Remote Sensing of Environment 158 (2015) 126–139

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Advancing the uncertainty characterisation of cloud masking in passive satellite imagery: Probabilistic formulations for NOAA AVHRR data

K.-G. Karlsson *, E. Johansson, A. Devasthale

Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden

ARTICLE INFO

Article history:
Received 23 May 2014
Received in revised form 24 October 2014
Accepted 31 October 2014
Available online 2 December 2014

Keywords:
Satellite
AVHRR
Cloud masking
Cloud probabilities
CALIPSO

ABSTRACT

Two alternative methods for probabilistic cloud masking of images from the Advanced Very High Resolution Radiometer (AVHRR) sensor have been examined. Both methods are based on Bayesian theory and were trained using data from the Cloud–Aerosol Lidar with Orthogonal Polarization (CALIOP) lidar onboard the Cloud–Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) satellite. Results were evaluated by comparing to independent CALIPSO–CALIOP observations and to a one-year ground-based cloud dataset composed from five different remote sensing systems over the observation site in Cabauw in the Netherlands. In addition, results were compared to two different cloud masks; one derived from the geostationary Spinning Enhanced Visible and Infrared Imager (SEVIRI) sensor and one from the Climate Monitoring Satellite Application Facility Clouds (CMSAF), Albedo and Radiation dataset from AVHRR data (CLARA-A1). It was demonstrated that the probabilistic methods compare well with the referenced satellite datasets and for daytime conditions they provide even better performance than the reference methods. Among the two probabilistic approaches, it was found that the formulation based on a Naïve Bayesian formulation (denoted PPS-Prob Naïve) performed clearly superior to the formulation based on a linear summation of conditional cloud probabilities (denoted PPS-Prob SPARC) for daytime conditions. For the study based on the observations over the Cabauw site, the overall daytime Kuipers Skill Score for PPS-Prob Naïve was 0.84, for PPS-Prob SPARC 0.79, for CLARA-A1 0.74 and for SEVIRI 0.66. Corresponding results for night-time conditions were less favourable for the probabilistic formulations (Kuipers Skill Score 0.74 for PPS_Prob Naïve, 0.68 for PPS-Prob SPARC, 0.80 for CLARA-A1 and 0.79 for SEVIRI) but still relatively close to the reference dataset. The Cabauw distribution of cloudiness occurrences in different octa categories was reproduced very closely by all methods, including the probabilistic formulations. Results based on Cabauw observations were also largely in good agreement with results deduced from comparisons with the CALIPSO–CALIOP cloud mask.

The PPS-Prob Naïve approach will be implemented in an upcoming version of the Polar Platform System (PPS) cloud software issued by the EUMETSAT Nowcasting Satellite Application Facility (NWC SAF). It will also be used in the second release of the CMSAF CLARA cloud climate data record based on historic AVHRR GAC data (to be denoted CLARA-A2).

1. Introduction

Satellite data have become an indispensable part of the observational system for global environmental and climate monitoring. This concerns the monitoring of Earth surfaces as well as the coupled Earth–Atmosphere system (Aschbacher & Milagro-Perez, 2012; Barret & Curtis, 2013; Powell, Qu, & Sivakumar, 2013). The availability of persistent satellite sensors or sensor families, with observational records now reaching or even exceeding 30 years, is also gradually increasing the role and importance of satellite-based data records in climate change studies. Several individual datasets have now evolved from being just one of several series of available Environmental Data Records (EDR) to become true Climate Data Records (CDR). The latter category is based upon carefully homogenised and inter-calibrated radiances (such as those described by e.g. Chen, Cao, & Menzel, 2013; Heidinger, Straka, Molling, Sullivan, & Wu, 2010). Some examples of satellite-derived CDRs with data coverage over three decades are described by Foster & Heidinger, 2013; Karlsson et al., 2013; Peng, Meier, Scott, & Savoie, 2013; Tucker et al., 2005.

Subsequent steps in improving and extending the usefulness of these datasets deal not only with adding more years of measurements (e.g. as discussed by Loew, 2013) and improving the basic retrieval algorithms but also with a better characterisation of uncertainty and

* Corresponding author at: SMHI, Folkborgsvagen 17, 60176 Norrköping, Sweden. Tel.: +46 114958407.
E-mail addresses: karl-goran.karlsson@smhi.se (K.-G. Karlsson), erik.johansson@smhi.se (E. Johansson), abhay.devasthale@smhi.se (A. Devasthale).

http://dx.doi.org/10.1016/j.rse.2014.10.028

stability. This paper addresses the question whether it is possible to better characterise cloud occurrence in passive satellite imagery. At least concerning datasets based on short-wave (visible, VIS) and long-wave (infrared, IR) spectral radiances, the dependence on a correct cloud screening is fundamental for the quality of the defined dataset. This holds true for datasets consisting exclusively of parameters related to the Earth surface (where presence of even small amounts of clouds prohibits an accurate estimation) as well as for datasets describing cloudiness and other atmospheric parameters. Up until now the use of fixed cloud masks or cloud masks with a small set of quality flags has been the most common way of dealing with the cloud occurrence problem. Here, we will examine two different formulations of a probabilistic cloud mask and compare results to the results of common cloud masks and to independent cloud observations (both space- and ground-based).

This study builds upon long experience of using methods for retrieving cloudiness and cloud properties from Advanced Very High Resolution (AVHRR) imagery from the polar orbiting NOAA and Metop satellites. These previous methods have been comprehensively described by Musial et al. (2014), and will not be repeated here. Instead, in Section 2 we first elaborately discuss some general principles of cloud screening and basic probabilistic theory followed by a description of two different methods for estimating the probability of cloud occurrence in multispectral AVHRR imagery. Both methods are based on Bayesian probabilistic theory but they utilise different approaches for simplifying the complexity implied if intending to use the full set of available image feature information characterising the complete multispectral and spatial context. This is done in order to allow the method to be used in massive processing applications, e.g., in climate data reprocessing events. In Section 3 we then outline the approach to evaluate the two methods and we explain how results are verified against both space-based and ground-based observations. Results are then presented in Section 4 followed by a summary and some concluding remarks in Section 5.

2. Methodology—introduction of two probabilistic cloud masking formulations for AVHRR cloud screening

2.1. General AVHRR cloud screening aspects

This study will focus on cloud screening methods applied to the NOAA AVHRR sensor (Cracknell, 1997). However, the principles and methods could also be applied in a modified form to any other multispectral sensor. The AVHRR sensor measures in six spectral channels according to Table 1 which also shows the sensor evolution in time (for all satellites) since the launch on Tiros-N in 1978 until present time (2014).

Basically, the first three channels measure exclusively reflected solar radiation (visible or VIS), although channel 2 is often denoted near-infrared (NIR) and channel 3A short-wave infrared (SWIR). Channel 3B (denoted medium-wave infrared—MWIR) is situated in an overlapping zone measuring both reflected solar radiation and emitted thermal radiation. Channels 4 and 5 measure exclusively long-wave infrared radiation, i.e. thermal radiation emitted from Earth surfaces and from clouds, aerosols and some gases (primarily water vapour) in the atmosphere. The latter two channels are often called split-window IR channels since they sub-divide the atmospheric window region between 10 and 13 μm in two parts with slightly different characteristics regarding the absorption/emission from atmospheric water vapour and thin ice clouds.

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

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

Standard cloud screening methods utilise essentially five basic properties that separate the behaviour of clouds and Earth surfaces in passive satellite imagery:

1. Clouds appear bright (i.e., having high Top of Atmosphere (TOA) reflectance) in VIS and NIR channels as opposed to ice-free water surfaces and vegetation-covered Earth surfaces
2. Clouds consisting of liquid cloud particles (not ice crystals) reflect strongly in SWIR and MWIR channels while Earth surfaces (including snow and ice) appear dark
3. Clouds are generally colder than Earth surfaces meaning that they appear bright in IR channel imagery (if displayed in inverted form, i.e. with low radiances appearing bright and high radiances appearing dark)
4. Thin Cirrus clouds have a higher transmissivity in IR channel 4 than in channel 5 which enables Cirrus detection if using the split-window IR brightness temperature difference
5. Broken clouds give rise to a scattered pattern or texture in images over otherwise homogeneous surfaces (especially ice-free ocean)

It can also be noticed that the second property leads to the ability to detect low-level water clouds at night since these clouds are not behaving as perfectly emitting black-bodies (as most other clouds and Earth surfaces) due to their reflective behaviour in the MWIR channel.

To perform the cloud screening, most methods define thresholds in the analysed spectral channels or channel combinations. These thresholds may be static (empirically or climatologically derived) but most methods pre-calculate them dynamically by use of radiation transfer models (RTM) with calculations initialised with various ancillary data (e.g., satellite viewing and solar geometry information and prescribed surface temperatures and atmospheric profile data from numerical weather prediction models). Good examples of such AVHRR-based methods are given by Dybbroe, Karlsson, and Thoss (2005a,b) and Kriebel, Gesell, Kästner, and Mannstein (2003).

2.2. Treatment of cloud masking uncertainty—limitation of traditional approaches and recent progress

The success of cloud screening methods varies strongly depending on e.g. illumination conditions and the state of the surface and the atmosphere. Problems are most evident for clouds over bright surfaces (e.g. desert or snow cover), during night when no reflected sunlight is available and in cold winter situations when the surface and the lower troposphere are often colder than clouds. The frequent failure of cloud masking in certain situations has resulted in the definition of quality flags, e.g., as described by Dybbroe et al. (2005a). However, since this is basically still based on thresholding methods no continuous measure of the cloud masking uncertainty is provided. Some more advanced quality measures have been introduced for other related sensors by Platnick et al. (2003) for the Moderate Resolution Imaging Spectroradiometer (MODIS), by Hutchison et al. (2005) and recently by Kopp et al. (2014) for the Visible Infrared Imaging Radiometer Suite (VIIRS). However, despite the probabilistic approach in the definition of this uncertainty information the products are in the end still limited to provide only a set of discrete quality flags.

One reason for the limited progress in this field, despite the existence of a well-defined probabilistic theory (see next section), has been the problem to find an appropriate cloud observation reference. In other words, there is an urgent need of something that may represent the truth since we cannot yet model cloud occurrence accurately with invertible retrieval theory (like what may be done for other cloud

properties, e.g., cloud top height). Furthermore, the observation reference needs to have characteristics which are not too far away (e.g., in terms of viewing geometry and with the ability to provide global observations) from the conditions prevailing for a scanning radiometer on a space platform.

However, with the 2006 launch of the Cloud–Aerosol Lidar with Orthogonal Polarization (CALIOP) onboard the Cloud–Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) satellite, the situation has improved considerably. CALIOP offers global cloud observations with higher detection sensitivity than any other passive instrument (Winker et al., 2009). Furthermore, observations can be matched simultaneously in time (however, restricted to certain conditions) to observations by current operational AVHRR sensors. This has triggered numerous studies examining AVHRR-based cloud detection methods in detail (e.g., Karlsson & Dybbroe, 2010; Karlsson & Johansson, 2013; Stengel et al., 2013). It has also paved the way for more systematic attempts to provide cloud probabilities rather than fixed cloud masks as the final result of cloud screening (Heidinger, Evan, Foster, & Walther, 2012; Musial, Hüsler, Sütterlin, Neuhaus, & Wunderle, 2014).

2.3. Probabilistic approaches based on Bayesian theory

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

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

Despite its simple form, the solution of Eq. (1) is not easy to find in a situation with multispectral measurements (i.e., when the dimension of $\boldsymbol{F}$ is large). The estimation of parameters in the right hand side of Eq. (1) (especially $P(\boldsymbol{F}|cloudy)$) becomes increasingly difficult the more image features that are chosen. It then requires extraction of very large statistical training datasets to really describe the dependence on individual image features and, in addition, also the effect of their mutual correlation. What complicates things even further is that, even with one specific realisation of feature vector $\boldsymbol{F}$, probabilities may differ depending on different environmental situations (e.g. if the pixel measurement is made in winter or in summer, over land or over ocean, in mountainous terrain or over desert, etc.). Thus, the training process needs to take into account additional ancillary information for a correct description of environmental conditions. To reduce complexity of the problem some approximations may be utilised. One way to go could be the entirely empirical approach of estimating $P(cloudy|\boldsymbol{F})$ directly from predefined Lookup Tables composed during training with some stratification based on ancillary data. Such a method has been demonstrated by Musial et al. (2014). Alternatively, some simplifications and approximations can be made to Eq. (1). We will in the following Sections 2.6 and 2.7 describe two such methods whose results will be examined in detail in this paper.

2.4. Conditional probabilities for individual AVHRR image features

Fundamental to the methods under scrutiny is that one can estimate the conditional probability that it is cloudy given an individual feature f_i value ($P(cloudy|f_i)$), for a set of N image features (i.e., $i = 1…N$). Fortunately, these probabilities are relatively straight-forward to estimate if e.g., matching AVHRR measurements to CALIPSO–CALIOP cloud products. Let's demonstrate this for two image features with statistics based on 99 optimally matched CALIPSO/NOAA-18 orbits in the time period 2006–2009 (i.e., the same dataset as being used by Karlsson & Johansson, 2013). The optimal matching means that the observations at the simultaneous nadir observation (SNO) point (i.e., where the orbital tracks of the two satellites cross) do not differ in time more than 15 s. Since both satellites are afternoon satellites and in approximately the same orbital planes (only the altitude differs), it means that nowhere over the entire global orbit does the observation time difference exceed 3 min and the NOAA-18 satellite stays close to nadir over the full matched orbit (see Fig. 4 in Section 3 which summarises matching conditions).

The chosen image features for demonstration are two of the most commonly used image features by AVHRR cloud-screening methods: The visible reflectance of AVHRR channel 1 (here denoted Rvis) and the temperature difference between AVHRR channel 4 (brightness temperature) and a prescribed surface skin temperature (here denoted Tirdiff).

Fig. 1 shows the estimated cloud probabilities for the AVHRR VIS feature over low-latitude ocean surfaces and over high-latitude snow-covered mountain surfaces. The distinction between low- and high latitudes is made at ±45° latitude and mountainous terrain is

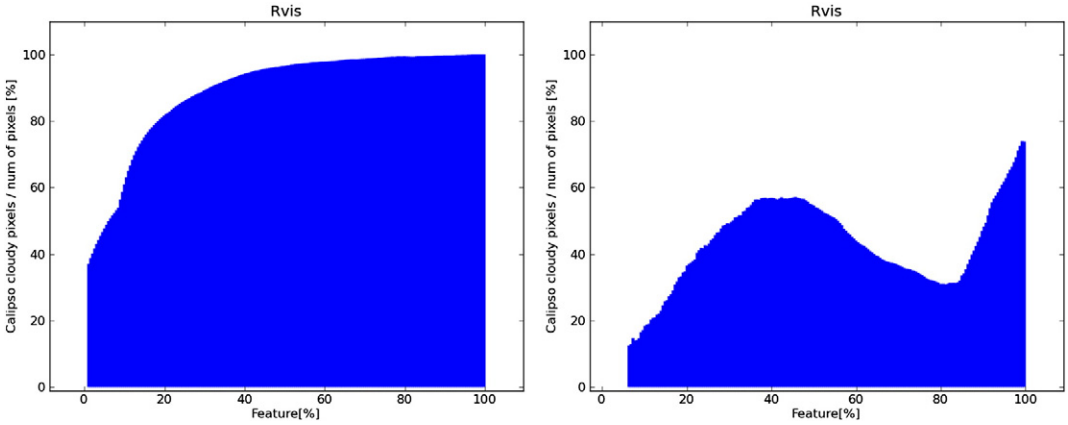

Fig. 1. Cloud probabilities estimated from CALIPSO–CALIOP cloud data in the period 2006–2009 as a function of AVHRR visible reflectances over Low Latitude ocean surfaces (left—defined in text) and over High Latitude mountain areas (right—defined in text).

defined as areas with topography above 500 m. Information on snow-cover is taken from the National Snow and Ice Data Center (NSIDC) dataset provided with the CALIPSO–CALIOP cloud product.

From Fig. 1 we conclude that cloud probabilities increase rapidly with reflectance over a very dark surface such as the ice-free ocean surface. Probabilities exceed 50% already at a very low reflectance value (at approximately 6% reflectance) and reach the 80% level at approximately 18% reflectance. Thus, conditions for cloud-screening appear almost ideal. This is not the case for the second situation in Fig. 1 (right) showing conditions over snow-covered ground in mountainous regions. Here, we hardly see any reflectance value where cloud probability exceeds 50% (which would be needed for this image feature to be useful for cloud screening purposes). This occurs only for moderately high reflectances close to 40% and for very high reflectances (approaching 100%). For the inter-mediate region of high reflectances probabilities are actually rather low which mainly is explained by the effect of non-isotropic reflection at very high solar zenith angles caused by illuminated snow-covered mountain sides.

A similar situation is seen over the same Earth surfaces in Fig. 2 for the infrared brightness temperature difference with regard to the surface skin temperature. Very good separability conditions are seen over low latitude ocean surfaces while they are very problematic over mountainous terrain. Notice in particular the effect of near-surface temperature inversions over mountainous terrain leading to a specific peak in cloud probability (although just slightly exceeding 50%) for negative values of the temperature difference (i.e., showing that clouds may then be frequently warmer than the surface temperature).

We conclude from Figs. 1 and 2 that conditions for efficient cloud screening may be drastically different depending on the geographic location and the prevailing illumination conditions (i.e., if it is day or night). This is one of the explanations for the very successful performance of simple bi-spectral VIS–IR cloud screening methods at low- to moderate latitudes (best exemplified by the results derived mainly from geostationary satellite data of the International Satellite Cloud Climatological Project—ISCCP—see Rossow et al., 1999). On the other hand, it also clearly illustrates the serious limitations for the same methods at high latitudes and over the Polar Regions.

2.5. Definition of a basic sub-set of constrained AVHRR image features

The probabilistic methods to be outlined more in detail in Sections 2.6 and 2.7 will utilise estimated conditional cloud probabilities (introduced in the previous section) for a sub-set of image features. These features are closely related to the five main cloud identification principles outlined previously in Section 2.1. However, rather than to define them in their purest form (as illustrated in Figs. 1 and 2) we have here chosen to define them linked to pre-calculated dynamic image feature thresholds used by one particular cloud screening method—the Polar Platform System cloud software package (PPS, see Dybbroe et al., 2005a,b). This software was developed by the Nowcasting Satellite Application Facility (NWC SAF) project which is organised by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT). The reason for linking image features to pre-calculated thresholds is that the latter have been defined in a way that takes a wide range of environmental conditions into account. This concerns image feature variability due to the following factors: Solar and satellite geometry (direct angular dependence and dependence on scattering angles), prevailing atmospheric profiles of temperature and humidity, climatological ozone and aerosol amounts, topography and land cover and spectral surface emissivities. If not taking all these factors into account when training the probabilistic classifier, results would risk being imprecise and most likely misleading under certain conditions or at certain geographic locations. We claim that it is better to piggy-back ride on existing prepared threshold information, composed from knowledge built over many years of experience of cloud thresholding, than to try to train a classifier from scratch with a need to create very large dimension Look-up Tables of statistical relations of cloudiness and image features and their respective dependencies on a wide range of environmental factors.

To illustrate the usefulness of this concept we consider one of the most commonly used AVHRR image features for detecting thin cirrus clouds (originally suggested by Inoue, 1987): The brightness temperature difference between AVHRR channels 4 and 5. The main principle used for Cirrus detection is that the cloud transmissivity for thin ice clouds is higher in AVHRR channel 4 than in AVHRR channel 5, thus creating a positive brightness temperature difference between AVHRR channels 4 and 5. Fig. 3 shows cloud probabilities as a function of this temperature difference but also as a function of the temperature difference relative to the corresponding PPS threshold.

We notice that in its original form (left panel of Fig. 3) we have two peaks in cloud occurrence where one is for differences close to zero K and the other for values exceeding approximately 4 K. The area between the peaks thus spans an interval of almost 4 K where cloud probabilities to a large extent are lower than 50%. In the alternative formulation (Fig. 3, right panel) results are much more distinctly organised and the range of probability values have been enlarged (which is favourable for the probabilistic classification process). The latter circumstance is especially true for the leftmost part of the distribution. We may interpret this as primarily an effect of being able to take into account the natural cloud-free contribution from atmospheric water vapour emission in the split-window channels. This emission is also able to create a discernible temperature difference in the absence of cirrus clouds explaining the broader and less decisive probability distribution in its original form (Fig. 3, left panel) for temperature differences below approximately 4 K. Resulting distributions after the coordinate transformation now

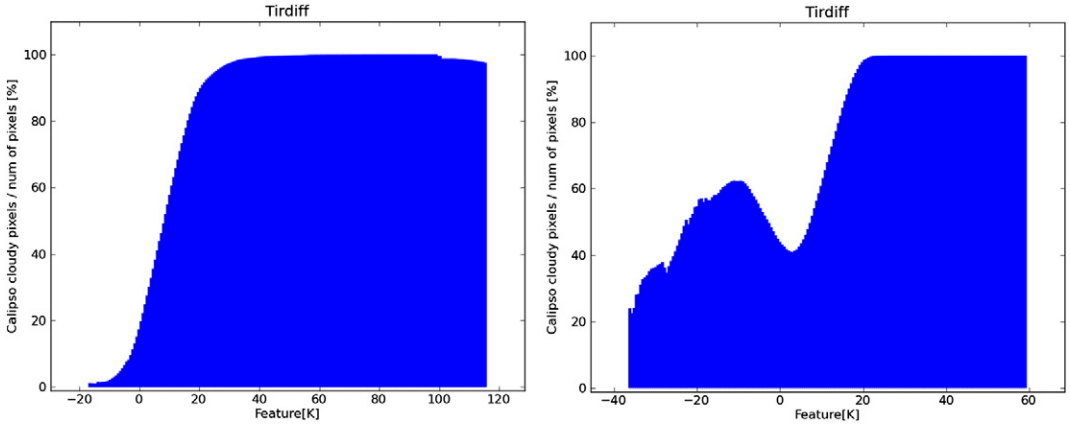

Fig. 2. Cloud probabilities estimated from CALIPSO–CALIOP cloud data in the period 2006–2009 as a function of AVHRR temperature differences between AVHRR channel 4 and the ERA-Interim (Dee et al., 2011) surface skin reference temperature over Low Latitude ocean surfaces during day (left) and over High Latitude mountain areas during night (right).

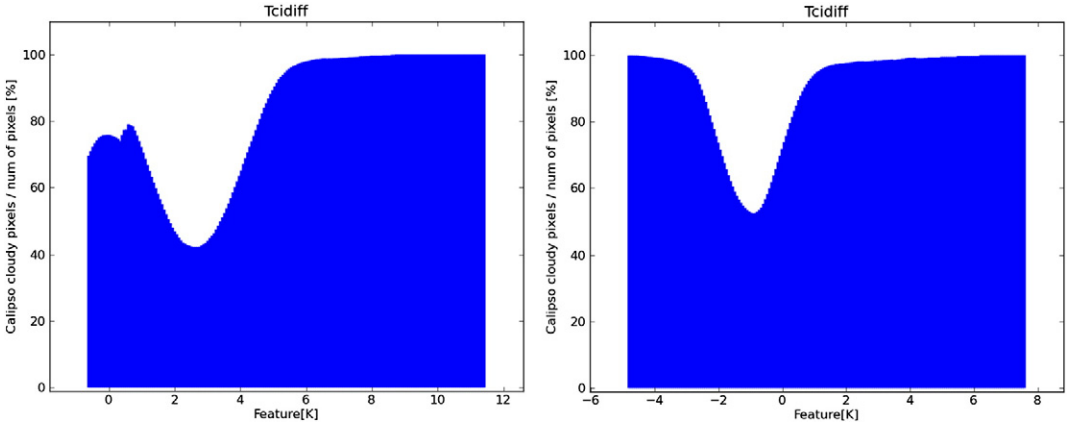

Fig. 3. Cloud probabilities estimated from CALIPSO–CALIOP cloud data in the period 2006–2009 as a function of AVHRR temperature differences between AVHRR channels 4 and 5 (denoted Feature in the plots) over Low Latitude ocean surfaces during night. Left panel shows results in original form and right panel if plotting results as a function of temperature differences related to PPS thresholds (consisting of dynamic threshold plus a tuning offset value).

clearly separates thin cirrus clouds to the right in the plot from the opaque clouds in the left part of the plot with cloud-free cases now concentrated around the transformed value of approximately −1 K. The fact that this value is not 0 K might indicate that the currently used PPS threshold is not optimal (at least, if taking the currently used CALIPSO dataset as reference). However, this is of no importance here since the correct (CALIPSO-derived) distribution relative to the possibly biased PPS threshold will be used anyway.

With this background we now list in Tables 2 and 3 a set of 8 transformed and constrained image features that will be used later for the definition of probabilistic cloud mask estimates. Four of them are selected for daytime conditions (i.e., solar zenith angles below 90°—Table 2) and four of them for night-time conditions (Table 3). However, one feature (Tirdiff) is used both day and night. Finally, in order to account for geographical and topographical differences we define 9 geographical regions over which we will train the probabilistic classifiers. These regions are listed in Table 4.

Snow and land use information were taken from National Snow and Ice Data Center (NSIDC) analyses and International Geosphere Biosphere Programme (IGBP) analyses, both of them provided together with the used CALIPSO–CALIOP cloud product (denoted Cloud and Aerosol Layer Information product version 3.01).

2.6. Linear aggregation of cloud probabilities—the PPS-Prob SPARC approach

The first of the two methods to be evaluated is inspired by the idea proposed by Khlopenkov and Thrishchenko (2007) for the AVHRR cloud processing over the Canadian region. They defined a concept, denoted Separation of Pixels Using Aggregated Rating over Canada (SPARC), where a cloud index was calculated as a linear sum of weighted contributions from various image features. We have adopted the same idea but instead of calculating a non-dimensional cloud index, originally placed in the interval −1 to +1 (going from completely clear to completely cloudy), we now approximate the total cloud probability $P(cloudy|\boldsymbol{F})$ in Eq. (1) directly as a linear sum of weighted contributions (with weights w_i) from conditional cloud probabilities $P(cloudy|f_i)$ for a set of N image features (i.e., $i = 1\ldots N$). Thus, we get the following two expressions for day and night using previous definitions in Tables 2 and 3:

$$P(cloudy|\boldsymbol{F})_{Day} = w_1 P(cloudy|Rswir) + w_2 P(cloudy|Tirdiff) + w_3 P(cloudy|Rmwir\_3b) + w_4 P(cloudy|Texture\_day) \quad (2)$$

$$P(cloudy|\boldsymbol{F})_{Night} = w_5 P(cloudy|Tirdiff) + w_6 P(cloudy|Tcidiff) + w_7 P(cloudy|Twdiff) + w_8 P(cloudy|Texture\_night) \quad (3)$$

What remains now is to estimate the appropriate weights w_i to be applied in the equations. This can be done in different ways. However, we note that one essential property of the chosen conditional cloud probabilities for an image feature is that it must be able to give probabilities that are away from the 50% level as frequently as possible. A probability of 50% just gives no or neutral guidance whether a pixel is cloudy

Table 2
Used transformed AVHRR image features for daytime probabilistic cloud masking.

Feature name	Definition	Main cloud detection ability
Rvis	Over land: AVHRR channel 1 TOA reflectances minus PPS thresholds Over ocean: AVHRR channel 2 TOA reflectances minus PPS thresholds	Identification of bright clouds over dark Earth surfaces
Tirdiff	AVHRR channel 4 brightness temperatures minus ERA-Interim (Dee et al., 2011) surface skin temperatures minus PPS thresholds	Identification of clouds which are significantly colder than the Earth surface
Rmwir_3b	(AVHRR channel 3b brightness temperatures minus AVHRR channel 5 brightness temperatures) minus PPS thresholds	Identification of clouds with significant reflection in the MWIR region (water clouds and thick multi-layered ice clouds), alternatively, clouds with significantly higher transmissivity in channel 3b than in channel 5 (thin ice clouds)
Texture_day	Over land: Not used (surface variability generally too large)! Over ocean: (Sum of local 3 × 3 pixel variances for AVHRR channel 1 TOA reflectances, AVHRR channel 3b brightness temperatures, AVHRR channel 4 brightness temperatures and AVHRR channel 3b and 5 brightness temperature differences) minus PPS thresholds	Identification of fractional or broken clouds over ocean

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

Feature name	Definition	Main cloud detection ability
Tirdiff	AVHRR channel 4 brightness temperatures minus ERA-Interim surface skin temperatures minus PPS thresholds	Identification of clouds which are significantly colder than the Earth surface
Tcidiff	AVHRR channel 4 brightness temperatures minus AVHRR channel 5 brightness temperatures minus PPS thresholds	Identification of thin cirrus clouds
Twdiff	(AVHRR channel 3b brightness temperatures minus AVHRR channel 4 brightness temperatures) minus PPS thresholds	Identification of water clouds
Texture_night	Over land: Not used (surface variability generally too large)! Over ocean: (Sum of local 3 × 3 pixel variances for AVHRR channel 4 brightness temperatures and AVHRR channel 3b and 5 brightness temperature differences) minus PPS thresholds	Identification of fractional or broken clouds over ocean

or not, thus we should strive for finding image features that are able to give better than neutral guidance (compare with examples in Figs. 1 and 2). One way to judge the actual usefulness of an individual feature f_i could then be to integrate the absolute difference $|P(cloudy|f_i) - 50\%|$ over the full feature domain ($f_{i,min} \le fi \le fi_{,max}$) according to the following equation:

$$P(f_i)_{diff\_integrated} = \int_{f_{i,min}}^{f_{i,max}} N(f_i)\{|P(cloudy|f_i) - 50|\}df_i \quad (4)$$

Here, $N(f_i)$ is the absolute frequency of f_i occurrences over a restricted interval (df_i) of f_i values.

The final weights can then be calculated as follows (with two realisations, one for day and one for night);

$$w_i = {}^{P(f_i)_{diff\_integrated}} \Big/ \sum\nolimits_i P(f_i)_{diff\_integrated} \quad (5)$$

Notice finally that weights w_i and $P(cloudy|f_i)$ will be estimated separately for all geographic regions according to Table 4.

In the remainder of this paper we will denote this probabilistic formulation "PPS-Prob SPARC".

2.7. The PPS-Prob Naïve Bayesian approach

The second probabilistic method is a simplified version of Eq. (1) where the right hand side has been reformulated following some approximations. If assuming that image features f_i are all independent (i.e., image features are uncorrelated), individual probabilities may now be multiplied following the fundamental statistical rule for "Compound Probability of Independent Events" when computing the total probability. Thus, Eq. (1) reduces to

$$P(cloudy|\boldsymbol{F}) = \frac{\overline{P(cloudy)}\prod_i P(f_i|cloudy)}{P(\boldsymbol{F})} \quad (6)$$

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

The problem has now been reduced to estimating individual probabilities $P(f_i|cloudy)$ and then multiplying them. We notice that there must be a mutual inter-dependence between $P(f_i|cloudy)$ and $P(cloudy|f_i)$. More clearly, if knowing the conditional probability that it is cloudy given a certain image feature value (which is how cloud probabilities were collected from CALIPSO data as illustrated in Sections 2.4 and 2.5), we can also calculate it the other way around from the same statistical training dataset (provided that both absolute and relative frequencies of cloud occurrences are stored). Remaining factors on the right hand side of Eq. (6) may also be calculated from training data. E.g., an estimation of the mean cloud occurrence $\overline{P(cloudy)}$ is possible and the factor $P(\boldsymbol{F})$ may be estimated by summing contributions from both cloudy and clear cases and then compute the overall frequency for which any particular realisation of vector $\boldsymbol{F}$ occurs.

The Naïve Bayesian approximation has been successfully applied to many scientific applications (e.g., Kossin & Sitkowski, 2009) and it has also recently been applied to the AVHRR cloud screening problem (Heidinger et al., 2012). We are here testing a similar approach but using a different concept in terms of the used image features, i.e., the constrained feature approach as described in Section 2.5.

In the remainder of this paper we will denote this method "PPS-Prob Naïve".

2.8. Training aspects and training datasets based on CALIPSO–CALIOP cloud data

For this study we have taken advantage of the previously collected dataset with optimally matched NOAA-18 and CALIPSO orbits described by Karlsson and Johansson (2013). This study and also several other studies (e.g., Stengel et al., 2013) have demonstrated that it is possible to collocate NOAA AVHRR data with CALIPSO data with comparable quality to what is achieved when matching with other internal datasets in the Aqua train (e.g. MODIS data). Some example results from this dataset have already been shown in Section 2.5. However, some important and necessary restrictions to the utilised information have been

Table 4
Geographical regions used when training the probabilistic classifiers.

Geographical region	Definition
Polar ocean	Ice-covered ocean at latitudes higher than 40°
High-latitude ocean	Ice-free ocean at latitudes higher than 40°
Low-latitude ocean	Ocean at latitudes lower than 40°
High-latitude snow-covered mountains	Mountain regions (topography exceeding 500 m) with snow-cover at latitudes higher than 40°
High-latitude snow-free mountains	Mountain regions (topography exceeding 500 m) without snow-cover at latitudes higher than 40°
High-latitude snow-covered land	Snow-covered land (topography below 500 m) at latitudes higher than 40°
High-latitude snow-free land	Snow-free land (topography below 500 m) at latitudes higher than 40°
Desert regions	Land areas without vegetation at latitudes lower than 40°
Low-latitude vegetated regions	Vegetated land areas at latitudes lower than 40°

applied during the training process. A great asset of the CALIPSO–CALIOP cloud products is the superior sensitivity for cloud detection compared to corresponding conditions for passive data like data from the AVHRR sensor. But this is also a problem when using this information as the basis for a statistical training of a probabilistic cloud masking method. More clearly, there is a risk for "over-training", i.e., that we force the method to try to detect clouds that are theoretically impossible to detect from AVHRR sensor data. As a result, the probabilistic cloud-screening method would then risk to systematically creating artificial clouds in truly cloud-free areas since the cloud-free signal cannot be confidently separated from the cloudy signal for these sub-visible cirrus clouds. Consequently, we need to find a way to restrict the used CALIOP-based cloud mask in the training process to include only those clouds which we believe are discernible also in AVHRR images. In other words, we need to define as accurately as possible the AVHRR cloud detection limit. On the other hand, applied training restrictions must not go too far so that they preclude detecting potentially detectable clouds which are not generally detected by today's cloud screening methods. We need to leave some margin for further improvement of cloud detection performance even if that margin probably is very small (when considering that the experience of AVHRR cloud detection is now based on more than 30 years of development).

We have again utilised the dataset collected by Karlsson and Johansson (2013) for finding the appropriate cloud detection limit. They concluded that the PPS method reached its optimal performance for clouds with optical thicknesses of 0.35. Below this value the method started to systematically miss clouds with increasing magnitude for smaller and smaller cloud optical depths. Further analysis of their data revealed that below a cloud optical thickness of approximately 0.2 the PPS loss of clouds exceeds 50%, i.e., less than 50% of the clouds with this optical thickness are detected. We have used this value (i.e., cloud optical thickness of 0.2) to represent the AVHRR cloud detection limit in the training of the probabilistic classifier. CALIPSO–CALIOP detected clouds below this threshold are treated as being non-existing and equivalent to cloud-free conditions. This compromise solution means that some clouds are still likely to be non-detectable by the probabilistic classifier but some of the currently non-detected clouds in the cloud optical thickness interval 0.2–0.35 may potentially be identified. As a consequence, our probabilistic classifier might still over predict cloud probability to some extent which may have some consequence for the final use of the results (e.g., when creating new fixed cloud masks based on the probabilistic results). We will address this aspect further in upcoming Sections 3 and 4.

The final training dataset consists of the same matched global NOAA-18 and CALIPSO orbits (99 orbits in total) as being used by Karlsson and Johansson (2013). It spans the period 2006–2009 and provides a reasonable global coverage over all seasons during that period. All in all it comprises almost 1 million matches of AVHRR Global Area Coverage (GAC) and CALIOP pixels/samples at approximately 5 km horizontal resolution. The constrained training (i.e., now being related to PPS threshold information) is based on results from the PPS software version 2010 with some extensions for AVHRR GAC processing. This is a much advanced PPS version compared to the original method described by Dybbroe et al. (2005a,b). The main new features of the method concerns adaptations to global processing (e.g., over desert and Polar Regions) and a systematic use of prescribed MODIS-derived surface emissivity information during night-time conditions. It was used by the EUMETSAT Climate Monitoring Satellite Application Facility (CMSAF, described by Schulz et al., 2009) as the basic cloud screening method for the definition of the CDR CMSAF Clouds, Albedo and Radiation dataset from AVHRR data (CLARA-A1, described by Karlsson et al., 2013).

3. Validating probabilistic results using space- and ground-based observations

For the evaluation of the results we will use three types of independent observations in addition to the natural inter-comparison with the standard PPS/CLARA-A1 results (for more information on how CLARA-A1 relates to other datasets like PATMOS-x and MODIS, see Karlsson et al., 2013; Stengel et al., 2013). The first type is again CALIPSO–CALIOP cloud observations but now taken from a period outside of the training period 2006–2009. We will compare PPS-Prob SPARC and PPS-Prob Naïve results from 78 global orbits in the year 2010 from the NOAA-18 and NOAA-19 satellites matched with CALIPSO–CALIOP data following the same optimal matching criteria as for original data being used for training.

However, to broaden the evaluation and to come away from the very close link to the CALIPSO–CALIOP cloud information we will also compare results to one surface-based observational dataset and one additional satellite-based dataset. An additional specific reason for this is that the training of the methods have utilised satellite data almost exclusively taken from a near nadir observation point (i.e., viewing angles less than 20° as illustrated in Fig. 4) whereas the final results have been produced for entire swaths with very variable satellite viewing angles. Thus, we want to study if results appear to be useful also outside of the near-nadir observation mode. We recall that one specific reason for using the constrained image feature approach (as outlined in Section 5) was for taking care of this particular dependence and we need to check if this concept is successful or not.

We have chosen to use a quite advanced ground-based reference observation dataset compiled from a set of combined remote sensing instrument over one particular measurement site; the Cabauw Experimental Site for Atmospheric Research (CESAR, 51°58′N, 4°55′E) in the Netherlands (Fig. 5). The remote sensors are based on active as well as passive systems and use either a hemispheric or columnar remote sensing technique.

The dataset (described by Boers et al., 2010) includes observations from the following five remote sensing techniques optimised for cloud detection:

1. Two cloud lidars (ceilometers) with different vertical ranges (4 km and 7 km)
2. A Degreane 35 GHz cloud radar with maximum vertical range of 12 km.
3. A NubiScope hemispheric scanning pyranometer with the method for cloud detection from infrared measurements described by Wauben, Bosveld, and Klein Baltink (2010).
4. Infrared pyrgeometers used for the Baseline Surface Radiation Network (BSRN). The method for interpreting cloud occurrence was the Automated Partial Cloud Amount Detection Algorithm (APCADA, described by Dürr & Philipona, 2004).
5. A TSI-440 Total Sky Imager (digital camera pointed downward at a hemispheric mirror). The algorithm for cloud detection is described by Boers et al. (2010).

Measurements from these different techniques were weighted according to the cloud base height reported by ceilometers at the observation time. The algorithm was able to provide fractional cloudiness observations every 10 min for 99.92% of the total period of 12 months (15 May 2008 to 14 May 2009). The high temporal resolution was one of the main reasons for choosing this dataset as a reference. This is a big advantage compared to conventional manual surface observations (synoptical observations) which are only made every 3 h and consequently very difficult to match in time with satellite overpasses. We also believe that this dataset is more objective (or at least more consistent) than conventional surface observations (often denoted SYNOP). Especially, remote sensors are not suffering from particular problems during night due to lack of sunlight which is an obvious and well-known weakness of SYNOP observations. Another difficulty of using the SYNOP observation is the way that SYNOP values 1 and 7 octas (i.e., cloud fraction expressed as eights) are defined. The value 1 octa is used as soon a small cloud is observed and 7 octas is used as soon as a small hole in a cloud deck is observed. This is done even if in reality values 0 and 8 octas should have been used instead (i.e., if averaging

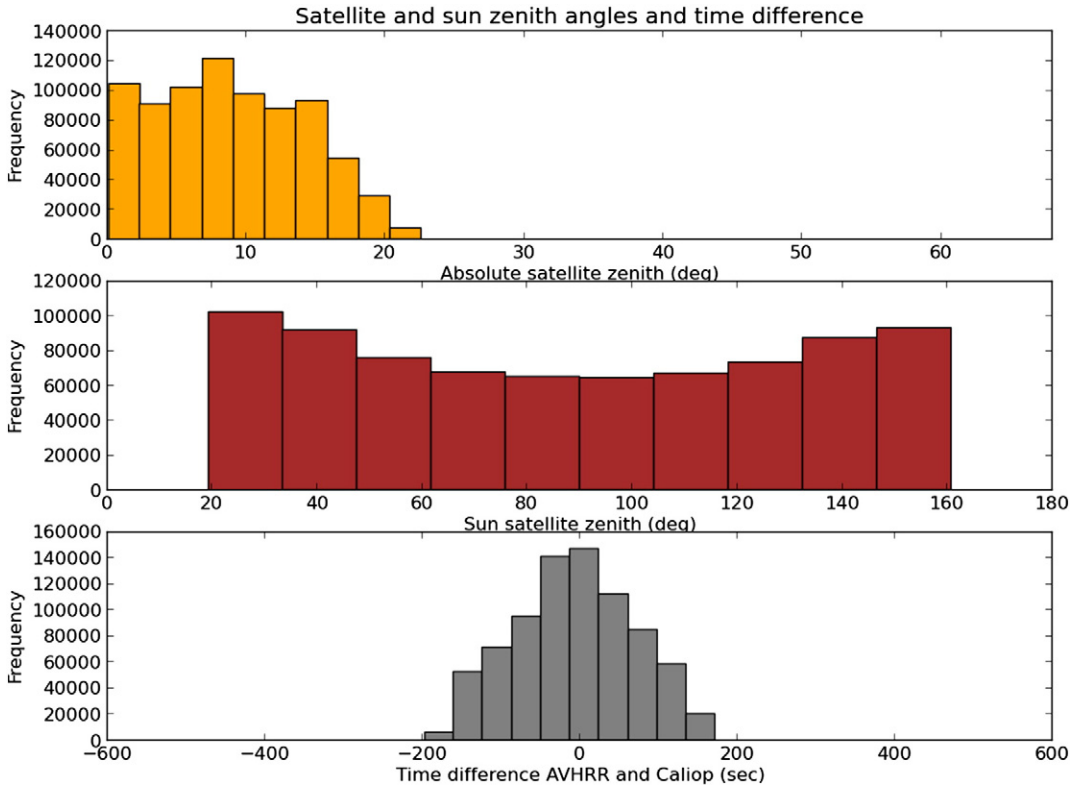

Fig. 4. Summary of viewing and time matching characteristics for NOAA-18 AVHRR observations matched to corresponding CALIPSO–CALIOP observations (from nadir) for 99 global orbits in the time period 2006–2009 (courtesy of Adam Dybbroe, SMHI).

hemispherical fractional cloudiness correctly). Nevertheless, Boers et al. (2010) compared the combined remote sensing observations with a 30-year SYNOP observer climatology and found rather good agreement with some exceptions (including a clearly different performance at night).

For this study, we selected randomly two NOAA-18 overpasses per day observing the Cabauw position during the period 15 May 2008 until 14 May 2009 where cloud probability results were compared to the combined remote sensing observation. One overpass was an afternoon overpass (ascending node) and one a night overpass (descending node). The reason for only selecting two overpasses per day was to avoid too much of correlation between subsequent overpasses. The random selection (i.e., not only selecting the best overpass as close as possible to the zenith position) was applied in order to get a wide variety of different satellite viewing angles when observing the Cabauw site.

As the last reference we have chosen to compare with cloud masks derived from the Spinning Enhanced Visible and Infrared Imager (SEVIRI) sensor carried by the geostationary Meteosat-8 and Meteosat-9 satellites operated by EUMETSAT. This sensor provides also measurements with high temporal resolution (15 min), thus allowing easy inter-comparison with NOAA-18 overpasses overflying the SEVIRI field of view. Cloud masks were generated by the NWC SAF Meteosat Second Generation

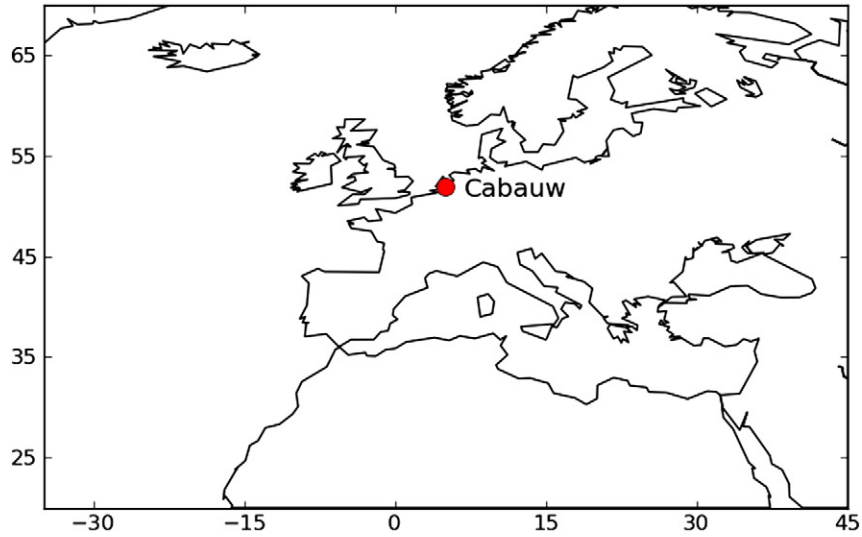

Fig. 5. Position of observation site in Cabauw, The Netherlands (at position 52°N and 5°E) for the inter-comparison of PPS-Prob SPARC, PPS-Prob Naïve, CLARA-A1 and SEVIRI cloud datasets (see text).

cloud software package. The original cloud masking algorithms were described by Derrien and LeGléau (2005) but in this experiment we have used a greatly upgraded version of the software (denoted NWC SAF MSG Cloud Software version 2012). Improvements concern mainly the treatment of clouds observed during twilight conditions (Derrien & Le Gléau, 2010). We have limited the inter-comparison to the same one-year NOAA-18 dataset being extracted for the comparison with ground observations at the Cabauw site.

For the matching of AVHRR GAC and SEVIRI results to observations at the Cabauw site, we first aggregated results for 3 × 3 pixels with the center pixels' containing the Cabauw coordinates. With approximately 5 km pixel size for both AVHRR GAC and SEVIRI this means that we approximate the Cabauw observation to be valid for an area of about 225 km^2. Also larger window sizes were tested but results were found to be best correlated when using a window size of 3 pixels. The alternative to use only one pixel was not found reasonable due to potential remaining uncertainties in pixel navigation and also because of the knowledge of the Cabauw observation being partly composed by hemispherical observations (e.g. techniques 3–5 in the list above).

Results were compiled using standard statistical measures such as those described by Karlsson and Johansson (2013). These include Probability of Detection Clear or Cloudy (POD), False Alarm Rate for Clear and Cloudy (FAR), Hit Score (HS) and Kuipers' Skill Score (KSS). Even if some of the reference results were provided as true fractional cloud estimations (e.g., the combined Cabauw observation was provided like that) we did not compare them to the PPS-Prob results directly. For reasons explained earlier in Section 2.8 (i.e., the existing 'over-training' of the probabilistic classifiers), we first transformed the PPS-Prob results to binary cloud masks using a threshold value before comparing directly to the CALIPSO–CALIOP cloud mask and before aggregating results over the Cabauw site. After testing various thresholds it was found that the best results were achieved when using a threshold of 60% probability. For lower values of this threshold it was clear that too many truly cloud-free regions were misclassified as cloudy because of the remaining 'over-training' of the classifier. The same optimal threshold of 60% was consistently found for comparisons with both the CALIPSO dataset and the ground-based dataset.

4. Results

4.1. Demonstration of resulting PPS-Prob SPARC and PPS-Prob Naïve results

Before summarising results in statistical scores, an image example of the achieved cloud probabilities is shown in Fig. 6 for one particular GAC scene. Results from the two methods are shown together with a colour composite image of the same scene from three original AVHRR channel radiance images. A quick visual inspection verifies that resulting cloud probabilities for both methods are highly correlated with obviously cloudy areas in the colour composite image. However, noteworthy is that thin and broken cloud fields over the ocean surfaces are much more highlighted in probability images than in the colour composite.

This is mainly explained by the added cloud information coming from features Rmvir_3b and Texture_day described earlier in Table 2. These features contain information from the 3.7 and 12 μm channels which is information that is not covered by the colour composite in the leftmost panel of Fig. 6. This example indicates also that for this particular case the PPS-Prob Naïve representation appears to be a better representation of the true cloudiness situation than PPS-Prob SPARC since the latter appears to get a relatively high contribution to the cloud probabilities also from cloud-free land areas.

4.2. Inter-comparison with independent CALIPSO–CALIOP data

Figs. 7 and 8 show results for POD, HR and KSS scores based on the 2010 dataset with 78 NOAA-18 and NOAA-19 scenes visualised using the same plotting method as in the study of Karlsson and Johansson (2013). This method plots results as a function of thresholded cloud optical thicknesses which means that all CALIOP-detected clouds below the shown cloud optical thickness on the x-axis is treated as being cloud-free (i.e., thinner clouds are filtered out). Consequently, original unfiltered results are seen for a cloud optical thickness value of 0.0.

In Fig. 7 we notice that both probabilistic methods improve the probability of detecting cloudy conditions (blue solid lines) compared to CLARA-A1. However, for clear conditions PPS-Prob SPARC results are

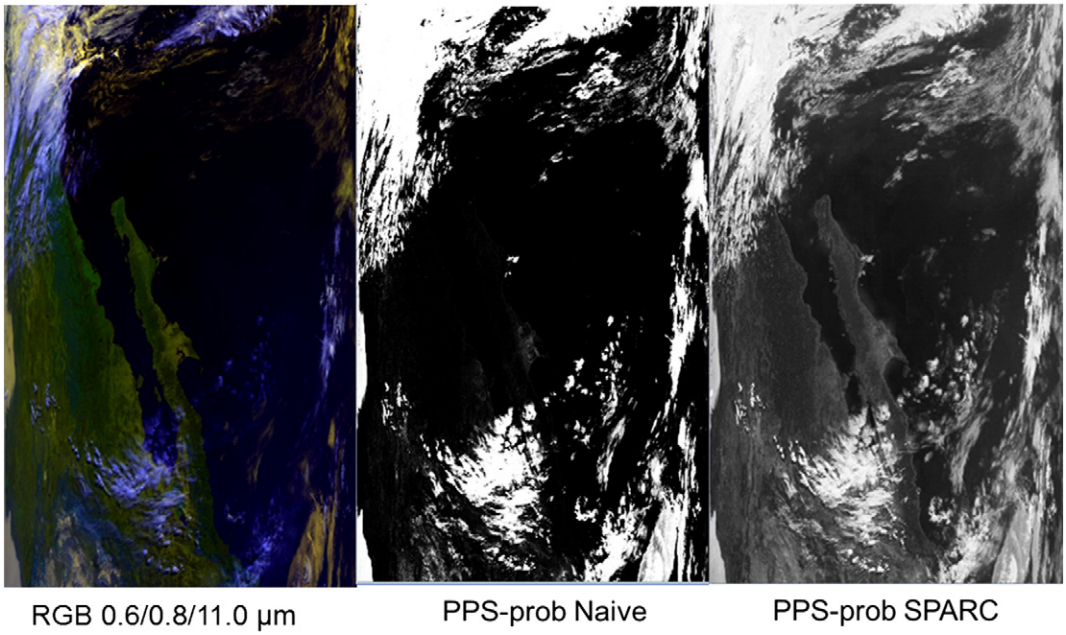

Fig. 6. Part of an original NOAA-18 AVHRR GAC scene in satellite projection over the North American west coast (with Gulf of California and Baja California in the center) registered in ascending mode (i.e., North is down, South is up) from 26 January 2010. Left: Colour composite with AVHRR channel 1 (red), channel 2 (green) and channel 4 (blue). Middle: Corresponding PPS-Prob Naïve cloud probabilities (as greyscale image with range 0–100%) Right: Corresponding PPS-Prob SPARC cloud probabilities (as greyscale image with range 0–100%).

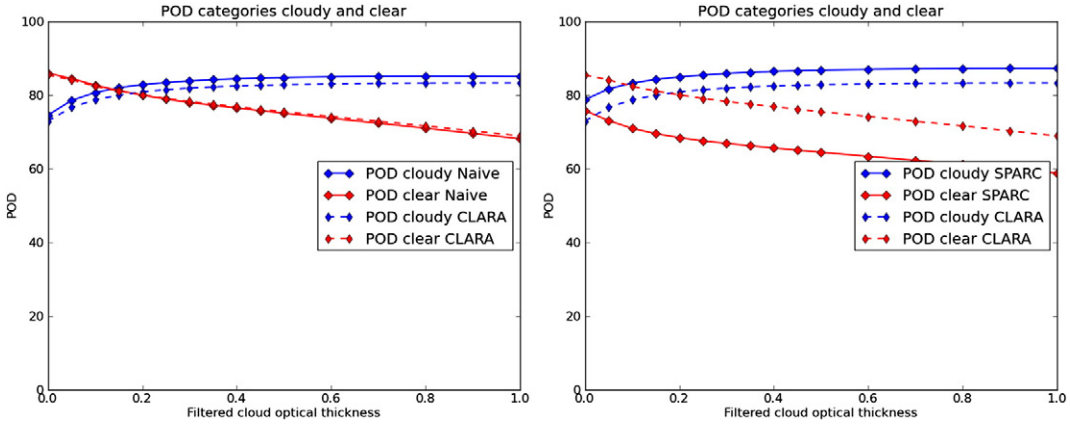

Fig. 7. Probability of detection (POD) for cloudy and clear conditions plotted as a function of filtered CALIPSO–CALIOP cloud optical thickness (explained in text). Results for PPS-Prob Naïve are shown in left panel and for PPS-Prob SPARC in the right panel. Both results are compared to results for CLARA-A1 (dashed lines). All results were derived for 78 NOAA-18 and NOAA-19 orbits in 2010.

significantly worse (red solid line) than CLARA-A1 while for PPS-Prob Naïve results are more or less identical. The total effect of this is clearly summarised by overall HR and KSS scores in Fig. 8 which are clearly better for PPS-Prob Naïve than for PPS-Prob SPARC. It seems that the linear summation of conditional probabilities for PPS-Prob SPARC generates too many mis-classified truly cloud-free pixels. Fig. 6 already indicated this problem showing quite high cloud probabilities over cloud-free land areas. Thus, the PPS-Prob Naïve method appears better in separating clear from cloudy conditions which is especially highlighted by the KSS score.

A comment on the actual POD values for cloudy conditions is relevant here. It seems from Fig. 7 that these are never exceeding 80–85% which may seem rather low, i.e., indicating that more than 15% of all clouds still remain undetected even after filtering out the thin cloud cases. This is only partly explained by remaining temporal and spatial mismatches between CALIPSO–CALIOP and NOAA AVHRR measurements. The major part of it comes from problems with cloud detection over the Polar Regions and especially during the Polar night. During these conditions even optically thick clouds may remain undetected because of having cloud top temperatures very similar to ground and ice-cover temperatures. Also, these clouds often consist of mixed water and ice particles leading to problems in efficiently using the Tcidiff and Twdiff features listed in Table 3. We conclude that cloud detection over the Arctic and Antarctic regions remains very challenging. On the other hand, if excluding the Polar Regions (e.g., if only considering latitudes between 65°S and 65°N) POD values for cloudy conditions increases to 90–95% or even higher for most regions (not shown).

4.3. Inter-comparison with combined ground-based remote sensing observations in Cabauw

Fig. 9 shows all statistical scores for four cloud datasets (i.e., the two probabilistic methods and the CLARA-A1 and SEVIRI methods) compared to the combined remote sensing cloud observations at Cabauw. Results have here been subdivided into daytime (ascending node) and night-time (descending node) portions for the studied NOAA-18 AVHRR observations. To raise the confidence in the results we have here discarded the partly cloudy cases (cloud fraction 25–75%) in the Cabauw observation. The reason is that in the inter-mediate range the risk increases for mismatched cloud observations due to remaining temporal and spatial effects. Thus, we are here only looking at the cases which we can consider as being Confidently Clear and Confidently Cloudy. We notice in the upper left corner of Fig. 9 that the basic cloud detection generally performs slightly better during daytime conditions compared to nigh-time for all methods although possibly with the exception of the SEVIRI dataset which shows good skills both night and day. The situation is not the same for clear conditions where the daytime and night-time figures have lower values and with a less distinct difference day and night. We conclude: All methods appear to be slightly clear conservative, i.e., to detect most clouds but to misclassify a certain portion of the cloud-free regions as being cloudy.

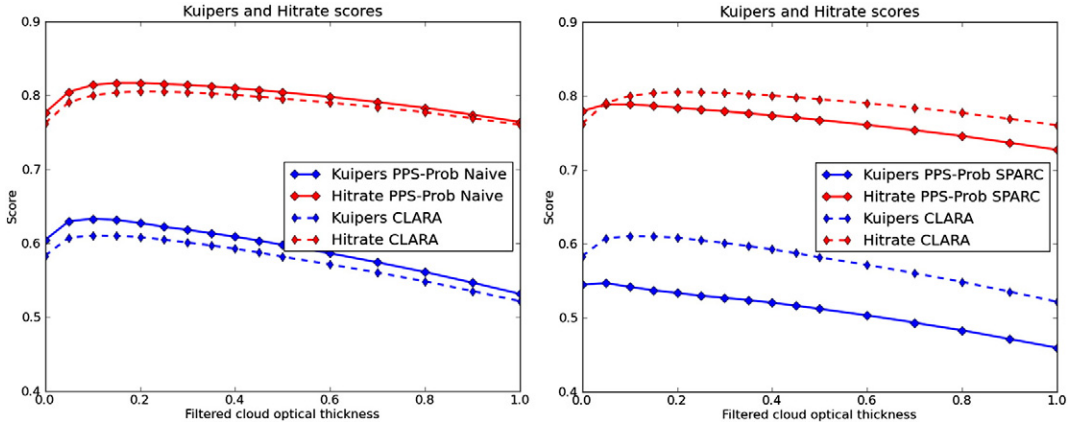

Fig. 8. Hitrate (HR) and Kuipers Skill Score (KSS) plotted as a function of filtered CALIPSO–CALIOP cloud optical thickness (explained in text). Results for PPS-Prob Naïve are shown in left panel and for PPS-Prob SPARC in the right panel. Both results are compared to results for CLARA-A1 (dashed lines). All results were derived for 78 NOAA-18 and NOAA-19 orbits in 2010.

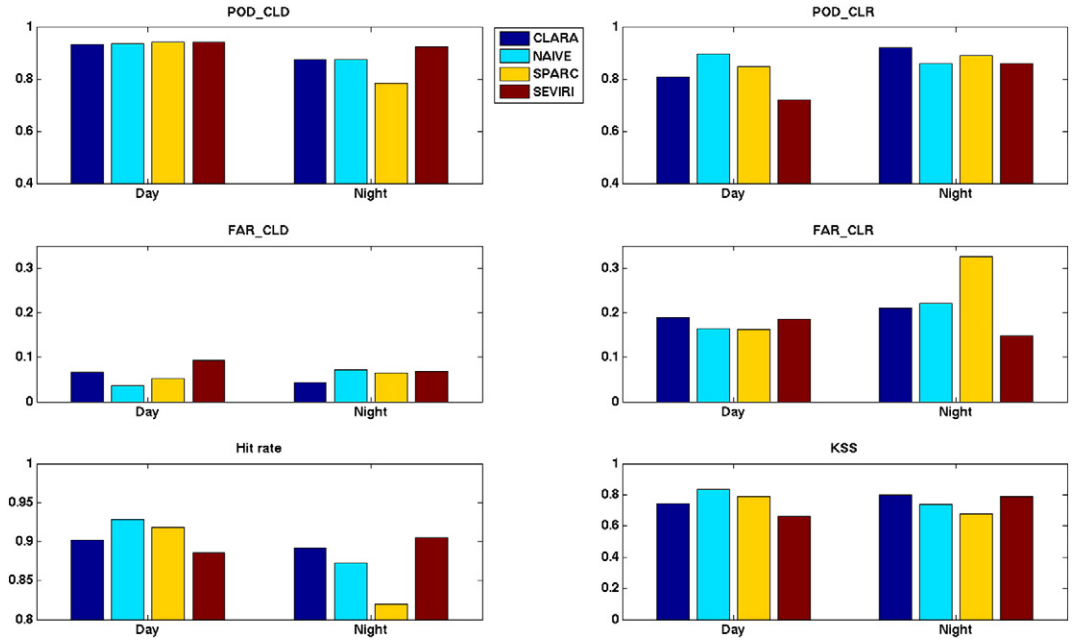

Fig. 9. Inter-comparison of four cloud masking results (CLARA-A1, PPS-Naïve, PPS-SPARC and SEVIRI—datasets specified in text), separated into daytime and night-time conditions, using the remote sensing observation reference at the Cabauw site over the period 15 May 2008 to 14 May 2009. POD for cloudy and clear conditions are shown in the upper panel, FAR for cloudy and clear conditions in the middle panel and Hitrate and Kuipers Skill Score (KSS) in the lower panel.

Moreover, both probabilistic methods give comparable or during daytime even better results than the two other satellite-based

datasets. This is most clearly seen in the aggregated Hit Rate and KSS scores where during daytime PPS-Prob Naïve gives the best results followed by PPS-SPARC. However, during night the situation is the opposite and the two reference datasets give better results. Despite this, differences are not large between the methods (especially not for PPS_Prob Naïve) and the conclusion is rather that all methods are quite comparable.

Fig. 10 shows the same results as in Fig. 9 but now separated into three different categories of satellite viewing (zenith) angles for the NOAA-18 satellite. We first notice the small but clearly noticeable general tendency to get higher POD values for cloudy conditions for higher viewing angles (except for SEVIRI results where the viewing angle from the geostationary position is obviously fixed) whereas the corresponding POD values for clear conditions show an opposite behaviour. This agrees well with the anticipated effect of how clouds would appear at larger observation angles (the parallax effect). Thin clouds appear thicker at higher viewing angles (i.e., the path length inside the cloud increases) and should consequently be easier to detect. Also, holes in the cloud deck are effectively hidden by existing clouds at increasing viewing angles. For cloud-free areas the increasing viewing angles

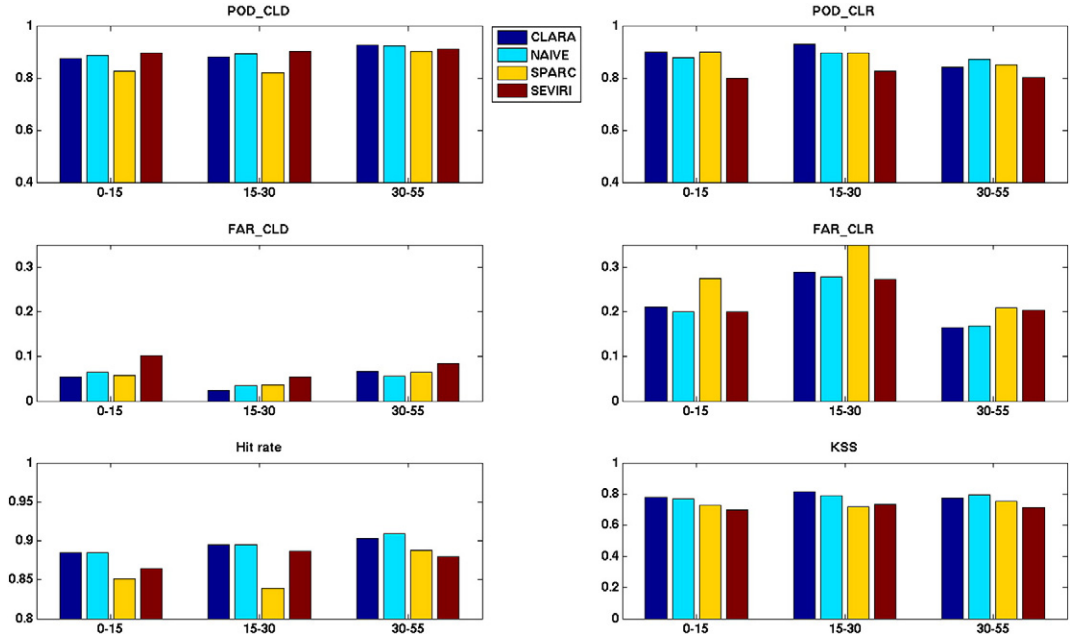

Fig. 10. Inter-comparison of four cloud masking results (CLARA-A1, PPS-Naïve, PPS-SPARC and SEVIRI—datasets specified in text), separated into three satellite viewing angle intervals, using the remote sensing observation reference at the Cabauw site over the period 15 May 2008 to 14 May 2009. Same presentation of skill scores as in Fig. 9.

increases emission contributions from e.g. atmospheric water vapour due to increasing atmospheric path lengths and the risk to misclassify this radiance as coming from clouds increases. This takes place despite the attempts made here to compensate for atmospheric effects. Also, high viewing angles increase the risk of mis-interpreting anisotropic reflection from Earth surfaces as being reflection from clouds. This is particularly serious when viewing bright surfaces towards the direction of the sun. We believe that the fact that we can see this general dependence on viewing angles in our results also increases the confidence in the remote sensing dataset compiled at Cabauw.

When looking at the results for the individual datasets, results are not very conclusive concerning the dataset showing the best performance. Rather, it seems that all datasets are more or less comparable. The main conclusion here is that there does not seem to be a very clear deterioration of results when we increase viewing angles. Especially, results for the probabilistic approaches definitely appear to be comparable with PPS results (CLARA-A1). Thus, the inbuilt compensation for varying viewing angles in the used constrained features appears to work reasonably well for the PPS-Prob methods.

Finally, Fig. 11 shows results for all involved cloud datasets in the Cabauw part of the study where all results (now obviously including also partial cloud cover cases) have now been sub-divided into octa categories, i.e. the representation of cloudiness used in standard SYNOP observations. We recognise the well-known U-shape of cloudiness with the most frequent cases for very low or very large cloud amounts.

Interesting is that all satellite-based methods reproduce the familiar U-shape but with larger values at the extreme ends and with smaller values in the intermediate range than what is given by the Cabauw observation. The most striking deviation here is seen for octa category 7 where all satellite-based methods give much lower frequencies. A similar appearance is seen for octa category 1. We suspect that this might be an effect of the way the satellite and ground-based observations are composed rather than a sign of a true mis-representation of the cloudiness distribution. The Cabauw observation should to some extent be affected by some contributions from hemispherically scanning sensors sensing cloudiness in positions outside of the 3x3 pixel domain being studied from satellite. This can act to even out the distribution over Cabauw slightly by increasing the chances of getting category 1 in cloud-free conditions in the nearest vicinity to Cabauw. The same effect can act to increase the frequency of category 7 in case of overcast conditions in the nearest vicinity to Cabauw.

We finally conclude from Fig. 11 that all satellite-based methods agree reasonably well with the Cabauw observations with no method showing significantly better results than any other. Ideally, this aspect of this study should have been advanced even further, e.g., investigating the performance day and night and with calculation of more comprehensive statistical performance scores for each octa category. However, the rather limited extension of the dataset (twice daily observations during one entire year) did not permit such studies with statistical significance.

5. Summary and concluding remarks

In this study we have investigated two different ways of transforming traditional AVHRR cloud masking results into a probabilistic formulation. Methods have been based on Bayesian theory but expressed in a simplified form to allow easy implementation and use in applications demanding fast processing of huge data amounts, e.g., the processing in Nowcasting applications or the processing of the entire historic AVHRR GAC dataset (~20 TB) for climate data record generation. For example, the PPS-Prob Naïve method was demonstrated to be executable within the same time limits as the original PPS method in the same computer environment.

An important part of the methodology was the use of a set of constrained core image features in the training of the cloud classifier. These features, chosen to cover the most important pieces of cloud information provided by the AVHRR spectral channels, were defined in a compact way to take into account a series of different environmental parameters affecting the AVHRR measurement. This included basic viewing and solar angle dependencies as well as the impact of atmospheric absorption. In addition, the method took into account additional

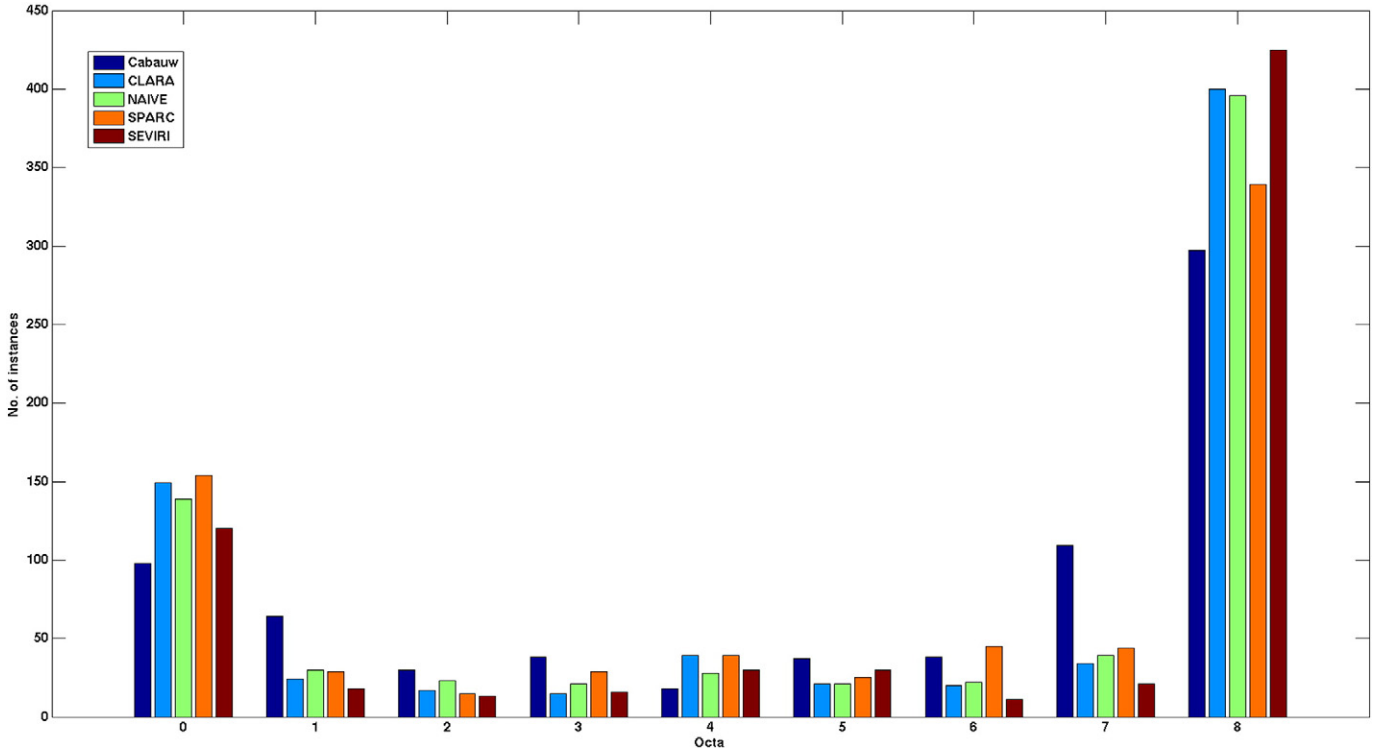

Fig. 11. Inter-comparison of all five cloud observation datasets (Cabauw remote sensing observation, CLARA-A1, PPS-Naïve, PPS-SPARC and SEVIRI—datasets specified in text) separated into octa intervals over the period 15 May 2008 to 14 May 2009.

variation explained by geographic location and land topography by operating it over 9 separate geographical categories. The method was trained based on detailed cloud information provided by the CALIPSO–CALIOP sensor in the period 2006–2009.

Results were evaluated in two ways: 1. Based on independent CALIPSO–CALIOP data from 2010 and compared to the original CLARA-A1 cloud climate data record. 2. Using a one-year (15 May 2008 to 14 May 2009) ground-based cloud dataset composed from remote sensing instruments over the observation site in Cabauw in the Netherlands. For the second evaluation, results were also compared to the original CLARA-A1 cloud climate data record but, in addition, to cloud masks derived from the METEOSAT SEVIRI sensor. It was demonstrated that the probabilistic methods compare well with the referenced satellite datasets and for daytime conditions even providing better performance than the referenced methods. It was also shown that, because of the use of the constrained image features, results were found reasonably useful also for high viewing angles.

When comparing the two probabilistic approaches, it was found that the formulation based on a Naïve Bayesian formulation (denoted PPS-Prob Naïve) performed clearly superior to the formulation based on a linear summation of conditional cloud probabilities (denoted PPS-Prob SPARC). Especially, a much more dynamic range of probabilities were produced by the former method which reproduced observed distributions better than results from the alternative method. For the study based on the observations over the Cabauw site, the overall daytime Kuipers Skill Score (considered as the most critical score for evaluating the cloud detection performance) for PPS-Prob Naïve was 0.84, for PPS-Prob SPARC 0.79, for CLARA-A1 0.74 and for SEVIRI 0.66. Corresponding results for night-time conditions were less favourable for the probabilistic formulations (Kuipers Skill Score 0.74 for PPS_Prob Naïve, 0.68 for PPS-Prob SPARC, 0.80 for CLARA-A1 and 0.79 for SEVIRI) but still relatively close to those from the reference datasets. In addition, the Cabauw distribution of cloudiness occurrences in different octa categories was reproduced very closely by all methods, including the probabilistic formulations. Finally, overall results for Cabauw were also largely in very good agreement with results given by the comparisons with the CALIPSO–CALIOP cloud mask, although scores for the latter were somewhat reduced reflecting the global character of the study (e.g., affected by the challenging conditions over the Polar regions). We interpret this as verifying the high quality of the Cabauw remote sensing dataset.

Some concern about the theoretical limitations of the PPS-Prob Naïve method could be raised and then especially the assumption of independent and uncorrelated image features. We know that such correlations exist in reality. For example, for convective clouds we often observe that when clouds grow thicker they also get a colder cloud top which would then lead to simultaneous increases in both cloud reflectances and the inverted cloud brightness temperatures. However, this does not appear to be a very strong rule (i.e., lots of opposite cases exist) as apparently the achieved results are equal or even better than those from more traditional methods. It seems as the probabilistic approach itself adds a better representations of cloud distributions in the image feature space compared to the results produced by traditional thresholding methods. Nevertheless, this aspect needs to be more deeply investigated in future studies as well as inter-comparing with results from applications based on a more complete representation of the original Bayesian approach (e.g., Merchant, Harris, Maturi, & MacCallum, 2005; Musial et al. 2014). Another remaining topic for continued studies is the definition of the AVHRR cloud detection limit which determines how accurately one can train a statistical cloud classification method and, in particular, how to finally get truly representative cloud probabilities. Finally, some additional development is also needed for dealing with data from the additional 1.6 μm channel (denoted 3a) of the AVHRR/3 sensor being carried by morning orbit satellites (NOAA-17 and the Metop series of satellites). Here, matching with CALIPSO data is only possible at higher latitudes which calls for additional development efforts using other reference data (e.g. from VIIRS and/or MODIS).

The PPS-Prob Naïve approach will be implemented in an upcoming NWC SAF version of the PPS cloud software (to be denoted PPS version 2016) which a scheduled release in 2017. A first version of the method will also be used in the second release of the CMSAF cloud climate data record based on historic AVHRR GAC data (to be denoted CLARA-A2) with a scheduled release in 2016. For both applications we expect further upgrades of the performance compared to what has been shown here which should follow from the continuous improvement of the pre-calculated PPS threshold information and the continued work in the finding of the optimal image features for AVHRR cloud detection. With this alternative formulation of cloud masking results we believe that the PPS cloud software will allow a more flexible use for different applications. We also believe that the ability to improve the formulation of uncertainty characteristics of the CLARA-A2 climate data record will be greatly enhanced. Additional studies on this topic are foreseen, including inter-comparison with other similar datasets, e.g.,,MODIS Collection 6, a new version of the PATMOS-x dataset and upcoming datasets from the European Space Agency in the Climate Change Initiative project (CCI, see Hollmann et al., 2013; Stengel et al., 2013).

Acknowledgements

This work was carried out within the framework of the Climate Monitoring SAF project and was co-sponsored by EUMETSAT and the Swedish Meteorological and Hydrological Institute (SMHI).

The used CALIPSO–CALIOP data were obtained from the NASA Langley Research Center Atmospheric Science Data Center. The authors are very grateful to Reinout Boers at the Royal Netherlands Meteorological Institute (KNMI) for providing the cloud remote sensing dataset from Cabauw and to Anke Kniffka at the German Meteorological Service (DWD) for providing the SEVIRI cloud masks.

References

Aschbacher, J., & Milagro-Perez, M. P. (2012). The European Earth monitoring (GMES) programme: Status and perspectives. *Remote Sensing of Environment*, *120*, 3–8, http://dx.doi.org/10.1016/j.rse.2011.08.028.

Barret, E. C., & Curtis, L. F. (2013). *Introduction to environmental remote sensing.* Routledge (480 pp).

Boers, R., de Haij, M. J., Wauben, W. M. F., Klein Baltink, H., van Ulft, L. H., Savenije, M., et al. (2010). Optimized fractional cloudiness determination from five ground-based remote sensing techniques. *Journal of Geophysical Research*, *115*, D24116, http://dx.doi.org/10.1029/2010JD014661.

Chen, R., Cao, C., & Menzel, W. P. (2013). Intersatellite calibration of NOAA HIRS CO2 channels for climate studies. *Journal of Geophysical Research*, *118*, 5190–5203, http://dx.doi.org/10.1002/jgrd.50447.

Cracknell, A. P. (1997). *The Advanced Very High Resolution Radiometer (AVHRR).* Taylor and Francis Ltd0-7484-0209-8.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597, http://dx.doi.org/10.1002/qj.828.

Derrien, M., & Le Gléau, H. (2010). Improvement of cloud detection near sunrise and sunset by temporal-differencing and region-growing techniques with real-time SEVIRI. *International Journal of Remote Sensing*, *31*, 1765–1780, http://dx.doi.org/10.1080/01431160902926632.

Derrien, M., & LeGléau, H. (2005). MSG/SEVIRI cloud mask and type from SAFNWC. *International Journal of Remote Sensing*, *26*, 4707–4732.

Dürr, B., & Philipona, R. (2004). Automatic cloud amount detection by surface longwave downward radiation measurements. *Journal of Geophysical Research*, *109*, D05201, http://dx.doi.org/10.1029/2003JD004182.

Dybbroe, A., Karlsson, K. -G., & Thoss, A. (2005a). NWCSAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modeling. Part I: Algorithm description. *Journal of Applied Meteorology*, *44*(1), 39–54.

Dybbroe, A., Karlsson, K. -G., & Thoss, A. (2005b). NWCSAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modeling. Part II: Tuning and validation. *Journal of Applied Meteorology*, *44*(1), 55–71.

Foster, M. J., & Heidinger, A. K. (2013). PATMOS-x: Results from a diurnally corrected 30-yr satellite cloud climatology. *Journal of Climate*, *26*, 414–425, http://dx.doi.org/10.1175/JCLI-D-11-00666.1.

Heidinger, A. K., Evan, A. T., Foster, M., & Walther, A. (2012). A Naïve Bayesian cloud detection scheme derived from CALIPSO and applied within PATMOS-x. *Journal of Applied Meteorology and Climatology*, *51*, 1129–1144.

Heidinger, A. K., Straka, W. C., Molling, C. C., Sullivan, J. T., & Wu, X. Q. (2010). Deriving an inter-sensor consistent calibration for the AVHRR solar reflectance data record. *International Journal of Remote Sensing*, *31*, 6493–6517.

Hollmann, R., Merchant, C. J., Saunders, R., Downy, C., Buchwitz, M., Cazenave, A., et al. (2013). The ESA climate change initiative: Satellite data records for essential climate variables. *Bulletin of the American Meteorological Society*, *94*, 1541–1552, http://dx.doi.org/10.1175/BAMS-D-11-00254.1.

Hutchison, K. D., Roskovensky, J. K., Jackson, J. M., Heidinger, A. K., Kopp, T. J., Pavolonis, M. J., et al. (2005). Automated cloud detection and typing of data collected by the Visible Infrared Imager Radiometer Suite (VIIRS). *International Journal of Remote Sensing*, *20*, 4681–4706.

Inoue, T. (1987). A cloud type classification with NOAA 7 split-window measurements. *Journal of Geophysical Research*, *92*, 3991–4000, http://dx.doi.org/10.1029/JD092iD04p03991.

Karlsson, K. -G., & Dybbroe, A. (2010). Evaluation of Arctic cloud products from the EUMETSAT Climate Monitoring Satellite Application Facility based on CALIPSO–CALIOP observations. *Atmospheric Chemistry and Physics*, *10*, 1789–1807.

Karlsson, K. -G., & Johansson, E. (2013). On the optimal method for evaluating cloud products from passive satellite imagery using CALIPSO–CALIOP data: example investigating the CM SAF CLARA-A1 dataset. *Atmospheric Measurement Techniques*, *6*, 1271–1286, http://dx.doi.org/10.5194/amt-6-1271-2013.

Karlsson, K. -G., Riihelä, A., Müller, R., Meirink, J. -F., Sedlar, J., Stengel, M., et al. (2013). CLARA-A1:The CMSAF cloud and radiation dataset from 28 yr of global AVHRR data. *Atmospheric Chemistry and Physics*, *13*, 5351–5367, http://dx.doi.org/10.5194/acp-13-5351-2013.

Khlopenkov, K. V., & Thrishchenko, A. P. (2007). SPARC: New cloud, snow, and cloud shadow detection scheme for historical 1-km AVHHR data over Canada. *Journal of Atmospheric and Oceanic Technology*, *24*(3), 322–343, http://dx.doi.org/10.1175/JTECH1987.1.

Kopp, T. J., Thomas, W., Heidinger, A. K., Botambekov, D., Frey, R. A., Hutchison, K. D., et al. (2014). The VIIRS Cloud Mask: Progress in the first year of S-NPP toward a common cloud detection scheme. *Journal of Geophysical Research*, *119*, 2441–2456, http://dx.doi.org/10.1002/2013JD020458.

Kossin, J. P., & Sitkowski, M. (2009). An objective model for identifying secondary eyewall formation in hurricanes. *Monthly Weather Review*, *137*, 876–892.

Kriebel, K. T., Gesell, G., Kästner, M., & Mannstein, H. (2003). The cloud analysis tool APOLLO: Improvements and validations. *International Journal of Remote Sensing*, *24*, 2389–2408, http://dx.doi.org/10.1080/01431160210163065.

Loew, A. (2013). Terrestrial satellite records for climate studies: How long is long enough? A test case for the Sahel. *Theoretical and Applied Climatology*, http://dx.doi.org/10.1007/s00704-013-0880-6.

Merchant, C. J., Harris, A. R., Maturi, E., & MacCallum, S. (2005). Probabilistic physically-based cloud screening of satellite infra-red imagery for operational sea surface temperature retrieval. *Quarterly Journal of the Royal Meteorological Society*, *131*, 2735–2755.

Musial, J. P., Hüsler, F., Sütterlin, M., Neuhaus, C., & Wunderle, S. (2014). Probabilistic approach to cloud and snow detection on Advanced Very High Resolution Radiometer (AVHRR) imagery. *Atmospheric Measurement Techniques*, *7*, 799–822, http://dx.doi.org/10.5194/amt-7-799-2014.

Peng, G., Meier, W. N., Scott, D. J., & Savoie, M. H. (2013). A long-term and reproducible passive microwave sea ice concentration data record for climate studies and monitoring. *Earth System Science Data*, *5*, 311–318.

Platnick, S., King, M. D., Ackerman, S. A., Menzel, W. P., Baum, B. A., Riédi, J. C., et al. (2003). The MODIS cloud products: Algorithms and examples from Terra. *IEEE Transactions on Geoscience and Remote Sensing*, *41*, 459–473, http://dx.doi.org/10.1109/TGRS.2002.808301.

Powell, A. M., Qu, J. J., & Sivakumar, M. V. K. (2013). An introduction to satellite-based applications and research for understanding climate change. *Satellite-based Applications on Climate Change* (pp. 1–12). Netherlands: Springer, http://dx.doi.org/10.1007/978-94-007-5872-8_1.

Rossow, W. B., & Schiffer, R. A. (1999). Advances in Understanding Clouds from ISCCP. *Bulletin of the American Meteorological Society*, *80*, 2261–2288.

Schulz, J., Albert, P., Behr, H. -D., Caprion, D., Deneke, H., Dewitte, S., et al. (2009). Operational climate monitoring from space: the EUMETSAT Satellite Application Facility on Climate Monitoring (CM-SAF). *Atmospheric Chemistry and Physics*, *9*, 1687–1709, http://dx.doi.org/10.5194/acp-9-1687-2009.

Stengel, M., Mieruch, S., Jerg, M., Karlsson, K. -G., Scheirer, R., Maddux, B., et al. (2013). The Clouds Climate Change Initiative: Assessment of state-of-the-art cloud property retrieval schemes applied to AVHRR heritage measurements. *Remote Sensing of Environment*, http://dx.doi.org/10.1016/j.rse.2013.10.035.

Tucker, C. J., Pinzon, J. E., Brown, M. E., Slayback, D., Pak, E. W., Mahoney, R., et al. (2005). An Extended AVHRR 8-km NDVI Data Set Compatible with MODIS and SPOT Vegetation NDVI Data. *International Journal of Remote Sensing*, *26*, 4485–4498.

Wauben, W., Bosveld, F., & Klein Baltink, H. (2010). *NubiScope laboratory tests and field evaluation, paper IOM 105(TD 1546) presented at Technical Conference.* Helsinki: World Meteorol. Org.

Winker, D. M., Vaughan, M. A., Omar, A., Hu, Y., Powell, K. A., Liu, Z., et al. (2009). Overview of the CALIPSO mission and CALIOP data processing algorithms. *Journal of Atmospheric and Oceanic Technology*, *26*(11), 2310–2323.