amt-2022-10

" Detection of supercooled liquid water clouds with ceilometers: Development and evaluation of deterministic and data-driven retrievals"

Authors' response to comments from Reviewer #2

2 May 2022

We would like to thank the editor and the two reviewers for their very constructive comments on our manuscript. We received genuine insights, which have significantly contributed to increasing the manuscript quality and potential impact. To improve the clarity in our responses we have numbered the reviewers' comments: for example, the comment 1 from reviewer 1 is listed as R1C1 and will refer to these comments as such in the following.

Based on some of the comments from the reviewers and late feedback from our co-authors, we also made small additional changes to the original manuscript:

(1) Following internal discussion between the co-authors, we realised that our initial definition of Supercooled liquid clouds (SLW) was misleading as it included both mixed-phase and SLW clouds. With our technique, and the technique from T19, we can detect clouds containing SLW, e.g., both SLW clouds and mixed phase clouds, as we cannot distinguish between the two. We therefore decided to replace instances where we referred to both SLW and mixed phase as "Supercooled Liquid water Containing Clouds" or SLCC, including in the title.
(2) We have incorporated late feedback from one of our co-authors; these were mostly typo and added precisions in the manuscript and figures.
(3) We have applied our two retrievals to additional data from the same instrument (the ceilometer) that covers a full annual cycle including austral winter months. We have included a new section 3.4 in the manuscript and a new Figure 10. We think that this will add value to the paper, showing how this new retrieval can produce climatology of supercooled liquid water containing clouds in regions such as Antarctica where observations are scarce. We also added a few sentences putting these results into perspective and in the light of other observations in Antarctica and elsewhere.
(4) We made some cosmetic modifications in the Figures: In Figure 1a, the arrow was wrongly labelled "North", we removed the label to Progress 3 and removed the bathymetry. In Figure 5, we changed the x-axis units. Figures 7 and 8 saw the unit of the axis changes to km instead of m to simplify the axis unit labels.

Reviewer #2

The manuscript presents a new machine learning approach for the classification of supercooled liquid water from ceilometer observations. Based on three months observation ins the Arctic, the approach shows improved performance compared to a previous method which also analysis ceilometer profile data. As reference, the authors use a mask derived from a combination of radar and depolarisation lidar observations. The study is nicely presented with convincing scientific quality. To highlight applicability of the novel tool at the large number of ceilometer data being collected globally, the authors could give perspective on the expected performance in other geographical settings and maybe measurements from other ceilometer types. The manuscript can be published after a series of minor comments are addressed.

Line 180: What is the vertical and temporal resolution of the RMAN cloud classification? How is this alight to the 're-gridded' data of Radar and ERA5? How can a classification be 'interpolated' or 'averaged'?

R2C1: The "raw" data from the RMAN lidar is of variable temporal resolution generally ranging from about 30 to 90 seconds, and 15 m vertical resolution. The lidar data are then averaged at 2 min resolution, and the cloud classification is therefore produced at 2 min, 15 m resolution. The cloud radar has a "raw" resolution of 12 s, 25 m. To retain some of the higher cloud radar temporal resolution, the two products are combined to produce a 1 min, 15 m resolution cloud mask, interpolating the radar fields to the vertical resolution of the RMAN and re-sampling the RMAN lidar data to 1 min resolution. The RMAN lidar classification is re-sampled to 1min simply by duplicating the nearest 2min timestep. We added this sentence in the manuscript:

"The radar has a sensitivity of around – 50 dBZ at 1 km, a vertical resolution of 25 m and a sampling frequency of 12 s."

"The "raw" data from the RMAN lidar is of variable temporal resolution generally ranging from about 30 to 90 seconds, and 15 m vertical resolution."

"The original 2 min resolution RMAN lidar classification was re-sampled to 1 min simply by duplicating the nearest 2 min timesteps."

Line 208: Not clear what is meant by 'how to label the 50 m bins.' Please rephrase.

R2C2: We rephrased that sentence, and it now reads: "However, they did not specify how to allocate a classification to the bins at the height of the peak and the surrounding bins (below and above the peak). We decided that based on the above, only the altitude bin corresponding to the location of the peak (if found) was labelled as liquid water."

Line 209: So only one bin is classified as liquid water?

R2C3: This is correct. We did not want to move away from T19, as this serves as our reference mask. In any case, the comparisons are then done timestep to timestep, so the number of bins labelled as SLCC do not interfere in the evaluation.

Line 211: In line 195 you state that the cloud phase mask utilises Ceilometer data only. But now you state that SLW and liquid water is differentiated according to the reanalysis temperature profile. This seems contradictory.

R2C4: The sentence in line 195 has been changed to include discussion of ERA5 and now reads:

"The first cloud phase mask presented herein is based on ceilometer observations, following the work from T19, augmented with ECMWF ERA5 interpolated temperature fields to differentiate SLW from other liquid water."

Line 239: if you say minimum peak width is 50m , this means only one range bin as you are operating an a grid of 50m?

R2C5: We have changed the sentence to: "….least a width of 50 m (thus only one range bin since this is the lowest resolution of our post-processed ceilometer data), and a peak…"

Line 241: Where do you define the peak width? At half maximum or base?

R2C6: We added the equation that explained how the peak width height is calculated to the text:

"The height at which the peak width is measured is relative to its prominence, following eq. (1):

$$H_{peak\ width} = H_{height} - Prom \times RH \quad (1)$$

With $H_{peak\ width}$ the height of the peak at which the width is measured ($m^{-1}\ sr^{-1}$), H_{peak} the absolute height of the peak, $Prom$ is the prominence ($m^{-1}\ sr^{-1}$), and RH the relative height, which was set at 0.5."

Line 244: 'Lowest' peak defined by peak magnitude or altitude?

R2C7: it now reads:" …with the lowest peak in altitude taking the number '0'".

Line 245: Again, you are using ERA5 temperatures. I think you need to be careful calling the algorithm to utilise "ceilometer data only"?

R2C8: This has been changed in R2C4 and elsewhere in the manuscript (abstract and introduction) to include mention of ERA5 temperature.

Line 254: rephrase "For single peaks, SLW data-only were selected based on the Boolean condition defined using the radar-lidar cloud mask". What is meant by "data-only"?

R2C9: This has been re-phrased as:

"For single peaks, data for which SLCC were identified were selected based on the Boolean condition defined using the radar-lidar cloud mask."

Line 255: remove "arbitrary"? Your conditions have an empirical basis.

R2C10: Indeed! We replaced “arbitrary” by “empirically based”.

Line 257: what is meant by “width < 4”? Bins? Maybe better to use width in units of meters?

R2C11: It now reads: “the width of the peak must be < 4 bins (200 m)”.

Line 265: Why is the multiple-peak distribution so narrow? Should there not be a dependence on the order of peak in the profile? i.e. could be not expect the peak at the lowest altitude in a multiple-peak profile to resemble the signature of a single-peak? Do you account for the order or altitude of the peaks?

R2C12: Apologies, the terminology “multiple peaks” has been mis-leading. Please see response to R1C8 and R1C9 where we address these issues.

Line 274: how often do you find this mismatch between peak-criteria and cloud classification mask that leads to an adjustment of the “true” indicator? What does this mean physically?

R2C13: Thanks to our clarification regarding the terminology used for “multiple peaks” (responses R1C8 and R1C9), our statement here line 279 shall be seen with that clarified definition in mind. This case is when typically for a given timestep when the primary peak is identified as SLCC, and the secondary peak(s) is(are) not. The label for the secondary peak(s) should then be revised as non SLCC.

Caption Figure 4: Introduce meaning of ‘ts’.

R2C14: Done.

Line 330: This seems like an artificial problem. The masks are created based on higher-resolution data. Why would you create a 50m vertical resolution grid for the ceilometer-based mask if the observations have a resolution of 10m? would it not be more appropriate to map all data to the same vertical resolution in the beginning so they could now be compared more easily?

R2C15: The resolution of 5 min, 5 0m was found optimal in ALCF to reduce noise in the attenuated backscatter. Using a higher temporal and spatial resolution will increase the noise to signal ratio. Therefore, we decided to use that same re-sampling strategy. For the radar-lidar mask, the 15 m bin resolution followed the same approach as in Alexander and Protat (2020). Bin to bin comparisons will be affected by many factors and is not the objective of our work here.

Line 334: same question for the temporal resolution.

R2C16: See also R1C3. We added this sentence to the manuscript:

“Subsampling is mostly done to improve signal-to-noise ratio. The cloud masking usually benefits from subsampling to 5 min intervals and 50 m vertical resolution, because it decreases the number of misclassified bins;”

Line 352: introduce meaning of confusion mask indicators

R2C17: We have inverted the order of the definition in that section (2.5) and have introduced the confusion matrix indicators at the end, after the basic definitions have been provided.

Line 357: state clearly what you are referring to with the term "prediction". Is true negative the case when the mask correctly indicates the absence of SLW? Then why call this "wrong prediction"? Also, if "false positive" refers to the mask wrongly assigning SLW, then why would you call this "wrongly indicating a correct prediction". Please clarify this paragraph.

R2C18: Thanks, indeed that wasn't very clear… we have modified that paragraph and it now reads:

"A true positive (TP) is defined as a test result indicating a correct prediction (correctly predicting the occurrence of SLCC), a true negative (TN) is defined as a test result correctly predicting the absence of SLCC, while a false positive (FP) is defined as a test result wrongly predicting the presence of SLCC, and a false negative (FN) is a test result wrongly indicating the absence of SLCC."

Line 388: what is the number of samples in the training data? Is this stated in the methods section?

R2C19: It was implied in the method, as 3 k-fold cross validation by definition allocates two third of the total data for training and one third for testing. We have added a sentence in the methodological section (new lines 318 319) to explicitly state this:

"With the 3 k-fold cross validation, two third of the data are allocated to training, while the remaining one third of the data is used for testing."

Line 411: Of course figures should be explained when they are being discussed, but please avoid repeating content of figure captions in the text.

R2C20: This sentence now reads: "In Figure 7, the average peak properties for which peaks had been identified for the 6th of January 2019 are presented as joint distributions using kernel density estimation plots with peak β value as the y-axis". The introduction of Figure 8 in the text was also changed similarly. We also simplified the text for the introduction of Figure 9 to avoid repetition between figure captions and the text.

Line 424: How did you evaluate presence of fog?

R2C21: We refer here to the fog identified based on our algorithm as described in the methodological section, such as: "The same method as T19 was again used here, detecting fog layers by identifying values of backscatter above ß = 10^{-5} m^{-1} sr^{-1} for the lowest grid point (corresponding to 0-50 m above the surface) and a ß value 250 m above the instrument of ß < 3 x 10^{-7} m^{-1} sr^{-1} (to restrict the identification to fog, and exclude low-level thicker clouds)." We did not refer to independent observations to confirm or not its presence.

Line 444: Given you are using various different products, please use consistent labels throughout. E.g. in a similar way you are using T19, please use one label (such as XGBoost) for the "new algorithm".

Also, please use one consistent label for the reference data. Right now, the reader can get easily confused. E.g. here I am wondering if "a data-driven threshold approach" has already been introduced or if this is yet another method.

R2C22: We have decided to label our new approach "G22-Davis" and refer as such in the manuscript. This terminology is now introduced in line 333:

"To facilitate the discussion in the next sections of this study, we further refer to our algorithm as G22-Davis. The extension "-Davis" illustrates that our G22 model had been trained with data collected at Davis, and we can imagine that the same model could be applied to data collected elsewhere."

Subsequently, there are 12 instances where we have replaced the unclear "new algorithm" terminology by "G22-Davis".

Line 473: So the thresholds are not actually "arbitrary", but rather empirical values determined based on the previous analysis. The fact that they work well for your dataset is hence not surprising. It would now be the next step to assess whether these thresholds are more widely applicable, e.g. to perform SLW detection for a different time period or different location.

R2C23: Yes, that's correct. We replaced "arbitrary" by "empirical". Following your comment, we have added a sentence in the discussion section:

"Finally, in the absence of radar-lidar mask to train and test a model like in G22-Davis, there remains the possibility of using the classification approach based on thresholds derived from peak characteristics joint distributions. This method should be assessed for different periods at Davis or for other locations than Davis, to test of these threshold values are widely applicable."

Line 484: After presenting these values, please put results into context e.g. to performance of the other approaches.

R2C24: This now reads: "This is an improvement of 0.07 as compared to the accuracy of T19, which was equal to 0.84. For the dataset for which peaks were identified, the total dataset was made of 11,327 datapoints. The best testing score was of 0.81 (with learning rate = 0.01, max depth = 12, child weight = 8), for training accuracy scores (or f1) of 0.94. This is a substantial improvement of almost 0.1 as compared to the accuracy of T19, which was equal to 0.72."

Line 507: This seems to contradict your statement from line 453: "The value of β at peak is directly correlated to the peak width height, making that feature redundant." Please explain.

R2C25: in former line 453, we changed the text to: "The value of β at peak versus peak prominence is not presented as these are directly correlated. The value of β at peak is correlated to the peak width height, with some differences."

The redundant feature is peak prominence, rather than peak width height, as can be seen in Figure 9.

Line 509: If peak temperature is not an important predictor, would it be possible to omit the ERA5 data and work solely on ceilometer observations as input?

R2C26: We added that sentence in new lines 594 to 597:

“Given the low importance of air temperature for accurate prediction of SLCC, we could consider not using that feature as an input to G22-Davis and removing our dependency on ERA5 or other NWP inputs. However, at Davis, we might likely be in the specific case where air temperatures are often too negative to produce liquid water droplets (other than supercooled) as seen in the T19 cloud phase classification. For other climates with higher air temperatures, the air temperature feature might be more relevant.”

Line 562: You are using software and algorithms developed elsewhere, yet you are not intending to share the code? Especially as you are claiming your algorithm has better performance than an existing approach of T19, it would be important to the community to be able to test your algorithm and verify your findings.

R2C27: See also R1C15. This was the status at the time of the submission, but our approach has changed since, and we are preparing the algorithms to be included in the ALCF developed by co-author Peter Kuma. It will therefore be available and applicable easily by future users. The text of the manuscript has been changed to:

“The new G22-Davis ceilometer algorithm described herein as well as the original T1 algorithms are in the process of being included in ALCF and will therefore be open-source and publicly available.”

Figure 6: how is the “baseline” determined based on which you quantify the “peak prominence”?

R2C28: We added these two sentences in the figure caption:

“The baselines (lowest points on the green lines) were calculated as the lowest contour lines around the peak. To identify the peak characteristics, we used the signal processing tools of the python library SciPy.”

Figure 7 and Figure 8: these are not a “scatterplots” because the individual sample pairs are not shown. Rather you are comparing isolines for the two cases.

R2C29: See R2C20. We also changed the captions of Figure 7 and Figure 8 accordingly and refer to the plots as: “Joint distribution using kernel density estimation plots”.