Review of manuscript egusphere-2023-1085

Title: Supercooled liquid water cloud classification using lidar backscatter peak properties

Authors: Luke Whitehead, Adrian McDonald, and Adrien Guyot

Summary:

In this manuscript, Whitehead and coauthors examine a machine learning algorithm (an extreme gradient boosting or XGBoost model) trained to identify warm liquid clouds, supercooled clouds and ice containing clouds based on ceilometer observations and associated layer temperatures. The algorithm is trained against cloud phase determined using lidar volume depolarization ratio measurements, which traditional ceilometers do not measure.

Overview:

While I agree with the authors that a lidar/ceilometer-based technique that determines cloud (layer) phase without reliance on lidar depolarization would be useful, I think the current algorithm and analysis needs a bit more work.

Recommendation: Publish in with major revisions.

General Comments:

1) Too much ice cloud and too little water cloud

My largest concern is that the algorithm, as-is, clearly substantially overestimates the occurrence of ICC, while underestimating both SLCC and WLCC.

Looking at Figure 6, I would guess that 1/3 or so of the supercooled clouds near 1.5 km (where the occurrence is largest in the reference dataset) is missing. That is, the failed detection rate (see point #2 below) is significant. This doesn't make the algorithm useless by any means, but I think this needs to be quantified and summarized in the abstract and conclusions.

It also very striking the degree to which ICC is being overestimated, especially but not limited to, altitudes below 2 km. See below specific comments starting at lines 361 and 377. As described in the specific comments, I think the reasons for this problem need to be explored further and I suspect the situation can be improved rather easily.

2) False and Failed detection rates.

The analysis is largely focused on the f1 score and the "balanced accuracy" score which averages f1 across the phase categories (if I understand correctly). While this is OK, it doesn't remove the value of knowing (and need to provide) the false and failed detection rates for each category, where:

False detection rate = (# of false positive detections / # of positive detections). This is the same as 1 – precision.

and

Failed detection rate = (# of failed detections / true total # of events in category). This is the same as 1 - recall.

I can't reconstruct these rates given only the f1 values (but I can get f1 score given only these two).

I don't mind if you use precision and recall rather than False and Failed detection rates (different fields somewhat different terminology) but at a minimum you need to provide at least a table (similar to Table 1) giving the recall and precision numbers and nominally plot the vertical profiles of recall and precision (as per Fig 6c and 6e) for the three phase categories.

3) A summary of the problems and limitations also needs to go into the abstract.

In my view, the abstract also needs to talk about the problems, presenting a more balanced view of the performance. The conclusion is reasonably well balanced but the abstract is not.

4) Multiple scattering and the VDR cloud phase

As discussed at several points in the manuscript, the VDR phase determination is far from perfect, and in particular multiple scattering can often increase the VDR and result in a miss identification of liquid phase cloud particles and ice phase. One can account for the multiple scattering (at least for optically thicker clouds). See for example, Mace et al (2020). Ideally, I would like to have seen an approach such as Mace's used as the reference, but such is obviously a major task and perhaps beyond the scope of what can be done with regard to the present manuscript. But I would encourage such as part of any continued development.

Mace, G. G., Benson, S., & Hu, Y. (2020). On the frequency of occurrence of the ice phase in supercooled Southern Ocean low clouds derived from CALIPSO and CloudSat. *Geophysical Research Letters*, *47*, e2020GL087554. https://doi.org/10.1029/2020GL087554

Specific Comments:

Line 29. Perhaps change " .. satellite-based measurements of …" to "satellite-based identification …".

Line 29. Perhaps change SLW to "SLW and Mixed Phase".

Line 34. "Transmission" not "emission". Emission is the process of radiating, transmission is the process of transmitting.

Line 36. I have never heard the term "Automatic lidar" or the "ALC" acronym before this manuscript. What is a non-automatic lidar? As far as I can see this acronym is only used one other time (on the next line 38/37). Perhaps remove entirely from the manuscript and write simply, "In this study we use a Micropulse lidar which measures the linear …"

*Line 117-121. Why did you subsample the VDR rather than average? This is especially confusing to me since in the next set of lines (119 to 121) you appear to be averaging these sub-samples to reduce noise??

Line 119-121. Perhaps expand the description of the processing here. As is, it appears that (1) you averaged the ratio rather than averaging the individual parallel and perpendicular backscattering components that go into Equation 1. The latter (averaging backscatter components not the ratio) is more physically sound as one expects noise to affect the ratio in a non-linear way that can amplify the effect of the noise (When noise makes $P_{\|}$ small it amplifies error). (2) Did you threshold before averaging? Again, this is not typically a good idea as it will typically generate bias.

Line 126. Please expand on this comment. I think of mixed phase clouds as being a combination of small (cloud-sized) liquid droplets and large (precipitation-sized) ice particles. I presume you mean that within the cloud, the scattering is dominated by cloud droplets and so one can't easily determine if precipitation sized particles are present, nor determine their phase. However, one can determine the phase of precipitation that is falling below cloud base, and in this way, identified mixed phase clouds which contain supercooled cloud droplets and precipitation ice (below).

Line 128 & 374. " … since ice cannot exist above 0 ◦C." Well, it takes time for ice to melt and it is quite common to have precipitating ice above 0 C. I think it is fair to say that small cloud-ice does typically exist below 0 ◦C. Perhaps rephrase to be more technically correct.

Line 140. The term "peak height" is potentially confusing as one might take height to mean altitude rather than the strength of the backscatter. Why not refer to this as the "peak magnitude"?

*Line 141. "Peak width height" is also potentially quite confusing. Further the meaning of "baseline" is not clear to me since one expect the backscatter not have equal values on both sides of the peak. An illustration of all eight characteristic would be very helpful here!

Line 186. Are the 5-minute time samples (I think this is what you are using in your statistics) within the same day being partitioned into different folds (both training and test folds)? Or is data for entire days (chunks of data) going into one fold or another? If the former, I am not sure it makes sense to talk about the data being independent. (Don't get me wrong it is good that you keep test data separate from the training data, but the 5 minute data are going to be highly correlated, and ideally you would partition data in time-chunks that are large enough that temporal correlation between the chunks is small).

Figure 4b. This is not showing all VDR values, just those where peaks are present, yes? This is not a problem per se, but perhaps worth explaining. In my view it would be nice to see the full VDR field (to get a sense for multiple scattering).

Line 197-201. What does this imply? Why not use the smallest / simplest network in this case (for if no other reason than to minimized the potential of overfitting)?

Line 299-310. I'm confused. I thought "balanced" meant averaged across phase categories (SLCC, WLCC, ICC) but here you are only describing the (binary) SLCC mask or WLCC masks?

Line 330. Perhaps worth discussing here is the low level clouds after 18 UTC which is water according to the VDR but ice to both of G22-Davis and G22-Christchurch. (Upon review, I see that you discuss this later in the manuscript. Perhaps just note here you discuss this cloud layer at a later point.)

Line 330. As best I understand, the miss-classification of the low cloud will have no or little effect on the accuracy statistics of SLCC presented here because the column contains SLW. If yes, perhaps point this out.

Line 332. On a very minor point but perhaps comment on how do you know this is a "thick band" of cloud. Perhaps simply note the lidar is being fully or heavily attenuated (and one is not seeing the top of cloud) here.

Line 340. I wrote "WHY?" in big letters in the margins on my first reading. Again, perhaps note that this problem is discussed in more detail later in the manuscript.

**Line 361-366 and 389-391. While I'm sure the VDR does misclassify some WLCC as ice because of multiple scattering, both of your examples and Figure 6a seems to suggest this is a very small percentage of the data. And if so, shouldn't the algorithm have learned that such warm clouds near the surface are rarely ice and therefore it should guess that warm clouds are liquid (rather than infrequently occurring ice)? I would wager that if you look carefully at the backscatter characteristics or masking/peak detection issues (see comment line 377) that these will have a lot more to do with this overestimate. When use later use ONLY temperature as an input, a discussed later in the manuscript, what happened to these low clouds? I require more evidence to believe that a small percentage of bad data in the reference set is really the cause of this problem.

Line 365. You write “This demonstrates the sensitivity of the reference mask, and thus the G22-Christchurch model, to the VDR threshold δLCC, and suggests that δLCC should be increased.” I entirely disagree with this in part because of my comment above (Line 361). In general, doing so will increase the amount of time that ice is being called liquid in the reference dataset. Rather, I think a better solution would be use a reference technique that accounts for multiple scattering as part of the training (see general comment #4).

*Figure 6. This is an important figure. Panels (c) and (e) need more tick marks and need to be expanded (or something) so one can read values. As per general comment #2, please also add recall/precision rates (and discuss such in the text). In general, it would be good to plot values for WLCC and ICC (not just SLCC).

**Line 377. You wrote “… other parts of the cloud are identified as ICC by the mask because there are either no peaks detected (so the cloudy bins in that profile are conservatively labelled ICC by default).” So this seems like a large potential source of the ICC overestimation problem. But if I understand, you know what regions are peaks (and neighboring width pixels) and which are not. And if so, you should be able to easily establish the degree to which this is a problem. And it occurs to me that simple potential fixes might be: (1) To either assume points not associated with a peak are liquid, if they are warm and ice otherwise. I note here that clouds with top temperatures only a few degrees below zero (say warmer than about -5 C) tend to be liquid, so personally I would use cloud-top-temperature. AND/or (2) assign the phase based on the closest peak within the same cloudy layer, or near the same altitude in a neighboring column.

*Line 395. It looks to me that 1/3 or more of the SLCC might be missing.

Line 398. “occasionally” misclassified? I think Figure 6d shows that this is a lot more than just occasionally. Much of the WLCC is missing.

Line 406. You write “In order to be confident the models can be applied in future work to ceilometer datasets from a range of locations, …”. So I think being applicable to other locations is an excellent goal and something well worth testing. You leave this for future work in the conclusion and that is OK, but presumably you have the Davis station data from the Guyot (who is a co-author) and the same VDR reference data so it would be easy to at least see what happens when you apply the G22-Chirstchurch algorithm to the Davis data.

Line 487. I like your analysis in this section (though I would like to see a table for recall and precision – is the story the same for these metrics)? Perhaps it is beyond the scope of what you are willing to do here but using surface temperature or surface temperature and altitude might also work well and thereby avoid the dependence on radiosondes or NWP.

Line 489-90. This is a good point. But it seems to me that it is an open question as to whether the relationship between temperature and SLCC occurrence is constant (doesn’t vary with location). In general, to me, you results argue for using the <T, omega> or <T,

omega, Beta_prom> version of the code as your “default algorithm” (that I hope you would provided to other researchers).

Line 539-40. See general comment #4.