remote sensing

MDPI

Article

Probabilistic Cloud Masking for the Generation of CM SAF Cloud Climate Data Records from AVHRR and SEVIRI Sensors

Karl-Göran Karlsson [1,*], Erik Johansson [1], Nina Håkansson [1], Joseph Sedlar [2] and Salomon Eliasson [1]

[1] Swedish Meteorological and Hydrological Institute, Folkborgsvägen 17, 601 76 Folkborgsvägen, Sweden; erik.johansson@smhi.se (E.J.); nina.hakansson@smhi.se (N.H.); salomon.eliasson@smhi.se (S.E.)

[2] Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, and NOAA Earth Systems Research Laboratory Global Monitoring Division, Boulder, CO 08013, USA; joseph.sedlar@colorado.edu

* Correspondence: karl-goran.karlsson@smhi.se; Tel.: +46-11-4958407

Received: 17 January 2020; Accepted: 19 February 2020; Published: 21 February 2020

Abstract: Cloud screening in satellite imagery is essential for enabling retrievals of atmospheric and surface properties. For climate data record (CDR) generation, cloud screening must be balanced, so both false cloud-free and false cloudy retrievals are minimized. Many methods used in recent CDRs show signs of clear-conservative cloud screening leading to overestimated cloudiness. This study presents a new cloud screening approach for Advanced Very-High-Resolution Radiometer (AVHRR) and Spinning Enhanced Visible and Infrared Imager (SEVIRI) imagery based on the Bayesian discrimination theory. The method is trained on high-quality cloud observations from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) lidar onboard the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) satellite. The method delivers results designed for optimally balanced cloud screening expressed as cloud probabilities together with information on for which clouds (minimum cloud optical thickness) the probabilities are valid. Cloud screening characteristics over 28 different Earth surface categories were estimated. Using independent CALIOP observations (including all observed clouds) in 2010 for validation, the total global hit rates for AVHRR data and the SEVIRI full disk were 82% and 85%, respectively. High-latitude oceans had the best performance, with a hit rate of approximately 93%. The results were compared to the CM SAF cLoud, Albedo, and surface RAdiation dataset from AVHRR data–second edition (CLARA-A2) CDR and showed general improvements over most global regions. Notably, the Kuipers' Skill Score improved, verifying a more balanced cloud screening. The new method will be used to prepare the new CLARA-A3 and CLAAS-3 (CLoud property dAtAset using SEVIRI, Edition 3) CDRs in the EUMETSAT Climate Monitoring Satellite Application Facility (CM SAF) project.

Keywords: probabilistic cloud mask; climate data records; CM SAF; AVHRR; SEVIRI

1. Introduction

Cloud masking is an essential first processing step for most retrievals of geophysical parameters based on radiance measurements from passive satellite imagery. It is essential for many different applications, including the retrieval of surface parameters, atmospheric ('clear air') properties, and cloud properties. As the satellite observation records gradually grow in temporal length, some of them now covering more than four decades, climate monitoring is becoming another important application. The generation of climate data records (CDRs) is imposing special requirements on the generated datasets, which are quite different from the requirements put on the standard environmental data

records (EDRs) produced in real-time or near-real-time. However, despite different requirements for different purposes, the same cloud masking approaches are often used regardless of application. Examples of this are given in [1–5], providing climate monitoring applications based on real-time or near-real-time cloud screening methods. Some schemes introduce increased flexibility [6–10], but no method focuses exclusively on the climate monitoring task. In this paper, we address the disadvantages associated with using similar methods in both real-time and climate monitoring applications. We propose an alternative cloud screening method, which is more suitable for climate monitoring applications, but which is still flexible enough for also being used in real-time applications.

The most serious problem with the use of standard and well-proven (in real-time applications) cloud screening methods for climate monitoring purposes is that most methods have initially been developed to serve surface parameter estimation applications. The classic example is cloud screening methods for sea surface temperature (SST) estimations [11]. Here, any presence of thin (and cold) clouds in areas declared as being cloud-free will result in a cold SST bias. Consequently, the cloud screening methods used for this purpose have been tailored to reduce the risk of producing cloud-free areas that still contain clouds; in other words, they are "*clear-conservative*" methods. The endeavor to minimize the probability of falsely labeling cloudy skies as clear, unfortunately, leads to an increased likelihood of falsely retrieving clouds when they are not present. Notice also that this clear-conservative approach could unfortunately have negative impact on retrievals of other parameters than SST. For example, aerosol optical depths might risk being underestimated since areas with high aerosol loads might now risk being falsely classified as clouds [12].

We conclude that these clear-conservative methods generally cause an overestimation of mean cloudiness if they are used for cloud CDR production. Examples of such overestimations are reported in [3,5,13,14] and confirmed by the comparisons with surface-based (SYNOP) observations and cloud lidar data from space (CALIPSO-CALIOP data). We conclude that cloud screening methods optimized for surface parameter retrievals are not necessarily optimal methods to be used for climate monitoring of cloud occurrence and cloud properties.

Consequently, there is a need for methods that aim at reproducing the most accurate mean cloudiness, by reasonably balancing false labeling of both clear-sky and cloudy cases. This paper presents a method for cloud screening that can be optimized for climate monitoring purposes, while being flexible enough for serving surface parameter retrieval applications. The method uses the Bayesian theory, which provides cloud mask information based on cloud probabilities.

The usefulness of a cloud CDR increases if results are well-characterized and include uncertainty measures. For this reason, the method has been trained in a semi-automatic iterative manner to establish a detailed description of cloud detection capabilities over different Earth surfaces. Access to high-quality information from CALIPSO-CALIOP measurements, regarding cloud presence and cloud layer optical thicknesses, has been essential in achieving this goal.

Section 2 describes the satellite data used, the background theory, the data used to train the new cloud masking method, and the principles used to achieve an optimal cloud screening over various Earth surfaces. Section 3 describes achieved results and validation results based on independent data followed by a discussion and conclusion in Sections 4 and 5.

2. Data and Methods

2.1. Satellite Data

The passive imagery component from polar-orbiting platforms comprises data from the Advanced Very-High-Resolution Radiometer (AVHRR). This sensor provides measurements in six spectral bands: Two visible bands at 0.6 μm and 0.9 μm, one near-infrared band at 1.6 μm, one short-wave infrared band at 3.7 μm, and two infrared bands at 11 μm and 12 μm. The spatial resolution is 1.1 km at nadir, but here we used the reduced resolution (approximately 4 km at nadir) global area coverage (GAC)

version of these measurements, since only GAC measurements are available on a global scale over the whole observation period (since 1979).

The polar-orbiting NOAA and Metop satellites, carrying AVHRR, operate nominally in pairs with one satellite in a morning orbit (i.e., with daytime local equator crossing time in the morning) and one satellite in an afternoon orbit. This constellation gives approximately four observations per day at the equator but an increasing number of observations with latitude (due to overlapping swaths), reaching a maximum of 28 observations per day at the poles. However, orbital drift effects and a variable lifetime of the sensors and satellites have caused large deviations from this nominal number of observations per day. However, due to technical achievements resulting in longer lifetimes of individual satellites, the frequency of observations has increased substantially during the last two decades.

The geostationary imagery component used in this study is taken from the Spinning Enhanced Visible Infra-Red Imager (SEVIRI) carried by the METEOSAT satellites operated by the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT). This sensor is a 12-channel sensor with 11 narrow-band visible, near-infrared, short-wave infrared, and infrared channels and one broad-band visible channel, covering the spectral range 0.6 to 13 μm. In this study, we used the narrow bands corresponding to the six heritage AVHRR channels mentioned earlier plus the infrared channel at 8.7 μm. The horizontal resolution at the sub-satellite point (nominally at the equator at the Greenwich meridian) was 3 km, and the revisiting observation time interval was 15 min.

High-quality cloud observations are available from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) sensor. This sensor is carried by the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) satellite. We used CALIOP observations to train and validate the cloud screening method. Sections 2.3–2.9 describe the training procedure, the CALIOP-based products, and the full extent of the training dataset in more detail.

2.2. Basic Theory—The Naïve Bayesian Classifier

The developed method is a further extension of previous work described in [8,15]. Here, it is sufficient to recall the fundamental equation for the Naïve Bayesian classifier formulated as

$$P\left(cloudy\middle|\boldsymbol{F}\right) = \frac{\overline{P(cloudy)}\prod_i P(f_i\middle|cloudy)}{P(\boldsymbol{F})} \quad (1)$$

where

$\boldsymbol{F}$ is a vector of satellite radiances or image features (e.g., brightness temperature differences or reflectances),

$P(cloudy|\boldsymbol{F})$ is the posteriori conditional probability that it is cloudy when given $\boldsymbol{F}$,

$\overline{P(cloudy)}$ is the overall probability (climatological mean) that is cloudy,

$P(f_i|cloudy)$ is the conditional probability that image feature f_i (i.e., one specific component of $\boldsymbol{F}$) occurs given it is cloudy,

$P(\boldsymbol{F})$ is the overall probability that any given value of $\boldsymbol{F}$ would occur.

Equation (1) is a simplification of a more general expression for the calculation of probabilities, formally denoted as the Bayes' Theorem. This particular simplification is valid if no strong correlation exists between individual image features. In that case, Bayes' Theorem reduces to a multiplication of individual feature probabilities, which is computationally much easier to handle than the more general expression, including all feature covariances. The multiplied conditional probabilities can be estimated using appropriate reference cloud observations, in this case from CALIPSO-CALIOP. The applicability of Equation (1) for cloud screening in satellite imagery has been demonstrated by [3,6,8,15]. The following section describes a further extension of the methodology seeking an improved and optimal training of the classifier.

2.3. Constrained Training over Different Earth Surfaces

The aforementioned Bayesian methods have generally compiled their required spectral feature statistics over different geographical regions. The purpose has been to find an appropriate classification performance avoiding overly broad statistical feature distributions that otherwise would occur if using globally merged statistics. More clearly, this is necessary to allow for varying spectral behavior of certain cloud types depending on the underlying surface (e.g., semi-transparent cirrus clouds). However, previous methods have not (or only in a minimal way) taken into account that not all clouds are detectable in passive imagery. For instance, the sensitivity to very thin clouds is significantly better in the referenced CALIOP observations than in passive imagery. This limited sensitivity of AVHRR data is particularly important over very bright surfaces, such as over snow cover or deserts, where very thin clouds (especially ice clouds), with detectability below the threshold, would effectively have similar spectral signatures as the underlying surface in passive imagery. In practice, this means that the optically thinnest clouds detected over bright surfaces will not have the same optical cloud properties as the thinnest detectable clouds over dark surfaces, e.g., over oceans. Conditions are generally similar at night: Very thin clouds may be indistinguishable from the surface, especially if cloud temperatures are close to surface temperatures. Such conditions lead to substantial differences in cloud properties for the thinnest clouds detected over the ice-free ocean compared to the corresponding thinnest detectable clouds over cold land surfaces.

In this study, we used CALIPSO-CALIOP cloud observations and validation tools, previously developed for evaluating cloud products derived from passive imagery [16,17], to discern which clouds, based on their optical thickness, are detectable over various surfaces. This information is crucial for any quantitative use of the derived cloud information in cloud climate data records (e.g., for evaluation of simulated cloudiness in climate models). The significant advantage of using CALIOP observations as the reference is that they not only provide information about the existence of clouds, but also information about the optical thickness of clouds [18]. This information (although restricted to the cloud optical thickness interval 0–5) can then be used to examine the ability for methods based on passive imagery to detect clouds with a particular optical thickness over different Earth surfaces, as demonstrated in [16,17]. We showed that by utilizing the same methods and data, we specified optimal training conditions for statistical classifiers like the Naïve Bayesian method.

2.4. Finding the Thinnest Clouds with Acceptable Detectability

We first assumed that the efficiency of cloud detection is a monotonously increasing function of a cloud's optical thickness. This relationship appears obvious and perfectly valid over a dark surface, like the ice-free ocean in the absence of sunglint and an aerosol-loaded atmosphere. However, it is also a reasonable assumption over bright and cold surfaces, even though the relation should be weaker here. For example, at night, we might have some clouds that will remain undetected regardless of their cloud optical thickness if they have similar thermal characteristics as the underlying surface. Nonetheless, most of the clouds will still show some detection dependency on the cloud optical thickness.

We also assumed that the minimum requirement, or expectation, for a cloud masking method is that it should at least detect more clouds than it misses. In other words, the probability of detection (*POD*) of a real cloud should be higher than 50%. Recalling that the efficiency of cloud detection should be a monotonously increasing function of cloud optical thickness, we should then try to locate the lowest cloud optical thickness value where *POD* is equal or larger than 50%. This particular optical thickness value has previously been defined as the Cloud Detection Sensitivity parameter (*CDS*) in [17]. We used this parameter in an iterative way to find the optimal performance over different Earth surfaces of our Naïve Bayesian cloud screening method. We can formalize this further by introducing the Hit Rate, *HR*, as described by Table 1 and Equation (2):

$$HR(\tau_{thr}) = \frac{a+d}{a+b+c+d} \quad (2)$$

Table 1. Contingency matrix for the number of Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO)-Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) cloudy and cloud-free observations versus predicted observations from any cloud masking method.

	Algorithm Predicted		
	Scenario	**Cloud-Free**	**Cloudy**
CALIPSO- CALIOP Observed	Cloud-free (or $\tau < \tau_{thr}$)	a	b
	Cloudy ($\tau > \tau_{thr}$)	c	d

Notice here that we use a generic definition of cloudy and cloud-free cases from CALIOP observations when we decide on cloud occurrence depending on the clouds' minimum optical thickness, τ_{thr}. In other words, cloudy cases can be either all clouds observed by CALIOP or a subset of all clouds, where clouds are relabeled as cloud-free if the optical depth is lower than a threshold value.

Furthermore, using the notation introduced in Table 1, we can define the *POD* quantity as

$$POD = \frac{d}{c+d} \tag{3}$$

Equation (3) describes the general case based on all cases described in Table 1. We can also restrict this quantity to be valid only for clouds with a specific cloud optical thickness τ:

$$POD(\tau) = \frac{d'}{c'+d'} \tag{4}$$

In this case, the primed quantities exclusively describe CALIOP-observed clouds with a particular value of τ:

The *HR* and *POD* quantities are our primary measures of cloud detection efficiency. In particular, we want to find that particular cloud optical thickness τ_{min} for which $POD(\tau)$ first equals or exceeds 0.5 (50%) when investigating the performance over increasing cloud optical thicknesses (expressed by an increasing τ_{thr} value) over a particular Earth surface. In other words, we want to define the reduced CALIOP cloud mask, compared to the unrestricted cloud mask with every optical thickness represented, that best fits the cloud detection capability of our method.

The cloud optical thickness threshold τ_{min} used for that particular reduced CALIOP cloud mask is then our *CDS* value. To make this clearer, we illustrate how we find this value in Figure 1. This figure plots the *HR* results of an arbitrary cloud detection method X as a function of the thresholded optical thickness value τ_{thr} of the reduced CALIOP cloud mask. We assume this curve to be valid when we have enough matchups with CALIOP. The leftmost position in Figure 1 (i.e., for zero value of τ_{thr}) defines *HR* results when using the original unrestricted CALIOP cloud mask for validation. We notice that when we start to reduce the CALIOP cloud mask, i.e., by increasing τ_{thr} values, the *HR* values first increase rapidly. The reason for this is that clouds with optical thicknesses below this threshold value, which were previously missed by the method, are now revalued from missed clouds to correctly determined clear cases by the CALIOP cloud mask reference (i.e., the latter now redefined as representing cloud-free conditions). At the same time, there is a revaluation of correctly classified cloudy cases now being labeled as false cloudy cases, but further inspection reveals that these occurrences are very few and have a very small impact on the results. Thus, the total effect is increasing *HR* values.

However, at a certain point, increasing the optical thickness threshold no longer increases the *HR*. For even higher optical thickness values, *HR* starts to drop again so that there is a peak *HR* value of the resulting curve. The maximum *HR* based on all matchups is close to 0.83 in this example case (Figure 1). The reason for the *HR* values decreasing for a further increase in optical thickness threshold is that an

increasing number of correctly identified clouds are now revalued as falsely identified clouds. Where the *HR* starts to decrease, the 'erroneously' revalued clouds now outnumber the undetected clouds. In practice, it means that precisely at the peak of the HR curve, we have 50% correctly classified and 50% misclassified cloudy cases for a cloud with the optical thickness given by the corresponding optical thickness threshold. In other words, $POD(\tau_{min})$ is 50%. Thus, here we get the best representation of the (restricted) CALIOP cloud mask according to our validation method and, consequently, the highest *HR* value for the investigated cloud detection method. This corresponding value of the thresholded optical thickness is the *CDS* value.

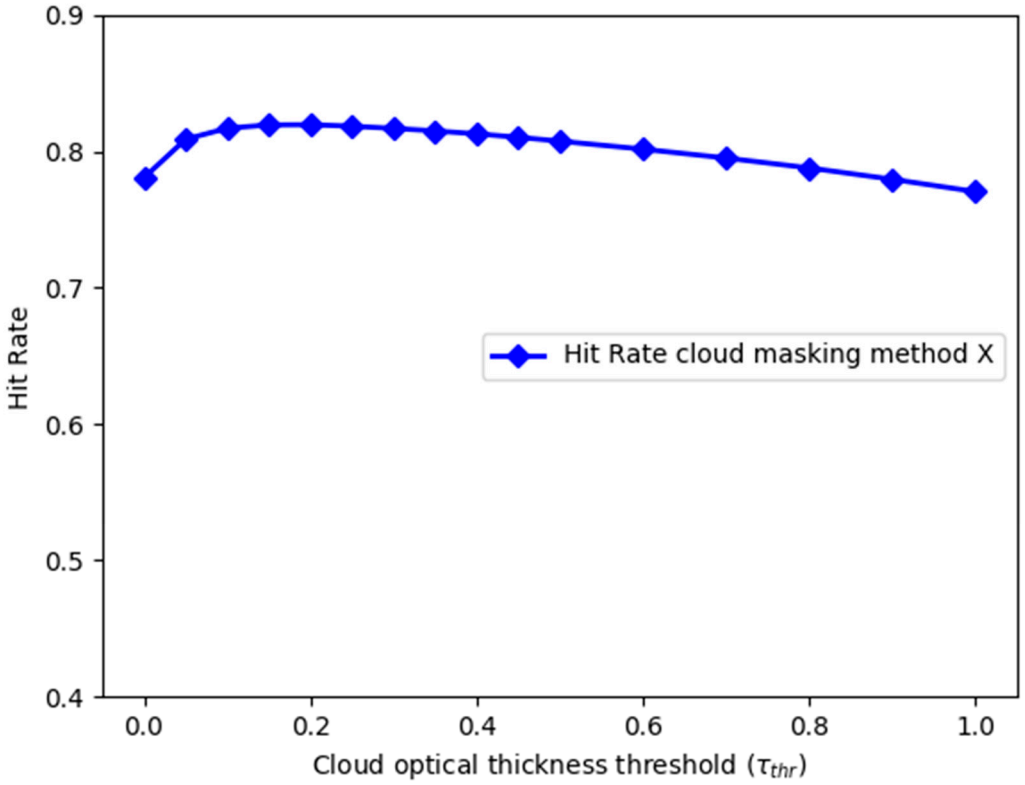

Figure 1. Cloud detection efficiency, as described by the Hit Rate (*HR*, giving the percentage of correct clear and cloudy cases compared to all cases) for an arbitrary cloud detection method X. These results are expressed as a function of CALIOP-filtered cloud masks with discrete optical thickness (τ_{thr}) thresholding steps. The results are estimated on cloud optical thickness intervals of 0.05 up to an optical thickness of 0.5. Above this value, the interval distance increases to 0.1.

2.5. Constraining the Training of a Statistical Classifier Using the CDS *Parameter*

Unfortunately, it is not enough to know the *CDS* value for making our results more easily understandable and more quantitatively useful when evaluating a probabilistic cloud masking method. The problem is that when training our method using any particular restricted CALIOP cloud mask, i.e., a mask resulting after applying a specific cloud optical thickness threshold, we would always be able to find a certain *CDS* value of that method when validating against all possible restricted CALIOP cloud masks. Thus, it is still not obvious which restricted CALIOP cloud mask to choose for training in order to ultimately get the best results.

We have chosen to solve this problem by requiring the following:

$$CDS \approx \tau_{thr}\_training \tag{5}$$

This means that the validation-derived *CDS* value has to agree (exactly or at least approximately) with the cloud optical thickness threshold τ_{thr} in the restricted CALIOP cloud mask being used for

training. In other words, we want to use that restricted CALIOP cloud mask in our training that is best reproduced (by peaking HR values) by our resulting image-based cloud mask.

In the ideal case, the method would then perfectly reproduce this restricted cloud mask and achieve the highest possible *HR* value, $HR = 1.0$, at this particular τ_{thr}. However, in reality, peak *HR* values seldom exceed 0.9. Notice also that if Equation (5) is valid (or approximately valid), it means that we have not only found that particular CALIOP cloud mask, which is best reproduced by our method. We have also ensured that a cloud probability threshold of 50% is always applicable over this particular Earth surface to create a binary cloud mask (which is required by many applications) to get optimal results. The latter is a consequence of the meaning of the peak in the *HR* curve in Figure 1 and the assumption of a monotonously increasing *POD* as a function of cloud layer optical thickness τ. Notably, in this context, a 50% cloud probability threshold has also been used to validate the cloud mask (i.e., for computing the *HR*-value) during this iterative process.

The use of this kind of cloud mask (i.e., cloud probabilities converted to a binary cloud mask at cloud probability 50%) would be optimal in the climate monitoring sense since it minimizes, in a balanced manner, misclassifications of both cloudy and clear cases. Such a cloud mask would be a definite improvement for users who otherwise may struggle with the problem of having to adjust their cloud probability threshold over different Earth surfaces for various applications when creating a binary cloud mask. Additionally, the probabilistic cloud mask product is flexible, such that applications requiring a more restrictive clear or cloud conservative cloud mask, can still adjust the threshold to either a lower or higher value than the optimal value at 50%.

By applying this strategy, we have also reduced the risk of overfitting our method, i.e., erroneously including truly sub-visible clouds. More clearly, if we would have trained against a CALIOP cloud mask that included clouds impossible to detect with passive imagery features, we would likely have produced a cloud mask very different from the one used for training. The reason would be that, for clouds with the lowest optical thicknesses in the CALIOP cloud mask, the spectral signature would be very similar to the cloud-free surface. Results would then be noisy, and the distribution of resulting cloud-free and cloudy pixels in this cloud optical thickness interval would become unrealistic. Consequently, such a CALIOP cloud mask is not likely to be selected as our optimal *CDS* value if following our selection strategy. Nevertheless, this reasoning is only valid for regions where sub-visible clouds are in the minority compared to all other clouds. Exceptional cases might occur in the tropics where sub-visible clouds can be relatively frequent, and where our method may still show some overfitting behavior just because of the high frequency of sub-visible clouds. This exception will be discussed further in Section 4.

We conclude that the ability to tie results from this method to a specific *CDS* value that is unique for each Earth surface could facilitate the quantitative use of the results in various applications. The resulting optimal *CDS* values over different surfaces are presented in Section 3.1.

2.6. Specification of Different Earth Surface Categories

An essential part of this method is to define homogeneous surface categories over the globe to achieve the best possible cloud screening results. The ideal method would be to train every individual grid point or pixel but limited global CALIPSO-CALIOP coverage (only observing in nadir) requires a limited number of surface categories in order to achieve statistical significance for the monitored spectral signatures. Also, training for individual grid points would certainly be risky in climate monitoring applications since it assumes that conditions are more or less static. For example, a position may change from permanent ice-cover to ice-free conditions later on (after the period with training data), which leads to problems because of a lack of the appropriate spectral statistics for that position. Table 2 gives an overview of all 28 surface categories used for training and final cloud screening.

Table 2. The surface categories used for training and 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 °C
Ocean high latitude north	G4	Ice-free ocean in NH with SSTs in the range 5–12 °C
Ocean mid latitude north	G5	Ocean in NH with SSTs in the range 12–22 °C
Ocean tropical	G6	Ocean with SSTs above 22 °C
Ocean mid latitude south	G7	Ocean in SH with SSTs in the range 12–22 °C
Ocean high latitude south	G8	Ice-free ocean in SH with SSTs in the range 5–12 °C
Ocean polar south	G9	Ice-free ocean in SH with SSTs below 5 °C
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 °C
Land homogeneous extra tropical seasonal snow	G12	Homogenous land with vegetation, seasonal snow cover, and LSTs below 12 °C
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 °C
Land rough extra tropical seasonal snow	G16	Land with vegetation over rough terrain with seasonal snow and LSTs below 12 °C
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 °C
Land rough tropical	G19	Land over rough terrain with vegetation and LSTs above 12 °C
Ocean polar north sunglint	G20	Arctic ice-free ocean with no sunglint and SSTs below 5 °C
Ocean high latitude north sunglint	G21	Ice-free ocean in NH with no sunglint and SSTs in the interval 5–12 °C
Ocean mid latitude north sunglint	G22	Ocean in NH with no sunglint and SSTs in the interval 12–22 °C
Ocean tropical sunglint	G23	Tropical ocean with sunglint and SSTs above 22 °C
Ocean mid latitude south sunglint	G24	Ocean in SH with sunglint and SSTs in the interval 12–22 °C
Ocean high latitude south sunglint	G25	Ice-free ocean in SH with sunglint and SSTs in the interval 5–12 °C
Ocean polar south sunglint	G26	Ice-free ocean in SH with no sunglint and SSTs below 5 °C
Coast extra tropical	G27	Coastal areas with LSTs below 12 °C
Coast tropical	G28	Coastal areas with LSTs above 12 °C

The definition of the surface categories had to be performed with care. We wanted the spectral signatures of clouds and land surfaces to differ significantly over each chosen surface while they should also preferentially differ from corresponding signatures over other surfaces. For the surface definitions, we used the following data, tools, and criteria:

- The 1 km land cover characterization from the United States Geological Survey (USGS, [19]) was used to separate land and ocean surfaces.

- The Global 30 arc seconds topography database GTOPO30 (https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30?qtscience_center_objects=0#qt-science_center_objects) was used to separate homogeneous and rough (mountainous) terrain.
- The products OSI-450 and OSI-430b from the Ocean and Sea Ice Satellite Application Facility (OSISAF) project [20] provided sea ice concentrations.
- The ERA5 reanalysis dataset [21] provided snow occurrence and land surface temperature.
- The separation between dry and vegetated land surfaces was made based on surface emissivity information provided by the Moderate Resolution Imaging Spectroradiometer (MODIS) MYD11C3 product (https://lpdaac.usgs.gov/products/myd11c3v006/). Results were aggregated into climatologies with a monthly temporal resolution before being used during training.
- The coastal zone was defined using the fraction of land calculated from the USGS 1 km land cover information in an 11 × 11 USGS pixel neighborhood. Results were remapped to the closest AVHRR or SEVIRI pixel and pixels with fraction of land larger than 25% but less than 90% were regarded as coastal. Notice also that areas in the vicinity to rivers or small lakes may be classified as coastal.

Observe that the ambition is to be able to process the entire AVHRR data record (starting in 1979) with this method. Thus, the used ancillary datasets should preferably cover this period without any gaps or discontinuities. This explains why, generally, datasets with better horizontal resolution but limited temporal coverage were not chosen (with the exception of the emissivity dataset where long-term datasets were missing).

Crucially, we entirely avoided fixed geographical boundaries (e.g., defined by latitude and longitude coordinates) when defining regions. Instead, we let existing thermal conditions decide whether we had tropical, sub-tropical, extratropical, or Arctic/Antarctic conditions. Although the choice of temperature limits was somewhat arbitrary, we believe that they give a realistic description of how temperature zones vary dynamically on both the short and long time scales. Using regions based on temperature limits instead of geophysical coordinates avoids potential discontinuities in the retrieved cloud climatologies. This make conditions within the defined Earth surface categories more homogeneous, which improves the applicability of the cloud screening method.

Sunglint conditions have been estimated based on the criterion for Phong specular reflection [22]. Sunglints appear when the angular difference between the satellite viewing angle and the reflection angle for Phong specular reflection is smaller than 27 degrees (empirically derived).

Retrievals along the coast are problematic. As opposed to all other regions where the underlying surface is either land or sea, many pixels in coastal areas contain information from both. The mixed composition within a pixel, together with uncertainties in navigation, and in the land mask used, may lead to spurious and false cloud patterns along coastlines or rivers. The use of coastal categories mitigates the problems mentioned above, but the disadvantage is that cloud probabilities are now generally more uncertain (i.e., closer to 50%) along coastlines and rivers. However, this correctly reflects a higher uncertainty in the results over these surfaces.

2.7. Selected Image Features in AVHRR and SEVIRI Imagery

The choice of image features used for cloud screening from the original spectral channels of the AVHRR and SEVIRI sensors was based on many years of experience with cloud masking from passive imagery [1,23,24]. We largely followed the approach outlined in [15], but some extensions were introduced, especially for the adaptation to SEVIRI data. Tables 3 and 4 list the different image features used (day and night, respectively), the associated spectral channels, and the main contribution to cloud screening from each image feature.

Table 3. Chosen image features for **daytime** cloud screening of Advanced Very-High-Resolution Radiometer (AVHRR) and Spinning Enhanced Visible and Infrared Imager (SEVIRI) scenes and their main contributions. Spectral channels are described with their central wavelengths.

Image Feature Name	Spectral Channels AVHRR	Spectral Channels SEVIRI	Composition and Importance for Cloud Screening
Rvis	0.63 μm (land) or 0.86 μm (ocean)	0.63 μm (land) or 0.81 μm (ocean)	Visible reflectances. Identification of bright clouds over dark Earth surfaces
Rswir_3a	0.63 μm and 1.61 μm	0.63 μm and 1.64 μm	Reflectance quota between the two channels. Identification of clouds with significant reflection in the visible near-infrared infrared region (in particular water clouds and thick multi-layered clouds over snow-covered surfaces)
Rvis37	3.74 μm	3.92 μm	Short-wave infrared reflectances. Identification of clouds with significant reflection in the short-wave infrared region (water clouds and thick multi-layered clouds)
Rswir_3b	3.74 μm and 12.0 μm	3.92 μm and 12.0 μm	Difference between brightness temperatures in the two channels. Identification of thin cirrus clouds.
Tirdiff	10.8 μm	10.8 μm	Difference between brightness temperatures and surface (skin) temperatures from ERA5. Identification of clouds colder than Earth surfaces
Tmirdiff	-	8.70 μm and 10.8 μm	Difference between brightness temperatures in the two channels. Contributes to the identification of fog due to different behavior of liquid water clouds and Earth surfaces in the two channels.
Texture_day	0.63 μm and 10.8 μm	0.63 μm and 10.8 μm	Sum of local variances (in 3×3 pixel windows) of reflectances and brightness temperatures. Identification of small cloud elements (fractional cumulus or cirrus) **exclusively** over ocean surfaces (i.e., well away from coasts and islands).

Table 4. Chosen image features for **night-time** cloud screening of AVHRR and SEVIRI scenes and their main contributions. Spectral channels are described with their central wavelengths.

Image Feature Name	Spectral Channels AVHRR	Spectral Channels SEVIRI	Composition and Importance for Cloud Screening
Tirdiff	10.8 μm	10.8 μm	Difference between brightness temperatures and surface (skin) temperatures from ERA5. Identification of clouds colder than Earth surfaces
Twdiff	3.74 μm and 10.8 μm	3.92 μm and 10.8 μm	Difference between brightness temperatures in the two channels. Contributes to identification of water clouds.
Tcidiff	10.8 μm and 12.0 μm	10.8 μm and 12.0 μm	Difference between brightness temperatures in the two channels. Contributes to identification of thin ice clouds.
Tmirdiff	-	8.70 μm and 10.8 μm	Difference between brightness temperatures in the two channels. Contributes to identification of fog due to different behavior of liquid water clouds and Earth surfaces in the two channels.
Texture_night	3.74 μm, 10.8 μm and 12.0 μm	3.92 μm, 10.8 μm and 12.0 μm	Sum of local variances (in 3×3 pixel windows) for the brightness temperature at 10.8 μm and for the difference between 3.74 (or 3.92) μm and 12.0 μm brightness temperatures. Identification of small cloud elements (fractional cumulus or cirrus) **exclusively** over ocean surfaces.
2D_Tirdiff_Twdiff	3.74 μm and 12.0 μm	3.92 μm and 12.0 μm	Two-dimensional combination of the **Tirdiff** and **Twdiff** image features. Secures a constrained use of the two features (see text for explanation).

A lack of the 8.7 µm channel and the alternating availability of 1.61 µm and 3.74 µm for the AVHRR sensor means that only 4 or 5 image features can be used simultaneously during the daytime and nighttime, while for SEVIRI, all described features can be used. For SEVIRI data, this can easily be handled theoretically (see Equation (1)) and may lead to further improved results as long as features do not show strong correlations. The method has been applied successfully to data from other sensors (e.g., the MODIS and VIIRS sensors, [25,26]), but these results are not covered here.

Notice that the **Tirdiff** and **Twdiff** features can be used either independently during the day or constrained during the night according to the **2D_Tirdiff_Twdiff** feature. The reason for the coupling of the two features at night emerged after observing that these features often neutralize or contradict each other in cold situations. One disadvantage of the Naïve Bayesian classifier is that if one single feature has near-zero probabilities, it may dominate image features to yield low probabilities even if all other image features would give high cloud probabilities. In this particular case, the problem often occurs for inversion-capped boundary layer water clouds over very cold (often snow-covered) land surfaces in the winter season. Even if the **Twdiff** feature gives high probabilities for a cloud in this situation, it can be completely neutralized by the **Tirdiff** feature since these clouds are often substantially warmer than the surface in this feature (i.e., clouds are generally colder than the surface). By introducing this two-dimensional feature, we can reduce this problem.

2.8. Solving Problems with AVHRR Data Representativity in the Available CALIPSO Matchup Dataset

The access to reference observations from the CALIPSO-CALIOP lidar is a tremendous asset for the development and validation of cloud screening methods from passive imagery (e.g., as demonstrated in [16,17]). However, there are two critical aspects of the CALIOP observations to take into account for the development of a successful cloud processing method that aim to be used operationally under all conditions:

1. CALIOP matchups with AVHRR and SEVIRI observations are only possible during nadir overpasses (from the CALIPSO perspective);
2. CALIPSO is orbiting in an afternoon orbit (with equator crossing time in the ascending mode near 13:30 Local Time).

Both limitations lead to the fact that collocations and matchups with global coverage can only be realized with data from polar satellites orbiting in an afternoon orbit similar to that of CALIPSO. Consequently, polar satellites in any other orbit can only be collocated with CALIPSO at very high latitudes (near 70 degrees of latitude). This aspect has three serious consequences when it comes to the prospects of training a cloud screening method to be applied to AVHRR data from polar-orbiting sun-synchronous satellites:

1. Training of cloud screening methods utilizing data from the 1.6 µm channel (i.e., channel 3a) can only be done at very high latitudes since AVHRR sensors with this channel mainly operates in morning orbits.
2. Since near-nadir viewing angles dominate the global AVHRR-CALIOP matchups, training data are mostly missing for medium to high viewing angles.
3. Training of sunglint signatures is severely restricted since global CALIPSO matchups require small satellite viewing angles for the AVHRR sensor at a local observation time several hours apart from noon (thus excluding the occurrence of specular reflection of the sun in the field of view of collocated data).

The following sub-sections describe our strategies for solving or mitigating the problems listed above. Importantly, the angular viewing conditions for matching CALIOP observations with SEVIRI data are much more advantageous due to the high temporal resolution of imagery from the geostationary platform. Consequently, all three aspects above can easily be dealt with when collocating with SEVIRI data. However, it is clear that with CALIPSO being a satellite in the A-train constellation, with its

main observations made in the early afternoon and the middle of the night, it will be challenging to cover all illumination conditions, regardless of whether training with polar or geostationary data. The twilight cases will be poorly represented over low-to-moderate latitudes and not at all considered in the tropical regions.

2.8.1. Training with 1.6 μm Channel Data

The main problem here is that, due to condition 1 above, we cannot train our method efficiently over dry surfaces (i.e., deserts and other regions with sparse vegetation) where we know that this channel measures high surface reflectances, which may be confused with cloud reflectances for some type of clouds. However, another sensor operating in a polar orbit similar to NOAA and Metop satellites and measuring in this spectral band is the MODIS sensor. Furthermore, if we restrict our interest to the MODIS sensor on the Aqua satellite (officially known as the EOS PM-1 satellite), we find that MODIS measurements are already closely tied to the CALIPSO orbit in the A-train constellation. Matchup conditions are excellent with an approximate 1-min time difference between the MODIS and CALIOP observation. It is also clear that the spectral response functions for this channel on the AVHRR and MODIS sensors are quite similar (Figure 2). Even though they are not identical, they are both well within a region with high atmospheric transmissivity (atmospheric window) in the spectral region 1.5–1.8 μm [27], which means that only surfaces with rapidly changing spectral reflectances may differ noticeably in appearance in the two channels. Consequently, we have assumed that MODIS measurements can reasonably well represent AVHRR measurements in this channel and we have collocated MODIS with CALIPSO-CALIOP data for the global training of the spectral signatures associated with this particular channel (i.e., linked to feature **Rswir_3a** in Table 3).

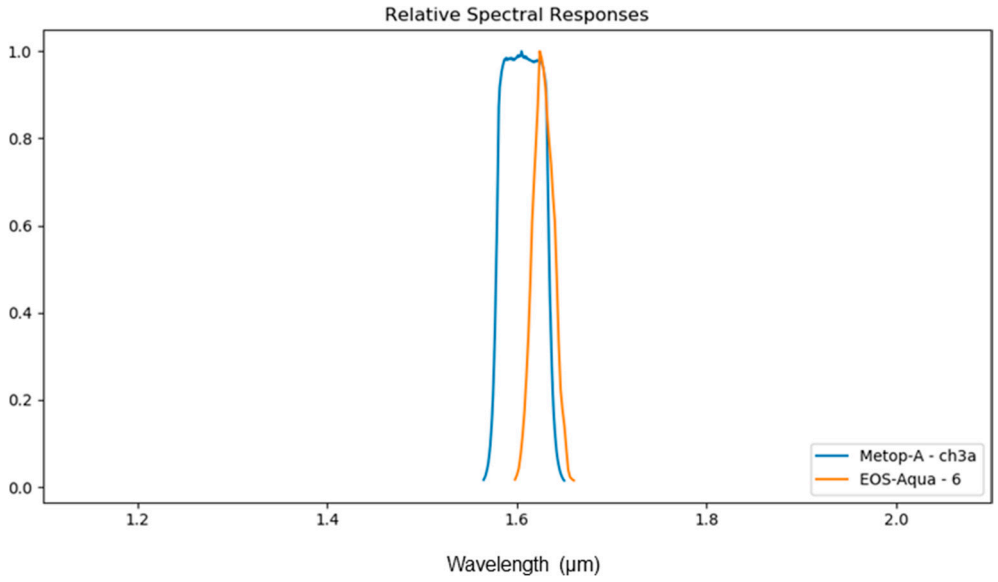

Figure 2. Spectral responses in the spectral interval 1.1–2.1 μm for the 1.6 μm channels of AVHRR (denoted channel 3a and here represented by the AVHRR of the Metop-A satellite) and MODIS (channel 6 on the MODIS/AQUA satellite).

2.8.2. Representing Measurements at High Viewing Angles

Two dominant effects make the spectral measurements at high viewing angles differ from corresponding measurements in nadir:

1. Surfaces or clouds may reflect differently (anisotropically) at high angles;
2. The cloud-free atmosphere and thin clouds look more opaque at high angles.

The first effect is difficult to handle (however, see discussion about sunglints in the next section), but the impact is generally relatively small except at very high viewing angles near the swath edge. Consequently, we have not compensated for this effect in the current version of the method.

The second effect (or at least the part related to the changed appearance of the cloud-free atmosphere) is simulated by reformulating the image feature of interest so that it includes an implicit dependency on the viewing angle. The method used here is the one introduced in [15], which takes into account the impact of increasing viewing angles. Basically, the impact is largely the same as the impact seen when measuring in an increasingly humid atmosphere. Thus, increasing viewing angles leads to an increasing amount of absorbing gases (mainly water vapor) along the line of sight. For infrared channels, it generally leads to lower measured radiances (i.e., lower brightness temperatures), while for visible channels the effect can be neglected in most cases (except in the presence of heavy aerosol loading in the troposphere). The reduction of temperatures also means that the probabilities of detecting a cloud will change compared to the nadir case with warmer temperatures. The approach here is to relate the image feature value for infrared features to a quantity that is a function of atmospheric humidity. In this way, we would also get an implicit dependency on the viewing angle.

The probabilistic classifier, being developed as an alternative method to the cloud masking method provided by the well-established Polar Platform System (PPS) software package for cloud processing developed by the EUMETSAT Nowcasting Satellite Application Facility (NWC SAF) project, uses the same prepared input data as PPS (from versions PPS 2014 and higher). The pre-calculated PPS thresholds for the different infrared image features are all functions of the viewing angle, and thus of the total moisture content in the measurement path through the atmosphere. Consequently, image features involving the infrared channels in Tables 3 and 4 were all reformulated as a difference to the pre-calculated PPS thresholds, which then enables an implicit dependency on viewing angles and atmospheric water vapor. We illustrate the effect of this reformulation of image features in Figure 3 for the **Tcidiff** feature in Table 4 (with statistics taken from the training dataset described in more detail in Section 2.9). The figure shows the relation between the brightness temperature differences in the 10.8 μm and 12.0 μm channels (**Tcidiff**) and the cloud occurrence deduced from CALIPSO-CALIOP observations.

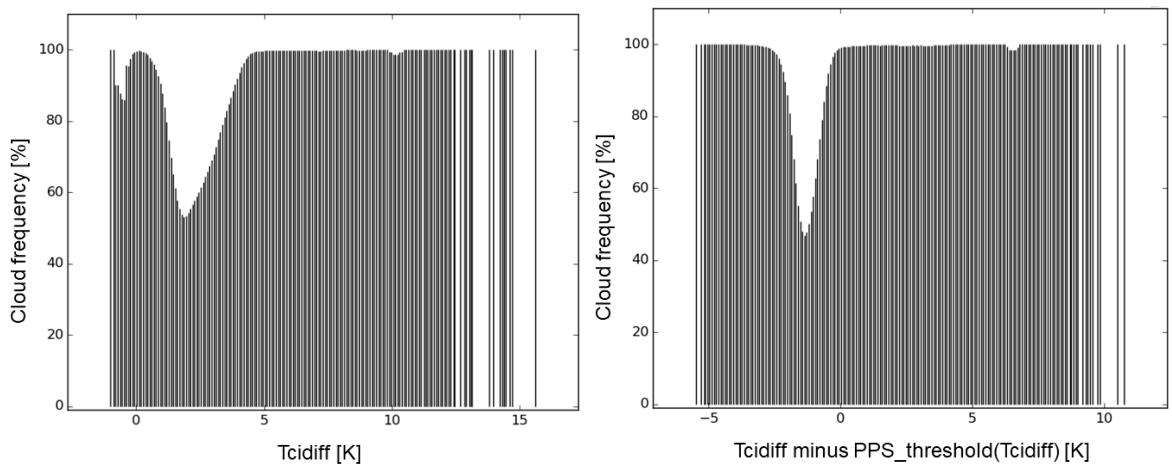

Figure 3. Cloud occurrence (or cloud frequency) histograms as a function of the original **Tcidiff** feature (left) compared to the case of a reformulated **Tcidiff** feature formed as a difference to the PPS threshold (right). The histograms are valid at night over tropical ocean surfaces. The **Tcidiff** value on the x-axis in the left figure is positive when brightness temperatures at 11 μm are larger (warmer) than at 12 μm. See text for further explanation.

When using the original form of this feature (Figure 3, left), we get near 100% cloud occurrence for differences close to zero and differences that exceed approximately 4 K. Thus, the area with lower cloud frequencies between the peaks spans an interval of almost 4 K. In the reformulated feature (Figure 3,

right), the interval reduces to about 2 K, and the range of probability values are slightly enlarged, which is favorable for the probabilistic classification process. The enlarged range of probabilities is especially true for the leftmost part of the distribution. We interpret this as a primary effect of better taking into account the cloud-free contribution from atmospheric water vapor emission in the split-window channels. This emission creates a discernible temperature difference in the absence of cirrus clouds, which explains the generally lower original cloud probabilities for temperature differences below approximately 4 K in Figure 3 (left). The resulting distribution after the small coordinate change in Figure 3 (right) now clearly separates thin cirrus clouds to the right in the plot (for reformulated **Tcidiff** values above zero) from the opaque clouds in the left part of the plot with cloud-free cases now concentrated around the threshold-subtracted value of around −1 K. Notice that cloud occurrences from CALIPSO-CALIOP in this figure contain all CALIOP-observed clouds (thus, including sub-visible clouds not detected from AVHRR measurements) over tropical ocean surfaces, explaining why minimum cloud frequencies are still relatively high and not zero.

2.8.3. Treatment of Sunglints

As mentioned previously, it is difficult (or nearly impossible) to find AVHRR-CALIOP matchups that include sunglint effects at the same time as requiring access to global collocations. The reason is that when NOAA and CALIPSO satellite orbits are very closely aligned, the observations are always near-nadir. Only very weak sunglints can be encountered primarily in windy conditions over the ocean in these collocations, but the strong sunglints centered around the position for specular reflection will always be well separated geometrically from matchup locations. For SEVIRI data, specular reflection effects can always be included in the CALIOP matchups due to the high temporal resolution imagery and the different viewing conditions from the geostationary platform.

However, there are special conditions under which stronger sunglints may appear in a small subset of AVHRR/CALIOP matchups. These occur when the NOAA satellite orbits start drifting, thus deviating from the orbital plane of CALIPSO. In the case of NOAA satellites, the satellites start to drift at the time when new satellites are being launched and declared as prime satellites instead of the old ones (since only prime NOAA satellites will undergo orbit corrections keeping the orbit stable).

During the short period of orbital decline, usually lasting a few years, AVHRR/CALIOP matchups are possible at high AVHRR viewing angles since the CALIOP observations match the AVHRR near the swath edges. As time progresses, the orbit drifts further, and the matchups become confined to high latitudes only (similar to the conditions for morning orbit satellites). Thus, over a limited time during this orbit transition process, it will be possible to encounter conditions with strong sunglint in some AVHRR/CALIOP matchups.

We have been utilizing this 'window' of possible sunglint observations in matchups for the two satellites NOAA-18 and NOAA-19. For NOAA-18, orbital drift started in 2010 (shortly after the launch of NOAA-19 in 2009), producing matchups with high viewing angles in the following two years. Similarly, NOAA-19 started drifting in 2013 (shortly after the launch of the Suomi-NPP satellite in late 2011), producing matchups with high viewing angles in the following two years. We used these time windows with high viewing angle conditions in matchups to train a sunglint signature for the probabilistic classifier applied on afternoon orbit data.

For AVHRRs in the morning orbit, we had no chance of training for the sunglint signature at low to middle latitudes. MODIS data cannot be used here either, since MODIS and CALIOP collocations are both fixed to near-nadir conditions in the A-train constellation. Therefore, for necessity, we exclusively used thermal features (here, the **Tirdiff** feature in Table 3) in sunglint conditions for the morning satellites.

2.9. Final Training Dataset and Selection of Independent Validation Data

The following sub-sections provide details on training and validation data for the involved AVHRR and SEVIRI sensors.

2.9.1. AVHRR Training and Validation

We chose to concentrate AVHRR training to the period when global matchups based on the afternoon NOAA-18 and NOAA-19 satellites were still possible. Therefore, collocations with CALIPSO-CALIOP data started in October 2006 and ended in December 2015 when the number of global AVHRR matchups was significantly reduced because of orbital drift.

The collected training and validation dataset provide reasonable global coverage for all seasons during that period. Data from afternoon orbits make up of 6319 NOAA-18 and NOAA-19 AVHRR GAC orbits and CALIOP pixels/samples at approximately 5 km horizontal resolution. The deduced spectral signatures in the training dataset for afternoon orbit satellites were also used when applying the method to morning orbit satellite data. However, they were here complemented with training statistics derived from 4827 MODIS Aqua granules matched with CALIPSO-CALIOP in order to cover also reflectances in the 1.6 μm channel, which were missing in the afternoon orbit dataset. This particular MODIS channel showed some problems with missing data from broken detectors but enough of training data could nevertheless be compiled. These MODIS-CALIOP collocations are from one year (2010). A small subset of all theoretically possible MODIS-CALIOP matchups, from the 1st and 14th of each month, was used for this purpose. We used the CALIPSO-CALIOP Cloud Layer (CLAY) product version 4.10 in all collocations and a more detailed description of this product is given in [28]. The constrained training (i.e., with image feature values being linked to PPS threshold information rather than being used as standalone feature values) was based on results of the PPS software version 2018 [29]. Observation time differences for AVHRR and CALIPSO collocations were limited to 5 min. Matchups between MODIS and CALIPSO-CALIOP had a fixed time difference of approximately one minute.

All the AVHRR-CALIOP collocations from 2010 constituted the independent validation dataset. Thus, these data were not used for the training of the method. The reason for choosing this particular year was that it offered global matchup data for both the NOAA-18 and NOAA-19 satellites before being subject to significant orbital drift. It also allowed an independent but restricted (to high latitudes) validation of data from one morning satellite (Metop-A).

In total, 29.8 million AVHRR-CALIOP matchups trained the method, complemented with 9.5 million MODIS-CALIOP matchups for the training based on the 1.6 μm channel measurements. The corresponding numbers for the selected validation datasets were 5.7 million matchups from 822 afternoon orbits and 0.2 million matchups from 232 morning orbits.

2.9.2. SEVIRI Training and Validation

The likelihood of finding useful collocations between SEVIRI and CALIOP observations is much better than for AVHRR due to the high repetition rate (15 min) of SEVIRI. This means that collocations are possible every day instead of every third day for AVHRR. Also, collocations are possible for several SEVIRI slots each day compared to only one or two orbits every third day for AVHRR. Consequently, after using a maximum time difference of 7.5 min, we collected matchups for every single CALIOP observation passing over the SEVIRI field of view.

We selected two years for training with SEVIRI data: 2010 from METEOSAT-9 and 2015 from METEOSAT-10. As for MODIS data, we first extracted collocations from every 1st and 14th day in 2010 for METEOSAT-9. However, to increase the statistical significance, the training dataset was then further extended with data from METEOSAT-10 in 2015 but now with data from every third day each month, resulting in a total of approximately 3 million matchups extracted from 6529 SEVIRI scenes.

As an independent validation dataset, we picked one full month of SEVIRI data from July 2012. This dataset contains 673,675 matchups from 1477 SEVIRI scenes.

3. Results

In the following sub-sections, we provide results from estimation of *CDS* values, demonstration of final cloud probability products, and from validation activities.

3.1. Determination of Cloud Detection Sensitivities over Different Earth Surfaces

The training of the probabilistic classifier included finding the appropriate optimal $\tau_{thr}\_training$ (*CDS*) values over different Earth surfaces in accordance with the previous discussion in Section 2.5. The process of finding this *CDS* value means that we had to train our method iteratively by successively shrinking CALIOP cloud masks as a consequence of increasing the filtering thresholding τ_{thr} value. Based on all individual validation results, the optimal *CDS* solution was found for the case when *CDS* equaled τ_{thr} ($=\tau_{min}$) for a particular restricted CALIOP cloud mask (i.e., satisfying Equation (5)). In some cases, this condition was satisfied for several adjacent restricted CALIOP cloud masks. In this scenario, the chosen *CDS* was the lowest τ_{min} value provided where peak *HR* values are not decreasing.

This concept for selecting the optimal *CDS* value and for finding the best representation of the CALIOP cloud mask worked best over dark and ice-free ocean surfaces. Figure 4 shows the daytime (i.e., for solar zenith angles below 80 degrees) results for the category *Ocean Polar South* (category G9 in Table 2). The different colored curves show *HR* results when the classifier was trained with different restricted CALIOP cloud masks. It is clear that optimal results, i.e., when the τ_{thr} value at the peak in *HR* equaled the τ_{thr} value for the restricted CALIOP cloud mask used for training, occurred for τ_{min} value 0.05. Thus, for this training condition, we got the highest peak in *HR*, and the *CDS* value was 0.05. It means that over the Southern Hemisphere ice-free polar ocean, our method is capable of detecting clouds down to an optical thickness of 0.05 with a detection efficiency of at least 50%. These results are also valid at night and for the corresponding category on the northern hemisphere (not shown). Notice that we got the highest *HR* peaks (near 95%) of all available categories for the two Ocean Polar categories (i.e., Ocean Polar south and Ocean Polar north). It is clear that for an ice-free ocean surface with a clean atmosphere (i.e., low aerosol and moisture contents), cloud detection works very efficiently, mainly leaving only errors caused by collocation errors and sub-pixel cloud effects (as discussed in [17]).

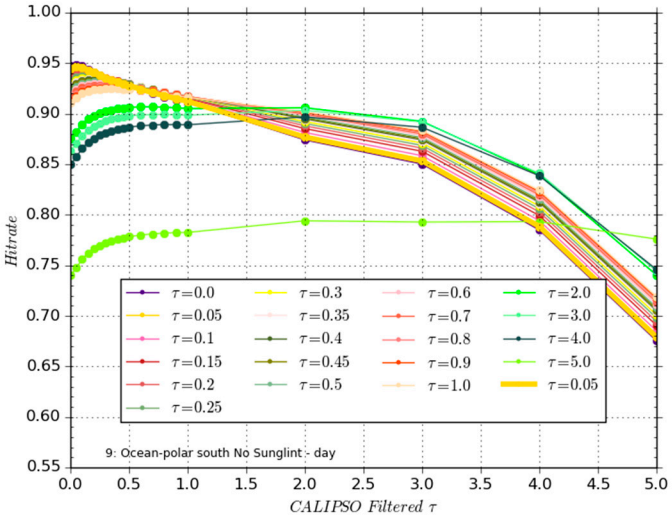

Figure 4. Validation results (Hit Rate) for the probabilistic classifier as a function of successively reduced (i.e., filtered with value τ_{thr}) CALIOP cloud masks for surface category *Ice-free ocean in SH with SSTs below 5 °C* (category G9 in Table 2) during daytime. The curve with a thick yellow line shows the training condition satisfying the selection criteria in Equation (5) for finding τ_{min} (see the text for further explanation).

Results were also promising giving the same τ_{min} value for other ocean categories (including ocean categories with sunglint—see Table 5) but peak *HR* values slowly decreased with latitude. An exception was the *Tropical Ocean* category for which τ_{min} decreased to 0.1. This is illustrated in Figure 5 for surface G6 *Tropical Ocean* at night. The humid atmospheric conditions here can explain the slight degradation of results. The high moisture content in the Tropics risks being misinterpreted as thin cirrus, and the moister the atmosphere, the more significant the risk. Results at daytime for this category were similar (Table 5), but here it was clear that factors (e.g., varying aerosol loads) other than simply cloud optical thickness also played a role since peak *HR* values were often as high or higher for both lower and higher τ_{thr} values than the actual value used for training.

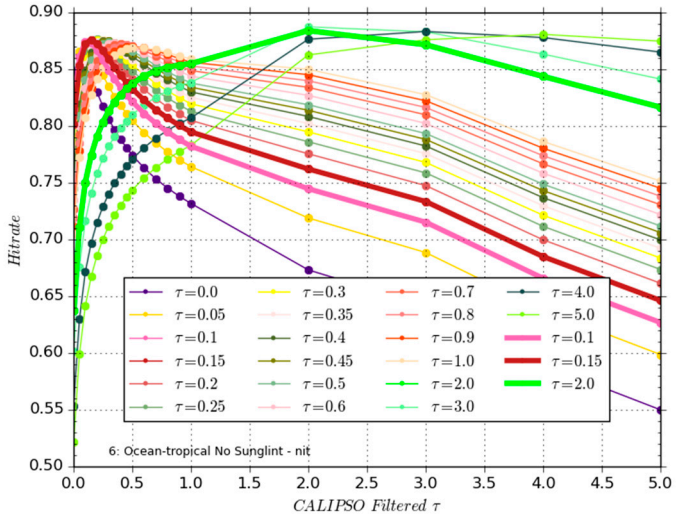

Figure 5. Validation results (Hit Rate) for the probabilistic classifier as a function of successively reduced (i.e., filtered with value τ_{thr}) CALIOP cloud masks for surface category *Tropical Ocean with SSTs above 22 °C* (category G6 in Table 2) during the night. The curves with thicker widths show training conditions with possible τ_{min} values satisfying the selection criteria in Equation (5) (see the text for further explanation).

Considering land surface categories, we can first study the daytime and night-time results for vegetated (non-dry) extra-tropical land areas (Figure 6) as defined by category G11 in Table 2. Similar to the G6 results during the daytime, the curves (Figure 6, left) failed to achieve distinct peaks in *HR* for approximately similar τ_{min} values. Our current results indicate that there is a solution for τ_{min} 0.45, but many curves in the τ_{thr} interval 0.2–0.4 are also close to being chosen. However, at night (Figure 6, right), we find a robust solution at τ_{min} 0.25, also yielding a slightly higher peak *HR* values than the daytime results. The conditions where a good solution is more difficult to find during daytime may relate to the fact that this surface category was not as homogeneous as the previously discussed ocean categories. For example, this surface characterization can contain substantial variability of surface and aerosol characteristics, and both of these factors significantly affect visible spectral channels. The direct dependence of results on varying cloud optical thicknesses may, therefore, weaken. Despite these potential consequences, we chose to apply the same τ_{min} value (0.25) for both day and night over this surface type since this would lead to the detection of more thin clouds during daytime compared to using the suggested value 0.45.

Dry surfaces like deserts and adjacent dry semi-vegetated regions produced interesting results. During the daytime (Figure 7, left), *HR* results exhibited a similar behavior to those found for vegetated areas in Figure 6, but the optimal *CDS* value was now even higher (τ_{min} = 1.0). It is clear that higher surface reflectances made it more difficult to detect very thin clouds, although thermal contrasts between surfaces and clouds may have increased (e.g., for cirrus). However, for night-time results, a solution was found for much thinner clouds with a τ_{min} value of 0.25.

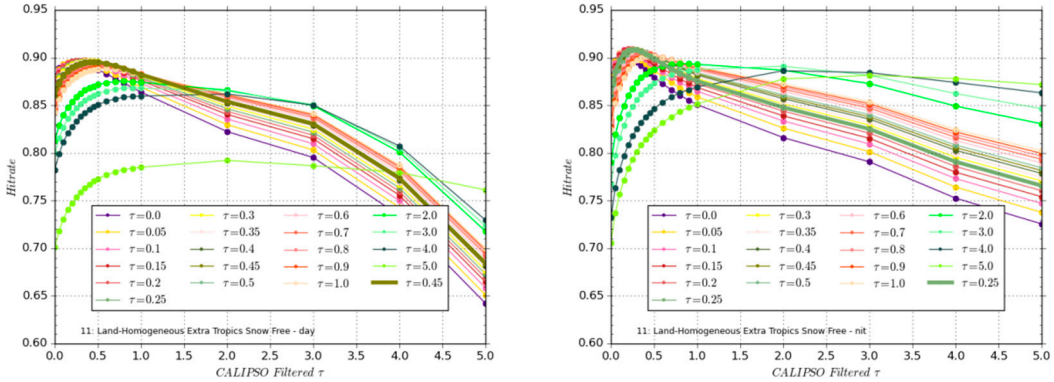

Figure 6. Validation results (Hit Rate) for the probabilistic classifier as a function of successively reduced (i.e., filtered with value τ_{thr}) CALIOP cloud masks for surface category *Land homogeneous extra tropical* (category G11 in Table 2) during daytime (left) and during night (right). The curves with thicker widths show training conditions with possible τ_{min} values satisfying the selection criteria in Equation (5) (see the text for further explanation).

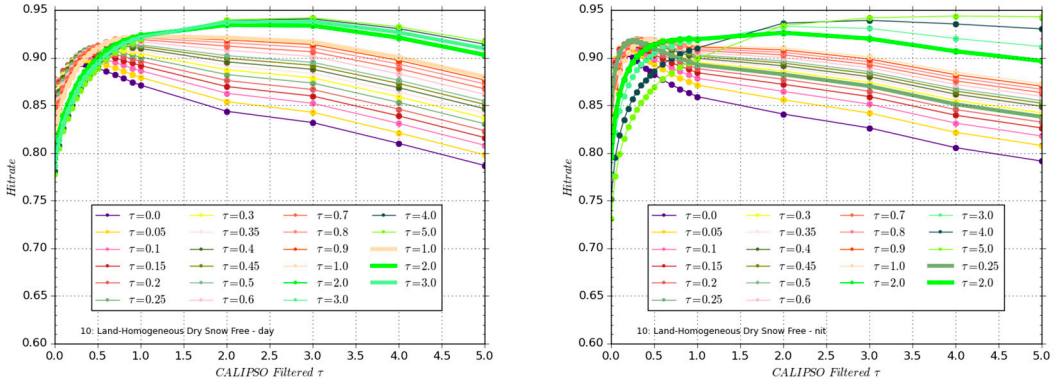

Figure 7. Validation results (Hit Rate) for the probabilistic classifier as a function of successively reduced (i.e., filtered with value τ_{thr}) CALIOP cloud masks for surface category *Homogeneous dry land* (category G10 in Table 2) during day (left) and night (right). The curves with thicker widths show training conditions with possible τ_{min} values satisfying the selection criteria in Equation (5) (see the text for further explanation).

The degraded performance during daytime was remarkable, considering that more spectral information should be available from the AVHRR instrument compared to the conditions at night. It is clear that during these surface and illumination conditions, our method was not able to accurately reproduce any restricted CALIOP cloud mask. The suggested solutions for τ_{min} 1.0 and higher were not convincing since the degradation of *HR* was prolonged after passing the peak value. Physically, this means that for τ_{thr} values higher than τ_{min}, we will not detect significantly more than 50% of all existing clouds, indicating that the performance of the cloud mask under these conditions is not exclusively dependent on the cloud optical thickness. Other factors, such as the varying surface emissivity/reflectivity and the reflectance contribution from aerosols over the desert, play essential roles for the cloud detection ability over these land surfaces. However, *HR* values were still quite high here for many different τ_{thr} values. Consequently, a compromise could be to select the same τ_{thr} value 0.25 as during night. Although *HR* values decreased slightly at the peak, it yielded a solution with clearly decreasing *HR* values to the right of the peak (indicating a *POD* higher than 50%), which is preferable. The drawback is that we cannot confidently claim that our τ_{min} value was actually equal to 0.25 for daytime illumination conditions.

It is not surprising to find that the most challenging conditions for passive imagery cloud masking occurred over surfaces with permanent snow cover, especially during the night (Figure 8). Overall,

HR values were considerably lower than for other surfaces (especially at night), indicating more challenging conditions for cloud detection. We failed to identify any cloud optical thickness training condition that adequately fulfilled our criterion in Equation (5) for finding the optimal *CDS* during the night. Daytime conditions seemed somewhat similar to conditions over dry surfaces at night, suggesting an approximate *CDS* value of 1.0. Nighttime conditions showed no suitable solution, even indicating that a τ_{min} value higher than 5.0 should be explored. However, such a solution would mean we have no detection skill at all for thin clouds, which disagrees with our current experience. A better solution here would be to select a much lower τ_{thr} value where we could still find a distinct peak in *HR* values. Although *HR* values were lower with this solution, it would still indicate decent detection ability also for thinner clouds compared to what the τ_{min} suggestion says. Here we chose values 0.3 and 0.5 for daytime and nighttime, respectively, to acknowledge the more difficult detectability over these surfaces while still allowing reasonable detection of thin clouds. The reason for the above-described breakdown of our method for finding the optimal detection conditions is most likely due to that several other factors, besides varying cloud optical thickness, start to influence cloud screening abilities. The most prominent factor here is that thick clouds can remain undetected during cold winter night conditions due to thermal resemblance with the underlying surface. This and other reasons will be discussed further in Section 4.

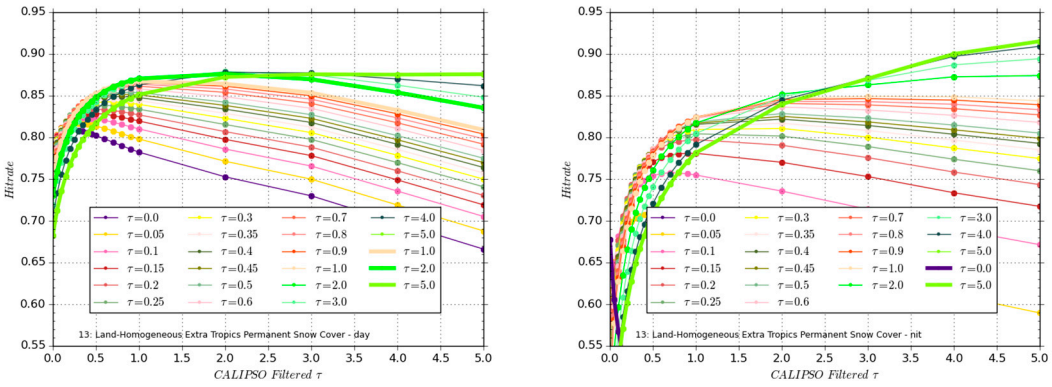

Figure 8. Validation results (Hit Rate) for the probabilistic classifier as a function of successively reduced (i.e., filtered with value τ_{thr}) CALIOP cloud masks for surface category *Homogeneous land with permanent snow cover* (category G13 in Table 2) during day (left) and night (right). See text for further explanation.

Finally, we present all estimated *CDS* values for all investigated Earth surfaces in Table 5. Notice that the twilight category currently re-used either the day or the night values based on a day-night test examining reflectances in the 0.6-micron channel.

Table 5. Estimated Cloud Detection Sensitivity (*CDS*) values for all surfaces defined in Table 2 sorted in three illumination categories.

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	0.35	0.25	0.35/0.25
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

3.2. Demonstration of Achieved Cloud Masking Products

Figures 9 and 10 below demonstrate the cloud probability product based on the Naïve Bayesian probabilistic cloud mask over the SEVIRI full disk for one daytime case (SEVIRI slot at 12:00 UTC) and one night-time case (SEVIRI slot at 00 UTC) from 16 July 2010. Figure 11 shows the same cloud probability product from three consecutive global NOAA-19 orbits observing at approximately the same time as the SEVIRI scenes. The product examples were taken from the Northern Hemisphere summer with rather typical weather patterns for the season. The night-time results in Figure 10 and the results in near-Antarctic regions in Figure 9 illustrate conditions when the spectral information was limited to purely infrared channels. Observe that cloud probabilities for the SEVIRI case were restricted to be valid only for satellite viewing angles up to 70 degrees. Results indicate dry and cloud-free conditions over most land areas in Africa and Europe (except near the equator) while oceanic regions were cloudier. Notice the high cloud probabilities in oceanic areas with broken cloud fields, which are not so obvious to interpret from visual inspection of the multispectral images.

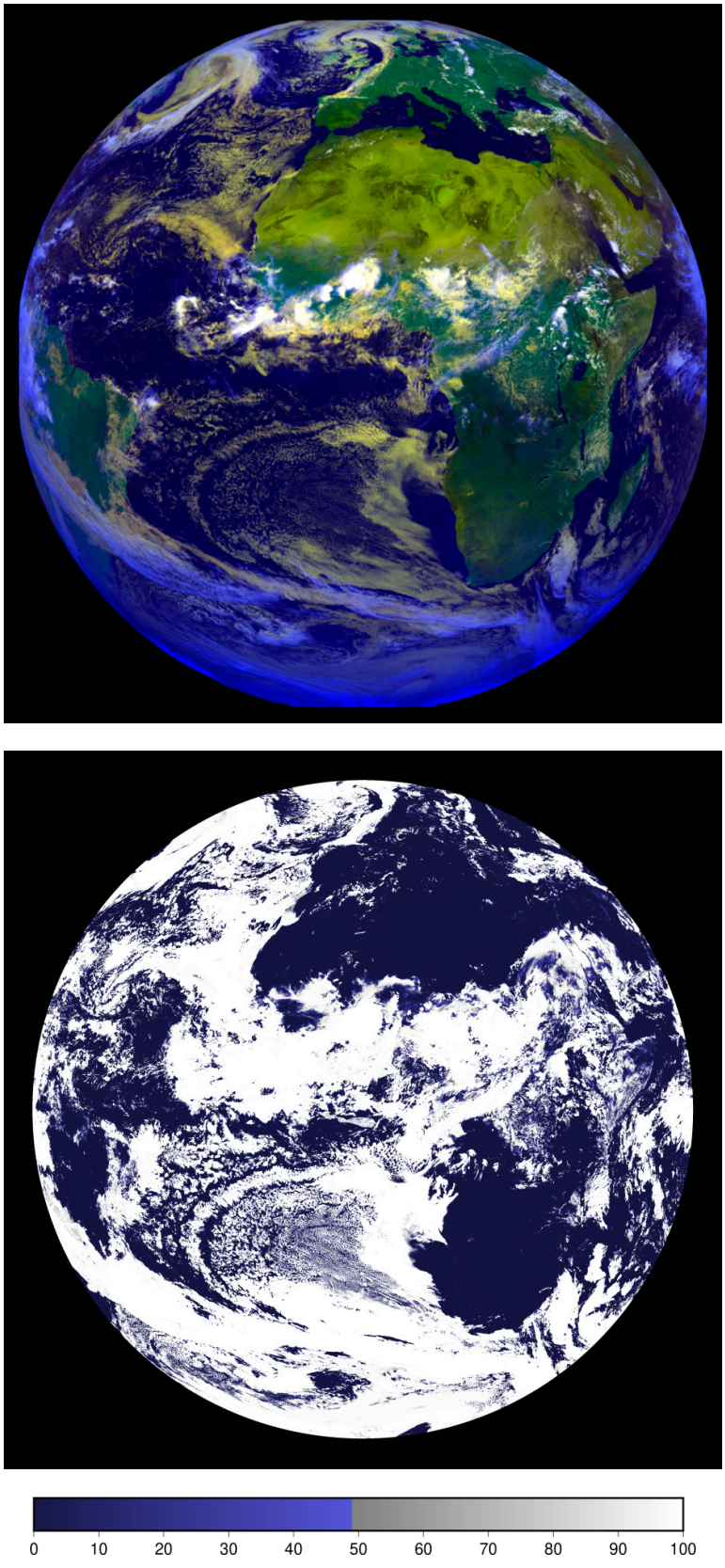

Figure 9. Top: Color composite of a METEOSAT-9 SEVIRI scene from 16 July 2010 at 12:00 UTC including information from SEVIRI channels at 0.6 µm, 0.9 µm, and 11 µm. Bottom: Corresponding cloud probability product (in %).

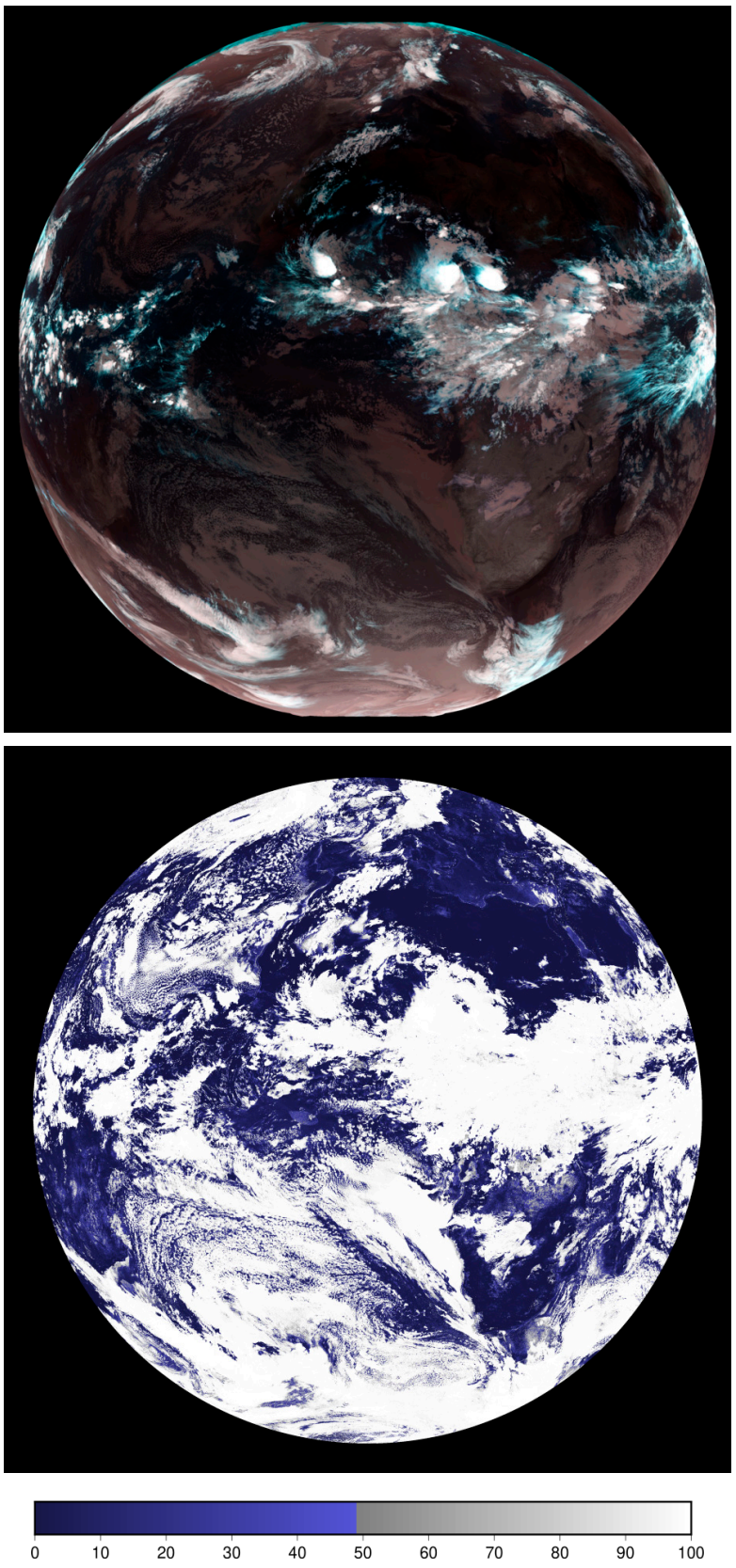

Figure 10. Top: Color composite of a METEOSAT-9 SEVIRI scene from 23 July 2010 at 00:00 UTC including information from SEVIRI channels at 3.9 µm, 11 µm, and 12 µm. Bottom: Corresponding cloud probability product (in %).

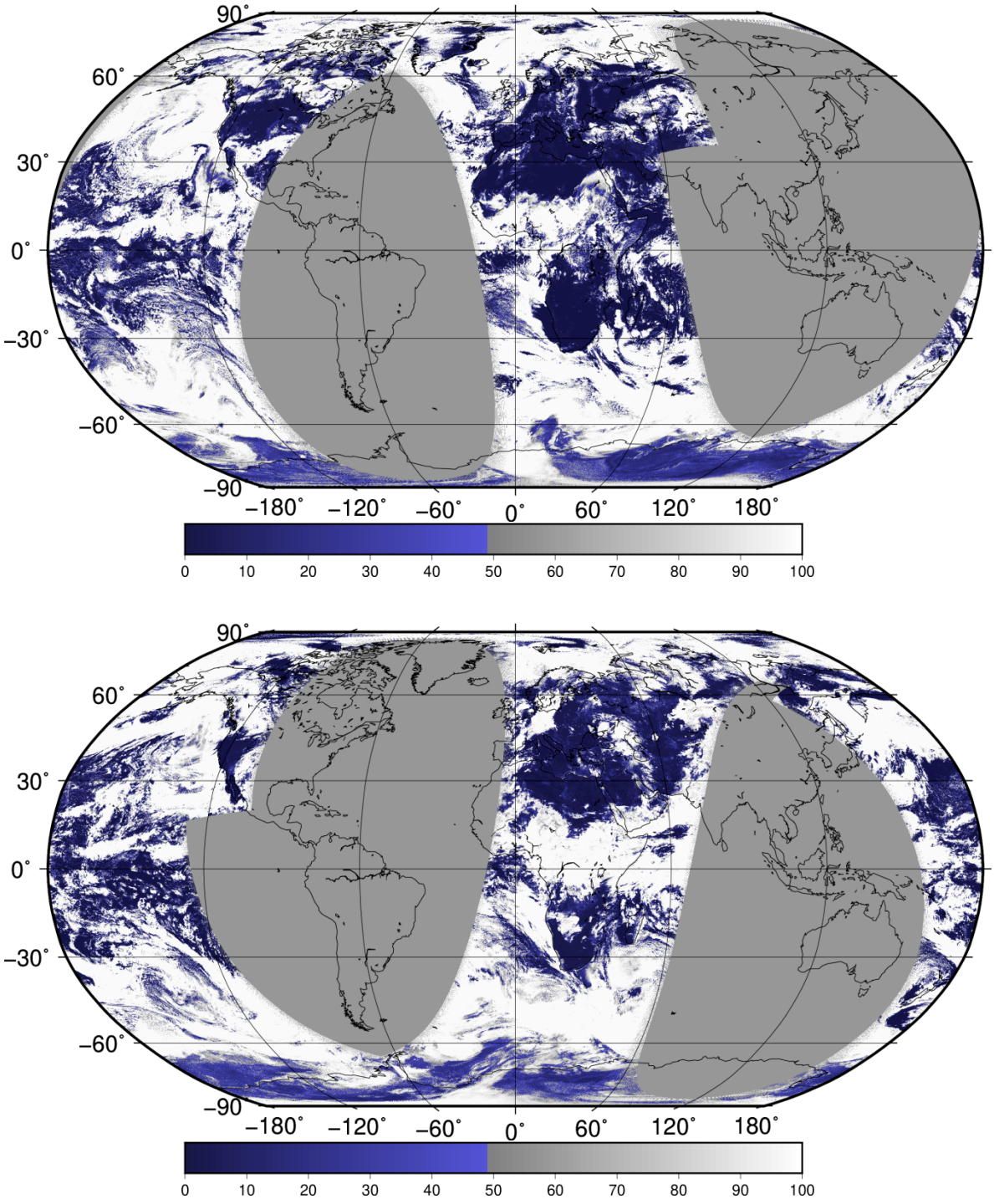

Figure 11. Top: Cloud probabilities (in %) from three consecutive NOAA-19 orbits 16 July 2010 close in time to the SEVIRI scene shown in Figure 9. First scanline is from 08:49 UTC and last scanline from 14:08 UTC. Bottom: Cloud probabilities (in %) from three consecutive NOAA-19 orbits 22–23 July 2010 close in time to the SEVIRI scene shown in Figure 10. First scanline is from 22 July 21:18 UTC and last scanline from 23 July 02:38 UTC.

.

The AVHRR product examples in Figure 11 largely support the SEVIRI-results, which shows the consistency of the applied method. Existing differences came mainly from time differences, i.e., SEVIRI

results were shown at one given time while AVHRR results were compiled from three consecutive orbits spanning a time interval of more than 4 h.

In Table 5, we previously summarized the smallest detectable cloud optical thicknesses (*CDS* values) with at least 50% probability of detection for all investigated Earth surfaces. In Figure 12. we visualize these values geographically for the two particular SEVIRI cases in Figures 9 and 10. Here we can see that the best conditions (lowest *CDS* values) were found over ocean surfaces at high latitudes. More problematic conditions were found over dry desert areas, in sunglint conditions (Figure 12 top), and near Antarctica. Observe also that less favorable conditions existed along some coastlines (including near rivers and lakes). Corresponding results for AVHRR (Figure 11) are not shown here but they weeree very similar with only small departures due to temporal differences.

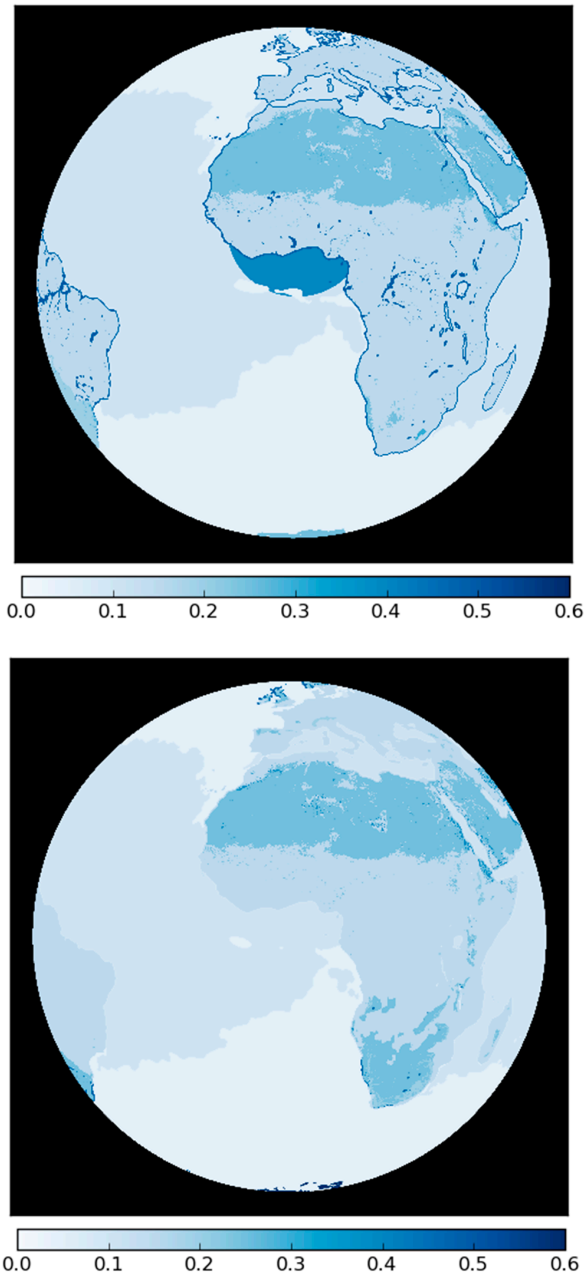

Figure 12. Visualization of the smallest detectable cloud optical thicknesses (*CDS* values) with at least 50% probability of detection for all investigated Earth surfaces. Results are valid for the SEVIRI case in Figure 9 (top) and Figure 10 (bottom).

3.3. Validation Results from Independent Data

In Tables 6–8 below, we present validation scores based on independent data, as outlined in Section 2.9. The scores were globally averaged quantities, as well as regionally (zonally) assessed quantities. Notice that for SEVIRI, global results mean the full SEVIRI disk. The primary validation score was the *HR* quantity (Equation (2)) with all CALIPSO-detected clouds included (i.e., with τ_{thr}= 0.0). Even though we know that scores would be improved if using a non-zero threshold (due to the concept used for training), the use of original unfiltered data for validation may still be considered as the best objective way of evaluating the overall performance. However, since it is also clear that the *HR* quantity may be distribution dependent (i.e., the most frequent category can bias the results), we also show results for the Kuipers' Skill Score (*KSS*), which is defined (using the same notations as in Equation (2)) by

$$KSS(\tau_{thr}) = \frac{ad - cb}{(a + b)(c + d)} \quad (6)$$

where $-1 \leq KSS \leq 1$.

Table 6. Hit Rate (*HR*) and Kuipers' Skill Score (*KSS*) as global and regional (zonal) averages for independent validation data for NOAA-18 and NOAA-19 satellites in 2010. Results are compared to results for the CLARA-A2 CDR [17].

Area/Region	*HR*	*KSS*	*HR CLARA-A2*	*KSS CLARA-A2*
Global	0.82	0.69	0.80	0.65
Tropical	0.78	0.67	0.79	0.63
Sub-Tropical	0.85	0.72	0.84	0.70
High-Latitude	0.85	0.71	0.83	0.68
Polar	0.69	0.50	0.67	0.47

Table 7. Hit Rate (*HR*) and Kuipers' Skill Score (*KSS*) as global and regional (zonal) averages for independent validation data for the Metop-A satellite in 2010. Results are compared to results for the CLARA-A2 CDR [17]. Notice that results at low and tropical latitudes are missing due to lacking matchups with CALIOP.

Area/Region	*HR*	*KSS*	*HR CLARA-A2*	*KSS CLARA-A2*
Global	0.75	0.53	0.73	0.56
Tropical	-	-	-	-
Sub-Tropical	-	-	-	-
High-Latitude	0.79	0.57	0.76	0.59
Polar	0.70	0.46	0.68	0.49

Table 8. Hit Rate (*HR*) and Kuipers' Skill Score (*KSS*) as full disk and regional (zonal) averages for independent validation data for the SEVIRI sensor in July 2012. Notice that results in the Polar area are missing due to lacking coverage from SEVIRI.

Area/Region	*HR*	*KSS*
Full SEVIRI disk	0.85	0.70
Tropical part of SEVIRI disk	0.79	0.65
Sub-Tropical part of SEVIRI disk	0.85	0.70
High-Latitude part of SEVIRI disk	0.90	0.71
Polar part of SEVIRI disk	-	-

The *KSS* score addresses the question of how well the estimation separates the cloudy events from the cloud-free events. A value of 1.0 in this respect describes the situation of a perfect discrimination, while the value −1.0 describes a complete discrimination failure. This score is less distribution dependent. It has been found to be very useful for identifying failures in cloud detection in situations

where cloudy cases are rare (e.g., over desert regions). A good rule is that a genuinely improved cloud screening method should show improvements in both quantities. If only the *HR* improves, a possible cause could be a different distribution of validation cases and not necessarily an improved cloud masking methodology.

The geographical differences in cloud screening capabilities (see Tables 2 and 5) motivate the addition of regionally summarized results. However, the results shown below are not compiled according to the regions specified in Table 2. These regions were not static since they followed varying thermal characteristics rather than fixed geographic coordinates. To avoid too many regional details but still be able to identify some regional variations, we show overall results for four geographically fixed zonal regions where each includes contributions from both hemispheres: Polar (above latitude 75°), High-latitude (latitude interval 45–75°), Sub-Tropical (latitude interval 10–45°), and Tropical (within latitude 10°, here representing only the innermost part of the tropical region). Finally, for the AVHRR-based results, we also compared them with corresponding validation results derived for the CM SAF cLoud, Albedo, and surface RAdiation dataset from AVHRR data–second edition (CLARA-A2) [17]. Even if the validation dataset is not the same (i.e., the CLARA-A2 validation dataset covers the full period 2006–2015 which could influence *HR* values to some extent), we believe that, at the very least, a comparison to their *KSS* values is justified.

According to Table 6, the developed cloud screening method produced improved results for AVHRR GAC data compared to earlier results presented in a comprehensive validation effort of the CLARA-A2 CDR [17]. There was a definite improvement for the *KSS* score, which is a measure that tests how well the separation of cloudy and clear cases works regardless of the true distribution of cloudy and clear cases. The corresponding improvements for the morning orbit results in Table 7 were evident in the *HR* scores but not necessarily in the *KSS* scores. This may be explained by the access to a limited (i.e., small in number of matchup samples) validation dataset over an area with very high surface variability (near 70° latitude).

Similar, or in some regions even further improved results, were achieved when applying the method on geostationary SEVIRI imagery (Table 8). These improvements may emerge due to relatively larger sample population of cases in tropical and sub-tropical regions, where cloud detection is easier than over high latitude land regions. However, the good performance may also partly be explained by the use of additional spectral information not available in the AVHRR case, e.g., the simultaneous use of 1.6 μm and 3.9 μm data, and the use of the 8.6 μm channel.

4. Discussion

In the following, we will discuss several strengths and limitations of the described cloud screening method as well as potential future development paths.

4.1. Flexibility of the Cloud Screening Method

A key improvement compared to traditional binary cloud mask realizations is that our derived cloud probability methodology may be used more flexibly since it can provide a clear-conservative or cloud-conservative masks after just adjusting the cloud probability threshold. In the current study, we used the neutral threshold of 50% cloud probability for the validation reported in Tables 6–8, which we claim is the most appropriate setting for climate monitoring applications.

In this context, it is also necessary to emphasize that the cloud probability product only considers the uncertainty in cloud detection when provided with a specific set of image feature values and ancillary data as input. Thus, the probabilities only represent the retrieval uncertainties assuming perfect input data. Additional work, probably requiring comprehensive Monte Carlo simulations to cover the full range of random and systematic errors [30,31], is required to also take into account uncertainties in input data.

4.2. Improved Quantification of Cloud Screening Results

The most significant improvement of the current cloud masking product is that the method offers a way of quantifying the exact meaning of cloud probabilities over different Earth surfaces. More clearly, the method is tuned so that the 50% cloud probability level over a particular Earth surface links to a specific minimum cloud optical thickness, i.e., the *CDS* value, and thus, we have a 50% chance of detecting a cloud with this specific minimum cloud optical thickness. For thicker clouds, probabilities will be higher than 50%. This specific cloud optical thickness will be different over different Earth surfaces. The different *CDS* values are in line with our experience that a cloud with a particular cloud optical thickness is much easier to detect over dark and warm surfaces (e.g., ocean surfaces) compared to over bright and cold surfaces (e.g., permanent snow).

The *CDS* value is estimated for various Earth surfaces in Table 5. We anticipate that this information can be utilized together with the primary satellite-derived cloud probability in various quantitative applications. An example here could be the use of *CDS* information in cloud dataset simulators for evaluating cloud simulations in climate models such as the Cloud Feedback Model Inter-comparison Project (CFMIP) Observation Simulator Package (COSP, [32–34]). The current simulators only have a simplified description of the cloud screening method limitations (i.e., normally based on a single globally fixed threshold of the detectable cloud optical thickness). A new candidate simulator for the CLARA-A2 CDR is demonstrated in [35], presenting a first attempt at treating cloud retrievals from the CLARA-A2 CDR differently over varying Earth surfaces. The aim is to use the cloud screening method described in this paper for the successor to CLARA-A2 to be named CLARA-A3. A tentative CLARA-A3 simulator could, in principle, make use of the *CDS* information directly without the need to perform extensive validation efforts after its release. The concept of the CLARA-A3 simulator would also allow the use of it in the future (i.e., beyond the CLARA-A3 observation period), provided that the temporal and spatial evolution of the specified Earth surface categories in Table 5 can be described. This is an important further development of the method presented in [35] since it does not depend on the continuous availability of high-quality reference measurements for estimating the *CDS* values.

4.3. The Impact of Spatial and Temporal Variations

Across many regions of Earth, cloud screening is still very challenging and, therefore, in these locations, very substantial improvements in results are hard to achieve. For this particular methodology, the primary cause of the encountered problems is the difficulty of finding reasonably homogeneous surface conditions.

Although we still claim that there must be a straightforward relationship between a cloud's optical thickness and the capability of detecting it over a specific surface type, the success of this particular method depends on whether the defined surface categories can be considered homogeneous (i.e., reasonably invariant) or not. The majority of Earth's surfaces fulfill this condition, especially over every oceanic surface and many land surfaces at low and middle latitudes, where the method works quite efficiently. The exceptions are found over high latitude land surfaces and in the Polar Regions. Here, the idealized model for optimal cloud detection (as outlined in Sections 2.4 and 2.5) does not seem to be fully applicable. Therefore, as a consequence, solution compromises were necessary (e.g., see discussion related to Figure 8 in Section 3.1).

Across certain regions, complications can be related to the fact that these surfaces vary considerably in both space and time regarding both visible and infrared characteristics. For example, the extra-tropical land categories G11-G12 in Table 2 cover a wide range of surface vegetation conditions and surface temperatures. Extreme surface temperature variations are typical for category G13 (permanent snow cover), which also complicates the retrieval conditions. Another example is category G10 (desert surfaces) for which surface parameters such as infrared surface emissivity adds to the variability. A finer subdivision of Earth surfaces (e.g., utilizing MODIS-based land cover type classifications) could improve the situation. However, this would also lead to smaller matchup samples and potentially

reduce the statistical robustness of the relationships. Future studies with access to larger reference datasets could address these limitations in an improved manner.

4.4. The Problem of Overfitting

The described methodology does not necessarily provide us the absolute best match with the referenced 'ground truth' from CALIPSO-CALIOP observations when comparing to the unfiltered CALIOP results (which was the validation concept used for the results shown in Tables 6–8). Other machine learning or statistical regression methods could potentially find even better agreement with CALIOP observations.

A more in-depth investigation of the results presented in Figures 4–8 in Section 3.1 also indicates this effect for the new cloud screening method. In some of these figures, the results when training with unfiltered CALIOP data (blue curve for τ = 0.0) yielded higher *HR* values for the unfiltered case (at *CALIPSO Filtered* τ = 0.0) than the selected results based on curves with optimal *CDS* values. These details are hard to see in the current figures (because of too many overlaid curves) but appear in a close-up of results in Figure 5 valid over tropical ocean surfaces at night. Table 9 lists achieved *HR* scores for this case when training our method using the five largest CALIOP cloud masks, i.e., original cloud masks being thresholded at optical thicknesses 0.0, 0.05, 0.10, 0.15, and 0.20. According to Figure 5, we get the highest *HR* values when validating with a restricted CALIOP cloud mask with τ_{thr} = 0.1 (which is then our selected *CDS* value). The three rightmost columns in Table 9 give corresponding *HR* scores when validating with restricted CALIOP cloud masks thresholded at optical thicknesses 0.0, 0.05, and 0.10. We recognize the maximum *HR* value of 0.874 for the peak seen in Figure 5 for τ_{thr} = 0.1. However, we notice that we actually fulfill the *CDS* criterion in Equation (5) also for τ_{thr} = 0.0 and τ_{thr} = 0.05. The reason for selecting *CDS* to be 0.10 is that resulting *HR* scores are higher here than for the two mentioned alternatives. Thus, for *CDS* set to 0.10, we have the best representation or reproduction of the CALIOP cloud mask in AVHRR imagery. We also claim that the *CDS* alternative 0.0 is theoretically problematic since it suggests that the unfiltered CALIOP cloud mask is better reproduced than any other filtered CALIOP cloud mask. This violates the generally accepted fact that the CALIOP measurements are more sensitive to thin clouds than AVHRR. We have rejected this possibility and interpreted it as an effect of statistical overfitting.

Table 9. Hit Rate (*HR*) scores for tropical ocean surfaces at night showing a close-up of results close to the Y-axis of Figure 5. Results are shown based on the five largest CALIOP cloud masks in the validation process (as defined by the leftmost column corresponding to the X axis in Figure 5). *HR* scores are presented for cases being trained on three different restricted CALIOP cloud masks, which were filtered using three different optical thickness thresholds (values 0,00, 0.05, and 0.10, corresponding to *HR* results of the first three curves in Figure 5).

τ_{thr} (CALIOP Mask) Used at Validation	*HR* ($\tau_{thr\_training}$ = 0.00)	*HR* ($\tau_{thr\_training}$ = 0.05)	*HR* ($\tau_{thr\_training}$ = 0.10)
0.00	0.846	0.863	0.857
0.05	0.830	0.866	0.872
0.10	0.815	0.861	0.874
0.15	0.801	0.853	0.872
0.20	0.789	0.845	0.864

The real world generally does not provide us with equal distributions of cloudy and clear cases. Instead, cloudy conditions dominate in many regions of the world. For instance, sub-visible and thin cirrus clouds dominate the tropical ocean surface category (see Figure 5). If cloudy conditions dominate for sub-visible (to AVHRR) clouds, AVHRR-based methods should no longer know the difference between the surface and the cloud. However, statistical methods would still tend to prefer the cloudy solution since this would maximize the validation scores used in the optimization process. Thus, in this way, the method could be overfitted in the sense that the method is now trying to describe, what

rightfully should be regarded as noise, since there is no true detection ability. The risk of overfitting is serious since, if it happens, it would thn give the impression that AVHRR can observe optically thinner clouds than what is actually possible. A misinterpretation here can lead to inconsistencies if results are used in downstream applications (e.g., in COSP simulators). Consequently, we have prioritized finding solutions where we can get as high validation scores as possible, while at the same time specifying for which clouds these results are indeed valid. We believe that this is the most sensible way to describe the capability of the AVHRR sensor for cloud screening purposes.

4.5. The Importance of Limited Temporal and Spatial Sampling of CALIOP Observations

The fact that the CALIPSO satellite only operates in an afternoon orbit leads to restricted sampling of the diurnal variation of global cloud conditions. More clearly, CALIOP observations only cover local times close to 01:30 p.m. and 01:30 a.m. in most places (except near the poles where a transition from night to day and vice versa occurs). Various ways of dealing with this has already been described in Section 2.9.

This restriction might appear quite serious if one has the ambition to apply the method to all possible illumination and thermal conditions. However, we claim that since we train over the whole year and actually over almost an entire 10-year period, we will still cover the majority of varying solar zenith angles and thermal conditions for every location on the Earth. This is an effect of the seasonal changes. Such a training dataset should describe average daytime and night-time conditions adequately over most places. Furthermore, the decision to also use matchups with satellites being subject to orbital drift (described in Section 2.8.3) has also helped in getting coverage of a wider range of existing solar zenith angles and solar azimuth angles. Thus, the strength of our method is that our training dataset covers a very long period capable of describing the seasonal variability.

4.6. Limitations of the Naïve Bayesian Method

A final limiting factor for the achieved results is also the basic Naïve Bayesian approach, which forms the basis for our statistical training. This simplification of the true Bayesian formulation is only valid if the image features utilized are uncorrelated. Unfortunately, many of the image features used show obvious signs of correlation. For example, for gradually thicker and colder clouds, we also find gradually higher reflectances and decreasing brightness temperatures. At the same time, other clouds (for example, thin cirrus clouds) do not show this strong feature covariation. Another problem is the vulnerability to features showing near-zero probabilities in the multiplication of individual probabilities in Equation (1). We have tried to decrease this vulnerability by introducing two-dimensional image features (see Table 4) at night, but this does not remove all problems related to this effect. Thus, in general, the Naïve Bayesian method produces reasonably good results, but some defects are unavoidable. Consequently, future developments should aim to introduce true Bayesian approaches (e.g., as suggested in [36,37]) or more advanced machine learning methods (e.g., as suggested in [10,38]). However, it is imperative to use the principles from this study to ensure that the different conditions for cloud detection over different Earth surfaces are taken into account for such approaches as well. Importantly, an effort to quantitatively describe, as precisely as possible, which clouds are truly detectable must be made, as well as performing an adequate uncertainty analysis.

5. Conclusions

A new concept for cloud screening in AVHRR and SEVIRI imagery has been described based on the Naïve Bayesian theory and utilization of 'ground truth' data from the CALIPSO-CALIOP cloud lidar in the period 2006–2015. The purpose has been to construct a method providing a balanced or neutral cloud screening that simultaneously minimizes the false clear and false cloudy occurrences. Such a method would better serve climate monitoring applications than many of the methods presently used, which often overestimate cloudiness due to the use of clear-conservative approaches.

The results of the method have been evaluated using independent CALIPSO-CALIOP cloud observations (with all clouds observed) for one year (2010) with validation results expressed as *HR* and *KSS* scores. A global *HR* value of 0.82 was found for AVHRR data while the corresponding value for SEVIRI, over the SEVIRI full disk, is 0.85. A closer look at validation results showed that the best performance for AVHRR was found over ice-free oceanic regions at high latitudes with *HR* scores of 0.93 and *KSS* scores of 0.72; similar *HR* results were found in the same region for SEVIRI data. Due to mainly poor performance during the polar night, the algorithm achieved less-satisfying results in the Polar Regions with *HR* scores of 0.69 and *KSS* scores of 0.50.

Results were also compared to similar validation results for the CLARA-A2 CDR, which was previously evaluated thoroughly in [17]. General improvements were found, especially for the *KSS* score. The improvement of the *KSS* score confirms that, besides seeing an overall improvement of scores, an improved balance between falsely clear and falsely cloudy cases has also been achieved. This aspect was one of the primary goals of this study. Modest improvements were found in the Polar Regions underlining the problem with very challenging conditions here, especially during the Polar night. Nevertheless, the introduction of probabilistic cloud mask information must be seen as a step forward also for the Polar Regions compared to the previous dominant use of binary cloud mask products. The probabilistic information provides a much better view of the uncertainties of the results in this problematic region.

Finally, a major additional achievement of this study is the provision of fundamental information about the cloud detection capabilities in AVHRR and SEVIRI data over varying Earth surfaces. Table 5 describes the estimated minimum cloud optical thickness of a cloud layer, which can be detected with at least 50% probability by the present method. Figure 12 visualizes how this may look for individual SEVIRI scenes. We argue that this information is essential for potential applications making use of these cloud screening results. More clearly, due to the concept of this method, the derived cloud probabilities are only relevant for clouds having a minimum total optical thickness as listed in Table 5 over various surfaces. Such information is crucial for quantitative use in applications, such as COSP simulators of satellite-based cloud climate data records. In our view, a probabilistic cloud mask lacking this additional information is problematic to use efficiently in various applications since it leaves the open question: For which clouds is the probability actually valid?

The described cloud screening method will be used in the preparation of the next editions of the CLARA (AVHRR-based) and CLAAS (SEVIRI-based) CDRs. The AVHRR-based CDR will be named CLARA-A3 with a tentative release by the EUMETSAT CM SAF project in the 2021–2022 timeframe. The SEVIRI-based CDR will be named CLAAS-3 (CLoud property dAtAset using SEVIRI, Edition 3), and the planned release is foreseen in 2021. The method will also be added (as an updated product named CMAPROB) to the PPS cloud processing software for polar orbiting satellite data provided by the EUMETSAT NWC SAF project.

Author Contributions: Conceptualization, K.-G.K.; methodology, K.-G.K.; software, K.-G.K., E.J., J.S., and N.H.; validation, K.-G.K. and N.H.; formal analysis, K.-G.K., N.H., J.S., and S.E.; investigation, K.-G.K.; resources, K.-G.K.; data curation, K.-G.K.; writing—original draft preparation, K.-G.K.; writing—review and editing, K.-G.K.; visualization, E.J., J.S., and S.E. supervision, K.-G.K.; project administration, K.-G.K.; funding acquisition, K.-G.K.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by EUMETSAT, through the Climate Monitoring SAF project in cooperation with the national meteorological institutes of Germany, Sweden, Finland, the Netherlands, Belgium, Switzerland, France and the United Kingdom.

Acknowledgments: CALIPSO-CALIOP datasets were obtained from the NASA Langley Research Center's Atmospheric Science Data Center Surface Meteorological and Solar Energy (SSE) web portal, supported by the NASA LaRC POWER Project.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Dybbroe, A.; Thoss, A.; Karlsson, K.-G. NWCSAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modelling—Part I: Algorithm description. *J. Appl. Meteorol.* **2005**, *44*, 39–54. [CrossRef]
2. Karlsson, K.-G.; Riihelä, A.; Müller, R.; Meirink, J.F.; Sedlar, J.; Stengel, M.; Lockhoff, M.; Trentmann, J.; Kaspar, F.; Hollmann, R.; et al. CLARA-A1: A cloud, albedo, and radiation dataset from 28 yr of global AVHRR data. *Atmos. Chem. Phys.* **2013**, *13*, 5351–5367. [CrossRef]
3. Karlsson, K.-G.; Anttila, K.; Trentmann, J.; Stengel, M.; Meirink, J.F.; Devasthale, A.; Hanschmann, T.; Kothe, S.; Jääskeläinen, E.; Sedlar, J.; et al. CLARA-A2: The second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data. *Atmos. Chem. Phys.* **2017**, *17*, 5809–5828. [CrossRef]
4. Derrien, M.; LeGleau, H. MSG SEVIRI cloud mask and type from SAFNWC. *Int. J. Remote Sens.* **2011**, *26*, 4707–4732. [CrossRef]
5. Benas, N.; Finkensieper, S.; Stengel, M.; van Zadelhoff, G.-J.; Hanschmann, T.; Hollmann, R.; Meirink, J.F. The MSG-SEVIRI-based cloud property data record CLAAS-2. *Earth Syst. Sci. Data* **2017**, *9*, 415–434. [CrossRef]
6. Heidinger, A.K.; Evan, A.T.; Foster, M.J.; Walther, A. A Naïve Bayesian cloud-detection scheme derived from CALIPSO and applied within PATMOS-x. *J. Appl. Meteor. Climatol.* **2012**, *51*, 1129–1144. [CrossRef]
7. Heidinger, A.K.; Foster, M.J.; Walther, A.; Zhao, Z. The Pathfinder Atmospheres Extended (PATMOS-x) AVHRR climate data set. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 909–922. [CrossRef]
8. Heidinger, A.; Botambekov, D.; Walther, A. A Naïve Bayesian Cloud Mask delivered to NOAA Enterprise, Algorithm Theoretical Basis Document, NOAA NESDIS Center for Satellite Applications and Research, Version 1.2. 2016. Available online: https://cimss.ssec.wisc.edu/patmosx/documents/Cloud_Mask_Enterprise_ATBD_v1.2_2016.pdf (accessed on 20 February 2020).
9. Frey, R.A.; Ackerman, S.A.; Liu, Y.; Strabala, K.I.; Zhang, H.; Key, J.R.; Wang, X. Cloud Detection with MODIS. Part I: Improvements in the MODIS Cloud Mask for Collection 5. *J. Atmos. Ocean. Technol.* **2008**, *25*, 1057–1072. [CrossRef]
10. Stengel, M.; Stapelberg, S.; Sus, O.; Finkensieper, S.; Würzler, B.; Philipp, D.; Hollmann, R.; Poulsen, C.; Christensen, M.; McGarragh, G. Cloud_cci Advanced Very High Resolution Radiometer post meridiem (AVHRR-PM) dataset version 3: 35 year climatology of global cloud and radiation properties. *Earth Syst. Sci. Data* **2020**, *12*, 41–60. [CrossRef]
11. Barale, V.; Gower, J.F.R.; Alberotanza, L. (Eds.) *Oceanography from Space*; Springer Science & Business Media B.V.: Berlin/Heidelberg, Germany, 2010; pp. 213–216. ISBN 978-90-481-8680-8. [CrossRef]
12. Yang, Y.; Zhao, C.; Sun, L.; Wei, J. Improved Aerosol Retrievals Over Complex Regions Using NPP Visible Infrared Imaging Radiometer Suite Observations. *Earth Space Sci.* **2019**, *6*, 629–645. [CrossRef]
13. Tzallas, V.; Hatzianastassiou, N.; Benas, N.; Meirink, J.F.; Matsoukas, C.; Stackhouse, P.; Vardavas, I. Evaluation of CLARA-A2 and ISCCP-H Cloud Cover Climate Data Records over Europe with ECA&D Ground-Based Measurements. *Remote Sens.* **2019**, *11*, 212. [CrossRef]
14. Wang, Y.; Zhao, C. Can MODIS cloud fraction fully represent the diurnal and seasonal variations at DOE ARM SGP and Manus sites? *J. Geoph. Res. Atmos.* **2016**, *122*, 329–343. [CrossRef]
15. Karlsson, K.-G.; Johansson, E.; Devasthale, A. Advancing the uncertainty characterization of cloud masking in passive satellite imagery: Probabilistic formulations for NOAA AVHRR data. *Remote Sens. Environ.* **2015**, *158*, 126–139. [CrossRef]
16. Karlsson, K.-G.; Johansson, E. 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.* **2013**, *6*, 1271–1286. [CrossRef]
17. Karlsson, K.-G.; Håkansson, N. 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.* **2018**, *11*, 633–649. [CrossRef]
18. Winker, D.M.; Vaughan, M.A.; Omar, A.; Hu, Y.; Powell, K.A. Overview of the CALIPSO mission and CALIOP data processing algorithms. *J. Atmos. Ocean. Technol.* **2009**, *26*, 2310–2323. [CrossRef]
19. Eidenshink, J.; Faundeen, J. The 1 km AVHRR global land data set-first stages in implementation. *Int. J. Remote Sens.* **1994**, *15*, 3443–3462. [CrossRef]

20. Lavergne, T.; Sørensen, A.M.; Kern, S.; Tonboe, R.; Notz, D.; Aaboe, S.; Bell, L.; Dybkjær, G.; Eastwood, S.; Gabarro, C.; et al. Version 2 of the EUMETSAT OSI SAF and ESA CCI sea-ice concentration climate data records. *Cryosphere* **2019**, *13*, 49–78. [CrossRef]
21. Copernicus Climate Change Service (C3S). ERA5: Fifth Generation of ECMWF Atmospheric Reanalyses of the Global Climate. Copernicus Climate Change Service Climate Data Store (*CDS*). 2017. Available online: https://cds.climate.copernicus.eu/cdsapp#!/home (accessed on 20 February 2020).
22. Phong, B.T. Illumination for computer generated pictures. *Commun. ACM* **1975**, *18*, 311–317. [CrossRef]
23. Derrien, M.; LeGleau, H. Improvement of cloud detection near sunrise and sunset by temporal-differencing and region-growing techniques with real-time SEVIRI. *Int. J. Remote Sens.* **2010**, *31*, 1765–1780. [CrossRef]
24. Musial, J.P.; Hüsler, F.; Sütterlin, M.; Neuhaus, C.; Wunderle, S. Probabilistic approach to cloud and snow detection on Advanced Very High Resolution Radiometer (AVHRR) imagery. *Atmos. Meas. Technol.* **2014**, *7*, 799–822. [CrossRef]
25. NWC SAF 1. Algorithm Theoretical Basis Document for Cloud Mask Probability of the NWC/PPS, NWC/CDOP3/PPS/SMHI/SCI/ATBD/CloudProbability, v 1.0. 2018. Available online: http://www.nwcsaf.org/web/guest/scientificdocumentation#NWCSAF/PPS%20Basic%20Documents (accessed on 20 February 2020).
26. NWC SAF 2. Science and Validation Report for the Cloud Product Processors of the NWC/PPS, NWC/CDOP3/PPS/SMHI/SCI/VR/Cloud, v 2.0. 2018. Available online: http://www.nwcsaf.org/web/guest/scientificdocumentation#NWCSAF/PPS%20Basic%20Documents (accessed on 20 February 2020).
27. Clark, R.N. Chapter 1: Spectroscopy of Rocks and Minerals, and Principles of Spectroscopy, in Manual of Remote Sensing. In *Remote Sensing for the Earth Sciences*; Rencz, A.N., Ed.; John Wiley and Sons: New York, NY, USA, 1999; Volume 3, pp. 3–58.
28. Winker, D. *CALIPSO LID L2 5km Standard HDF File—Version 4.10 [Data set]*; NASA Langley Research Center Atmospheric Science Data Center DAAC: Hampton, VA, USA, 2016. [CrossRef]
29. NWC SAF. *Algorithm Theoretical Basis Document for the Cloud Mask of the NWC/PPS, NWC/CDOP3/PPS/SCI/ATBD/CloudMask*; EUMETSAT, Version 2.1; NWC SAF: Canberra, Australia, 2018.
30. Merchant, C.J.; Paul, F.; Popp, T.; Ablain, M.; Bontemps, S.; Defourny, P.; Hollmann, R.; Lavergne, T.; Laeng, A.; de Leeuw, G.; et al. Uncertainty information in climate data records from Earth observation. *Earth Syst. Sci. Data* **2017**, *9*, 511–527. [CrossRef]
31. Merchant, C.J.; Holl, G.; Mittaz, J.; Woolliams, E. Radiance Uncertainty Characterisation to Facilitate Climate Data Record Creation. *Remote Sens.* **2019**, *11*, 474. [CrossRef]
32. Bodas-Salcedo, A.; Webb, M.J.; Bony, S.; Chepfer, H.; Dufresne, J.-L.; Klein, S.A.; Zhang, Y.; Marchand, R.; Haynes, J.M.; Pincus, R.; et al. COSP: Satellite simulation software for model assessment. *Bull. Am. Met. Soc.* **2011**, *92*, 1023–1043. [CrossRef]
33. Pincus, R.; Platnick, S.; Ackerman, S.A.; Hemler, R.S.; Hofmann, R.J.P. Reconciling Simulated and Observed Views of Clouds: MODIS, ISCCP, and the Limits of Instrument Simulators. *J. Clim.* **2012**, *25*, 4699–4720. [CrossRef]
34. Swales, D.J.; Pincus, R.; Bodas-Salcedo, A. The Cloud Feedback Model Intercomparison Project Observational Simulator Package: Version 2. *Geosci. Model Dev.* **2018**, *11*, 77–81. [CrossRef]
35. Eliasson, S.; Karlsson, K.-G.; Willén, U. A simulator for the CLARA-A2 cloud climate data record and its application to assess EC-Earth polar cloudiness. *Geosci. Model. Dev.* **2019**, *13*, 297–314. [CrossRef]
36. Musial, J.; Karlsson, K.-G. VEOR probabilistic cloud mask—A prototype for the generation of the CM SAF climate data records. In Proceedings of the 5th User Workshop Satellite Application Facility on Climate Monitoring, Mainz, Germany, 3–5 June 2019.
37. Musiał, J.; Bojanowski, J. AVHRR LAC satellite cloud climatology over Central Europe derived by the Vectorized Earth Observation Retrieval (VEOR) method and PyLAC software. *Geoinf. Issues* **2017**, *9*, 39–51.
38. Gomis-Cebolla, J.; Jimenez, J.C.; Sobrino, J.A. MODIS probabilistic cloud masking over the Amazonian evergreen tropical forests: A comparison of machine learning-based methods. *Int. J. Remote Sens.* **2019**. [CrossRef]