Dear Reviewer,

thank you for taking the time to review our manuscript. Your feedback is greatly appreciated and was helpful in improving the quality of this research. We value your constructive criticism and thoughtful comments, which have helped to identify areas that require further clarification and refinement.

We carefully considered your suggestions and incorporated them into the revised manuscript to address the issues raised, as specified below (referee comments in blue; our answers in black).

General comments

One thing that is mentioned in the abstract and in the introduction is the low/inadequate accuracy of traditional cloud typing methods in detecting cloud top phase. Perhaps a comparison with at least one another product available online for cloud type retrieval (could be from geostationary or low earth orbit instruments) could be interesting. It could be briefly shown as direct comparison or discussed. It could be for example the NWC-SAF Cloud Type product (even though it does not explicitly report the cloud phase, though the cloud types can be broadly related to cloud phase).

Comparing the ProPS method with another established cloud phase detection product is a very valuable suggestion. We have taken your advice and made a thorough comparison between ProPS and the CM SAF CLoud property dAtAset using SEVIRI - Edition 3 (CLAAS-3) product. We have included the main results of this comparison in the paper and added a detailed discussion to the appendix of the paper, as follows:

“In order to better characterize ProPS, we conduct a comparison to the CM SAF CLoud property dAtAset using SEVIRI - Edition 3 (CLAAS-3) product, which was released in 2022 (Meirink et al., 2022). This new edition of the CLAAS product offers an extended phase classification system, distinguishing between clear sky, liquid, supercooled, and various ice cloud types, which we condensed into one ice cloud category for simplification.

The method of the CLAAS-3 cloud detection, called CMA-prob, is shows some similarities to ProPS, especially because it uses a Bayesian approach based on the CALIPSO/CALIOP (but not on CloudSat/CPR) cloud mask as ground truth and a selection of visible and infrared SEVIRI channels as inputs (Karlsson et al., 2017). While this probabilistic methodology is similar for ProPS and CMA-prob, their tactics differ slightly: CMA-prob does not use conditions (except for surface types) for the probabilities but instead subtracts pre-calculated image feature thresholds from each channel (combination). These thresholds are dynamic, depending for instance on satellite geometry and atmospheric conditions. In contrast to ProPS, CMA-prob assumes independence of the different channel (combinations). Another deviation from ProPS is that CMA-prob excludes thin ice clouds with optical thickness smaller than 0.2 to prevent overfitting. For the pixels classified as cloudy by their initial procedure CMA-prob, CLAAS-3 employs a (separated) cloud-top phase determination. It relies on a series of threshold tests utilizing SEVIRI channels at wavelengths of 3.8, 6.3, 8.7, 10.8, 12.0 and 13.4 μm, as well as clear- and cloudy-sky simulated IR radiances and brightness temperatures. Additionally, consistency with the cloud optical thickness and particle effective radius retrieval from solar and NIR channel combinations is demanded (Meirink et al., 2022).

To compare ProPS and CLAAS-3, we use 12 SEVIRI scenes sampled at different seasons and different times of day. Figure B1 shows one such scene. The circumstances in which ProPS and CLAAS-3 differ

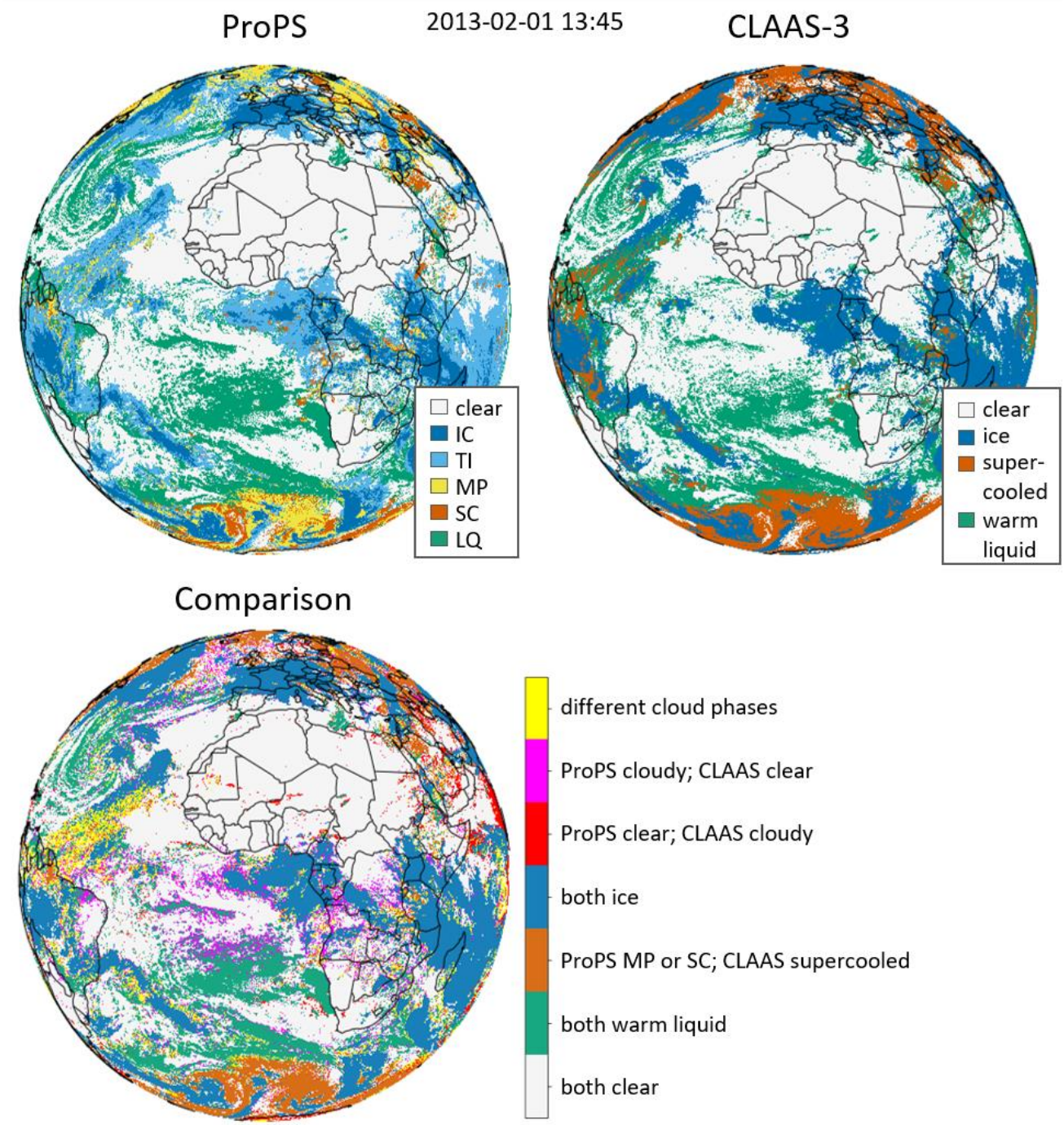

Figure B1. Comparison of ProPS with the CM SAF CLoud property dAtAset using SEVIRI - Edition 3 (CLAAS-3) for one exemplary SEVIRI scene. The upper row shows the results of both methods. The lower row shows the comparison of the ProPS and CLAAS-3 results.

in the figure are similar for the other scenes used in the comparison. Figure B2 is a statistic over all 12 scenes, comparing the classification of CLAAS-3 and ProPS.

Overall, the figures show that there is a good general agreement between the two methods. In Fig. B1, the positions and phases of the clouds generally agree well when looking at the "big picture". However, there are differences in the details.

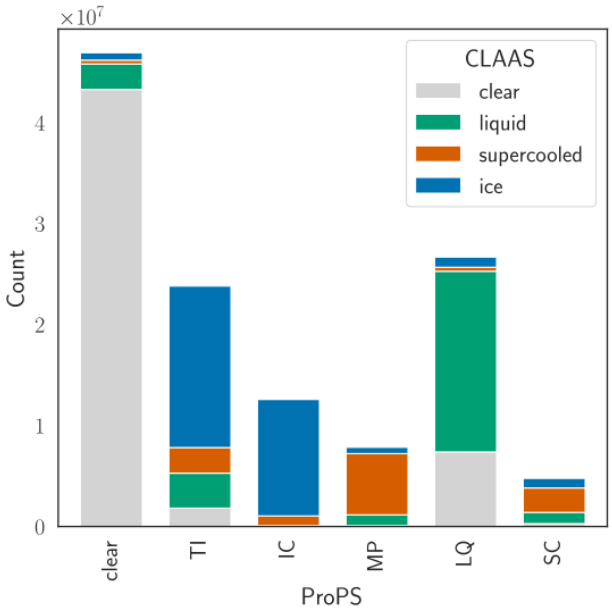

Figure B1. Statistic of the comparison of ProPS with CLAAS-3 over 12 SEVIRI scenes sampled at different seasons and different times of day.

For cloud detection, discrepancies between ProPS and CLAAS-3 could stem on the one hand from differences in the "training" datasets (ProPS employing DARDAR, while CLAAS-3 utilizes data from CALIPSO). On the other hand, there are some differences in the selection of SEVIRI channels and the conditions/thresholds employed, as well as the implementation of the Bayesian approach. These nuances likely contribute to the observed differences in cloud and phase detection.

We find that ProPS classifies more pixels as cloudy than CLAAS-3: For the 12 scenes, ProPS classified 62% of all pixels as cloudy, while CLAAS-3 classified 57% as cloudy. The differences between ProPS and CLAAS-3 are often found at the cloud edges, especially for small scale warm cumulus and thin cirrus clouds, both in general difficult cloud types to detect (e.g. in Fig. B1 the pink areas in the Tropics and the cumulus deck West of Africa). The agreement is better during the day than during the night, as expected. Especially low, warm clouds are difficult to distinguish from the surface using IR channels alone, leading to the larger discrepancies between ProPS and CLAAS-3 during the night compared to the day: During the day, ProPS and CLAAS-3 agree on the classification of 81% of all pixels; during the night they agree on 78% of all pixels. For thin ice clouds, the difference between the two methods might come (partly) from the exclusion of clouds with OT smaller than 0.2 in CLAAS-3. In general, ProPS tends to overestimate rather than underestimate the amount of cloud (as discussed in Section 6), i.e. it is a clear sky conservative algorithm, whereas CLAAS-3 seems to be a cloud conservative algorithm. Exceptions are high satellite zenith angles (> 70°) and bright surfaces (deserts, ice, snow), where CLAAS-3 has higher cloudiness values compared to ProPS.

Next, we take a look at the phase categorization of both methods. ProPS has an additional phase category, namely MP, which has no direct correspondence in CLAAS-3. We find that clouds classified as MP by ProPS are mostly categorized as supercooled by CLAAS-3; almost no ProPS MP clouds are classified as ice by CLAAS-3. The CLAAS-3 supercooled clouds are also the largest contribution to the ProPS SC category. The main differences in phase detection (as the cloud detection) are found at cloud edges or at the transition regions between different phases (in Fig. B1 for instance at the transition between supercooled and warm liquid clouds over the Southern Ocean). The phase category of ProPS which differs the most from CLAAS-3 are thin ice clouds (see TI bar in Fig. B2):

ProPS categorizes more pixels as thin ice than CLAAS-3. In most cases, ProPS and CLAAS-3 agree on the existence and position of thin ice clouds, however they often have a larger extent in ProPS (see the yellow regions in Fig. B1 at ice cloud edges). These differences might be due to the mentioned exclusion of clouds with OT smaller than 0.2 in CLAAS-3. The high sensitivity of ProPS to thin ice might however also lead to false alarms. CLAAS-3 categorises parts of the SC and MP categories of ProPS as warm liquid (green parts of MP and SC bars in Fig. B2), suggesting a tendency towards warmer cloud types in the CLAAS-3 classification scheme compared to ProPS."

In the discussion in chapter 8.1 it is mentioned that a difficult case is represented, perhaps unsurprisingly, by overlapping cloud layers with a high thin ice cloud over a low liquid cloud. Does the measure of certainty provide an extra information that can be used to isolate these cases? More in general, it would be interesting also to show an example of how the certainty measure looks like for a typical example of the product as in Fig. 7. This is very useful especially when the first phase choice is only marginally more certain than the second choice (as it is mentioned in chapter 8.2 in relation to the POD of MP and SC types). This discussion could be added to chapter 8.3 which currently is fairly limited in content.

We agree that a more detailed discussion of the certainty measure and its meaning is valuable. We have therefore added a certainty measure panel to the example curtains in Fig. 7 as suggested (see figure below). We have added a brief discussion of the meaning of the certainty measure in the example figures in the discussion of challenging situations for the ProPS retrieval in section 8.1 as follows:

"Often, the ProPS q^* is spatially slightly shifted against the DARDAR results, especially in the high latitude example in Fig. 7(a) where q^* is often slightly shifted to the left relative to q_{dardar}. This is most likely due to the different viewing geometries of the two instruments. Further, as SEVIRI looks at the clouds under a given angle, a high cloud can cover a neighbouring lower cloud from SEVIRI's perspective. In addition, the cloud cover in the rest of the SEVIRI 2D pixel can be different from that in the overflight swath of the polar orbiting satellite, and there can be a time difference of up to 7.5 minutes between the satellites. These effects could explain some of the differences between the ProPS and DARDAR classifications, especially for high certainty pixels where we expect the classification to be correct. However, these effects are difficult to account for in a quantitative evaluation (see Sect. 8.2) and lead to lower probabilities of detection.

The example figures also demonstrate that the cloud situation is often complex, with multi-layered clouds at different altitudes, cloud phase changes on small scales, and other atmospheric factors such as aerosols. The certainty parameter can be an indicator of the complexity of the scene: Complicated cloud scenes, such as multi-layered clouds or rapidly changing phases on small scales, tend to have lower certainty values compared to simpler scenarios. For example, the certainty drops from almost one to lower values in Fig. 7(a) to the left and right of the thick ice cloud, where it becomes thinner with underlying liquid layers."

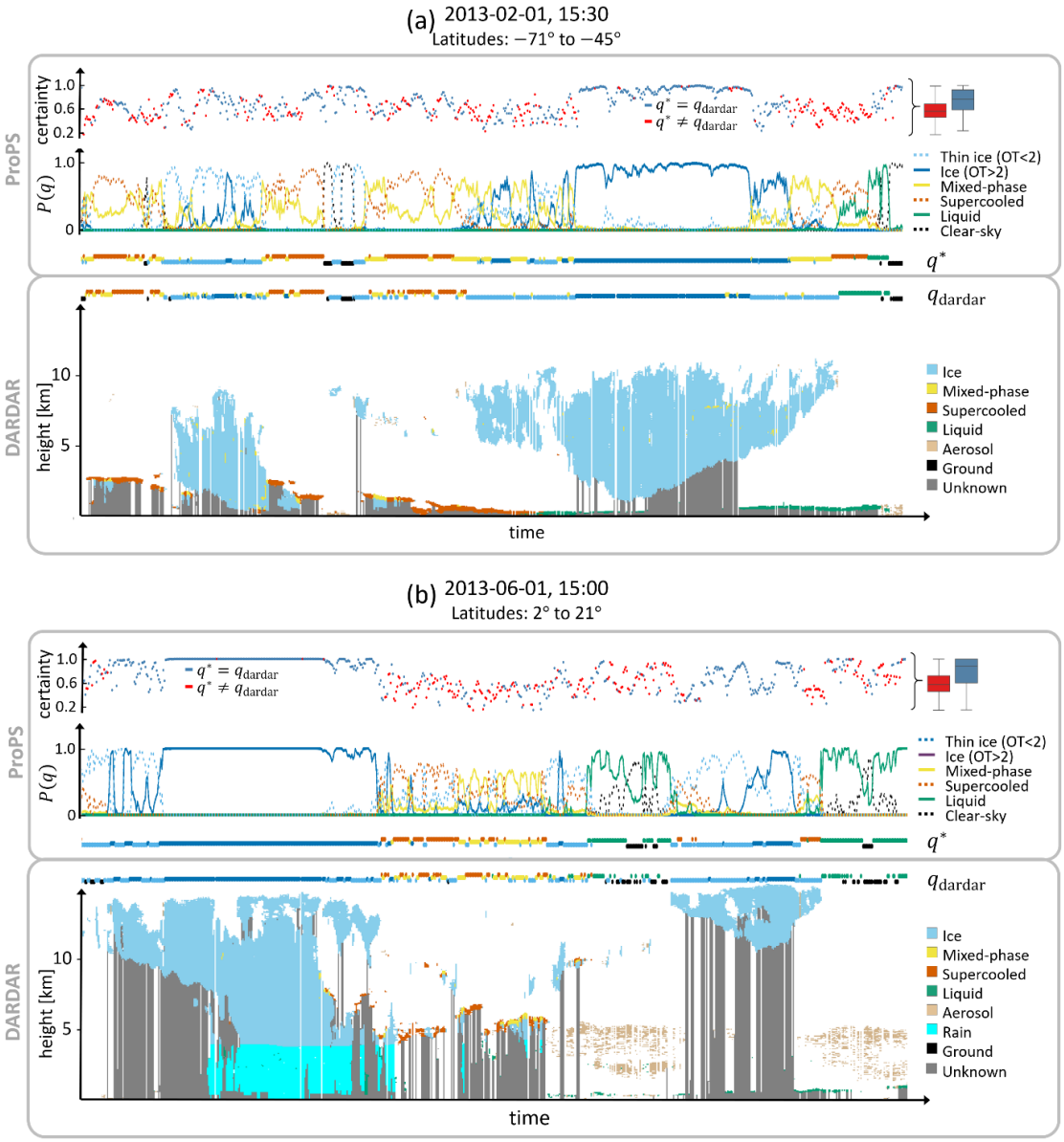

Figure 7. Example application of ProPS to DARDAR tracks in (a) high latitudes and (b) low latitudes. The bottom panel of each sub-figure shows the DARDAR curtain coarsened to SEVIRI resolution; the corresponding results of the ProPS algorithm (probabilities P(q)) are shown in the panels above. The cloud state retrieved from DARDAR, q_{dardar}, and the most likely cloud state from ProPS, q^*, along the track are shown in between (in the same colour code as P(q)). Above the P(q) panels, the corresponding certainty of the ProPS results are shown, with the color code indicating whether q^* agrees with q_{dardar}. The box plots on the right show the quartiles of the certainty measure for disagreement ($q^* \neq q_{dardar}$; red) and agreement ($q^* = q_{dardar}$; blue).

Additionally, we have added the probability of detection (POD) and counts for each cloud phase as functions of the certainty parameter and their weighted mean in Figure 9 (see below). We have also added the false alarm rates (FAR) as functions of the certainty parameter for cloud detection and for each cloud phase to provide a more detailed analysis of the performance of ProPS and the interpretation of the certainty parameter. We have extended the discussion on the relation to the certainty parameter in section 8.3 as follows:

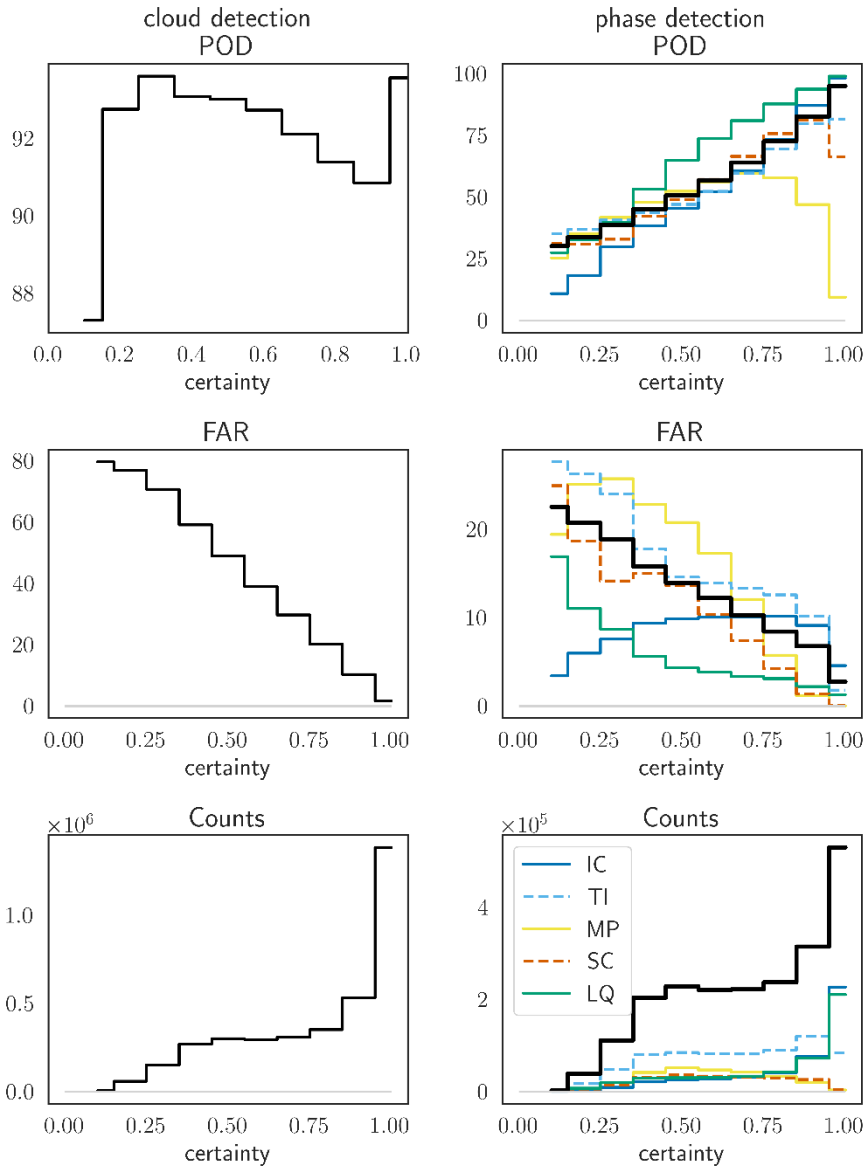

Figure 9. Top row: POD of cloud and phase detection (given that a cloud was detected) for each phase separately (in colour) and their weighted average (in black) as a function of the certainty parameter. Middle row: FAR for cloud and phase detection. Lower row left: Number of occurrences of certainty values. Lower row right: Number of occurrences of certainty values given a cloud was detected.

"One of the advantages of the Bayesian approach is the certainty parameter for the retrieval (see Sect. 3.3). For the example curtains in Fig. 7, the mean certainty values are shown on the right for pixels where ProPS and DARDAR agree or disagree. Where ProPS and DARDAR agree, the average certainty is higher, indicating that the certainty measure is meaningful. However, as the examples in Fig. 7 show, this is only true on average - there are still cases with a low level of certainty that are correctly identified, and vice versa.

Figure 9 gives an overview of the relation to the certainty parameter for the six months of validation data for the ProPS day version. It shows the POD and false alarm rate (FAR) for cloud detection and phase determination (given a detected cloud) for each phase separately and their average (weighted by the counts of each phase) per certainty bin of width 0.1. The two lower panels show the number

of occurrences of the certainty values. The average POD for cloud detection is high (> 90 %) for almost all certainty values; the FAR decreases monotonically with increasing certainty. This means that ProPS tends to overestimate cloud amount at low certainty values, as also mentioned in Sect. 8.2, but has an increased detection accuracy at higher certainty values. For phase determination, the average POD increases monotonically with the certainty parameter, while the average FAR decreases. Hence, the certainty parameter is a useful tool to decide whether to trust a result.

From the number of occurrences of certainty values (lower panels in Fig. 9) and examples as in Fig. 7 we see that the most unambiguous cases are clear sky, IC and LQ clouds (if their spatial extent is large enough to fill whole SEVIRI pixels). MP, SC and TI clouds have on average lower certainty values than the other cloud states."

A clarification regarding the certainty measure: how should it be interpreted when the probabilities of the two most probable states are both about 0.5 (one slightly higher than the other), while for all the other states probabilities are ~0? In this case has the certainty measure the same meaning as for a case where for example P(q*|M,A)=0.5 and the remaining 5 states have all P=0.1?

As this comment correctly describes, the certainty measure cannot capture all the information contained in the posterior P(q | M,A): Compare for instance the MP and SC clouds approximately in the middle of both scenes in Fig.7. In 7(a), only probabilities for MP and SC are above zero, meaning that the algorithm is fairly certain that it is one or the other. In 7(b) also TI and IC have probabilities above zero, which mirrors the more complicated cloud scene in this case where the MP and SC clouds are sometimes interrupted by ice in between. If both cases have similar certainty values, one would need to look at the posterior probabilities to get more detailed information.

Minor and technical comments

Abstract, line3: "mainly distinguished between" -> mainly identified/detected

We changed the phrase to "mainly detected".

Section 3.1, line 124: the sentence "These probabilities…" is repeated twice.

Thank you for noticing this mistake, we deleted it.

Section 4.1: not clear where the information content for the prior is shown in Fig 3. I imagine is the first panel, but from the figure caption is not clear.

We agree that the caption was unclear in this point. We rewrote the caption to explain Fig. 3 in more detail:

"First panel: Mutual information *I* between the latitude and the cloud state q (first row), cloudy/clear, abbreviated as c/c, (second row) and cloud phases (third row) for different sets of conditions *C*. This represents the information content of the different priors we considered, where latitude is a fixed condition, i.e. P (q | lat, *C*). Other panels: Mutual information *I* between SEVIRI channel (combinations) and cloud state q, cloudy/clear and cloud phases for different sets of conditions *C*. Empty spaces for *C* mean no condition, i.e. the starting point of *I* before conditions are introduced. The different mutual information values for q, cloudy/clear and phase indicate whether a channel (combination) contributes more to cloud or phase detection. The blue boxes indicate the

sets of conditions selected for ProPS."

Section 4.3, Line 229: this applies also to fractional cloud cover?
Yes, that is correct. We added this aspect to the sentence.

Section4.3: given the strong dependency on the surface emissivity at wavelength around 8.7 μm, should an emissivity map also be taken into account, or the surface type is enough to account for the effect?

We agree that an emissivity map could provide additional information for the BTD (10.8 - 8.7). However, we believe that the surface type constraint covers most of the emissivity differences for the channels. The advantages of the surface type variable are that 1) it can be used in several of the probabilities and only needs to be read in once, 2) it is (most of the time) a temporally fixed quantity that does not need to be retrieved and regridded for each time step, thus saving computational cost, and 3) it is a categorical variable rather than a continuous variable, which means that the data requirements for calculating probabilities using surface type as a condition are less than for using emissivity. For these practical reasons we chose surface type as a condition for the probability.

Section 4.5: The surface type can be used as a proxy for surface albedo, but does this also include the spectral variation of the albedo? Are significant changes in surface albedo between 0.6 μm and 1.6 μm (e.g. over snowy surfaces) important in this context or the surface type is enough?

Since the probabilities for the reflectance ratio are computed from the measured DARDAR data conditioned on the surface type, they already implicitly include the information on the spectral variation of the albedo for the given surface type. Only when the spectral variation of the albedo changes between pixels of the same surface type this is not included in the computed probabilities. However, we believe that in most situations this effect plays a minor role compared to the other conditions used. For this reason, and for practical reasons (mainly the limited amount of DARDAR 'training' data available), we have chosen not to include albedo.

Section 4.6: why the mutual information between RR1.6/0.6 and q seems to benefit by the inclusion of the surface type or the R1.6 but not from the use of both together?

This is an interesting question, but it is not easy to answer. One would need to study the information content of the channels and the different conditions in more detail, which is beyond the scope of this paper. In general, however, the mutual information does not have to increase when more conditions are added. For example, if we condition the mutual information between variables X and Y on a parameter C that is a confounder or mediator of X and Y, the mutual information will typically decrease (if there are no other parameters that lead to opposite effects).

Section 8.1, line 388: "by nature/DARDAR to it" not very clear, please rephrase. Also, missing bracket in "(see Fig 7".

We made the statement clearer by rephrasing to "…highlight the strengths of the ProPS retrieval and the challenges posed by, for example, complex cloud scenes or the different viewing geometries of polar orbiting and geostationary satellites"

Section 8.1, line 418: “spatially shifted against the DARDAR clouds” please clarify.

To make this clearer, we have added the following sentence, referring to examples in Fig. 7: “Often, the ProPS q^* is spatially slightly shifted against the DARDAR results, especially in the high latitude example in Fig. 7(a) where q^* is often slightly shifted to the left relative to q_{dardar}.”

Section 8.2, line 438: “less sensitive to optically thin clouds that->than Lidar”
Thank you for noticing this mistake, we corrected it.

Section 8.2, line 480: As discussed in the text, is it clear that SC and MP types are often difficult to distinguish, and that the certainty computed by ProPS is often marginally higher for either of the two types. Also, what is the confidence in the DARDAR supercooled water classification?

As the referee correctly describes, MP and SC are the two cloud states which are most difficult to distinguish. This can be seen from the example curtains in Fig. 7, from the POD of ProPS in Fig. 8 and from the on average lower certainty values of MP and SC compared to other cloud states (Fig. 9).

Regarding the confidence of DARDAR in supercooled water classification, unfortunately there is no uncertainty of the retrieved classification of DARDAR available in the product. In general, the phase classification in DARDAR is done using temperature information from ECMWF, the lidar backscatter, radar reflectivity and cloud layer thickness as criteria (Ceccaldi et al., 2013): Cloud layers containing supercooled water are identified by their strong lidar backscatter and subsequent attenuation in temperature ranges between 0 ◦C and -40 ◦C. A further distinction into pure supercooled water without ice crystals is made using the absence of radar return, since the diameter of cloud droplets is mostly below the CloudSat sensitivity (Hogan et al., 2003). If the layer is thicker than 300 m in the temperature range 0 ◦C to -40 ◦C, it is assumed to be fully glaciated.

Section 9: It would be interesting to briefly discuss how a new sensor (e.g. MTG FCI) would impact such an algorithm, both in terms of the information contained in the new channels and in the surface resolution.

We agree that the application of ProPS to new sensors is an interesting point. We therefore extended the discussion on further developments, focusing on the initial difficulty that new channels would need collocated active data in order to use them:

“In terms of further development of the ProPS method, the algorithm can be extended to other satellites with few modifications using for instance spectral band adjustment factors, as proposed by Piontek et al. (2023), since similar channels as used for ProPS are available in most current operational polar and geostationary satellite passive imagers. However, in order to incorporate and use channels not available to SEVIRI that contain phase information, such as additional channels in the near infrared of the Flexible Combined Imager (FCI) aboard the follow-on satellite of MSG (Meteosat Third Generation – MTG, launched on 13 December 2022 (Durand et al., 2015)), one first needs to collect a data set of collocated active observations to compute the necessary probabilities. In the future, this could be done with the EarthCARE satellite (Wehr et al., 2023) (planned launch May 2024).”

Figure 3: the caption could be clearer. Also, at first I did not understand that the first column of each panel (apart from the third) represents the starting point of each information content before the introduction of each new condition.

We have rewritten the caption of Figure 3 to clarify the unclear points (see above in response to the comment on Section 4.1).

Figure 6: perhaps adding another RGB composite helps a better comparison with the categorization as many of the high thin cirrus are lost in the RGB shown in the current figure.

Unfortunately, in many of the RGB composites in which thin ice are better visible, the low clouds are not very well visible. We therefore decided to keep the "natural color" RGB composite as a compromise for visibility of both high and low clouds.

References

Ceccaldi, M., Delanoë, J., Hogan, R. J., Pounder, N. L., Protat, A., and Pelon, J.: From CloudSat-CALIPSO to EarthCare: Evolution of the DARDAR cloud classification and its comparison to airborne radar-lidar observations, Journal of Geophysical Research: Atmospheres, 118, 7962–7981, https://doi.org/10.1002/jgrd.50579 , 2013

Durand, Y., Hallibert, P., Wilson, M., Lekouara, M., Grabarnik, S., Aminou, D., Blythe, P., Napierala, B., Canaud, J.-L., Pigouche, O., Ouaknine, J., and Verez, B.: The flexible combined imager onboard MTG: from design to calibration, https://doi.org/10.1117/12.2196644, 2015.

Hogan, R. J., Francis, P. N., Flentje, H., Illingworth, A. J., Quante, M., and Pelon, J.: Characteristics of mixed-phase clouds. I: Lidar, radar and aircraft observations from CLARE'98, Quarterly Journal of the Royal Meteorological Society, 129, 2089–2116, https://doi.org/10.1256/rj.01.208 , 2003

Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Meirink, J. F., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., Benas, N., van Zadelhoff, G.-J., Schlundt, C., Stein, D., Finkensieper, S., Håkansson, N., Hollmann, R., Fuchs, P., and Werscheck, M.: CLARA-A2: CM SAF cLoud, Albedo and surface RAdiation dataset from AVHRR data - Edition 2, https://doi.org/10.5676/EUM_SAF_CM/CLARA_AVHRR/V002 , 2017.

Meirink, J. F., Karlsson, K.-G., Solodovnik, I., Hüser, I., Benas, N., Johansson, E., Håkansson, N., Stengel, M., Selbach, N., Marc, S., and Hollmann, R.: CLAAS-3: CM SAF CLoud property dAtAset using SEVIRI - Edition 3, https://doi.org/10.5676/EUM_SAF_CM/CLAAS/V003 , 2022.

Piontek, D., Bugliaro, L., Müller, R., Muser, L., and Jerg, M.: Multi-Channel Spectral Band Adjustment Factors for Thermal Infrared Measurements of Geostationary Passive Imagers, Remote Sensing, 15, 1247, https://doi.org/10.3390/rs15051247 , 2023.

Wehr, T., Kubota, T., Tzeremes, G., Wallace, K., Nakatsuka, H., Ohno, Y., Koopman, R., Rusli, S., Kikuchi, M., Eisinger, M., Tanaka, T., Taga, M., Deghaye, P., Tomita, E., and Bernaerts, D.: The EarthCARE mission – science and system overview, Atmospheric Measurement Techniques, 16, 3581–3608, https://doi.org/10.5194/amt-16-3581-2023 , 2023.