggested.

Line 182 – CALIPSO followed the Aqua spacecraft
Corrected, as suggested.

Line 188 – only the 5 km product

Corrected, as suggested.

Line 258 – The balancing of the sample …

Corrected, as suggested.

Line 261 – ‘more accurate results’, ‘more insightful results’, rather than ‘more reliable’?

Corrected, as suggested.

Line 417 – indicates a very high level …

Corrected, as suggested.

Line 419 – high values of POD are observed ?
Corrected, as suggested.

Reference: Winker, D., X. Cai, M. Vaughan, A. Garnier, B. McGill, M. Avery and B. Getzewich, 2024: “A Level 3 monthly gridded ice cloud dataset derived from 12 years of CALIOP measurements”, Earth Syst. Sci. Data, 16, 2831–2855, https://doi.org/10.5194/essd-16-2831-2024.