Overall: This paper presents results from a unique and valuable dataset. The two main contributions are cloud classification of this dataset into clear skies, ice cloud, and mixed phase cloud, and an algorithm that can quickly do this classification (which is presumably applicable elsewhere). However, we believe there are a number of serious problems with this paper. Most importantly, there is insufficient evidence that the authors are classifying cloud phase. Instead, it appears likely that they are grouping views into 3 types: 1) clear sky, 2) colder, optically thinner clouds, and 3) warmer, optically thicker clouds. Given this, it is not clear what value the algorithm adds to the literature, given that other methods exist that can classify phase that also classify optical depth and hydrometeor effective radius. The authors need to determine and report what they are actually classifying views into, e.g. using simulated data. More details follow.

Respectfully, we disagree with the reader's deductions. Regarding the two points ("problems") raised:

1) The category Mixed phase (in the training set, in the test set and in the entire dataset) comprises clouds with optical depths spanning from about 0.1 to the order of some unities.
2) The article deals with the application of the identification and classification method. The algorithm's features and edge on other methods are discussed in previous articles and only summarized here.

More details follow.

A better review of the literature and comparison to existing methods are needed

Referencing of the lidar instrument and how phase is determined is insufficient.

More details are provided in a new version of the article, also in response to another community comment.

Referencing of recent work on Antarctic cloud properties and similar cloud property retrievals is insufficient. Reading this paper, it would seem that there have been no surface-based studies of Antarctic clouds after 2012. The authors should reference recent papers by Lachlan-Cope et al 2016, Silber et al 2018, Lubin et al 2020, etc.

The current version of the article contains references to studies presented in 2019 and 2020. Anyway, the suggested literature will be taken into consideration for the final version of the work (we already knew some of the suggested works) and we are grateful for the advice. New references will be eventually inserted in the article in accordance with the goal and focus of the present work.

Machine learning concepts need to be referenced. Due to a complete lack of such references, it is unclear what are established methods (PCA, confusion matrix, hit rate, etc) and what was invented by the authors. E.g. are there references for the method of using a test set and an extended test set? Summing subtracted eigenvectors? Such references would be very helpful to fill in gaps and help understand what is novel.

Suggestion accepted.

The paper should compare this new method to existing methods for retrieving cloud phase from infrared radiances. For example, they reference Cox et al 2014, who retrieve cloud properties from Arctic infrared radiances, but do not compare to this work. They should also reference and discuss comparison to Rowe et al 2019 & Lubin et al 2020, which includes development and application of a cloud property retrieval, including phase, to clouds over McMurdo, Antarctica. Simulated datasets exist which could be used for an inter-comparison of methods. See, e.g. Cox et al 2016, Earth System Science Data, 8(1), 199–211.

We don't perform a retrieval but a cloud identification and classification. The cited references are all regarding retrievals. The CIC classification methodology was tested against synthetic and real data in previous publications and master thesis. In the paper by Di Natale et al. (2020) the reader can evaluate the results of a retrieval process considering of a large subset of the same dataset analyzed here.

Examination of the data in a real-world context is needed

The authors report the common occurrence of cloud with a liquid base and an ice layer at the top, which is contrary to what has been reported previously, both in the Arctic and Antarctic. This difference from previous work calls for some justification. This also underscores the need for a better explanation of the lidar design and methodology for determining cloud phase. What is meant by determining cloud layers from lidar by "human intervention?" Is this objective and repeatable? Why can't it be automated? Overall, using lidar as truth is not properly justified.

The procedure describing the usage of lidar data for the definition of the 3 classes will be better explained in the final version of the paper. See also reply to Reviewers 1 and 2. Moreover, the definition of mixed phase cloud will be better detailed. The lidar data are perfectly appropriate for the definition of the 3 classes of observation (i.e Ricaud 2020, 2017), especially for the Antarctic Plateau region conditions in which optically and geometrically thin clouds are very frequent.

The authors use Principal Component Analysis (PCA), but they never explore, plot, or discuss the associated eigenvalues and eigenvectors. The retrieval is blind in the sense that it does not take into account the atmospheric state in terms of temperature,

humidity, CO2 concentration etc. This would be ok if it was shown that the retrieval works without taking these into consideration, including some exploration of how it works, but this has not been done.

There is a paper (see references) describing the CIC algorithm and its main features. One of the strengths of the algorithm (as described in the reference) is that it is based on the signal only and doesn't need any other ancillary information to perform the classification. The present work is an application of the method and not a repetition of the description of the algorithm.

It should be noted that almost all the variance, and thus the strongest PCs, will be associated with cloud temperature and optical depth, not phase. Which PCs are associated with phase? Why use all PCs believed to be above the noise level?

We agree with the generic statement of the reader. Nevertheless, a one to one relation among the PCs and cloud features cannot be generalized. The reader is invited to review the algorithm methodology and in particular the metric defining the classification. We never analyze a spectrum singularly, but only in addition to all the training set elements.

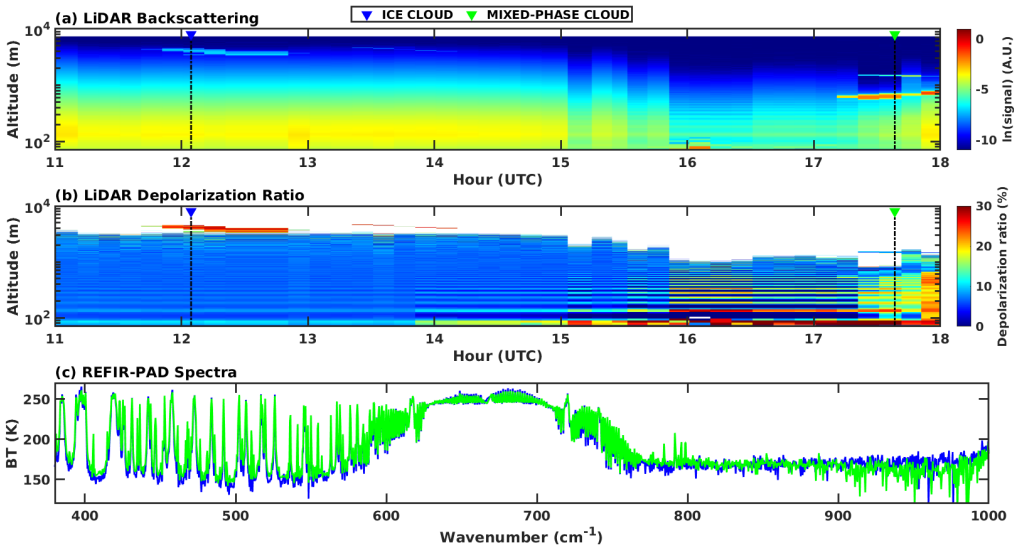
It seems likely that the classification is not based on cloud phase at all, but rather that scene views are subdivided into: 1) clear sky, 2) colder, optically thinner clouds, and 3) warmer, optically thicker clouds. They call category 2 "ice" and category 3 "mixed phase." It is possible these classifications are often correct, since ice clouds tend to be optically thinner and colder, and liquid clouds tend to be optically thicker and warmer on the Antarctic Plateau. However, this needs to be characterized, addressed and discussed, including errors and caveats. Several lines of evidence support the idea that they are not classifying cloud phase but rather optically thick and warm vs optically thin and cold clouds. First, looking at Fig. 2, it is unlikely that it is possible to determine phase from the green spectrum. This spectrum looks saturated, which means phase will have no influence on it - that is, there is no information about phase. It does, however, indicate that the cloud is optically thick. The authors could assess for which cases phase cannot be retrieved, using simulated spectra. Instead, are all such cases classified as "mixed phase" by the algorithm?

The deduction is erroneous and somehow unjustified.

What shown in Figure 2 are just two examples of spectra. We agree that maybe two different spectra should be selected as examples to avoid possible confusion. In the new version of the paper this point will be made clearer.

The training sets of ice and mixed-phase clouds are both composed of thin and thick clouds. Below one example of a thin ice cloud and one example of a thin mixed-phase cloud. Note that the spectra reported in the figure below are from the same day considered in Figure 2.

See also Di Natale et al. 2020 which shows cloud properties retrievals (including OD) from a subset of the CIC classified spectra presented in this work. In the cited paper the range of values over which the cloud optical depths span can be evaluated both for ice and mixed phase clouds



Second, as the authors point out, it has been shown that the far IR is critical for determining phase. Yet Fig. 6 suggests that a wavenumber range that excludes the far IR altogether would be equally good as one that includes it: the threat score is close to 1 for a range of just above 560 cm-1 to ~1020 cm-1. Indeed, the authors find the best range to be 540-1020 cm-1 for mixed phased clouds (it is unclear how they determine this), excluding essentially all of the far IR.

The methodology for the selection of the best range is described in the text. Probably the plots do not evidence enough the advantage of exploiting the FIR that is, anyway, clear from the numerical result concerning the Threat Score of the test set (for all the classes). A 3-d plot could be used in the new version of the article which better highlights the advantage of using FIR channels down to 380 cm-1. The scale can also be adapted to highlight the enhancement in the threat score. Note that an increase of few cents in the threat score means a large increase in the number of spectra correctly classified when dealing with the entire dataset (87960)

Typically, mixed-phase clouds are associated to more humid conditions than ice clouds and also, as described in the new text, to precipitation of thin ice crystals. For these reasons, the inclusion of the smallest wavenumbers (associated to the less transparent part of the FIR) did not bring significant enhancement in the classification.

Note that the classification of the dataset spectra is performed using the interval producing the best ThS over all classes: 380-1000 cm^{-1}.

Third, in the cold macro-season the algorithm does not retrieve cloud phase at all; instead all clouds are assumed to be ice.

Correct

Given the above, the authors should report the results of testing their method on simulated data, as has been done for other methods in the literature. This would allow them to test whether they truly have a cloud phase categorizer or if they are categorizing by cloud temperature / optical thickness. They could also determine and define characteristics of each category in terms of temperature, optical depth and phase ranges. This would also allow exploration of how errors propagate.

Tests against simulated (and measured) data were performed in the papers introducing the CIC (see literature). We also recall to the reader that the CIC is the official cloud classificator in the ESA FORUM E2E simulator and it is severely tested everyday by many scientists in the community.

Also, it is mentioned in the text that sensitivity tests. applied to synthetic stratified clouds with constant total OD, provide different classifications according to the relative amount of liquid/ice water content. The tests prove that CIC does not rely on a single parameter (i.e. optical depth) but on the entire spectrum characterization.

In the current paper we don't perform any retrieval. Nevertheless, the results of the classification include thin clouds both in the ice and mixed phase category. A large subset of the identified spectra is analyzed in the recent work by Di Natale et al. (2020). Di Natale et al. (2020) perform a cloud optical and microphysical retrieval. Please check the results. For example, you can evaluate the retrieved OD: it ranges from about 0.1 to about 4 in case of ice (figure 9) and about 0.1 to 10 in case of mixed phase (figure 10).

The authors ignore previous work on the temperature dependence of the single-scattering parameters (SSPs) of liquid water, which indicate that the SSPs of supercooled liquid water are intermediate between those of liquid and ice (Rowe et al 2013 and 2020 and references therein). In particular, Rowe et al (2020) indicates that uncertainties are large in the far IR.

We don't ignore it. The paper was brought to our attention by the reader via email 2 days before we submitted the present article. Anyway, we recall to the reader that the classification is a "discretized" result. We are aware that there are conditions "in the middle" not only due to the SSP, but also because of the physical structures of the observed scenes. It is noted, in the new version of the article, that what we call mixed phase cloud frequently occurs as a precipitating ice layer plus a thicker layer with low depolarization ratio values plus an upper layer with increasing depolarization ratio.

Questions and Concerns about Methodology and use of Machine Learning

The authors need to justify why they used the method they developed. It is not clear why PCA is used, or why the SID is used. Why isn't a simpler method tried, or at least compared to, to justify the more complicated method used?

The methodology is accurately described in previous papers. We regret that the details of the methodology applied are not clear to the reader and in particular the CIC ability in identifying optically thin clouds that are missed by simplistic (i.e. based on thresholds) methodologies. We also note that in Maestri et al. (2019) a methodology (CCREF) based on a combined linear discriminant analysis and support vector machine method is used on a similar dataset. Please refer to the cited literature. Some effort will be performed to better resume the CIC potentialities in the new version of the current paper.

Fig. 5 suggests that only one wavenumber is needed to distinguish cloudy from clear skies. Such a cloud mask has been reported in the literature but is not referenced or noted here (e.g. Weaver et al 2017, Atmos. Meas. Tech., 10, 2851–2880, 2017, Appendix). Classification using a single wavenumber would be sufficient for all of the cold macro-season data. Why is a considerably more complicated method used?

Fig 5 is used for illustrative goals only and shows the Training set mean BTs and their standard deviations. The figure just demonstrates what the reader is stating: "ice clouds tend to be optically thinner and colder, and liquid clouds tend to be optically thicker and warmer on the Antarctic Plateau". Nevertheless, the mean spectra are not used in any part of the classification process. Otherwise, the plotted quantities suggest that a large variability exists within the elements composing the classes (note the magnitude of 1 standard deviation).

As far as very simple methodologies based on single channels or BT difference thresholds, it has been demonstrated that they fail in detecting optically thin clouds which are very frequent in the considered experimental conditions.

To distinguish ice cloud from mixed-phase cloud, how many PCs are needed? Is PCA justified? Also, it seems odd to first divide cases into clear sky vs ice cloud and confusing that these each include mixed-phase. Why not divide first to clear and cloudy? Then subdivide cloudy into ice and mixed-phase. Such important details are left unexplored by the authors.

The number of the PCA used comes from the empirical function, called the factor indicator function (IND), defined by Malinokowski (1977, 2002) and reported in Turner et al (2006). It is usually of the order of ten in our case and it represents the number of PCs that allows the correct identification of the elements of each training set.
About the flow of the comparison, a team in one country may decide to do things quite differently than a team in another country. There is much to be learned from comparisons and discussion.

The authors use PCA to remove noise (Eqns 3-4) using an established method. However, Antonelli et al (2004, J Geophys Res 109, D23102), who should be referenced, state that the size of the training set should be greater than the number of spectral elements (M>N) to most accurately reconstruct the atmospheric signal and most efficiently remove noise. Here it appears that M<<N. How does this affect the noise reduction and signal reconstruction?

We know the work of Paolo, nevertheless we don't want to accurately reconstruct the full radiometric signal of each selected spectrum. Otherwise, CIC defines a metric based on the changes in the main PCs characterizing a set of spectra (the training set spectra) when a new element (the analyzed spectrum) is added to the set. The evaluation of the change (the modified information content) is assessed always as a comparison with respect to the change obtained when a different training set is considered. This is a change of metric with respect to previous methods. Please refer to references cited in the text.

Antonelli et al (2004) also state that if some spectra are not well-represented by the set of spectra used for noise reduction, a larger number of PCs may be needed to properly represent those spectra. This seems likely to be the case when the input spectrum is not a member of the training set in Eq. (6). How is this handled and how does it impact the results?

This is one point that justify the methodology used. Please, see the two answers above.

The authors reduced the dimensionality of the observations by modifying the spectral interval of the test set members and re-running the algorithm. Given that the authors are already using PCA, and that PCA is typically used for dimensionality reduction, why isn't PCA used for this dimensionality reduction?

PCA, as well as other techniques, could be used to reduce the dimensionality in this sense. We have used linear discriminant analysis in Maestri et al. (2019) to this goal. However, such a method would select specific wavenumbers along the spectrum, according with the highest eigenvectors elements. This is not what we are interested in at the moment. For general purpose and to maintain the methodology simplicity, we want to select continuous portions of the spectrum. This led the procedure to remove the smallest and largest measured wavenumbers which are affected by the largest instrument noise. In such a way the maximum amount of information is anyway passed to the PCA used in the CIC.

Furthermore, it is not good practice to use the test set to select features (wavenumbers) to use. Using the test set to optimize the algorithm exaggerates the accuracy of the method and can lead to overfitting. Model development should be done using training or validation sets. See, e.g. Ripley, B.D. (1996) Pattern Recognition and Neural Networks,

Cambridge: Cambridge University Press, p. 354. The data with known labels should be split into training, validation, and testing sets. The testing set should be held apart and only used to estimate the accuracy of the method. None of the training, testing, or validation data should then be included in the analysis. The authors need to clarify which data is being used in each step and ensure they are following established practice.

We are aware of the procedure described by the reader and we have used it in the past. Nevertheless, the selection process of the input wavenumber interval cannot be properly defined as a hyper-parameter setting. It simply reduces the amount of channels ingested by the CIC. This is not a parameter that directly affects the behaviour of the algorithm. The algorithm operates exactly the same way independently of the selected interval. In this sense, a proper validation set is, therefore, not necessary.

Indeed, splitting what is currently defined as "test set" into two distinct groups of equal size (a "validation set" and a "test set") leads to the same exact results reported in the current version of the paper. We can provide these results if needed. The same wavenumber interval is selected on the validation set (380-1000 cm-1), and the same hit rates are found for the test set, implying the same results on the entire dataset.

More detail is needed to allow the analysis to be repeated.

It is not clear how the authors handle erroneous data points. One of the reviewers pointed out that a data point at the center of the CO2 band is erroneous, at 667 cm-1. This is typical with such instruments because calibration is impossible at such wavenumbers (see Rowe et al 2011, Optics Express, 19(6), 5451–5463, and Optics Express 19(7), 5930-5941). There are many other erroneous brightness temperatures evident - for example, none of the BTs below 200 cm-1 appear useable, as well as many between 200 and ~350 cm-1, where BTs are very high. How did the authors handle such points in their analysis? Were they included or omitted? The authors should briefly explain the instrument error characterization and point to a reference with more detail.

Correct. The bad calibration points fall in spectral regions that are not considered in the analysis. This is explained in the answers to reviewers 1 and 3 and implemented in the new text.

The algorithm description could use some clarification. The development should proceed linearly from training to testing to implementation. It seems that what is meant by the input spectrum on line 164 varies; this needs to be clarified. For example, it seems the SIDs and the CSIDs are developed from the training set first (to get Fig. 4)? How is this done?

The request of clarification is not clear to us. The development proceeds exactly as described. The CSID definition is part of the optimization process. See the text and references.

Finally, the utility of this algorithm seems likely to be specific to the unique conditions on the Antarctic Plateau. The authors should discuss whether it would be applicable elsewhere.

We respectfully note that the CIC methodology, in less than 3 years, has been applied to:

- Spectral radiance simulations over the globe
- The airborne TAFTS data
- The airborne ARIES data
- Satellite FORUM simulated radiances within the ESA End2End simulator (two different instances of sensor are considered plus another ideal sensor)
- the ground based REFIR-PAD

It is all in the cited literature. At the moment we are applying the CIC to IASI and FIRMOS data.

References

Di Natale, G.; Bianchini, G.; Del Guasta, M.; Ridolfi, M.; Maestri, T.; Cossich, W.; Magurno, D.; Palchetti, L. Characterization of the Far Infrared Properties and Radiative Forcing of Antarctic Ice and Water Clouds Exploiting the Spectrometer-LiDAR Synergy. Remote Sensing, 2020, 12 (21), 3574. https://doi.org/10.3390/rs12213574.

Maestri, T.; Arosio, C.; Rizzi, R.; Palchetti, L.; Bianchini, G.; Del Guasta, M. Antarctic Ice Cloud Identification and Properties Using Downwelling Spectral Radiance From 100 to 1,400 Cm −1. J. Geophys. Res. Atmos., 2019, 124 (8), 4761–4781. https://doi.org/10.1029/2018jd029205.

Maestri, Tiziano; Cossich, William; Sbrolli, Iacopo, Cloud identification and classification from high spectral resolution data in the far infrared and mid-infrared, «ATMOSPHERIC MEASUREMENT TECHNIQUES», 2019, 12, pp. 3521 - 3540. DOI: 10.5194/amt-12-3521-2019

Magurno, D.; Cossich, W.; Maestri, T.; Bantges, R.; Brindley, H.; Fox, S.; Harlow, C.; Murray, J.; Pickering, J.; Warwick, L.; Oetjen, H. Cirrus Cloud Identification from Airborne Far-Infrared and Mid-Infrared Spectra. Remote Sens. 2020, 12(13), 2097; https://doi.org/10.3390/rs12132097.

Malinowski, E. R. Determination of the number of factors and the experimental error in a data matrix. Anal. Chem., 1977, 49 , 612–617. https://doi.org/10.1021/ac50012a027

Malinowski, E. R. Factor Analysis in Chemistry. 3d ed. Wiley and Sons, 2002, 414 pp.

Ricaud, P.; Bazile, E.; del Guasta, M.; Lanconelli, C.; Grigioni, P.; Mahjoub, A. Genesis of Diamond Dust, Ice Fog and Thick Cloud Episodes Observed and Modelled above Dome C, Antarctica. Atmos. Chem. Phys., 2017, 17 (8), 5221–5237. https://doi.org/10.5194/acp-17-5221-2017.

Ricaud, P.; Del Guasta, M.; Bazile, E.; Azouz, N.; Lupi, A.; Durand, P.; Attié, J. L.; Veron, D.; Guidard, V.; Grigioni, P. Supercooled Liquid Water Cloud Observed, Analysed, and

Modelled at the Top of the Planetary Boundary Layer above Dome C, Antarctica. Atmos. Chem. Phys., 2020, 20 (7), 4167 4191. https://doi.org/10.5194/acp-20-4167-2020.

Turner, D. D.; Knuteson, R. O.; Revercomb, H. E. Noise Reduction of Atmospheric Emitted Radiance Interferometer (AERI) Observations Using Principal Component Analysis, J. Atmos. Ocean. Tech., 2006, 23, 1223-1238. https://doi.org/10.1175/JTECH1906.1