Atmospheric Chemistry and Physics Discussions

Atmos. Chem. Phys. Discuss., community comment CC1
https://doi.org/10.5194/acp-2021-97-CC1, 2021

Comment on acp-2021-97

Penny Rowe

Community comment on "Ice and mixed-phase cloud statistics on the Antarctic Plateau" by William Cossich et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2021-97-CC1, 2021

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.

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.

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.

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.

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.

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

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.

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.

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? 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? 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 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?

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?

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?

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.

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.

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?

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.

Penny Rowe and Steven Neshyba