EUMETSAT Satellite Application Facility on Climate Monitoring

Algorithm Theoretical Basis Document for the Cloud Probability product of the NWC/PPS package

NWC/CDOP3/PPS/SMHI/SCI/ATBD/CloudProbability, Issue 2, Rev. 0

Applicable to SAFNWC/PPS version 2021

Applicable to the following PGEs:

Acronym	Product ID	Product name	Version number
CMa-prob	NWC-154	Cloud Probability	1.1

Prepared by Swedish Meteorological and Hydrological Institute (SMHI)

REPORT SIGNATURE TABLE

Function	Name	Signature	Date
Prepared by	SMHI		26 April 2021
Reviewed by	SAFNWC Project Team EUMETSAT		 23 April 2021
Authorised by	Anke Thoss, SMHI *SAFNWC PPS Manager*		26 April 2021

Document Signature Table

	Name	Function	Signature	Date
Authors	Karl-Göran Karlsson Erik Johansson Salomon Eliasson Nina Håkansson Sara Hörnquist Joseph Sedlar	CM SAF scientists		26/04/2021
Editor	Rainer Hollmann	Science Coordinator		
Approval	Rainer Hollmann	Science Coordinator		
Release	Rainer Hollmann	Project Manager		

Distribution List

Internal Distribution	
Name	**No. Copies**
DWD Archive	1
CM SAF Team	1

External Distribution		
Company	**Name**	**No. Copies**
PUBLIC		1

Document Change Record

Issue/ Revision	Date	DCN No.	Changed Pages/Paragraphs
1.0	24/04/2015	SAF/CM/SMHI/ATBD/GAC/CMA-prob	Initial version, submitted for PCR 2.2
1.1	27/05/2016	SAF/CM/SMHI/ATBD/GAC/PBCM	Version submitted for DRR 2.2
1.2	02/08/2016	SAF/CM/SMHI/ATBD/GAC/PBCM	Updated version for DRR 2.2 Close-Out
1.3	20/12/2016	SAF/CM/SMHI/ATBD/GAC/PBCM	Version submitted for NWC SAF PPS 2018 PCR
1.4	17/10/2018	SAF/CM/SMHI/ATBD/GAC/PBCM	Version prepared for NWC SAF PPS 2018 ORR (equivalent to document NWC/CDOP3/PPS/SMHI/SCI/ATBD/1c)
1.0d	17/10/2018	NWC/CDOP3/PPS/SMHI/SCI/ATBD/CloudProbability	Same as previous issue but transferred from CM SAF versions to NWC SAF
1.0	13/12/2018	NWC/CDOP3/PPS/SMHI/SCI/ATBD/CloudProbability	Updates after NWCSAF/PPS v2018 DRR. RID-001, 002, 003, 004, 005, 007, 008, 010 Other changes: Added a comment on no-data in twilight.
2.0d	22/03/2021	NWC/CDOP3/PPS/SMHI/SCI/ATBD/CloudProbability	Updates for PPS v2021.
2.0	26/03/2021	NWC/CDOP3/PPS/SMHI/SCI/ATBD/CloudProbability	Updates after PPS v2021 RR: RID-7, 53: Defined ‘technically running’ RID-020, 043, 047: editorial RID-42: updated CM-SAF information

Table of contents

List of Figures

List of Tables

1 The EUMETSAT SAF on Climate Monitoring

The importance of climate monitoring with satellites was recognized in 2000 by EUMETSAT Member States when they amended the EUMETSAT Convention to affirm that the EUMETSAT mandate is also to “contribute to the operational monitoring of the climate and the detection of global climatic changes". Following this, EUMETSAT established within its Satellite Application Facility (SAF) network a dedicated centre, the SAF on Climate Monitoring (CM SAF, http://www.cmsaf.eu).

The consortium of CM SAF currently comprises the Deutscher Wetterdienst (DWD) as host institute, and the partners from the Royal Meteorological Institute of Belgium (RMIB), the Finnish Meteorological Institute (FMI), the Royal Meteorological Institute of the Netherlands (KNMI), the Swedish Meteorological and Hydrological Institute (SMHI), the Meteorological Service of Switzerland (MeteoSwiss), and the Meteorological Service of the United Kingdom (UK MetOffice), and the Centre National de la Recherché Scientifique (CNRS). Since the beginning in 1999, the EUMETSAT Satellite Application Facility on Climate Monitoring (CM SAF) has developed and will continue to develop capabilities for a sustained generation and provision of Climate Data Records (CDR’s) derived from operational meteorological satellites.

In particular the generation of long-term data sets is pursued. The ultimate aim is to make the resulting data sets suitable for the analysis of climate variability and potentially the detection of climate trends. CM SAF works in close collaboration with the EUMETSAT Central Facility and liaises with other satellite operators to advance the availability, quality and usability of Fundamental Climate Data Records (FCDRs) as defined by the Global Climate Observing System (GCOS). As a major task the CM SAF utilizes FCDRs to produce records of Essential Climate Variables (ECVs) as defined by GCOS. Thematically, the focus of CM SAF is on ECVs associated with the global energy and water cycle.

Another essential task of CM SAF is to produce data sets that can serve applications related to the new Global Framework of Climate Services initiated by the WMO World Climate Conference-3 in 2009. CM SAF is supporting climate services at national meteorological and hydrological services (NMHSs) with long-term data records but also with data sets produced close to real time that can be used to prepare monthly/annual updates of the state of the climate. Both types of products together allow for a consistent description of mean values, anomalies, variability and potential trends for the chosen ECVs. CM SAF ECV data sets also serve the improvement of climate models both at global and regional scale.

As an essential partner in the related international frameworks, in particular WMO SCOPE-CM (Sustained COordinated Processing of Environmental satellite data for Climate Monitoring), the CM SAF - together with the EUMETSAT Central Facility, assumes the role as main implementer of EUMETSAT’s commitments in support to global climate monitoring. This is achieved through:

- Application of highest standards and guidelines as lined out by GCOS for the satellite data processing,
- Processing of satellite data within a true international collaboration benefiting from developments at international level and pollinating the partnership with own ideas and standards,
- Intensive validation and improvement of the CM SAF climate data records,

- Taking a major role in data set assessments performed by research organisations such as WCRP. This role provides the CM SAF with deep contacts to research organizations that form a substantial user group for the CM SAF CDRs,
- Maintaining and providing an operational and sustained infrastructure that can serve the community within the transition of mature CDR products from the research community into operational environments.

A catalogue of all available CM SAF products is accessible via the CM SAF webpage, www.cmsaf.eu/. Here, detailed information about product ordering, add-on tools, sample programs and documentation is provided.

2 Introduction

This Algorithm Theoretical Basis Document (ATBD) describes a new probabilistic cloud masking product - denoted CMa-prob – which has been developed by CM SAF during the CDOP-2 and CDOP-3 phases. Based on Bayesian theory, it is complementary to the standard SAFNWC PPS CMa cloud mask with the latter being used for the CM SAF CLARA-A1 and CLARA-A2 data records (i.e., forming the basis of the Fractional Cloud Cover product CM-11011). For upcoming CLARA versions (i.e., CLARA-A3, and beyond), the probabilistic cloud mask will be used instead to improve the error characterisation of cloud masking and its influence on any downstream cloud, surface radiation and surface albedo products. For CLARA-A2, a preliminary demonstration product was provided for users to become acquainted with the product capability and for preliminary evaluation.

The ATBD generally follows the description of the method in Karlsson et al. (2015) but also describes several extensions which were introduced to improve the usefulness of the CMa-prob product. One of the most central new features is that the CMa-prob value of 50 % has been tuned to give optimal cloud detection everywhere regardless of underlying surfaces and observation conditions. Karlsson et al. (2020) describe the final design of the methodology.

Since the CMa-prob methodology is tightly linked to the NWC SAF PPS software package (i.e., with a shared use of image features and threshold datasets), the ATBD is now formally considered as one single NWC SAF document even if the product development took (and will continue to take) place in the CM SAF project. In this way, duplication of work associated with formal reviews in the two SAF projects is avoided.

2.1 Applicable documents

Reference	Title	Code	Version	Date
AD 1	NWC SAF Product Requirements Document	NWC/CDOP3/SAF/AEMET/MGT/PRD	1.4e	23/11/2020

2.2 Reference documents

Reference	Title	Code	Version	Date
RD 1	Algorithm Theoretical Basis Document for the Cloud Mask of the NWC/PPS	NWC/CDOP3/PPS/SCI/ATBD/CloudMask	3.0	26/04/2021

Reference	Title	Code	Version	Date
RD 2	Scientific and Validation Report for the Cloud Product Processors of the NWC/PPS	NWC/CDOP3/PPS/SMHI/SCI/VR/Cloud	3.0d	15/03/2021

3 Theoretical description of the CMa-prob method

3.1 Background – problems with traditional cloud masking and suggested new approaches

For many years, the definition of fixed cloud masks or cloud masks with a small set of quality flags has been the most common way of solving the cloud screening problem in applications based on passive multispectral satellite imagery. Many examples of this exist in the literature, e.g., Dybbroe et al., (2005a, 2005b), Kriebel et al. (2003), Derrien and LeGleau (2005), Frey et al. (2008) and Pavolonis et al. (2005). The use of a fixed cloud mask is straightforward for downstream applications (e.g. for Sea Surface Temperature (SST), surface albedo, clear sky radiance and NDVI vegetation index retrievals) which need to discard all cloudy pixels before the retrieval of the actual parameter. However, the drawback is that no, or very limited, information about the uncertainty in the cloud screening is generally available with these methods. Consequently, the error characteristics are generally unknown even if internal parameter-specific algorithm uncertainties may be known. Furthermore, a cloud mask is usually designed with a particular purpose or application in mind. Consequently, the performance may vary considerably from method to method depending on whether the cloud screening is executed in a clear-conservative way (i.e., defining clear pixels with high confidence) or in a cloud-conservative way (i.e., defining cloudy pixels with high confidence). Instead, the desire to define a more flexible cloud mask, suitable for any (or at least most) downstream applications, has become increasingly important. Such a cloud mask can either be expressed as a cloud index (as suggested by Khlopenkov and Thrishchenko, 2007) or a cloud probability (Merchant et al., 2005) meaning that any user should be able to define the most suitable mode of operation. In other words, it should be possible to use anywhere in the range from the clear-conservative mode to the cloud-conservative mode by just changing the tolerance level of the required cloud probabilities.

Although statistical and probabilistic (Bayesian) theory has been well established for decades (or even centuries), a problem has been to find appropriate observational references to represent the true global cloud occurrence from which a firm statistical cloud distribution database can be built. However, with the 2006 launch of the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) onboard the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) satellite, the situation has improved considerably. CALIOP offers global cloud observations with higher detection sensitivity than any other passive instrument (Winker et al., 2009). Furthermore, its measurements can be collocated (however, restricted to certain conditions) to observations by current operational AVHRR sensors. This has triggered numerous studies examining AVHRR-based cloud detection methods in detail (e.g., Karlsson and Dybbroe, 2010, Karlsson and Johansson, 2013 and Stengel et al., 2014). It has also paved the way for more systematic attempts to provide cloud probabilities rather than fixed cloud masks (Heidinger et al., 2012, and Musial et al., 2014), and the CMa-prob development described here is another example of this. Recently also improved versions of the CALIOP cloud datasets have been utilised for in-depth studies of the cloud detection limits for methods

based on AVHRR data (Karlsson et al., 2018). Their findings and their tools have been important for defining the final concept of the CMa-prob methodology (as reported by Karlsson et al., 2020 and further outlined in Section 3.6 and 3.7).

3.2 Bayesian theory

Let us first recapitulate some fundamentals of the probabilistic statistical theory. The theory is based on the pioneering work by Thomas Bayes. He formulated already in 1763 his famous theorem (nowadays referred to as Bayes' Theorem) for estimation of the posterior probability of an event as a function of likelihoods (conditional probabilities) and a priori probabilities of other events. In the context of analysis of radiance feature vectors measured by satellite sensors, we may express Bayes' Theorem as follows: If $\boldsymbol{F}$ is a vector of satellite radiances or image features (e.g., brightness temperature differences or reflectances), we may denote the posterior conditional probability that it is cloudy when $\boldsymbol{F}$ is given as $P(cloudy|\boldsymbol{F})$. In the same sense, we may denote the conditional probability that vector $\boldsymbol{F}$ occurs given it is cloudy as $P(\boldsymbol{F}|cloudy)$. If also introducing the overall probability (climatological mean) that it is cloudy as $\overline{P(cloudy)}$ and the overall probability that any given value of $\boldsymbol{F}$ occurs as $P(\boldsymbol{F})$, we may write Bayes' Theorem as follows:

$$P(cloudy|\boldsymbol{F}) = \frac{\overline{P(cloudy)}P(\boldsymbol{F}|cloudy)}{P(\boldsymbol{F})} \tag{1}$$

Despite its simple form, the solution to Eq. 1 is not easy to find in a situation with multispectral measurements (i.e., when dimensions of $\boldsymbol{F}$ are large). The estimation of parameters on the right-hand side of Eq. 1 (especially $P(\boldsymbol{F}|cloudy)$) becomes increasingly difficult as more image features that are chosen or added. It then requires extraction from huge statistical training datasets to fully describe the dependence on individual image features and, in addition, the effect of their mutual correlation. What complicates things even further is that even with one specific realisation of feature vector $\boldsymbol{F}$, probabilities may differ depending on different environmental situations (e.g. if the pixel measurement is made in winter or in summer, over land or the ocean, in mountainous terrain or over the desert, etc.). Thus, the training process needs to consider additional ancillary information for an accurate description of the environmental conditions.

To reduce complexity of the problem, one can utilise some approximations. One approach could be the entirely empirical approach of estimating $P(cloudy|\boldsymbol{F})$ directly from predefined Lookup Tables composed during training with some stratification based on ancillary data. Musial et al. (2014) demonstrated such a method. Alternatively, some simplifications and approximations could be made to Eq. 1. One such simplification is to use the *Naïve Bayesian approach*, as is done for the CMa-prob method.

3.3 The CMa-prob Naïve Bayesian approach

If assuming that individual image feature components f_i in $\boldsymbol{F}$ are all independent (i.e., image features are uncorrelated), a simplification is possible so that individual probabilities may multiply to get the total probability, following the fundamental statistical rule for "Compound Probability of Independent Events". Thus, Eq. 1 reduces to

$$P(cloudy|\boldsymbol{F}) = \frac{\overline{P(cloudy)}\prod_i P(f_i|cloudy)}{P(\boldsymbol{F})} \quad (2)$$

This approximation of Bayes' Theorem is denoted **the Naïve Bayesian approximation**.

The problem now reduces to estimating individual probabilities $P(f_i|cloudy)$ and then simply multiplying them. However, it must be emphasized that even if we have achieved a very simple equation for calculation of the probabilities, the big scientific challenge lies in defining and estimating the conditional probabilities on the right-hand side of the equation. This includes the very fundamental choice of appropriate image features f_i. These must be chosen in an optimal way, with each individual feature having documented capability to provide essential information about cloud occurrence. The following sub-sections will describe the methodology used and the choice of optimal image feature components f_i.

We also notice that there must be a mutual inter-dependence between $P(f_i|cloudy)$ and $P(cloudy|f_i)$. More clearly, if knowing the conditional probability that it is cloudy given a certain image feature value, we can also calculate it the other way around from the same statistical training dataset (provided that both absolute and relative frequencies of cloud occurrences are stored). The CMA-prob method utilises this fact in its definition. It is often more practical and easier to understand results for individual image features if compiling statistical distributions of $P(cloudy|f_i)$. It also provides a natural link to development efforts trying to find optimal thresholds in multispectral thresholding schemes like the standard CMa method in the PPS cloud processing package. The remaining factors on the right-hand side of Eq. 2 may also be calculated from training data. An estimation of the mean cloud occurrence $P(\overline{cloudy})$ is possible and the factor $P(\boldsymbol{F})$ may be estimated by summing contributions from both cloudy and clear cases and then compute the overall frequency for which any particular realisation of vector $\boldsymbol{F}$ occurs.

The Naïve Bayesian approximation has been successfully applied to many scientific applications (e.g., Kossin and Sitkowski, 2009) as well as to the AVHRR cloud screening problem (Heidinger et al., 2012). The main difference between CMa-prob and the latter method lies in the choice of image features and the used ancillary information.

3.4 Estimating conditional cloud probabilities from CALIPSO measurements

Having access to a system that can match and co-locate CALIPSO/CALIOP and NOAA/METOP AVHRR measurements makes it relatively straight-forward and simple to estimate conditional cloud probabilities, i.e. the frequency that it is cloudy at a certain image feature value $P(cloudy|f_i)$. However, as previously mentioned, this should only be done by using some restrictions on e.g. illumination conditions and the geographical coverage in order to avoid distributions either having a dynamical range of probabilities being too limited or too broad. If these restrictions are not considered, the final ability to separate cloudy from cloud-free radiances would be reduced (i.e., too often give cloud probabilities close to 50 %).

Figure 3-1 shows an example of estimated cloud probabilities as a function of the AVHRR visible reflectances from the 0.6 μm channel over tropical ocean surfaces (left panel) and high-latitude homogeneous surfaces with permanent snow-cover (right panel). Detailed information on all surface categories and the definition of AVHRR-based image features is given later in the next section (Table 3-1 to Table 3-4). The reference CALIPSO results are original CALIOP

cloud masks (CLAY product version 4.1) which have been used to calculate cloud probabilities as a function of individual image feature values (i.e., reflectances and brightness temperatures or combinations of channel parameters). All CALIOP-detected clouds were used to calculate these cloud probabilities, and this explains why the cloud probabilities never reaches the zero level (i.e., very thin clouds may exist despite very low measured reflectances in Figure 3-1). It should also be mentioned that daytime conditions in Figure 3-1 are valid for solar zenith angles below 80 degrees. Night-time conditions are defined for solar zenith angles above 89 degrees. The intermediate category Twilight is consequently defined by solar zenith angles between 80 and 89 degrees.

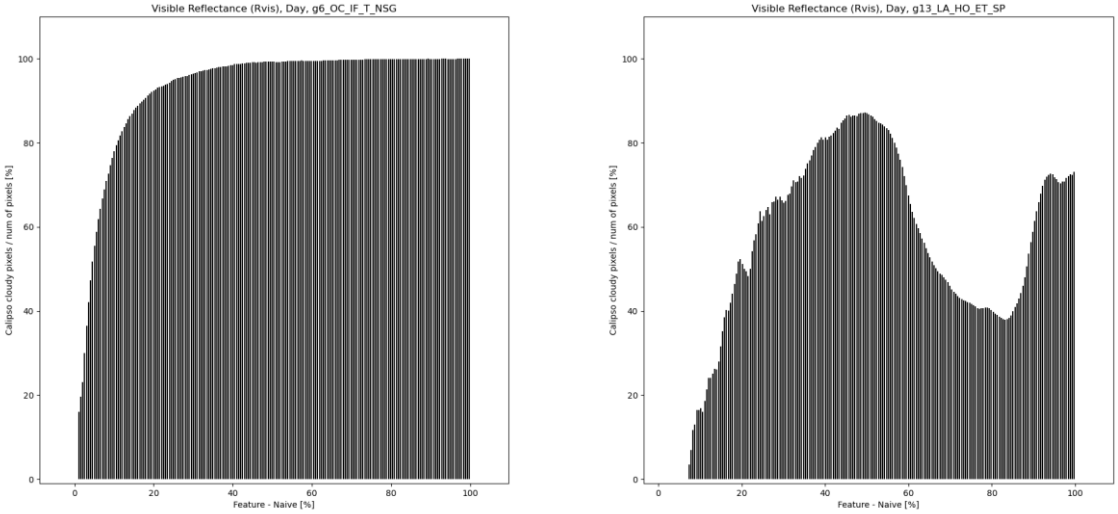

Figure 3-1 *Cloud probabilities estimated from global CALIPSO-CALIOP cloud data in the period 2006-2015 as a function of AVHRR 0.6 µm visible reflectances (Feature-Naïve in %, denoted Rvis).* ***Left:*** *Results over* ***Tropical ocean*** *surfaces (G6 in Table 3-4).* ***Right:*** *Results over* ***High Latitude surfaces with permanent snow-cover*** *(G13 in Table 3-4).*

From Figure 3-1, we conclude that cloud probabilities increase rapidly with reflectance over a very dark surface such as the ice-free ocean. Probabilities exceed 50 % already at a very low reflectance value (at approximately 5 % reflectance) and reach above the 80 % level near 12 % reflectance. Thus, conditions for cloud-screening appear almost ideal (i.e., applying a simple threshold at 5 % reflectance would give very satisfying results). This is not the case for conditions with the snow-covered ground in Figure 3-1 (right). Here, it is difficult to arrive at a robust, unique reflectance value where cloud probability exceeds 50 % (which would be needed for this image feature to be useful for cloud screening purposes). This occurs only for moderate to large reflectance values (30-60 % and above 90 %. For the intermediate region of high reflectances, probabilities are actually well below 50 % and even below 40 % for reflectivities between 85 % and 90 %. The reduced probability, especially in the interval 70-90 %, means that most cloud-free snow surfaces have high reflectances in this interval which are very similar to cloud reflectivities. Consequently, cloud-free conditions are more likely to occur than cloudy conditions in this reflectance interval. We also notice that zero reflectances are never reached over these bright surfaces.

A similar situation appears over the same surfaces for the difference between the infrared brightness temperature of the 11 µm channel and the surface skin temperature (Figure 3-2, difference denoted Tirdiff). In this figure, positive feature values mean that measured brightness temperatures are ***colder*** than the surface temperature. Notice also that these plots only show the relative cloud frequency or cloud probability for each feature value on the x-axis. Thus, there is no information about the occurrence density for each feature value here, meaning that spurious noisy values or some rare deviations, e.g., due to bad navigation (see outliers in the right part of Figure 3-2) appear to be more frequent than they really are. These outliers represent just a very small number of cases.

Excellent separability conditions are seen over tropical ocean surfaces where a threshold near 5 K would identify the vast majority of clouds. The situation is more problematic over snow-covered surfaces. In particular, notice the effect of near-surface temperature inversions (e.g., for negative temperature differences in Fig. 3-2 meaning that measurements are ***warmer*** than the surface temperature). We notice that clouds which are warmer than the surface may occur over the tropical ocean (e.g., a small fraction of marine stratocumulus) as well as over snow-covered surfaces. However, such clouds occur much more frequently over snow-covered surfaces. While the temperature difference (i.e., the inversion strength) is generally small over the ocean, it is generally much larger over snow-covered surfaces. The latter surfaces are dominated by cold and highly elevated surfaces in Antarctica and Greenland. It is clear that the use of a single threshold, e.g. at 5 K, would be able to detect the vast majority of clouds over the tropical ocean while the same threshold applied over snow-covered surfaces would instead miss a large fraction of all existing clouds.

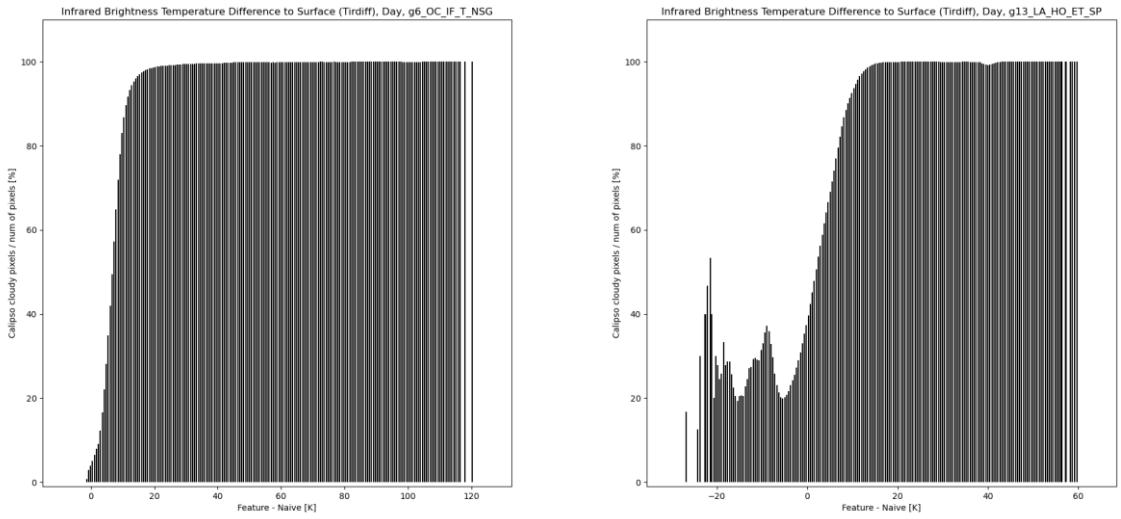

Figure 3-2 *Cloud probabilities estimated from CALIPSO-CALIOP cloud data in the period 2006-2015 as a function of the temperature difference between the ERA-5 (https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation) surface skin reference temperature and the AVHRR 11 µm brightness temperature (Feature_Naive in K, denoted Tirdiff).* ***Left:*** *Results over* ***tropical ocean*** *surfaces during day (G6 in* Table 3-4*).* ***Right:*** *Results over* ***High Latitude snow-covered*** *surfaces during day (G13 in* Table 3-4*).*

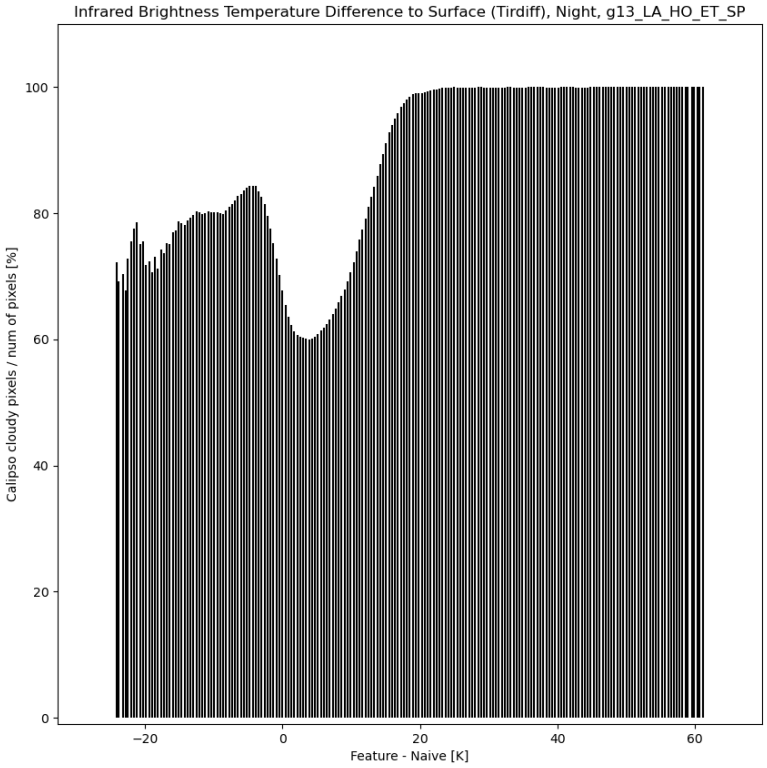

Figure 3-3 *Cloud probabilities estimated from CALIPSO-CALIOP cloud data in the period 2006-2015 as a function temperature difference between the ERA-5 surface skin reference temperature and the AVHRR 11 µm brightness temperature (Feature_Naive in K, denoted Tirdiff). Results are valid over* ***High Latitude snow-covered*** *surfaces (G13) during night and with strong near-surface inversions.*

Statistics from situations with very strong near-surface inversions, i.e., when surface temperatures are colder than temperatures in the 950 hPa level, over snow-covered surfaces at night in Figure 3-3 are also interesting. In such situations, using this temperature difference feature for cloud screening would be even worse. Clouds here may occur for any measured temperature difference but often with probabilities which are not far from 50 % (the level of maximum uncertainty). Only for clouds which are more than 20 K colder than the surface (generally medium- and high-level clouds) or possibly also clouds which are 5-15 K warmer than the surface (low-level inversion clouds) we get high enough cloud probabilities to detect clouds with certainty. We also notice a minimum in cloud probabilities at about 5 K instead of near 0 K which would be the anticipated value, i.e. a pixel that is neither warmer nor colder than the surface should be generally cloud-free. This indicates that the ERA-5 surface temperature reference probably has a small warm bias over these surfaces, i.e., that surface temperature inversions are probably somewhat stronger than ERA-5 is capable of showing. This is supported by, e.g., long-time experiences of a warm bias for ECMWF-forecasted minimum 2-meter temperatures in boreal forest regions at high latitudes during winter (Hogan et al., 2017). However, some cases of large positive differences in Fig. 3.3 may be linked to navigation errors in coastal areas with steep orography. Comparison of Figures Figure 3-3 and Figure 3-2 (right) reveal that cloud probabilities are generally higher over these surfaces at

(Polar) night conditions. However, a significant fraction of these clouds comes from very thin clouds that are not likely to be detected in passive satellite imagery.

We conclude from Figure 3-1 to Figure 3-3 that conditions for efficient cloud screening may be drastically different depending on the geographic location and the prevailing illumination conditions (i.e., day, night or twilight). This is one of the explanations for the very successful performance of simple bi-spectral VIS-IR cloud screening methods at low- to moderate latitudes (best exemplified by the results derived mainly from geostationary satellite data of the International Satellite Cloud Climatological Project – ISCCP – see Rossow et al., 1999 and Young et al., 2018). On the other hand, it also clearly illustrates potential serious limitations for the same methods over high latitudes and over the polar regions (as highlighted by Karlsson and Devasthale, 2018).

3.5 Definition of a basic sub-set of constrained AVHRR image features

The Naïve Bayesian CMa-prob method utilises estimated conditional cloud probabilities (introduced in the previous section) for a sub-set of image features. However, rather than to define them in their purest form (as illustrated in Figure 3-1 to Figure 3-3) we have chosen to define them linked to pre-calculated dynamic image feature thresholds used by the Polar Platform System cloud software package (PPS, see Dybbroe et al, 2005a, 2005b) and in this particular case PPS version 2021[RD 1]. The reason for linking image features to pre-calculated thresholds is that the latter are defined across a wide range of environmental conditions (see Dybbroe et al., 2005a for more details). This concerns image feature variability due to the following factors: Solar and satellite geometry (direct angular dependence and dependence on scattering angles), prevailing atmospheric profiles of temperature and humidity, climatological ozone and aerosol amounts, topography and land cover, and spectral surface emissivities. Without considering

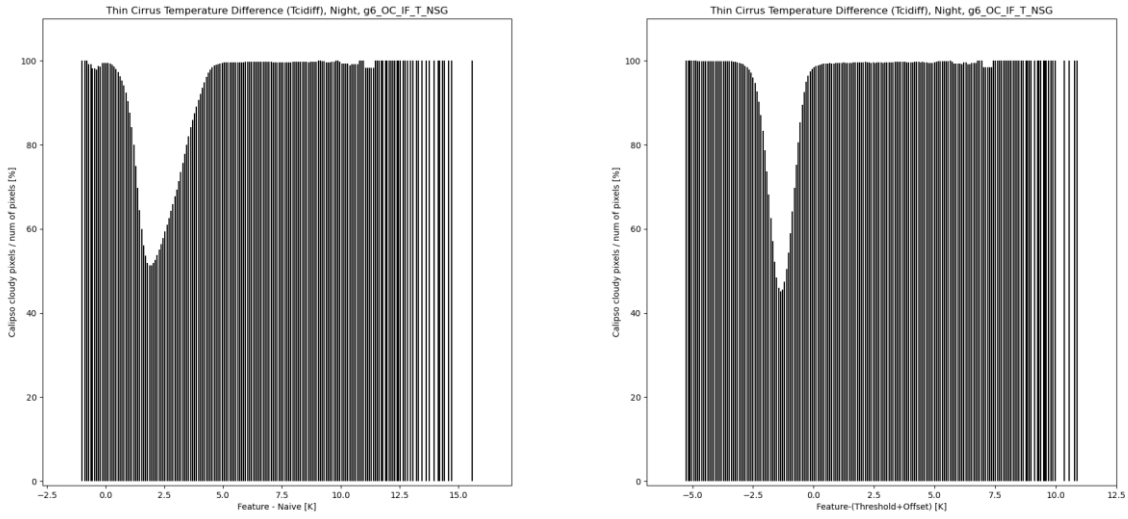

Figure 3-4 *Cloud probabilities estimated from CALIPSO-CALIOP cloud data in the period 2006-2015 as a function of AVHRR temperature differences between AVHRR channel 4 and 5 over Tropical ocean surfaces during night (surface group G6 in* Table 3-4*). **Left panel** shows results in original form and **right panel** if plotting results as a function of temperature differences subtracted with PPS thresholds (consisting of a dynamic threshold plus a tuning offset value).*

these factors when training the probabilistic classifier, the results risk being imprecise and, most likely, misleading. We claim that it is better to piggy-back on existing prepared threshold information, based on knowledge built on many years of cloud thresholding experience, than to try to train a classifier from scratch as a function of all these mentioned factors influencing the results. The latter would require the creation of very large dimension Look-up Tables of statistical relations of cloudiness and image features and their respective dependencies on a wide range of environmental factors.

To illustrate the usefulness of this concept we consider one of the most commonly used AVHRR image features for the detection of thin cirrus clouds, initially suggested by Inoue (1987): the brightness temperature difference between AVHRR channels 4 and 5 at 11 μm and 12 μm, respectively (denoted Tcidiff in Table 3-3). The main principle used for Cirrus detection is normally that the cloud transmissivity for thin ice clouds is higher in AVHRR channel 4 than in AVHRR channel 5, thus creating a positive brightness temperature difference between AVHRR channels 4 and 5. Figure 3-4 shows the cloud probabilities as a function of this temperature difference (Figure 3-4, left) but also as a function of the temperature difference relative to the corresponding PPS threshold (Figure 3-4, right).

In its original form (Figure 3-4, left), we have two peaks in cloud occurrence where one is for differences close to zero, and the other for values exceeding approximately 4 K. Thus, the area with lower cloud frequencies between the peaks spans an interval of almost 4 K. In the alternative formulation (Figure 3-4, right) results are much more distinctly organised where the interval with lower frequencies is now reduced to only about 2 K. We interpret this as primarily an effect of being able to take into account the natural cloud-free contribution from atmospheric water vapour emission in the split-window channels. This emission is also able to create a clear temperature difference in the absence of cirrus clouds explaining the broader less distinct probability distribution in its original form for temperature differences below approximately 4 K in Figure 3-4 (left). The threshold adjusted distributions after the coordinate change now clearly separate thin cirrus clouds to the right in the plot from the opaque clouds in the left part of the plot with cloud-free cases (although still allowing occurrence of very thin clouds) now concentrated around the x-coordinate value of around –1 K.

Table 3-1 *Spectral channels of the Advanced Very High Resolution Radiometer (AVHRR). Three different versions of the instrument are described as well as corresponding satellites.*

Channel Number	*Wavelength (μm) AVHRR/1 Tiros-N, NOAA-6,8,10*	*Wavelength (μm) AVHRR/2 NOAA-7,9,11,12,14*	*Wavelength (μm) AVHRR/3 NOAA-15,16,17,18 NOAA-19, Metop-A Metop-B, Metop-C*
1	0.58-0.68	0.58-0.68	0.58-0.68
2	0.725-1.10	0.725-1.10	0.725-1.10
3a	-	-	1.58-1.64
3b	3.55-3.93	3.55-3.93	3.55-3.93
4	10.50-11.50	10.50-11.50	10.50-11.50
5	Channel 4 repeated	11.5-12.5	11.5-12.5

Table 3-2 *Used AVHRR image features for* ***day illumination*** *probabilistic cloud masking.*

Feature name	Definition	Main cloud detection ability
Rvis	Over land: AVHRR channel 1 TOA reflectances minus PPS thresholds Over ocean: AVHRR channel 2 TOA reflectances minus PPS thresholds	Identification of bright clouds over dark Earth surfaces
Tirdiff	AVHRR channel 4 brightness temperatures minus ERA-5 surface skin temperatures minus PPS thresholds	Identification of clouds being significantly colder than the Earth surface
Rswir_3a *(morning orbit AVHRR/3)*	AVHRR channel 3a TOA reflectances divided by AVHRR channel 1 TOA reflectances	Identification of clouds with significant reflection in the visible near-infrared region (in particular water clouds and thick multi-layered ice clouds over snow-covered surfaces)
Rvis37 *(afternoon orbit all AVHRRs and morning orbit AVHRR/1 + AVHRR/2)*	AVHRR channel 3b TOA reflectances	Identification of clouds with significant reflection in the short-wave infrared region (water clouds and thick multi-layered ice clouds
Rswir_quota *(afternoon orbit all AVHRRs and morning orbit AVHRR/1 + AVHRR/2*	AVHRR channel 3b TOA reflectances divided by AVHRR channel 1 TOA reflectances	Identification of clouds with significant reflection in the short-wave infrared region (water clouds and thick multi-layered ice clouds), especially over ice and snow surfaces.
Rswir_3b (afternoon orbit all AVHRRs)	AVHRR channel 3b brightness temperatures minus AVHRR channel 5 brightness temperatures minus PPS thresholds	Identification of clouds with significant reflection in the visible near-infrared region (in particular water clouds and thick multi-layered ice clouds over snow-covered surfaces)
Texture_day	Over land: Not used (surface variability is generally too large)! Over ocean: Sum of local 3x3 pixel variances for AVHRR channel 4 brightness temperatures and AVHRR channel 1 TOA reflectances	Identification of fractional or broken clouds over ocean

With this background. we now list in Tables Table 3-2 and Table 3-3 a set of constrained image features (related to original AVHRR channels as described in Table 3-1) used for the definition of the CMa-prob probabilistic cloud mask estimates. Different image features are selected for two scene categories: day (solar zenith angle [SZA] below 80º); Tab. 3-2 and night (SZA ≥ 89º; Tab. 3-3). For the twilight case (80º ≤ SZA < 89º) either the day or night approach is used depending on if non-zero reflectances are detected in AVHRR channel 1. Notice that, because of the differences between the three AVHRR versions (Table 3-1) and the strategy of the switching of 3a and 3b channels for AVHRR/3 (i.e., switching active only for satellites in morning orbit), only a subset of the listed image features in Tables Table 3-2 and Table 3-3 are used simultaneously and, therefore, the number of used features varies from 3 for AVHRR/1 up to 6 for AVHRR/3).

Only one feature, Tirdiff, is used both day and night for all AVHRR versions. However, it should be noted that the underlying statistics compiled during the training of the method is here separated into day and night categories. One of the night features in Table 3-3, Tcidiff, is also used during day for AVHRR/3 when the 1.6 micron channel is active. The reason is that this feature contributes to the detection of thin ice (cirrus) clouds. This detection is difficult when using features involving the1.6 micron channel. Thus, by adding the use of Tcidiff in this case we are able to get approximately the same performance as we get from the daytime feature Rswir_3b in Table 3-2, a feature which is capable of detecting both ice and water clouds.

Overall, it is clear that the information in the 3.7 micron channel (channel 3b) is fundamental for the success of cloud detection based on AVHRR data. This fact clearly distinguishes AVHRR-based cloud datasets from cloud datasets derived from other historic imaging sensors. However, this channel has unfortunately also specific problems which must be accounted for. One problem that specifically occurs in data from AVHRR/1 and AVHRR/2 is the appearance of noise stripes in the images with varying intensity and impact. The noise seriously affects all image features based on channel 3b and especially those being used during night in Table 3-3 (since the noise is most serious for cold targets in the images). We have used a noise filtering method proposed by Karlsson et al, (2015) for minimising the effects of channel 3b noise.
Another limitation of channel 3b is that the channel easily saturates, i.e., reaching static maximum brightness temperatures, over very warm surfaces such as over deserts (as discussed by Trischenko et al., 2002). This is more serious for AVHRR/1 and AVHRR/2 versions since the saturation temperature is close to 322 K while the sensitivity has increased to approximately 335 K for AVHRR/3 (Trischenko et al., 2002). A consequence of this is that features based on channel 3b cannot be used properly when temperatures are close to or exceeding these saturation limits. For CMA-prob we have excluded the use of features derived from channel 3b for the AVHRR/1 and AVHRR/2 sensor versions if brightness temperatures in this channel exceed 320 K and if at the same time brightness temperatures in channel 4 exceed 305 K. No actions have been taken for AVHRR/3 since the increased sensitivity for this sensor has basically solved the problem of saturation.

To account for geographical and topographical differences, we defined 28 geographical surface categories over which we trained the probabilistic classifiers. These categories are listed in Table 3-4. Ice cover information is taken from OSI SAF ice concentration data and snow cover information (interpreted from the snow depth parameter) is taken from ERA-5 re-analyses. The land-sea mask is taken from USGS landuse dataset and the labelling of whether or not the surface is dry is based on land emissivity climatologies from MODIS Collection 6.1 (see RD 1 for more details on the use of ancillary datasets). The coastal categories (G27 and G28) were

defined as pixels where 25-90 % of the local surroundings (45x45 km for 5 km GAC data and 11x11 km för 1 km datasets, defining the base for the calculation of the Fraction of Land parameter) belonged to another Earth surface category (land or ocean) than the current pixel. Observe that we only use the specific coastal categories for satellite viewing angles below 58 degrees. For larger viewing angles the land-sea differences will be more clearly mixed up due to very large FOVs. To still not be too vulnerable to, e. g., small navigation errors outside of the threshold maximum viewing angle, the land-sea distinction is here specified by a 25 % limit of the Fraction of Land parameter. The classifier was trained using the CALIPSO-CALIOP cloud product, denoted Cloud and Aerosol Layer Information product version 4.1.

Several sunglint categories are included among the ocean surfaces in Table 3-4. However, generally the matching of CALIOP and AVHRR observations does not allow studying sunglint effects, thus no statistics for sunglint conditions can be compiled. We have overcome this problem by utilizing the fact that some sunglint conditions are actually encountered in CALIOP-AVHRR matches because of orbital drift effects of NOAA-18 and NOAA-19 satellites in the 2011-2015 period. This have led to increasing viewing angles in matched AVHRR data and consequently increasing chances of observing sunglints. Thus, statistics for weak and moderate sunglint have then been possible to compile and is now used by CMA-prob. Unfortunately, this could only be done for AVHRRs with channel 3b active during day since matchups with sunglint conditions cannot be found for AVHRRs with channel 3a active. Consequently, we use only night-time features according to Table 3-3 in sunglint conditions when channel 3a is active. The same restriction is also applied to data from AVHRR/1 since the sunglint statistics derived from AVHRR/3 with channel 3b active is heavily relying on information from AVHRR channel 5 which is missing in AVHRR/1. Thus, sunglints are treated exclusively with information from feature Tirdiff and Texture_night for AVHRR/1.

In summary, we trained the probabilistic classifier for 28 different surface regions, 2 illumination conditions and 3-6 AVHRR feature tests (depending on AVHRR version and satellite orbit mode) for each illumination class yielding a total of between 168 to 336 unique, individual probabilistic estimates (which are all linked to the same number of individual probability distribution functions, PDFs).

Table 3-3 *Used transformed AVHRR image features for* ***night illumination*** *probabilistic cloud masking.*

Feature name	Definition	Main cloud detection ability
Tirdiff	AVHRR channel 4 brightness temperatures minus ERA-5 surface skin temperatures minus PPS thresholds	Identification of clouds which are significantly colder than the Earth surface
Tcidiff (AVHRR/2 + AVHRR/3)	AVHRR channel 4 brightness temperatures minus AVHRR channel 5 brightness temperatures minus PPS thresholds	Identification of thin cirrus clouds (however, also used also during day for AVHRR/3 with the 1.6 micron channel active -see text)
Twdiff	AVHRR channel 3b brightness temperatures minus AVHRR channel 4 brightness temperatures minus PPS thresholds	Identification of water clouds
2D_Tirdiff_Twdiff	Joint and linked use of features Tirdiff and Twdiff in 2-dimensional lookup tables.	Applied over cold land surfaces and over ice-covered ocean for a more accurate identification of water clouds in very cold situations.
Texture_night (AVHRR/2 + AVHRR/3)	Over land: Not used (surface variability is generally too large)! Over ocean: (Sum of local 3x3 pixel variances for AVHRR channel 4 brightness temperatures and AVHRR channel 3b and 5 brightness temperature differences) minus PPS thresholds	Identification of fractional or broken clouds over ocean
Texture_night (AVHRR/1)	Over land: Not used (surface variability is generally too large)! Over ocean: Local 3x3 pixel variances for AVHRR channel 4 brightness temperatures minus PPS thresholds.	Identification of fractional or broken clouds over ocean

Table 3-4 *The surface categories used for training the method and used for final cloud screening. Used abbreviations: NH=Northern Hemisphere, SH=Southern Hemisphere, SST=Sea Surface Temperatures, LST=Land Surface Temperatures. See text for further details.*

SURFACE NAME	Surface id	Short description
Marginal sea ice high latitudes	G1	Sea ice concentrations in the range 15-90 %
Sea ice high latitudes	G2	Sea ice concentrations above 90 %
Ocean polar north	G3	Ice-free ocean in NH with SSTs below 5 ℃
Ocean high latitude north	G4	Ice-free ocean in NH with SSTs in the range 5-12 ℃
Ocean mid latitude north	G5	Ocean in NH with SSTs in the range 12-22 ℃
Ocean tropical	G6	Ocean with SSTs above 22 ℃
Ocean mid latitude south	G7	Ocean in SH with SSTs in the range 12-22 ℃
Ocean high latitude south	G8	Ice-free ocean in SH with SSTs in the range 5-12 ℃
Ocean polar south	G9	Ice-free ocean in SH with SSTs below 5 ℃
Land dry homogeneous	G10	Deserts and adjacent dry regions (no rough terrain or snow)
Land homogeneous extra tropical	G11	Homogenous land with vegetation and LSTs below 12 ℃
Land homogeneous extra tropical seasonal snow	G12	Homogenous land with vegetation, seasonal snow cover, and LSTs below 12 ℃
Land homogeneous extra tropical permanent snow	G13	Homogeneous land with permanent snow cover
Land dry rough	G14	Deserts and adjacent dry regions in rough terrain (no snow)
Land rough extra tropical	G15	Land with vegetation over rough terrain and with LSTs below 12 ℃
Land rough extra tropical seasonal snow	G16	Land with vegetation over rough terrain with seasonal snow and LSTs below 12 ℃
Land rough extra tropical permanent snow	G17	Extratropical land over rough terrain with permanent snow cover
Land homogeneous tropical	G18	Homogenous land with vegetation and LSTs above 12 ℃
Land rough tropical	G19	Land over rough terrain with vegetation and LSTs above 12 ℃
Ocean polar north sunglint	G20	Arctic ice-free ocean with no sunglint and SSTs below 5 ℃
Ocean high latitude north sunglint	G21	Ice-free ocean in NH with no sunglint and SSTs in the interval 5-12 ℃
Ocean mid latitude north sunglint	G22	Ocean in NH with no sunglint and SSTs in the interval 12-22 ℃
Ocean tropical sunglint	G23	Tropical ocean with sunglint and SSTs above 22 ℃
Ocean mid latitude south sunglint	G24	Ocean in SH with sunglint and SSTs in the interval 12-22 ℃
Ocean high latitude south sunglint	G25	Ice-free ocean in SH with sunglint and SSTs in the interval 5-12 ℃
Ocean polar south sunglint	G26	Ice-free ocean in SH with no sunglint and SSTs below 5 ℃
Coast extra tropical	G27	Coastal areas with LSTs below 12 ℃ and viewing angles (satellites zenith angles) below 58 degrees.
Coast tropical	G28	Coastal areas with LSTs above 12 ℃ and viewing angles (satellites zenith angles) below 58 degrees.

3.6 Training the classifier using CALIPSO-CALIOP cloud data with dependencies on CALIOP-estimated cloud optical thicknesses

To develop the CMa-prob classifier, we have taken advantage of the previously reported studies based on collocated NOAA/METOP AVHRR and CALIPSO orbits as described by Karlsson and Johansson (2013) and Karlsson and Håkansson (2018). In particular, the extension of the matchup database resulting from the study by Karlsson and Håkansson (2018) has been very important here. Some example results from this extended collocated dataset over 10 years (covering the period 2006-2015) have already been shown in the previous section. However, some necessary restrictions have since been incorporated to the probabilistic classifier to improve results further.

A great advantage of the CALIPSO-CALIOP cloud products is their superior cloud detection sensitivity compared to cloud products based on passive instruments such as the AVHRR sensor. However, this is also a problem when using this information as the basis for a statistical training of a probabilistic cloud masking method. There is a risk for "over-training", i.e., that we force the method to try to detect clouds that are theoretically impossible to detect from AVHRR sensor data. As a result, the probabilistic cloud-screening method would then risk to systematically creating false clouds in truly cloud-free areas, and vice versa, since the cloud-free signal cannot be confidently separated from the cloudy signal. For example, we might detect clouds usually denoted sub-visible cirrus clouds, but false clouds could then also be added in adjacent areas without such clouds. Consequently, we need to find a way to restrict the used CALIOP-based cloud mask in the training process to include only those clouds which are potentially discernible in AVHRR images. In other words, we need to define the AVHRR cloud detection limit as accurately as possible.

This task can also be formulated as follows: We need to find the proper restricted CALIPSO cloud mask that is most accurately reproduced by the AVHRR-based cloud masking method. This means that we have to filter out the thinnest CALIOP-detected clouds from the CALIOP cloud mask up to a certain limit or threshold in cloud optical thickness where we then can find the best resemblance with the AVHRR-derived cloud mask. This limit in optical cloud thickness can be denoted "Cloud Detection Sensitivity" for the method and it has previously been introduced as a useful concept by Karlsson and Håkansson (2018). We will use this concept here in the training and definition of the CMaprob method.

A practical solution to the desire of finding the optimal training of the method (according to the principles introduced above) can be found if also adding the following criterion:

> A binary cloud mask created by thresholding the cloud probability at the 50 % probability level should give maximum detection skill compared to a cloud mask derived by any other probability threshold.
>
> [1]

This criterion is not only necessary for solving the training problem but it also means that potential users of the product are given a clear recommendation on how to use the cloud probability results. Also, the above criterion should ideally (i.e., for practical reasons) be valid at every geographic location. A user could adjust the threshold making classification results to

be more clear-conservative (lowering the threshold) or more cloud-conservative (raising the threshold). However, the default threshold value at 50 % should be a reasonable starting point since the user would then know that by using this threshold should theoretically maximise the detection skill. For any other threshold, the rate of misclassifications will increase.

The most natural detection skill score to use in this context is the *Hitrate* which is simply the percentage of correct cloudy and clear predictions with respect to the total number of predictions.

To illustrate how we can use this skill score in the training process, we can study Figure 3-5 taken from Karlsson and Johansson (2013). They introduced a plot on how the *Hitrate* and other skill scores (in this example, the *Kuipers Skill Score)* could vary as a function of the filtered cloud optical thickness. Basically, this shows how well the results of a specific method (in Figure 3-5 cloud masks produced for the CLARA-A1 climate data record) agree with restricted CALIOP cloud masks compared to the original CALIOP cloud mask. Notice that filtering does not mean that data is removed from the validation dataset. Instead, clouds with optical thicknesses below the filtered cloud optical thickness value are now interpreted as non-existent (i.e., changed into cloud-free).

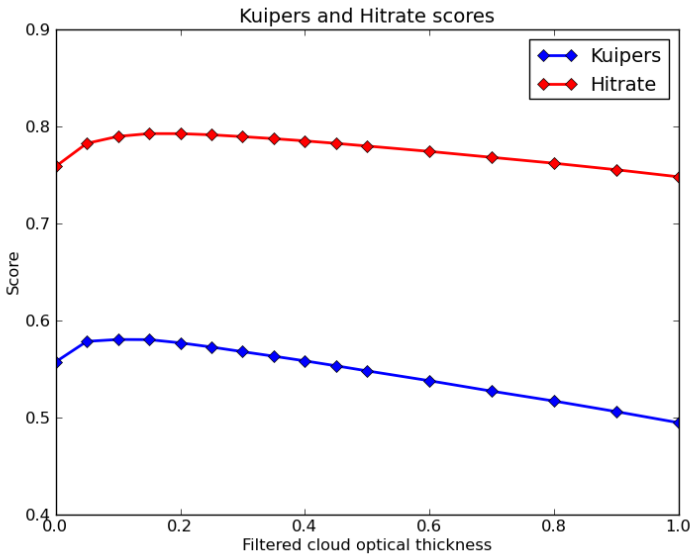

Figure 3-5 *Hitrate and Kuipers skill scores as a function of filtered CALIOP cloud masks (cloud optical thickness limits) for CLARA-A1/PPS 2014 cloud masks. Results derived from 99 collocated NOAA-18 and CALIPSO-CALIOP cloud masks in the period 2006-2009. (From Karlsson and Johansson, 2013).*

Figure 3-5 shows how skill scores first improves if filtering the thinnest (mostly non-detectable) clouds from the validation dataset but then decreases again after reaching a certain value of the filtered optical thickness. The decrease is explained by the fact that a larger fraction of correctly detected CALIOP-observed clouds is here converted to cloud-free (lowering the Hitrate) relative to how many missed clouds are converted into cloud-free (raising the Hitrate). At the maximum value for the Hitrate we get the best fit with a restricted filtered CALIOP cloud mask. In this example, this maximum Hitrate occurs for a filtered cloud optical thickness value of approximately 0.2 (or slightly lower for the Kuipers score). This optimal filtered value for

Hitrate is identical to the Cloud Detecting Sensitivity parameter introduced by Karlsson and Håkansson (2018). They calculated it for all locations on the globe when validating results of the CLARA-A2 climate data record (Karlsson et al., 2016).

An important feature of the Hitrate curve in Figure 3-5 is that it possesses important information about the probability of detecting a cloud with a certain optical thickness. If calculating the Hitrate gradient from the values at finite intervals of filtered cloud optical thicknesses, we can deduce where the probability of detection reaches 50 %. More clearly, it means that, when the gradient is zero at the point of maximum Hitrate, we have reached the situation when the probability of detection of clouds is precisely 50 %. If the Hitrate stays constant after filtering (i.e., converting CALIOP clouds to cloud-free) a certain amount of clouds, it means that 50 % of those clouds were initially correctly classified. For intervals with lower cloud optical thicknesses, the probability of detection is always lower, and on the other side of the Hitrate maximum, it is always higher. If we imagine that our finite intervals (shown in Figure 3-5 as the distance between individual values on the curve) decrease towards infinitesimal widths we can then also say that we have a probability of detection of 50 % for cloud layers with an optical thickness of exactly this filtered cloud optical thickness value. We will utilise this property in the training process for CMa-prob since it has a direct link to criterion [1] above.

To find a solution which optimizes our cloud probability results, i.e., which gives us as small Cloud Detection Sensitivity values as possible everywhere on the globe, we need to add another level of complexity. The trick is first to do repeated training of the classifier using the full range of restricted cloud masks, i.e. full range of filtered cloud optical thicknesses. In the next step, we then select the training statistics valid for that particular restricted cloud mask with a certain filtered optical thickness which coincides with the position of the Cloud Detection Sensitivity (i.e., position for maximum Hitrate). Then we can be sure that the 50 % probability of detection is valid for this particular Cloud Detection Sensitivity which means that we are fulfilling criterion [1] above.

We can consider two cases:

1. Hitrate peaks at smaller optical thicknesses than the filtered value used for the restricted cloud mask during training.
2. Hitrate peaks at larger optical thicknesses than the filtered value used during training.

Both cases mean that we are not succeeding in reproducing the restricted cloud mask that was used during training. The first case means that the used training statistics appears to be valid also to some extent for clouds with smaller optical thicknesses. Therefore, we should try to use a lower filtered optical thickness value during training. The second case means that a significant portion of the clouds is misclassified (missed), and we should therefore use a larger value for the optical thickness threshold of the restricted cloud mask.

In practice, this means that we have to carry out this evaluation separately for every defined surface or region according to Table 3-4. In this way, we will get different optimal Cloud Detection Sensitivities for each surface or region which reflects the different separabilities of cloudy and cloud-free conditions that exist and depend on the underlying surface characteristics. Consequently, even if the 50 % cloud probability is used everywhere (as a basis for a binary cloud mask), the probability applies to clouds with different layer optical thicknesses which will vary with surface and region. The ambition is that by this method we

will be able to identify clouds which are radiatively significant, i.e., being identified for having a sufficient contrast against the surface for each region. If combining all these results globally we would then get improved results compared to if we just unconditionally train our method against the original and unrestricted CALIOP cloud mask. To test if this is finally achieved, we must prove that the overall validation results when validating against the unrestricted CALIOP cloud mask are better than results based on training exclusively using the original cloud mask. The latter case should give a higher rate of misclassifications if the concept is working.

3.7 Resulting sets of optimal training statistics for different Earth surfaces

The collected training dataset spans the period 2006-2015 and provides good global coverage over all seasons during that period. However, one year (2010) has been excluded from the training dataset to constitute an independent validation dataset. The training dataset is collected from almost 3000 NOAA-18 and NOAA-19 AVHRR Global Area Coverage (GAC) orbits and CALIOP pixels/samples at approximately 5 km horizontal resolution. The CALIPSO-CALIOP Cloud Layer (CLAY) product version 4.1 has been used. A more detailed description of this product is given by Karlsson and Håkansson (2018). The constrained training (i.e., image features now being related to PPS threshold information) is based on results from the PPS software version 2021 [RD 1]. This PPS version is highly advanced compared to the original method described by Dybbroe et al. (2005). The main new features of the method concern adaptations to global processing (e.g., over desert and Polar Regions) and a systematic use of prescribed MODIS-derived surface emissivity information.

Figures Figure 3-6 – Figure 3-9 below illustrate the selection of optimal training datasets over different surfaces and regions. Hitrate scores are plotted here using different curves representing training constellations using different restricted CALIOP cloud masks (explained in the legend). Notice that the training constellations that give the giving highest cloud detection sensitivities (i.e., the capability to detect the thinnest cloud optical thicknesses) and which fulfil criterion [1] above, are highlighted by thicker lines (in both figure and legend). In Figure 3-6, we study conditions for the surface with the best overall results globally, namely ice-free ocean surfaces over the extra-tropics (categories G4 and G8 in Table 3-4). We deduce that over these surfaces, we can use training data from a slightly restricted CALIOP cloud mask (τ-filtering at 0.05) during the day and from the unrestricted original CALIOP cloud mask during the night. Thus, the cloud detection sensitivity will then be 0.05 or even smaller over these surfaces.

Conditions are more problematic over snow-free extratropical land surfaces (category G11 in Table 3-4) as shown in Figure 3-7. Best results are here found during daytime for a filtered cloud optical thickness of 0.6. Solutions for larger optical thicknesses also exist but the highest Hitrates are found for the lowest of the suggested solutions and; therefore, the solution with the smallest optical depth is chosen for this category. Nighttime results are less clear since we cannot find an obvious case where the Hit rate peak coincides with a particular trained filtered optical thickness. However, the highest scores are still given for almost the same chosen filtered optical thickness as for daytime conditions and we decided to use this value also at night. It should be noted that the perfect solution at night is maybe not covered by the different training constellations we tested. A finer resolution of restricted CALIOP masks could perhaps have suggested a cloud optical thickness value of 0.65.

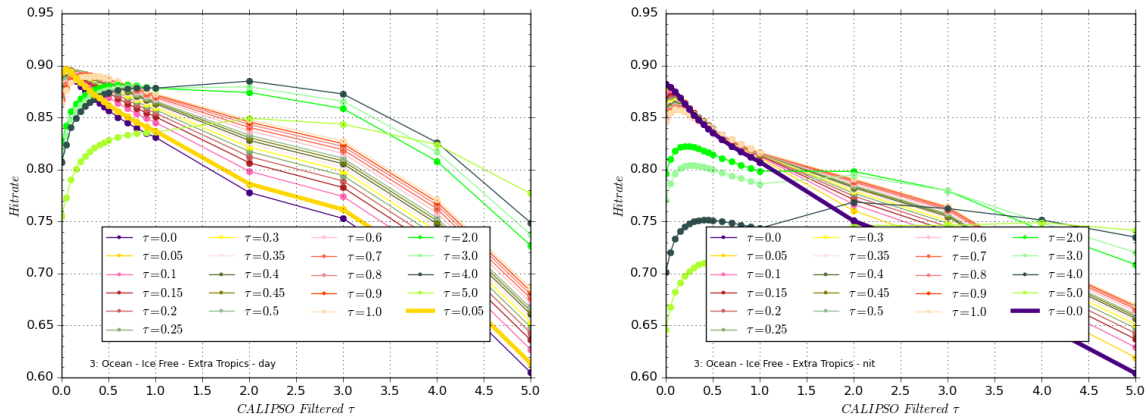

Figure 3-6 *Hitrate as a function of filtered cloud optical thicknesses of the CALIOP cloud mask for different training scenarios (coloured curves) over ice free extratropical ocean (categories G4 and G8 in Table 3-4) during day (left) and night (right). The coloured curves describe results based on different restricted CALIOP cloud masks used when training. Thick lines denote possible solutions fulfilling the criterion that Hitrate should be maximized for the same filtered optical thickness as was used during training.*

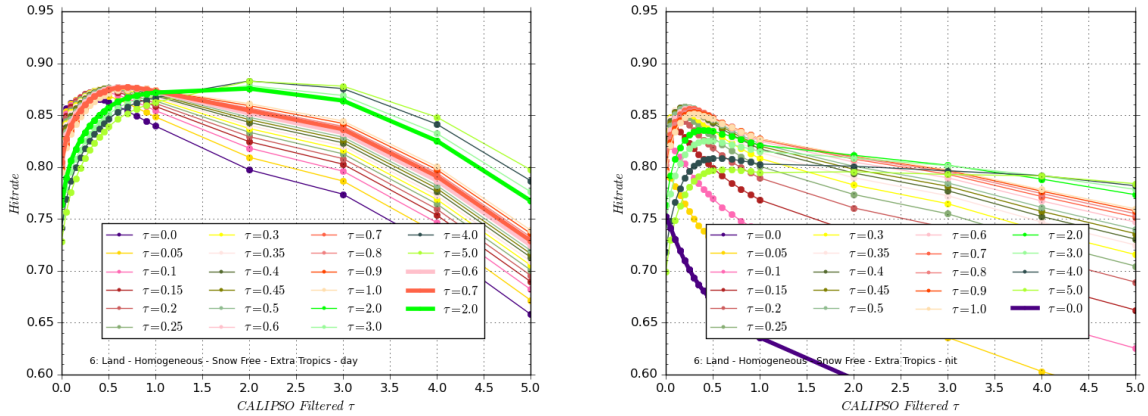

Figure 3-7 *Same as Figure 3-6 but for snow-free extratropical land surfaces (category G11 in Table 3-4) during day (left) and night (right).*

Looking at results over well-known problematic areas, such as areas with permanent snow-cover (category G13 in Table 3-4 including the dominant parts of Greenland and Antarctica) in Figure 3-8, the situation becomes less clear. Here, it is difficult to find any clear guidance on which solution to choose when training. No clear peaks in Hitrate seem to coincide with the selected filtering level used when training, indicating that, despite having seemingly homogeneous surface conditions, the true variability in results is very large over those surfaces. In addition, the variation seems to be much less linked to the cloud optical thickness of the clouds we intend to detect than for other surfaces. This can also be explained by existing contradicting results from different image features. If the necessary contrast needed for good cloud discrimination is found in one feature the situation can be completely the opposite for another image feature. This is a drawback of the Naïve Bayesian method, i.e. due to the

multiplication of individual probabilities, a near-zero value for one image feature will be able to compensate or neutralise results from other features.

The problem with the results in Figure 3-8 is that no distinct peak can be found that coincides with the trained filtered cloud optical thickness. For daytime conditions, it seems a cloud detection sensitivity of about 1.0 would be suitable. However, alarmingly, the Hitrates for this value do not clearly decrease for larger filtered optical thicknesses. Therefore, even for higher optical thicknesses, almost 50% of all clouds will be missed. A better solution would be to choose a lower trained optical thickness value which is having a distinct peak but at a somewhat lower Hitrate value. A reasonable compromise could be to use the value 0.3. The peak in Hitrrate is still reasonably high, and a clear capability to detect clouds with larger optical thicknesses is retained.

For night conditions in Figure 3-8 (right), the results are even harder to interpret. At first sight, a solution using the maximum filtered value of 5.0 is suggested. But in practice, it means that we will never reach 50 % probability of detection for any value of the layer optical thickness in the studied interval since the curve for the filtered value of 5.0 is monotonically increasing over the full range of filtered optical thicknesses. There are, however, distinct peaks in Hitrate for other filtered optical thickness (e.g. in the range 0.2- 0.8) but then having much lower Hitrates than what is indicated for the value 5.0 at the maximum filtering level. A compromise solution here could be to use the value 0.5 to acknowledge that situations should be more difficult than during daytime but not as extreme as suggested by the solution of using the value 5.0 for the restricted cloud mask. We conclude that, over the most problematic categories, we have to be careful using the general and idealized concept presented earlier since conditions are much more complex than over other Earth surfaces. To find the optimal reproduced CALIOP cloud mask seems very difficult here which forces us to use some compromise solutions.

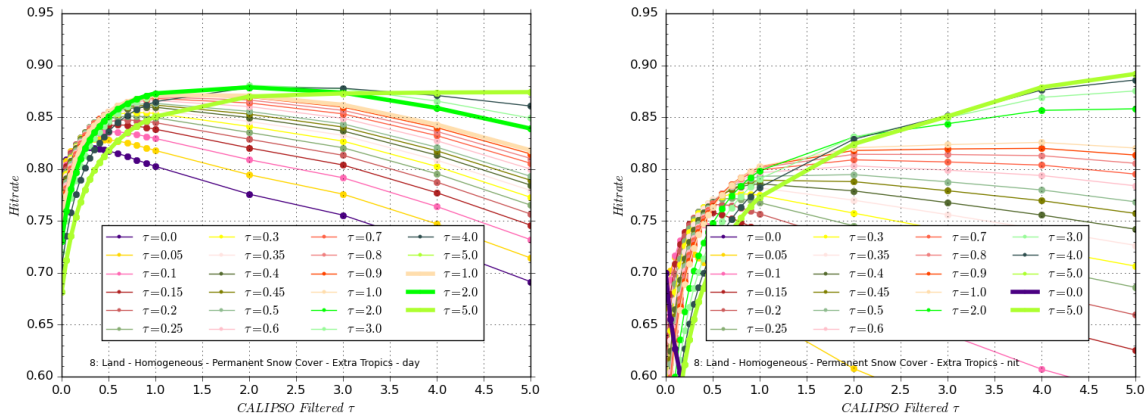

Figure 3-8 *Same as Figure 3-6 but for homogeneous land surfaces with permanent snow-cover (category G13 in Table 3-4) during day (left) and night (right).*

Finally, we will also take a closer look at dry surfaces with sparse vegetation (category G10 in Table 3-4) in Figure 3-9. We generally notice very high Hitrates during both day and night and that the curves are quite similar to those found for extra-tropical land surfaces in Figure 3-7. However, the fact that no clear solution exists for the optimal cloud detection sensitivity during night or day, results appear uncertain, which probably indicates that conditions over desert surfaces are again not as homogeneous and representative as initially assumed. In addition, the behaviour of the curves at night is very different compared to how it looks during day and over

other surfaces. One reason for this could be the varying surface emissivities (influenced by both soil moisture conditions and surface material characteristics) in the infrared channels causing "cloud-like" temperature differences also in clear situations. Another reason could be that our reference surface skin temperatures from ERA-5 do not capture minimum temperatures at night very well (which is also problematic for the polar winter surfaces described earlier in Figure 3-8 right). However, because of the high Hitrates found here, we will use similar optimal cloud optical thickness values as for the extra-tropical land surfaces.

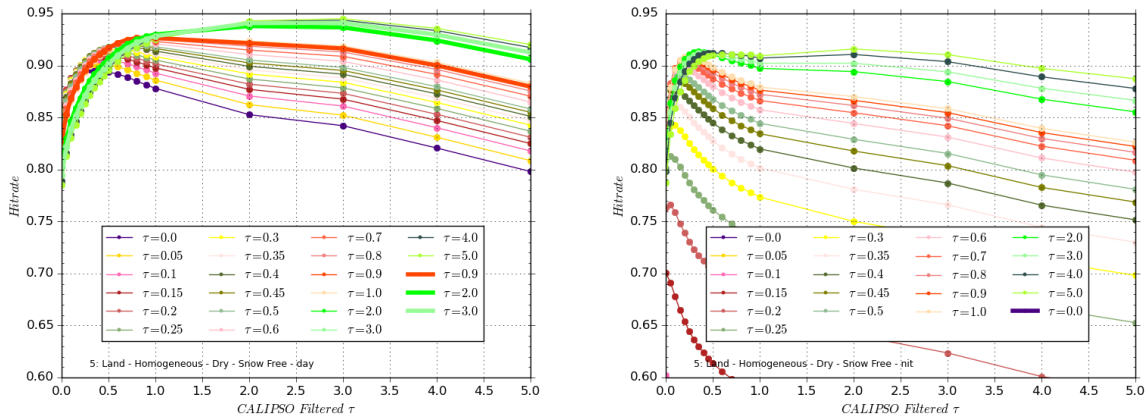

Figure 3-9 *Same as Figure 3-6 but for dry land surfaces (category G10 in Table 3-4, dominated by deserts) with sparse vegetation during day (left) and night (right).*

We summarize the estimated optimal cloud detection sensitivities for all investigated surfaces in Table 3-5. The final statistics for the CMa-prob classifier is compiled by choosing trained statistics for each surface category trained by the CALIOP cloud masks defined by the surface-specific optimal cloud detection sensitivities in Table 3-5.

For the twilight category in Table 3-5, the CMa-prob classifier selects either night-time or daytime statistics depending on if a detectable reflection signal can be found in AVHRR channel 1 at 0.6 microns. Thus, no specific twilight statistics is compiled but the used cloud detection sensitivity may still be different from the pure night and day case.

Table 3-5 *Estimated optimal Cloud Detection Sensitivies (i.e., lowest cloud layer optical thicknesses with probability of detection exceeding 50 %) for different Earth surfaces (defined in Table 3-4) and for different illumination categories. For these filtered optical thicknesses, the best resemblance is achieved with the CALIOP cloud mask over different surfaces.*

SURFACE DESCRIPTION	Surface id	DAY	NIGHT	TWILIGHT
Marginal sea ice high latitudes	G1	0.05	0.05	0.05
Sea ice high latitudes	G2	0.25	0.60	0.25/0.60
Ocean polar north	G3	0.05	0.05	0.05
Ocean high latitude north	G4	0.05	0.05	0.05
Ocean mid latitude north	G5	0.05	0.05	0.05
Ocean tropical	G6	0.10	0.10	0.10
Ocean mid latitude south	G7	0.05	0.05	0.05
Ocean high latitude south	G8	0.05	0.05	0.05
Ocean polar south	G9	0.05	0.05	0.05
Land dry homogeneous	G10	0.25	0.25	0.25
Land homogeneous extra tropical	G11	0.20	0.25	0.20/0.25
Land homogenous extra tropical seasonal snow	G12	0.20	0.35	0.20/0.35
Land homogeneous extra tropical permanent snow	G13	0.30	0.50	0.30/0.50
Land dry rough	G14	0.30	0.40	0.30/0.40
Land rough extra tropical	G15	0.30	0.20	0.30/0.20
Land rough extra tropical seasonal snow	G16	0.30	0.50	0.30/0.50
Land rough extra tropical permanent snow	G17	0.15	0.60	0.15/0.60
Land homogeneous tropical	G18	0.15	0.15	0.15
Land rough tropical	G19	0.15	0.15	0.15
Ocean polar north sunglint	G20	0.05	-	0.05
Ocean high latitude north sunglint	G21	0.05	-	0.05
Ocean mid latitude north sunglint	G22	0.05	-	0.05
Ocean tropical sunglint	G23	0.40	-	0.40
Ocean mid latitude south sunglint	G24	0.05	-	0.05
Ocean highlatitude south sunglint	G25	0.05	-	0.05
Ocean polar south sunglint	G26	0.10	-	0.10
Coast extra tropical	G27	0.20	0.50	0.20/0.50
Coast tropical	G28	0.50	0.15	0.50/0.15

4 Final implementation of CMa-prob and some demonstrated results

4.1 Demonstration of impact of using cloud detection sensitivity statistics instead of statistics based on original CALIOP cloud mask

Some results from selecting training statistics differently over various Earth surfaces are illustrated in this sub-section for some image features.

Figure 4-1 shows the achieved training statistics for the visible reflectance feature (Rvis in Table 3-2) over dry and homogeneous land surfaces (category G10 in Table 3-4). The left part of the figure shows the cloud probability distribution as a function of reflectance after training with the cloud detection sensitivity value 0.25 (according to Table 3-5). The right part shows statistics compiled when training using the original unfiltered CALIOP cloud mask. The latter shows that cloud occurrences are not negligible for low-to-moderate reflectivities even if the majority of clouds occur for higher reflectivities. However, after training against a restricted cloud mask, filtering clouds with optical thicknesses below 0.25, the cloud probabilities at low-to-moderate reflectivities are reduced (Figure 4-1 left). It means that some of the very thin clouds which previously were mixed up with the cloud-free reflectance are now treated as cloud-free cases. This improves the overall separability of cloudy and cloud-free cases. The separability would have improved even further if the suggested cloud detection sensitivity of 0.9 of Figure 3-9 had been used. However, this would also have led to a considerable loss of fully detectable clouds in the optical thickness range of 0.25-0.9. Thus, the final choice of the cloud detection sensitivity parameter (i.e., the resulting CALIOP cloud mask to train the method against) has to be a compromise between the wish to get as high Hitrates as possible with a restricted CALIOP cloud mask and the wish to detect as many clouds as possible without introducing too many false detections. The peak of cloud occurrences seen in Figure 4-1 at very low reflectances is most likely explained by shadows from high-level clouds cast on other lower-level clouds.

Figure 4-2 shows the same type of results as in Figure 4-1 but now for the reflectance in the 3.7 micron channel (short-wave infrared with image feature named Rvis37). Again, we can see how the removal of the thin clouds (i.e., meaning that we now interpret them as cloud-free) below optical thickness 0.25 increases the separability between cloudy and clear cases. Clouds in this spectral region are either weakly reflecting (ice clouds) or strongly reflecting (water clouds). A few cases of very high reflectivities can be noted, but these link to noisy pixels not associated with proper quality flag information in original data. The intermediate region in the figure (i.e., valid for reflectivities in the interval 15-35 %) is dominated by moderately reflecting desert surfaces. If not filtering out the thinnest clouds, they are easily mixed up with the Earth's surface as seen by the non-zero cloud frequencies here. But after filtering, the risk of misclassifying (remaining) clouds is reduced. We repeat that by filtering out the thin clouds we reduce the capability to detect any of these clouds, but on the other hand, we will now minimise the risk of creating false clouds with more or less the same spectral signature. The use of the cloud detection sensitivity parameter for guiding us with the filtering procedure will guarantee that we gain more than we lose by this filtering procedure.

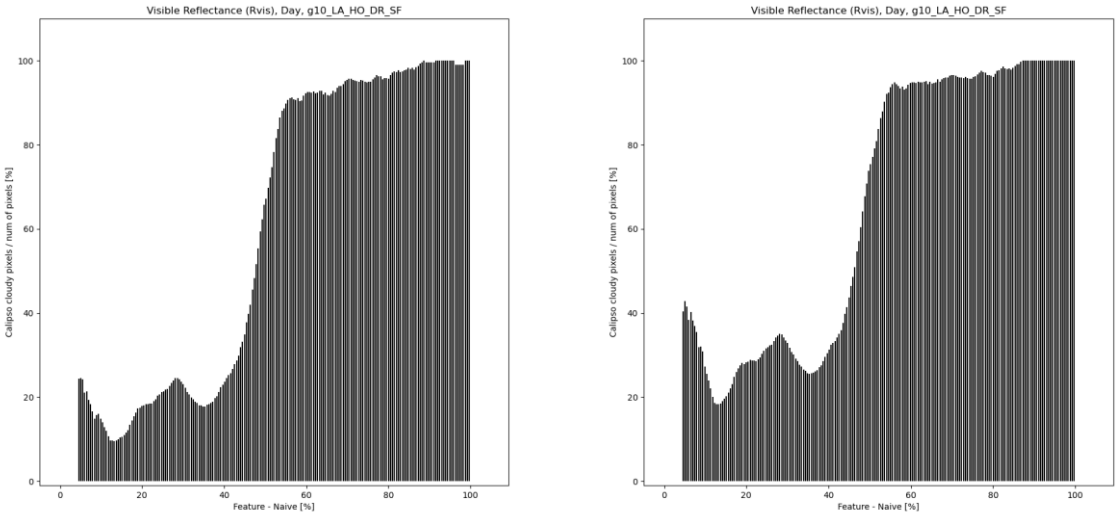

Figure 4-1 *Distribution of cloud occurrences (or cloud frequencies) as a function of the visible AVHRR reflectance at 0.6 micron (Rvis) over **dry homogeneous surfaces** (category G10 in Table 3-4 with predominantly desert surfaces). Left: Statistics based on training with a CALIOP cloud mask filtered at optical thickness 0.25. Right: Statistics based on training with the original unfiltered CALIOP cloud mask.*

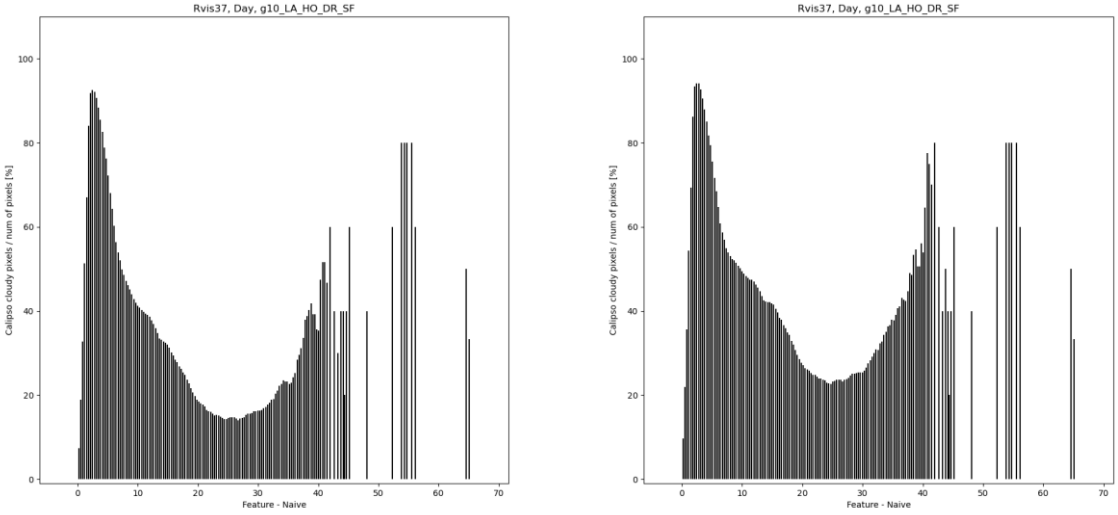

Figure 4-2 *Distribution of cloud occurrences (or cloud frequencies) as a function of the AVHRR reflectance at 3.7 micron (in %, image feature Rvis37 in Table 3-2) over **dry homogeneous surfaces** (surface G10 in Table 3-4 with predominantly desert surfaces). **Left**: Statistics based on training with a CALIOP cloud mask filtered at optical thickness 0.25. **Right**: Statistics based on training with the original unfiltered CALIOP cloud mask.*

A final example of the effect of filtering is shown in Figure 4-3 for the absolutely most problematic cloud detection condition encountered for AVHRR data: night-time cloud detection over cold and snow-covered surfaces (e.g., polar night over Greenland and Antarctica in surface group G13 in Table 3-4). The figure shows the cloud probability as a function of the difference between the AVHRR brightness temperature at 11 microns and the surface temperature from ERA-5 (image feature Tirdiff). The right part shows results when training against the unfiltered CALIOP cloud mask while the left part shows results when training against a CALIOP cloud mask filtered at optical thickness 0.5.

It is clear that for unfiltered training the probability of cloudy conditions is high for almost all the time, except for cases when the surface temperature is much colder than the measured 11 micron brightness temperature (i.e., for large negative differences). However, probabilities are generally uncertain (i.e., close to 50 %) and only for restricted parts of the distribution do we find high and low probabilities which could contribute favourably to the cloud detection process. After filtering with a cloud optical thickness of 0.5 (Figure 4-3, left), a large fraction of all clouds disappears (i.e., are now interpreted as cloud-free) and only clouds which are much colder than the ERA-5 surface temperature remain with cloud probabilities clearly above 50 %. It means that only optically thick and cold clouds are possible to detect with confidence over this surface type for this image feature.

The unfiltered results (Figure 4-3, right) indicate some skill in also identifying clouds which are warmer than the surface (e.g., “black stratus”) for temperature differences near -25 K. The probabilities for the detection of these clouds after the filtering has generally improved. Still, they are nevertheless close to 50 % (i.e., very uncertain). Successful identification of these clouds now depends on if also cloud probabilities are high enough for the other two infrared image features ((Twdiff and Tcidiff in Table 3-3). For the black stratus clouds, Twdiff is important since it normally shows a negative temperature difference between AVHRR channels at 3.7 microns and 11 microns as opposed to clear areas and ice clouds (the latter showing a positive temperature difference). But for very cold situations, this typical cloud feature becomes less reliable, which is caused by the transition from pure water clouds at higher temperatures to mixed phase clouds at cold temperatures. In addition, increasing radiometric noise in the 3.7 micron channel (leading to random positive and negative Twdiff values) will further decrease the usefulness of the Twdiff feature. In conclusion, cloud detection capabilities in AVHRR data over cold surfaces during the polar night remains as the most challenging of tasks for any cloud detection scheme.

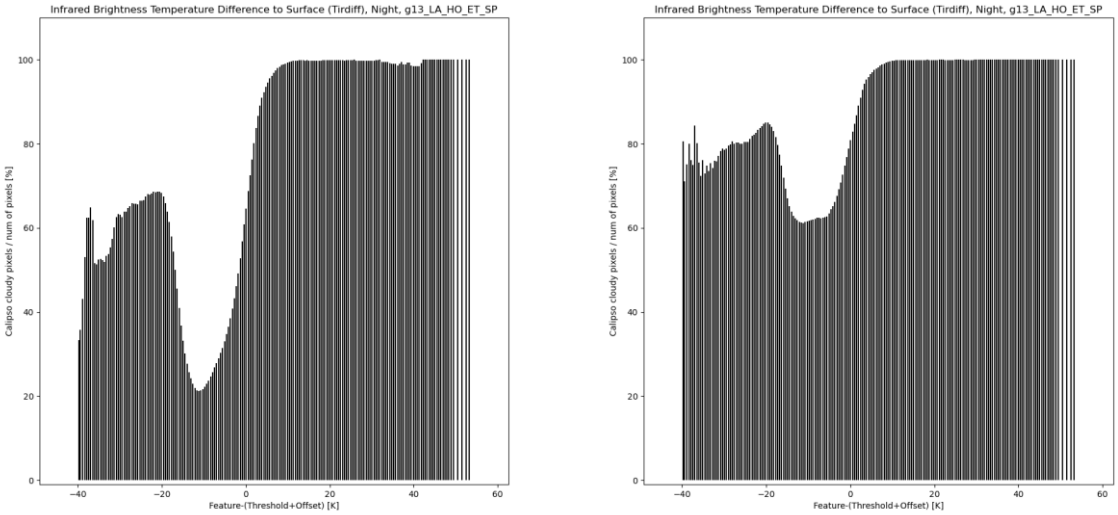

Figure 4-3 *Night-time distribution of cloud occurrences (or cloud frequencies) as a function of the difference between the AVHRR brightness temperature at 11 micron and the surface temperature from ERA-5 (image feature Tirdiff) over* ***permanently snow-covered surfaces*** *(group G13 in Table 3-4) at night.* ***Left****: Statistics based on training with a CALIOP cloud mask filtered at optical thickness 0.5.* ***Right****: Statistics based on training with the original unfiltered CALIOP cloud mask.*

An attempt trying to improve the detection of low clouds at night is to treat features Twdiff and Tirdiff jointly and not as independent single features. The latter often leads to that information in the two features cancel out each other (i.e., they are negatively correlated). More clearly, the Twdiff feature could indicate high cloud probabilities in situations when we have strong surface temperature inversions but such conditions are often associated with low cloud probabilities in image feature Tirdiff (i.e., clouds are predominantly colder than the surface). To avoid cancellation in this way, we have used feature 2D_Tirdiff_Twidiff which as introduced earlier in Table 3-3. Figure 4-4 illustrates cloud probabilities in this feature over snow-free land surfaces at night over relatively cold surfaces (category G11). Here we can clearly see that it is possible to get high cloud probabilities (higher than 50 %) also when the Tirdiff feature is strongly negative. So, by using this 2-D feature, we have reduced the risk of problems due to the negative correlation between these two features.

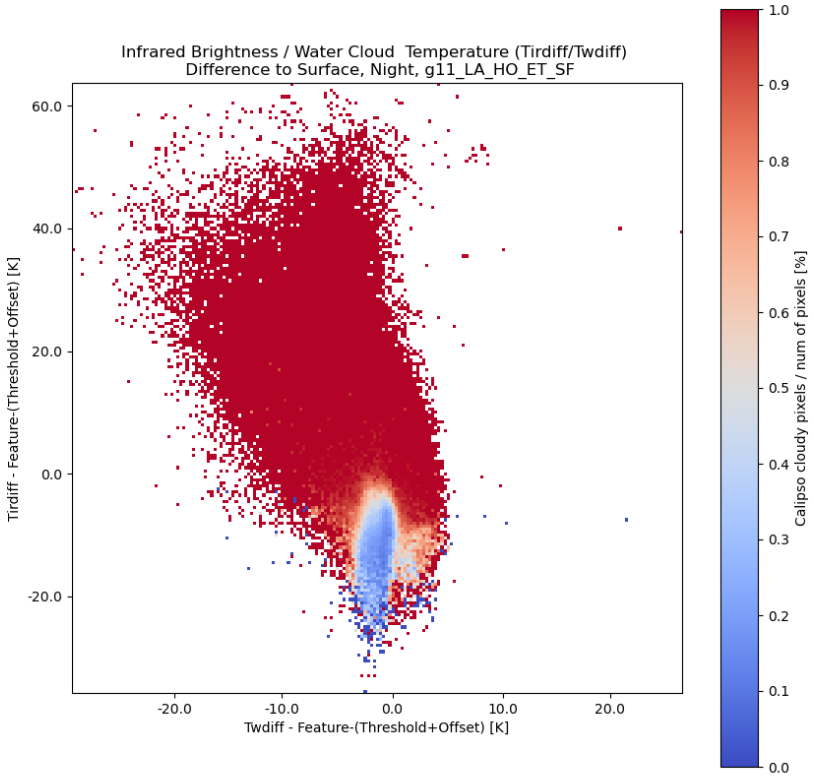

Figure 4-4 *Illustration of cloud probabilities in a 2-D histogram with respect to features Twdiff (x-axis) and Tirdiff (y-axis) over snow-free land surfaces at high latitudes (category G11 in Table 3-4) during night.*

4.2 Treatment of data from satellites with the 1.6 micron channel replacing the 3.7 micron channel

Table 3-2 lists also the image feature Rswir_3a which is based on the reflectivity in the AVHRR channel at 1.6 microns divided by corresponding reflectivities in the AVHRR channel at 0.6 microns. This feature is conceptionally similar to feature Rswir_quota in Table 3-2 (also used by CMa-prob but not illustrated here). Channel 3a is only available operationally (with only a few exceptions) from satellites operating in morning orbits (e.g., NOAA-17, METOP-A, METOP-B and METOP-C). Data from satellites in morning orbit can also be collocated with CALIPSO-CALIOP data but only for positions close to approximately latitude 70 degrees. This means that Rswir_3a statistics cannot be collected with global coverage. This is particularly problematic for the dry land surface categories (G10 and G14 in Table 3-4) when considering that the surface reflectivity at 1.6 microns is high over desert surfaces risking to be mixed up with corresponding cloud reflectivities.

We have overcome this problem by training CMa-prob against corresponding 1.6 micron radiances measured by the MODIS instrument carried by the Aqua satellite. This can be done since the spectral responses at this particular channel (and also for the channel at 0.6 microns

used to calculate the reflectance quota) are very similar for the AVHRR and MODIS sensors. The Aqua satellite is part of the A-train, and it thus offers almost simultaneous and continuous observations with CALIPSO-CALIOP. We used one year of MODIS data (2010) for training the CMa-prob method.

Figure 4-5 shows resulting daytime Rswir_3a probability distributions for desert surfaces (category G10) and for surfaces with permanent snow-cover (category G13). Results here are based on training with the unfiltered CALIOP cloud mask. We notice a very distinctive cloud signature in both cases (with the majority of cloud-free cases predominantly occurring at feature values below 0.2) showing that this image feature can be used with great confidence in the cloud screening process. This feature appears actually more reliable for detection of clouds over desert surfaces than the previously described Rvis37 feature (see Figure 4-2 right). Distributions for the latter has a much more serious mix between cloudy and clear radiances (e.g., in the interval 15-35 % reflectivity). The high cloud detection capability of this feature over snow-covered surfaces is also clearly seen in Figure 4-5 (right). This was the major reason for introducing this channel historically.

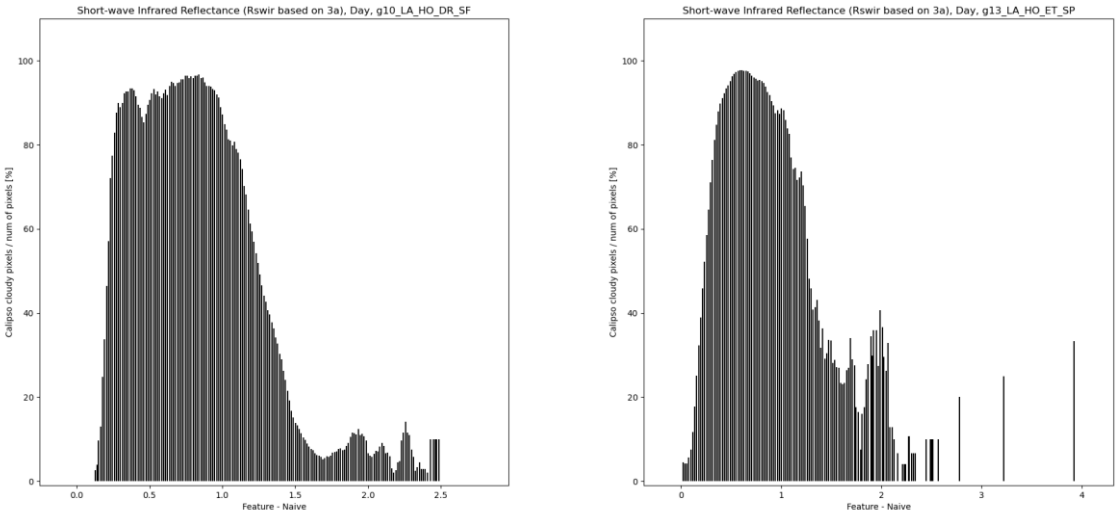

Figure 4-5 *Daytime distribution of cloud occurrences (or cloud frequencies) as a function of the reflectance quota (Rswir_3a in Table 3-2) between AVHRR-heritage channels at 1.6 micron and 0.6 microns based on Aqua Modis data.* ***Left****: Distribution over* ***dry surfaces*** *(surface category G10 in Table 3-4).* ***Right****: Distribution over* ***surfaces with permanent snow-cover*** *(surface category G13 in Table 3-4). All statistics are calculated from collocations with the original unfiltered CALIOP cloud mask.*

4.3 Product demonstration

The CMa-prob product is available in PPS version 2021 for both high resolution AVHRR data (HRPT, 1 km) and reduced resolution Global Area Coverage (GAC, 5 km) data. It is also possible to produce CMa-prob results for MODIS, VIIRS, and MERSI-2 data based on statistics derived from AVHRR and MODIS data. Representative training data for VIIRS from collocations with CALIPSO-CALIOP remains to be compiled in the future. It is also technically working[1], to process data from SLSTR and METimage instruments (for the latter only tested on simulated data so far); for PPS v2021 the processing of SLSTR and METimage is considered demonstrational.

We will here show two examples of the resulting products based on AVHRR GAC data. Figure 4-6 They are selected for their quite difficult observation conditions: High solar zenith angles over an area largely covered with snow and ice at the surface.

Figure 4-6 shows a NOAA-18 overpass over the Greenland area in ascending node. We have twilight conditions in a large part of the scene, especially in the lower portion (closest to the North Pole). The colour composite illustrates the difficulty in separating clouds from bright surfaces over the entire area. By using just visible and thermal infrared data, this separation is difficult. But the access to measurements in shortwave infrared channel 3b of AVHRR (not shown in the colour composite image) provides essential information for identifying low level water clouds. These clouds are identified with high confidence (high probability) in the CMa-prob image (Figure 4-6, right).

The situation is not as favourable for the identification of ice clouds since these clouds are not as reflective as in the shortwave infrared channel. This means that such clouds over the Greenland ice sheet (in the right part of the image) are hard to detect and they consequently get lower cloud probabilities, and even below 50 % as is indicated by the bluish colour. Similar thin clouds over the open water surfaces further south (in the upper part of the image) are, however, much easier to detect and readily get cloud probabilities near 100 % (white colour). Bare in mind that the cloud probabilities seen over Greenland is with respect to clouds with a minimum cloud optical thickness of 0.5 (i.e., thinner clouds will not be detected) while the corresponding reference over open water is clouds with minimum cloud optical thickness of 0.05 (see Table 3-5).

Figure 4-7 shows a similar situation but now with data from the NOAA-17 satellite in a morning orbit and in descending node. In this case, we no longer have access to the short-wave infrared channel at 3.7 microns but instead data from the near infrared channel at 1.6 microns (channel 3a, see Table 3-1). However, despite the channel shift, results look very similar to the previous case. Water clouds are easily detected, even over snow- and ice-cover. Notice in particular, the cloud detection performance over snow covered land in north-eastern Canada in the lower part of the image. The same limitations regarding the detection of thin ice clouds can also be noticed

[1] With technically working we mean: 1) Visual inspection shows product similar to what we get for supported sensors. 2) Visual inspection of RGBs compared to the product confirms that results are reasonable. 3) We have no reason to expect degraded results for this specific sensor. 4) *But* the product is not tuned or quantitatively validated with this specific sensor.

over the Greenland ice sheet. The human eye can see shadows on the ground and at the sides of these clouds, which reveal their existence, but to use this information in this context is very difficult. Nevertheless, we conclude that the CMa-prob concept seems to work with fully comparable results even with different input data from the AVHRR sensor.

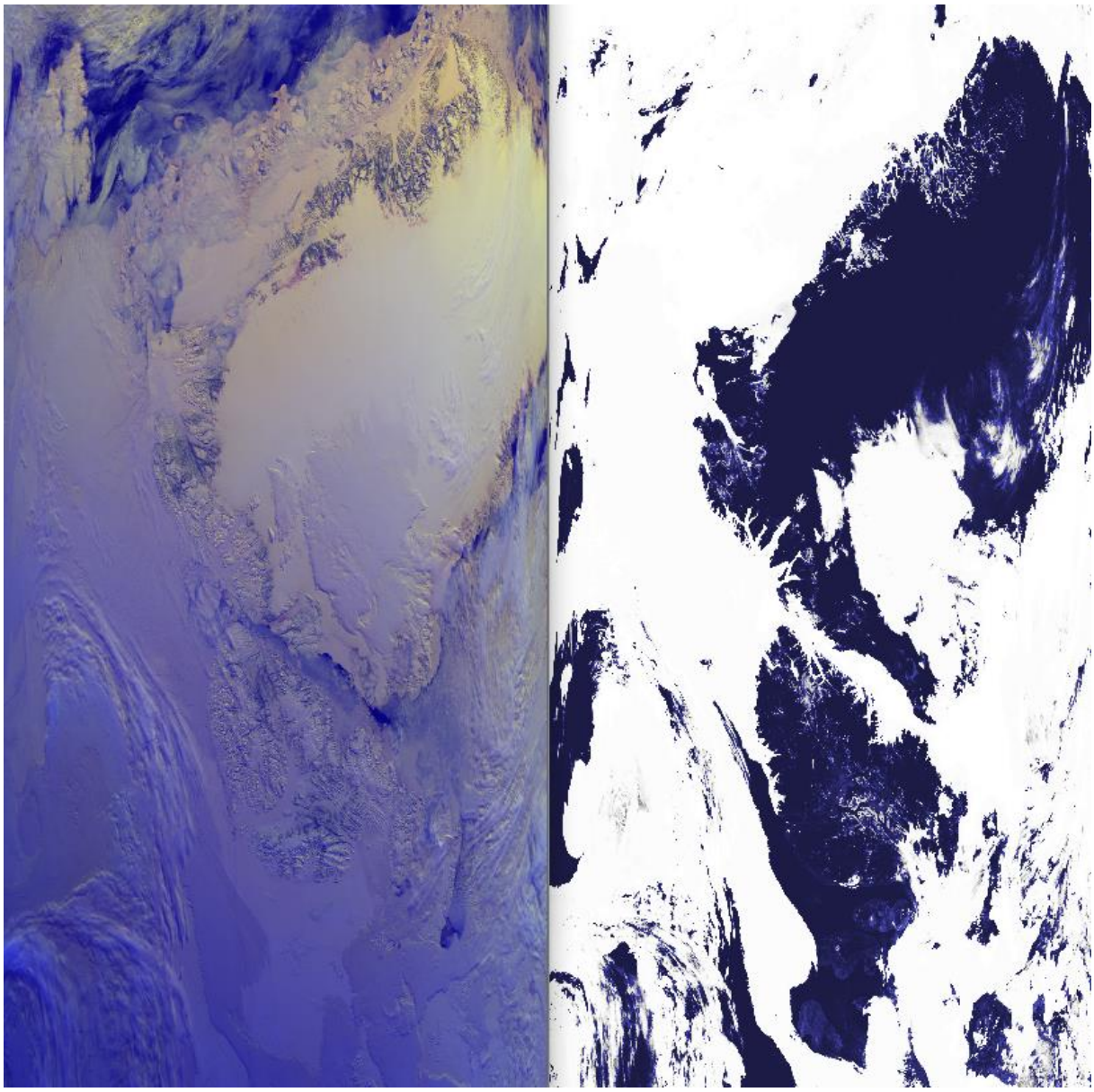

Figure 4-6 *Part of an original NOAA-18 AVHRR GAC scene in satellite projection over the Greenland area registered in ascending mode (i.e., North is down, South is up) from 16 May 2007 at 11:59 UTC.* ***Left****: Colour composite with AVHRR channel 1 (red), channel 2 (green) and channel 4 (blue).* ***Right****: Corresponding CMa-Prob cloud probabilities as greyscale image with range 0-100 %. Notice, however, that cloud probabilities below 50 % has a blueish colour.*

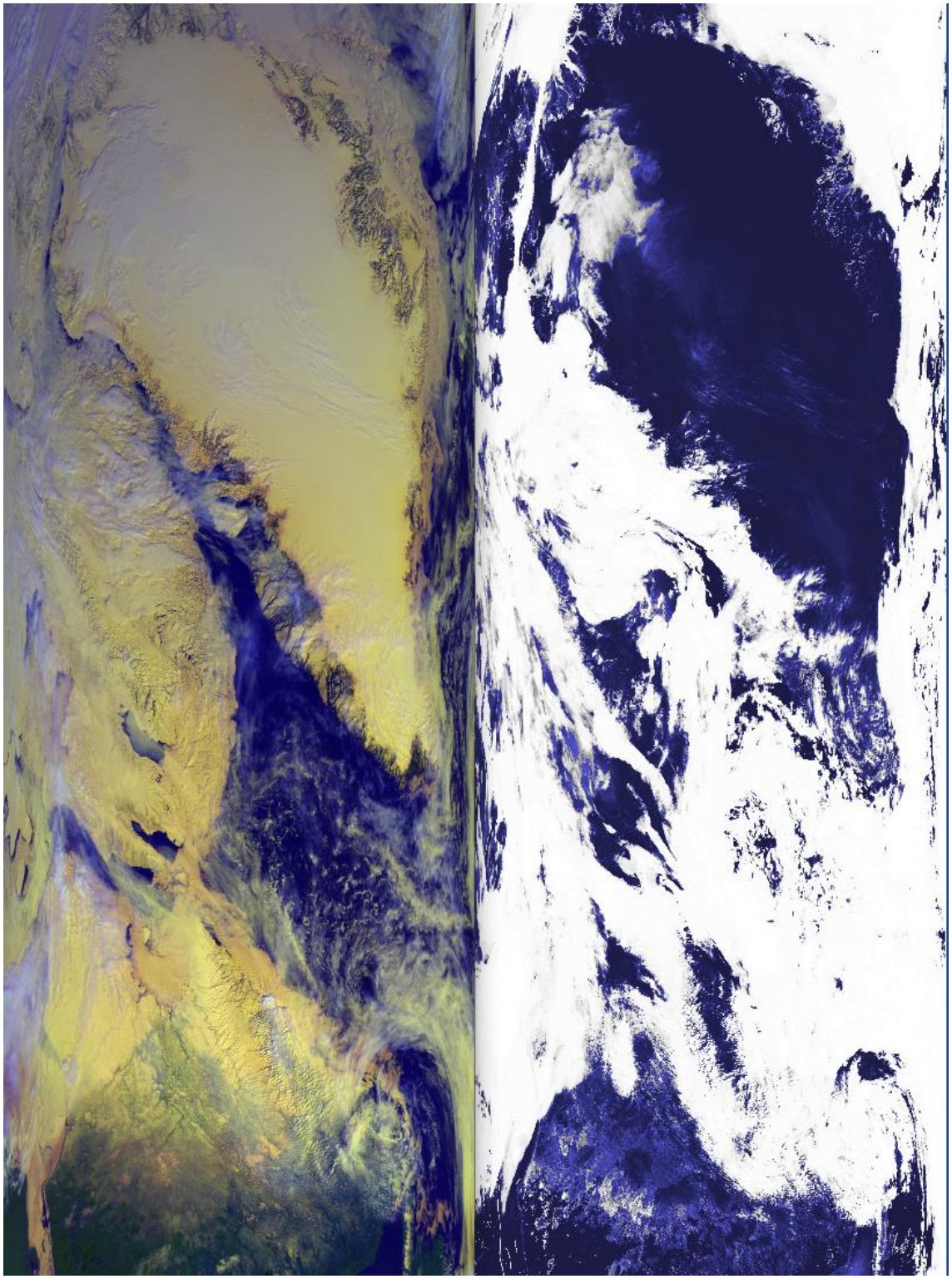

Figure 4-7 *Part of an original NOAA-17 AVHRR GAC scene in satellite projection over the Greenland area registered in descending mode (i.e., North is up, South is down) from 16May 2007 at 15:13 UTC.* ***Left****: Colour composite with AVHRR channel 1 (red), channel 2 (green) and channel 4 (blue).* ***Right****: Corresponding CMa-prob cloud probabilities as greyscale image with range 0-100 %. Notice, however, that cloud probabilities below 50 % has a blueish colour.*

5 Limitations and areas for future improvements

The CMa-prob method described in this document is the fourth prototype version and is this time based on dynamic threshold information from PPS version 2021. The first version was described in the paper by Karlsson et al. (2015) and was based on dynamic threshold information from PPS version 2010. The second version (briefly described by Karlsson et al., 2016) was based on dynamic threshold information from PPS version 2014. The third version was based on dynamic threshold information from PPS version 2018. Thus, it is clear that the method should be seen as an extension of the official PPS software and it cannot be run independently from PPS. Consequently, if continuing with this approach the method needs to be updated (new training) for every new release of the PPS method.

Statistical methods are always limited by the amount and quality of training data being used. It is clear from probability distributions shown in Sections 3.6 and 3.7, that the current training dataset needs to be further extended by providing very well-defined probability distributions over all Earth surfaces and for all conditions. Additionally, for some sensors like VIIRS, MERSI-2 and SLSTR, the method is currently executable based exclusively on statistics based on AVHRR and MODIS data. Thus, thorough training with, e.g., real VIIRS data has still to be done. In that respect, it is very encouraging that results are still as good as being documented in a previous validation report for an earlier version of CMa-prob (RD 2). However, it is very clear that an improved amount of training data would be beneficial for future versions of the method. Consequently, it is planned to extend the training material with considerably more data from CALIPSO up to present date. Also, ways of transferring the results between different satellite sensors will be considered (e.g., by use of Spectral Band Adjustment Factors (SBAFs, Bhatt et al., 2016).

6 Final remarks

This ATBD describes a method trying to address the quite serious problem that most cloud detection methods (including already existing probabilistic methods) do not describe in a quantitative way for which clouds the cloud mask or cloud probability is valid. This lack of information may mislead the user to believe that results are valid for any kind of cloud. The CMa-prob method provides additional information, in the sense that it describes for which clouds (i.e., their minimum cloud optical thickness) the cloud probability is valid. This minimum cloud optical thickness varies over different Earth surfaces since cloud separability is a function of both the cloud optical thickness and the characteristics of the Earth surface.

The key tool for adding this level of information to the cloud masking results has been the access to detailed cloud information from the CALIPSO-CALIOP measurements.

The concept used for CMa-prob assumes that cloud detectability is a function of cloud optical thickness and that this function can be described properly over homogeneous Earth surfaces. This has proven to be correct for many surfaces, e.g. over ocean surfaces without sunglint and ice cover. However, over many land surfaces and, especially, during cold surface conditions at night, application of the CMa-prob concept has shown to be problematic, leading to approximate solutions. The likely reason for this is that these surfaces do not provide reasonably homogeneous conditions which means that cloud detectability will depend on many other variables than just the cloud optical thickness. Further progress of methods applied over these

problematic surfaces has to include ways of stratifying data in a more efficient way to isolate the specific link to cloud optical thickness. This can possibly be achieved by use of more advanced machine-learning methods. The hope is that some of the findings from the work with CMa-prob methodology can be utilised in the development of such more advanced methods.

7 References

Bhatt, R., Doelling, D. R., Scarino, B. R., Gopalan, A., Haney, C. O., Minnis, P., & Bedka, K. M. (2016). A consistent AVHRR visible calibration record based on multiple methods applicable for the NOAA degrading orbits. Parts I&II; Methodology. J. Atmos. Ocean. Tech., 33(11), 2499-2515.

Derrien, M.and H. LeGléau, 2005: MSG/SEVIRI cloud mask and type from SAFNWC. *Int. J. Remote Sens.,* **26**, 4707-4732.

Dybbroe, A., A. Thoss and K.-G. Karlsson, 2005a: NWC SAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modeling - Part I: Algorithm description, *J. Appl. Meteor*, **44**, pp. 39-54.

Dybbroe, A., A. Thoss and K.-G. Karlsson, 2005b: NWC SAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modeling - Part II: Tuning and validation, *J. Appl. Meteor*, **44**, 55-71.

Frey, R. A., S. A. Ackerman, Y. Liu, K. I. Strabala, H. Zhang, J. R. Key, and X. Wang, 2008: Cloud detection with MODIS. Part I: Improvements in the MODIS cloud mask for collection 5, *J. Atm. Ocean. Tech*., **25**, 1057-1072.

Heidinger, A.K., A. T. Evan, M. Foster and A.Walther, 2012: A Naïve Bayesian Cloud Detection Scheme Derived from CALIPSO and Applied within PATMOS-x. *J. Appl. Meteor. Climatol.,* **51**, 1129-1144.

Hogan, R.J., M. Ahlgrimm, G. Balsamo, A. Beljaars, P. Berrisford, A. Bozzo, F. Di Giuseppe, R.M. Forbes, T. Haiden, S. Lang, M. Mayer, I. Polichtchouk, I. Sandu, F. Vitart and N. Wedi, 2017: Radiation in numerical weather prediction, ECMWF Technical Memorandum, **816**, ECMWF Research, Forecast and Copernicus Departments.

Inoue, T. 1987: A cloud type classification with NOAA 7 split-window measurements. *J. Geoph. Res.,* **92**, 3991–4000, DOI: 10.1029/JD092iD04p03991.

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., and Hollmann, R., 2017: CLARA-A2: The second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data, *Atmos. Chem. Phys.*, **17**, 5809-5828, www.atmos-chem-phys.net/17/5809/2017, doi: 10.519/acp-17-5809-2017.

Karlsson, K.-G. and A. Devasthale, 2018: Inter-comparison and evaluation of the four longest satellite-derived cloud climate data records: CLARA-A2, ESA Cloud CCI V3, ISCCP-HGM and PATMOS-x, *Rem. Sens.*, **10**, 1567, doi:10.3390/rs10101567.

Karlsson K.-G. and A. Dybbroe, 2010: Evaluation of Arctic cloud products from the EUMETSAT Climate Monitoring Satellite Application Facility based on CALIPSO-CALIOP observations. *Atmos. Chem. Phys.,* **10**, 1789–1807, 2010.

Karlsson, K.-G., and N. Håkansson, 2018: Characterization of AVHRR global cloud detection sensitivity based on CALIPSO-CALIOP cloud optical thickness information : demonstration of results based on the CM SAF CLARA-A2 climate data record. *Atmos. Meas. Tech*., **11**, 633–649. https://doi.org/10.5194/amt-11-633-2018

Karlsson, K.-G.; Håkansson, N.; Mittaz, J.P.D.; Hanschmann, T.; Devasthale, A. Impact of AVHRR Channel 3b Noise on Climate Data Records, 2017: Filtering Method Applied to the CM SAF CLARA-A2 Data Record. *Remote Sens.*, **9**, 568. https://doi.org/10.3390/rs9060568

Karlsson, K.-G. and E. Johansson, 2013: On the optimal method for evaluating cloud products from passive satellite imagery using CALIPSO-CALIOP data: example investigating the CM SAF CLARA-A1 dataset. *Atmos. Meas. Tech.,* **6**, 1271–1286, www.atmos-meas-tech.net/6/1271/2013/ , doi:10.5194/amt-6-1271-2013.

Karlsson, K.-G., E. Johansson and A. Devasthale, 2015: Advancing the uncertainty characterisation of cloud masking in passive satellite imagery: Probabilistic formulations for NOAA AVHRR data, *Rem. Sens. Env.* , **158**, 126-139; doi:10.1016/j.rse.2014.10.028.

Karlsson, K.-G., E. Johansson, N. Håkansson, J. Sedlar, S. Eliasson, 2020: Probabilistic Cloud Masking for the Generation of CM SAF Cloud Climate Data Records from AVHRR and SEVIRI Sensors. *Remote Sens.* **12**, 713. https://doi.org/10.3390/rs12040713

Kossin, J. P. and M. Sitkowski, 2009: An objective model for identifying secondary eyewall formation in hurricanes. *Mon. Wea. Rev.,* **137**, 876-892.

Kriebel, K. T., G. Gesell, M. Kästner and H. Mannstein, 2003: The Cloud Analysis Tool APOLLO: Improvements and Validations. *Int. J. Remote Sens*., 01/2003; **24**(2003):2389-2408. DOI:10.1080/01431160210163065.

Merchant, C. J., A. R., Harris, E. Maturi and S. MacCallum, 2005: Probabilistic physically-based cloud screening of satellite infra-red imagery for operational sea surface temperature retrieval, *Quart. J. Royal Met. Soc.,* **131**, 2735-2755.

Musial, J. P., F. Hüsler, M. Sütterlin, C. Neuhaus and S. Wunderle, 2014: Probabilistic approach to cloud and snow detection on Advanced Very High Resolution Radiometer (AVHRR) imagery. *Atmos. Meas. Tech.,* **7**, 799-822, www.atmos-meas-tech.net/7/799/2014/, doi:10.5194/amt-7-799-2014.

Pavolonis, M. J.; A. K. Heidinger and T. Uttal, 2005: Daytime global cloud typing from AVHRR and VIIRS: Algorithm description, validation, and comparisons. *Journal of Applied Meteorology*, **44**, Issue 6, 804-826.

Rossow, W. B., and R. A Schiffer, 1999: Advances in Understanding Clouds from ISCCP. Bull. Amer. Meteor. Soc., 80, 2261–2288.

Stengel, M., S. Mieruch, M. Jerg, K.-G. Karlsson, R. Scheirer, B. Maddux, J.F. Meirink, C. Poulsen, R. Siddans, A. Walther and R. Hollmann, 2013: The Clouds Climate Change Initiative: Assessment of state-of-the-art cloud property retrieval schemes applied to AVHRR heritage measurements, *Rem. Sens. Env.*, http://dx.doi.org/10.1016/j.rse.2013.10.035.

Trischenko, A. P., G. Fedosejevs, Z. Li, and J. Cihlar, 2002: Trends and uncertainties in thermal calibration of AVHRR radiometers onboard NOAA-9 to -16, *J. Geophys. Res.*, **107**(D24), 4778, doi:10.1029/2002JD002353.

Winker, D.M., M.A.Vaughan, A. Omar, Y. Hu, K.A. Powell, Z. Liu, W.H. Hunt and S.A. Young, S.A, 2009: Overview of the CALIPSO Mission and CALIOP Data Processing Algorithms. *J. Atm. Ocean. Tech*. **26**:11, 2310-2323.

Young, A. H., K.R. Knapp, A. Inamdar, W. Hankins and W.B. Rossow, 2018. The International Satellite Cloud Climatology Project H-Series climate data record product, Earth Syst. Sci. Data, **10**, 1-11, https://doi.org/10.5194/essd-10-583-2018

8 Glossary

ATBD	Algorithm Theoretical Baseline Document
AVHRR	Advanced Very High Resolution Radiometer
CALIOP	Cloud-Aerosol Lidar with Orthogonal Polarization (CALIPSO)
CALIPSO	Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations
CDOP	Continuous Development and Operations Phase
CLARA	CMSAF cLoud, Albedo and surface RAdiation dataset
CMa-prob	Cloud Mask (probabilistic)
CM SAF	Satellite Application Facility on Climate Monitoring
CPP	Cloud Physical Properties
DRI	Delivery Readiness Inspection
DWD	Deutscher Wetterdienst (German MetService)
ECMWF	European Centre for Medium Range Forecast
ECV	Essential Climate Variable
EPS	European Polar System

EPS-SG	EPS Second Generation
ERA-5	ECMWF Reanalysis, 5th edition https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FOV	Field of view
GAC	Global Area Coverage (AVHRR)
GCOS	Global Climate Observing System
IOP	Initial Operations Phase
ITCZ	Inter-Tropical Convergence Zone
KNMI	Koninklijk Nederlands Meteorologisch Institut
MERSI-2	Medium Resolution Spectral Imager onboard Chinese FYI-3 satellites
METimage	Meteorological imaging instrument for EPS-SG satellites
NASA	National Aeronautics and Space Administration
NDBC	National Data Buoy Center
NESDIS	National Environmental Satellite, Data, and Information System
NOAA	National Oceanic & Atmospheric Administration
NODC	National Oceanographic Data Center
NSIDC	National Snow and Ice Data Center
NWCSAF	Satellite Application Facility for Nowcasting
NWP	Numerical Weather Prediction
PPS	Polar Platform System
PRD	Product Requirement Document
PUM	Product User Manual
RMIB	Royal Meteorological Institute of Belgium
RMS	Root Mean Square
RSMAS	Rosenstiel School of Marine and Atmospheric Science
RSS	Remote Sensing Systems
SAF	Satellite Application Facility
SLSTR	Sea and Land Surface Temperature Radiometer onboard satellite Sentinel-3
SMHI	Swedish Meteorological and Hydrological Institute
SST	Sea Surface Temperature