Atmos. Meas. Tech., 6, 1271–1286, 2013
www.atmos-meas-tech.net/6/1271/2013/
doi:10.5194/amt-6-1271-2013

Atmospheric Measurement Techniques
Open Access

On the optimal method for evaluating cloud products from passive satellite imagery using CALIPSO-CALIOP data: example investigating the CM SAF CLARA-A1 dataset

K.-G. Karlsson and E. Johansson

Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

Correspondence to: K.-G. Karlsson (karl-goran.karlsson@smhi.se)

Received: 9 January 2013 – Published in Atmos. Meas. Tech. Discuss.: 1 February 2013
Revised: 22 April 2013 – Accepted: 22 April 2013 – Published: 15 May 2013

Abstract. A method for detailed evaluation of a new satellite-derived global 28 yr cloud and radiation climatology (Climate Monitoring SAF Clouds, Albedo and Radiation from AVHRR data, named CLARA-A1) from polar-orbiting NOAA and Metop satellites is presented. The method combines 1 km and 5 km resolution cloud datasets from the CALIPSO-CALIOP (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation – Cloud-Aerosol Lidar with Orthogonal Polarization) cloud lidar for estimating cloud detection limitations and the accuracy of cloud top height estimations.

Cloud detection is shown to work efficiently for clouds with optical thicknesses above 0.30 except for at twilight conditions when this value increases to 0.45. Some misclassifications of cloud-free surfaces during daytime were revealed for semi-arid land areas in the sub-tropical and tropical regions leading to up to 20 % overestimated cloud amounts. In addition, a substantial fraction (at least 20–30 %) of all clouds remains undetected in the polar regions during the polar winter season due to the lack of or an inverted temperature contrast between Earth surfaces and clouds.

Subsequent cloud top height evaluation took into account the derived information about the cloud detection limits. It was shown that this has fundamental importance for the achieved results. An overall bias of −274 m was achieved compared to a bias of −2762 m when no measures were taken to compensate for cloud detection limitations. Despite this improvement it was concluded that high-level clouds still suffer from substantial height underestimations, while the opposite is true for low-level (boundary layer) clouds.

The validation method and the specifically collected satellite dataset with optimal matching in time and space are suggested for a wider use in the future for evaluation of other cloud retrieval methods based on passive satellite imagery.

1 Introduction

The introduction of the A-Train (i.e. Aqua Train, or sometimes referred to as the Afternoon Train) series of satellites (Stephens et al., 2002) has been a major milestone for cloud research and for satellite meteorology in general. For the first time in history, a series of satellites and sensors are able to provide not only the global coverage of cloud fields and aerosols but also the vertical structure of clouds and aerosols and their respective properties. The vertical probing capability was in particular associated with the launch of the CloudSat (Stephens et al., 2002) and CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation; Winker et al., 2009) missions in 2006 where both satellites are carrying active sensors – a cloud profiling radar (CPR) on CloudSat and a cloud and aerosol lidar (CALIOP) on CALIPSO. These satellites have now produced more than six years of data. This means that, despite the relatively limited spatial sampling capability from the polar orbit (i.e. a consequence of both active sensors operating exclusively in nadir view), the long time series now offers enough measurements to become useful for studying mean conditions (approaching the estimation of climatologies) in parallel to the more obvious use in case-to-case process-oriented studies. Good examples of this and of methods and tools aiming at climatological studies are given by Stubenrauch et al. (2013), Cesana et al. (2012), Liu et al. (2012), Devasthale and Thomas (2011), Chepfer et al. (2010), and Delanoë et al. (2011). In addition, the improved statistical significance of the datasets (i.e. spanning over many years) now make them very useful as "ground truth" datasets for training of cloud algorithms (Heidinger et al., 2012). Similarly, an important application is the use for more thorough evaluation of cloud products from various algorithms based on data from other satellite platforms. This concerns especially those based on data from wider-swath scanning sensors measuring in visible, infrared, and microwave spectral regions (i.e. data from passive imagers) as demonstrated by Holz et al. (2008), Minnis et al. (2008), Reuter et al. (2009), Heidinger and Pavolonis (2009), and Karlsson and Dybbroe (2010).

The information from the CALIPSO-CALIOP lidar is particularly interesting since this sensor is undoubtedly much more sensitive to the presence of clouds in the atmosphere than any other spaceborne sensor at hand. Because of this it has the potential of being used for establishing a firm knowledge of the cloud detection limit for other cloud retrieval methods based on data from other satellite sensors. This aspect is of fundamental importance for securing an optimal and unambiguous use of the information from various cloud retrieval algorithms. An important application in this respect is comparing satellite-derived cloud parameters to information simulated by climate models and numerical weather prediction (NWP) models. To ensure an appropriate inter-comparison here specific tools have been developed. The most well-established tool is the Cloud Feedback Model Inter-comparison Project (CFMIP) Observational Simulation Package (COSP), which is described by Bodas-Salcedo et al. (2011). COSP may simulate cloud datasets from various satellites and sensors from the model state variables. However, this is done differently depending on the sensor. One important piece of information here is to simulate each sensor's ability to detect clouds so that clouds which should be considered as sub-visible for that particular sensor are not taking part in the comparison. Thus, information about cloud detection limits for each sensor needs to be included in COSP.

This paper focuses on the use of CALIPSO-CALIOP data for evaluating the cloud detection limitations of the methods used to derive one particular satellite-derived climate data record: the CLARA-A1 dataset. The acronym stands for the Climate Monitoring Satellite Application Facility (CM SAF – see www.cmsaf.eu and Schulz et al., 2009) Clouds, Albedo, and Radiation dataset from AVHRR data (Karlsson et al., 2013). It is based on global historic Advanced Very High Resolution Radiometer (AVHRR) data from the polar-orbiting NOAA satellites covering the period 1982 until 2009.

While performing the evaluation, several issues arose related to the interpretation of the CALIPSO-CALIOP cloud datasets. This was mainly triggered by the notification of some differences between CALIOP cloud datasets created at different spatial resolutions, differences which are directly related to the applied retrieval methodology. This behaviour differs to a large extent from the results of methods used to interpret clouds at different horizontal resolutions in passive imagery. We claim that these differences have not been accounted for in some previous studies using CALIPSO datasets for evaluating cloud datasets from passive imagery. Also, the philosophical question on how to define the upper boundary (cloud top) of a cloud needs specific attention. These issues may all be critical to the final results and we want to highlight these aspects in this paper.

Section 2 introduces the two datasets to be inter-compared, and Sect. 3 elaborates further on the problems associated with this comparison and suggests a method on how to deal with them. This is followed in Sect. 4 by the presentation of results on the performance of cloud detection, its regional dependency and the apparent cloud detection limit in terms of the thinnest (in the cloud optical thickness sense) clouds being detected. Section 5 presents results for the cloud top height determination, taking into account the deduced cloud detection limitations. Finally, Sect. 6 concludes and gives some further discussion on the optimal method to be used for cloud parameter validation.

2 Data

2.1 The investigated dataset: CLARA-A1

The CLARA-A1 dataset of global cloud products retrieved by CM SAF cloud retrieval methods is based on reduced resolution (approximately 4 km) global area coverage (GAC) AVHRR data spanning the time period 1982–2009. The total set of cloud products includes cloud fractional cover, cloud top level, cloud optical thickness, cloud phase, liquid water path, ice water path, and joint cloud property histograms. Here, we will concentrate on the evaluation of the first two products. For a full description of the dataset the reader is referred to Karlsson et al. (2013).

The cloud fractional cover (CFC) product is derived directly from results of a cloud screening, or cloud masking, method. CFC is defined as the fraction of cloudy pixels per grid square compared to the total number of analysed pixels in the grid square. Fractional cloud cover is expressed in percent. The product is calculated using the Nowcasting Satellite Application Facility (NWC SAF) PPS (Polar Platform System) cloud mask algorithm (see http://www.nwcsaf.org/ for details on the NWC-SAF project). The algorithm (detailed by Dybbroe et al., 2005) is based on a multi-spectral thresholding technique applied to every pixel of the satellite scene. Several threshold tests may be applied (and must be passed) before a pixel is assigned to be cloudy or cloud free. Thresholds are determined from present viewing and illumination conditions and from the current atmospheric state (prescribed by data assimilation products from numerical weather prediction models – here, the ERA-Interim dataset; see Dee et al., 2011, and http://www.ecmwf.int/research/era/do/get/era-interim). Also, ancillary information about surface status (e.g. land use categories and surface emissivities) is taken into account. Thus, thresholds are dynamically defined and therefore unique for each individual pixel.

The cloud top level (CTO) product is also derived using NWC SAF PPS algorithms. The product is abbreviated CTO because it can be expressed in three alternative forms: cloud top height (in meters), cloud top pressure (in hPa), and cloud top temperature (in Kelvin). In this paper we concentrate on the cloud top height version since this quantity is directly measured by the CALIPSO-CALIOP sensor. Consequently, we will refer to the product as either CTO or cloud top height.

Cloud top processing is subdivided using two separate algorithms, one for opaque and one for fractional and semi-transparent clouds, and it is applied to all cloudy pixels as identified by the PPS cloud mask product. The opaque algorithm uses simulated cloud-free and cloudy top of atmosphere (TOA) 11 µm radiances which are compared and matched to measured radiances. Cloudy radiances are simulated assuming "black-body" clouds at various levels. The semi-transparent algorithm (described by Korpela et al., 2001) is applied to all pixels classified as semi-transparent cirrus or fractional water cloud. This classification is based on the analysis of brightness temperature differences of the 11 µm and 12 µm (split window) channels, noting that this difference is generally small or negligible for opaque clouds. A histogram technique is applied based on the construction of two-dimensional histograms using AVHRR 11 and 12 µm brightness temperatures composed over larger segments (typically 32 × 32 pixels). By an iterative procedure a polynomial curve (simulating the arc shape) is fitted to the histogram-plotted values from which the cloud top temperature and pressure (taken from ERA-Interim profiles) are retrieved.

Obviously, only a small fraction of the CLARA-A1 dataset may be evaluated using CALIPSO-CALIOP data (limited to years 2006–2009). However, it is believed that results should largely be valid also for results before these years provided that a reasonably large number of collocations can be found and considering that the AVHRR instrument has not undergone drastic changes throughout the years. The basis for the comparison is the use of original PPS cloud mask and cloud top height products for full orbit swaths (about 13 000 scan lines) which are collocated with CALIPSO-CALIOP orbits using specific matching criteria (further described in Sect. 3).

In the remainder of the text we will use the notation CLARA-A1/PPS to emphasise that we examine the performance of the PPS cloud mask and PPS cloud top height products for the PPS version used for defining the CLARA-A1 dataset.

2.2 The reference validation dataset: CALIPSO-CALIOP cloud products

The Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite was launched in April 2006 together with CloudSat. The satellite carries the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) and the first data became available in August 2006. CALIOP provides detailed profile information about cloud and aerosol particles and corresponding physical parameters. CALIOP measures the backscatter intensity at 1064 nm, while two other channels measure the orthogonally polarised components of the backscattered signal at 532 nm. The horizontal resolution of each single field of view (FOV) is in practice 333 m (but the true FOV is actually not larger than about 100 m) and the vertical resolution is 30–60 m. The CALIOP cloud product reports observed cloud layers, i.e. all layers observed until signal becomes too attenuated. In practice the instrument can only probe the full geometrical depth of a cloud if the total optical thickness is not larger than a certain threshold (somewhere in the range 3–5). For optically thicker clouds only the upper portion of the cloud is sensed.

CALIOP products have been retrieved from the NASA Langley Atmospheric Science Data Center (ASDC, http://eosweb.larc.nasa.gov/JORDER/ceres.html). We have used the Lidar Level 2 Cloud and Aerosol Layer Information product version 3.01 and the associated information from the Lidar Level 2 Vertical Feature Mask product. Regarding the latter it is important to notice the use here of the categorisation of low-level, medium-level and high-level clouds introduced by the International Satellite Cloud Climatology Project (ISCCP). This categorisation uses pressure levels of 680 hPa and 440 hPa to separate the three categories. We will use this classification later for separating results of cloud top height determinations between the three vertical groups of clouds.

The CALIOP products are defined in three different versions with respect to the along-track resolution ranging from 333 m (individual footprint resolution), 1 km and 5 km. The two latter resolutions are consequently constructed from several original footprints/FOVs. This allows a higher confidence in the correct detection and identification of cloud and aerosol layers compared to when using the original high-resolution profiles. For example, the identification of very thin cirrus clouds is more reliable in the 5 km dataset than in the 1 km dataset since signal-to-noise levels can be raised by using a combined dataset of several original profiles. Also to notice is that in the derivation of the 5 km dataset subsequent averaging procedures also at coarser resolution (e.g. 20 km and 80 km) have been made, meaning that the thinnest clouds are found on these larger scales (20 km or 80 km), although correctly described in the 5 km resolution representation. For a full description of the retrieval methodology, see Vaughan et al. (2009) and Winker et al. (2009).

The natural choice of product resolution for the validation of 4 km AVHRR GAC products is to use the CALIOP 5 km dataset. The CALIOP 5 km dataset also offers estimation of cloud optical thicknesses of individual layers (not available for finer resolution FOVs), which is a very attractive feature since it means that this offers a possibility to analyse cloud detection limits quantitatively.

3 Methodology

3.1 Concern about differences of CALIOP and AVHRR 1 km and 5 km cloud datasets

One of the central features of CALIOP cloud retrieval algorithms (as outlined by Vaughan et al., 2009) is to take maximum advantage of the possibility to increase signal-to-noise levels by averaging results from high-resolution fields of view into coarse-resolution fields of view. By doing this it is possible to identify cloud layers which are too thin to be detected in the original fine FOV resolution of 333 m because of high noise levels. This means that in theory the optically thinnest cloud layers will be found from analysis of CALIOP data averaged even up to 80 km segment resolution. Since a lot of concern in climate research for many years has been given to the potential impact of thin and subvisible cirrus clouds (Stephens et al., 1990), this capability of the CALIPSO mission has been very much highlighted and numerous reports have been published on this subject (two examples are Haladay and Stephens, 2009, and Virts et al., 2010).

However, it is possible that the focus on the thin cloud identification may have led to some drawbacks for the prospect of evaluating the performance of other cloud algorithms, e.g. algorithms based on data from passive imagery. At least, it is important to notice that globally retrieved cloud results achieved when reducing the CALIOP horizontal resolution from, e.g., 1 km to 5 km, will behave very differently compared to the case of using AVHRR-based datasets with similar resolutions. AVHRR radiances at the 5 km scale (GAC resolution) are composed by linear averaging over several original 1 km FOVs, while the averaging of CALIOP measurements and the subsequent cloud retrieval is done in a more complicated manner (see Winker et al., 2009). For CALIOP the basic aim is to detect more of the really thin cloud layers when averaging at coarser resolutions. But during this process, contributions from some highly reflective boundary layer clouds at the original FOV resolution are removed in order to not completely dominate over weaker signals. Thus, we both add and lose clouds when shifting from the 1 km CALIOP datasets to the 5 km representation. Normally the total cloud fraction (abbreviated CFC below) summed over a full orbit of the 5 km dataset is higher than the 1 km dataset, but not for all cases. For these latter ones it means that the orbit then includes quite a high fraction of sub-resolution cloudiness detected only at the finest horizontal resolution. To exemplify, we have the following three possible cases:

1. $\mathrm{CFC}_{1\mathrm{km}} < \mathrm{CFC}_{5\mathrm{km}}$: several new thin cloud layers are detected at coarser resolutions with only a small loss of clouds at finer resolutions.
2. $\mathrm{CFC}_{1\mathrm{km}} = \mathrm{CFC}_{5\mathrm{km}}$: new cloud layers are detected at coarser resolutions but just balanced by the loss of clouds at finer resolutions.
3. $\mathrm{CFC}_{1\mathrm{km}} > \mathrm{CFC}_{5\mathrm{km}}$: higher loss of clouds at finer resolutions compared to newly detected thin cloud layers at coarser resolutions.

Then the question is, will this different behaviour of results have any consequences for the case of comparing them to retrievals from passive imagery, and then in particular to results in the AVHRR 1 km and 4 km (GAC) resolutions?

In the case of the newly added thin cloud layers at the CALIOP 5 km resolution it all depends on if the clouds are thick enough to be detected in the AVHRR GAC resolution. This can be investigated by comparing the degree of detection efficiency with the retrieved cloud optical thickness of those layers (we will do that as outlined in Sect. 3.3).

More serious is the fact that quite a large fraction of the highly reflective but sub-resolution cloud elements being removed when constructing the 5 km CALIOP dataset might still be detected at the 4 km AVHRR GAC resolution. This has to do with non-linear effects in the original 1 km AVHRR FOVs, meaning that a high total reflectivity can be achieved for that FOV even if the fractional coverage (i.e. geometric size) of the cloud element in the FOV is very small. This, in turn, may also then affect results averaged to the 4 km GAC resolution. The identification of such partly cloudy FOVs is important since it may affect the accuracy of, e.g., SST retrievals. Thus, the fact that such cloud elements are not included in the CALIPSO 5 km dataset might give an unwanted bias to the results. More clearly, we want to avoid labelling these cloudy or cloud-contaminated pixels wrongly as cloud-free pixels, which otherwise easily could be concluded if relying on the CALIOP 5 km dataset.

Because of these differences in how to delineate global cloudiness by use of data from active and passive sensors, we suggest that one should try to use the existing information from both the 1 km and 5 km CALIOP datasets in a combined way to estimate the cloud situation. Especially, contributions to global cloudiness from highly reflective boundary layer clouds must be taken into account also at scales compatible with the coarse AVHRR GAC resolution. This means that at least a part of the information about these clouds that was suppressed or lost when preparing the CALIOP 5 km resolution datasets needs to be restored. In the next subsection we introduce a method which we believe takes the best from both CALIOP datasets, thereby defining a better reference dataset for AVHRR GAC data than the nominal 5 km CALIOP dataset.

3.2 Proposed evaluation of cloud amounts using combined 1 km and 5 km CALIOP datasets

The principle for constructing an optimal cloud dataset (i.e. optimal for the inter-comparison with cloud datasets from passive imagery) from CALIOP 1 km and 5 km datasets is based on the fact that that thick (opaque) clouds are well described by the CALIOP 1 km dataset while thin clouds are best described by the 5 km dataset. Here, we also assume that the contribution from highly reflective boundary layer clouds (as detected at the original FOV resolution of 333 m) is correctly represented also in the 1 km dataset. In other words, we think that the potentially "lost" thick clouds in the 5 km dataset after averaging (as described in the previous section) are most likely included in the 1 km dataset. Thus, we can construct a new merged cloud dataset by going through the following rather simple post-processing steps:

- Step 1: Compute a preliminary cloud fraction (CFC$'$) at 5 km segment resolution from the 1 km segments.

  (Thus, CFC$'$ can now take the six discrete values of 0 %, 20 %, 40 %, 60 %, 80 % and 100 % for every 5 km segment.)

- Step 2a: Set 5 km data to CLOUDY if CFC$' > 50$ %.

  (If a cloud layer was missing in the 5 km dataset but covered more than 50 % of the involved 1 km segments, then a new layer will now be added.)

- Step 2b: Set 5 km data to CLOUD-FREE if CFC$' < 50$ %.

  (If only a few 1 km columns are cloudy, we should not consider the full 5 km segment as cloudy.)

- Step 3: If a cloud layer exists at 5 km segment resolution but NOT at any 1 km segment,

  ⇒ new thin layer detected!

  ⇒ set 5 km segment to CLOUDY (or, rather, keep the 5 km dataset unchanged).

By these simple steps we believe that we have achieved our goals even if steps 2a, 2b and 3 still mean that there are undetermined retrieved values of both CFC and cloud optical thickness. For example, in step 2a we might add (or restore) a new 5 km layer, but we have no way of giving this new layer a retrieved value of cloud optical thickness (since this quantity is only retrieved for the 5 km segment dataset and not for the 1 km segment dataset). We have "solved" this by prescribing the new value to optical thickness 1.0. This is just to show that we believe that this cloud layer should not belong to the category of very thin cloud layers, thus assuring that it will not be included in subsequent cloud detection limit studies focusing on clouds with low optical thickness values. Furthermore, step 2b means that there could be cases when at 1 km segment resolution there are only one or two cloudy columns while at 5 km segment resolution we have a cloud layer (implying that it covers the entire 5 km segment). This cloudy 5 km segment will now be removed, which maybe could be questioned. In some sense, we then say that in this case we rely more on the 1 km dataset than on the original 5 km dataset. This is at least partly justified for passive imagery not being capable of detecting very thin clouds. The ambiguity comes also from the consideration that it could theoretically be a cloud layer that is only partly detected in the 1 km segments within the 5 km segment, while in reality it is actually covering the entire 5 km segment. We simply have not enough information here to judge what the truth is, so we have to stay with the simple interpretation resulting from steps 2a and 2b. We actually think this uncertainty is marginal in comparison with the general uncertainty about the true CFC within the 5 km segment concerning the entirely "new" thin cloud layers appearing in the 5 km dataset that are detected after the averaging procedure for reducing the signal-to-noise levels. These new cloud layers are assumed to cover the entire 5 km segment, but there is actually no way of estimating the true CFC within the 5 km segment resolution. It is possible that these clouds only cover a fraction of the 5 km segment. In some sense, it is even possible that these interpreted thin cloud layers might be just broken cloud layers which are optically relatively thick but just cover a small fraction of the 5 km segment (as suggested by Abhay Devasthale, personal communication, 2012).

Despite these remaining ambiguities, we believe that the proposed approach yields reasonable results which are more consistent and robust than results based exclusively on either 1 km or 5 km segment data.

3.3 Method for evaluating cloud detection efficiency and the cloud detection limit

The merged new 5 km CALIOP dataset (compiled according to the method described earlier) now includes information about cloud layers at each 5 km segment, and for each cloud layer an estimated cloud optical thickness is given. However, it is important to remember that for the lowermost layer it might be only a minimum value since the entire cloud layer might not be penetrated by the lidar signal. We may now evaluate the cloud detection efficiency of the methods used to derive the CLARA-A1 dataset either in a direct inter-comparison (i.e. using all CALIOP-detected cloud layers) or by applying a filtering mode where cloudy columns having an integrated cloud optical thickness below a certain value are treated as being cloud free. In this way we should be able to quantify the cloud detection limit of the CLARA-A1 dataset. For this purpose we have filtered the CALIOP dataset in cloud optical thickness steps of 0.05 in the range 0.0–0.5 and in steps of 0.1 in the range 0.5–1.0.

For quantifying results we have used the following statistical scores:

1. mean error (Bias),

2. root mean square error (RMS),
3. probability of detection (POD) for both cloudy and cloud-free conditions,
4. false alarm rate (FAR) for both cloudy and cloud-free conditions,
5. hit rate (HR), and
6. Kuiper's skill score (KSS).

For the estimation of cloud occurrence or CFC, we have used a binary representation of the results (i.e. cloud cover = 1 for cloudy conditions and cloud cover = 0 for cloud-free conditions) for each individual pixel or FOV. Consequently, results are accumulated over the full matchup track to get a mean CFC (according to Eq. 1 below) and the associated Bias and RMS values. As a final step, all matchup results for all matched orbits are accumulated and averaged.

$$\mathrm{CFC} = \frac{\sum \mathrm{cloudy}}{\sum \mathrm{all\ pixels}} \tag{1}$$

For the remaining four quantities we have used the following definitions (referring to notations in the contingency matrix in Table 1):

$$\mathrm{POD}_{\mathrm{cloudy}} = \frac{d}{c+d} \tag{2}$$

$$\mathrm{POD}_{\mathrm{cloud\text{-}free}} = \frac{a}{a+b} \tag{3}$$

$$\mathrm{FAR}_{\mathrm{cloudy}} = \frac{b}{b+d} \tag{4}$$

$$\mathrm{FAR}_{\mathrm{cloud\text{-}free}} = \frac{c}{a+c} \tag{5}$$

$$\mathrm{HR} = \frac{a+d}{a+b+c+d} \quad \text{where} \quad 0 \leq \mathrm{HR} \leq 1 \tag{6}$$

$$\mathrm{KSS} = \frac{a \cdot d - c \cdot b}{(a+b)\cdot(c+d)} \quad \text{where} \quad -1 \leq \mathrm{KSS} \leq 1 \tag{7}$$

The POD and FAR quantities estimate how efficient CLARA-A1/PPS is in determining either cloudy or cloud-free conditions. Naturally, we want POD values to be as high as possible and FAR values to be minimised. The HR is a condensed measure of the overall efficiency of cloud detection. Finally, the KSS quantity is a complementing measure since the HR can sometimes be misleading because it is heavily influenced by the results for the most common category. For example, if a case is almost totally cloud free but all the few cloudy portions are misclassified as cloud free by CLARA-A1/PPS, then the HR score would still be high. A more reasonable measure in such a condition is the KSS that at least to some extent punishes misclassifications even if they are in a small minority of all the studied cases. The KSS tries to answer the question of how well the estimation separated 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.

Table 1. Contingency matrix for the two different satellite observations.

	CLARA-A1/PPS AVHRR		
CALIPSO-CALIOP	Scenario	Cloud-free	Cloudy
	Cloud-free	a	b
	Cloudy	c	d

The use of a wide range of statistical scores should be seen in the light of the fact that it is not obvious which of the scores that best describes the cloud detection limit of a cloud screening method. We hope that a closer look at the results would suggest which score or which combinations of scores are the most optimal for this particular aspect.

In addition, we have also separately studied the cloud detection efficiency over various regions of the Earth and the performance as a function of time of day. Here, we have used the twilight category defined as valid for solar zenith angles between 80 and 95 degrees with day and night categories either having lower or higher solar zenith angles, respectively. Concerning the study of the geographical variation, we have separated results according to geographical regions defined in Table 2. These results have also been further separated for land and ocean conditions using a land mask.

3.4 Method for evaluating accuracy of cloud top height products

Considering that there is a cloud detection limit (expressed in terms of minimum cloud optical thickness, τ_{min}), for a dataset such as CLARA-A1, the evaluation of corresponding cloud top height products must take this into account. It is of course trivial that clouds which are not detected cannot be given a valid cloud top height. But also for clouds that are detected, the effect of cloud detection limitations must be taken into account in some way. For example, if a very thin cloud layer (not detectable by CLARA-A1/PPS but present in the CALIOP dataset) is overlaying a thicker cloud layer (detected by CLARA-A1/PPS), then one should actually neglect this uppermost layer when doing cloud top height validation. Even in the case when we have just detected one single cloud layer, the uppermost part of that layer (with

Table 2. Definition of geographical sub-regions.

Region notation	Latitude band
TROPICAL	Latitudes (±) 0–15 degrees
MID-LATITUDE	Latitudes (±) 15–45 degrees
HIGH-LATITUDE	Latitudes (±) 45–75 degrees
POLAR	Latitudes (±) 75–90 degrees

integrated optical thickness of the minimum detection value) should theoretically be discarded. One could actually claim that a representative cloud top height would be even lower since the measured radiance for the AVHRR instrument is a mix of contributions from several altitudes below the cloud top unless the cloud is optically very thick. In other words, the AVHRR-representative cloud top is the radiatively efficient cloud top rather than the physical or geometrical cloud top. Thus, for an AVHRR-detected cloud layer a representative cloud top height should rather be the mid-layer altitude of the CALIOP-detected layer than the uppermost cloud layer boundary. This can also be motivated for the clouds that are not fully penetrated by the CALIOP lidar signal. When the cloud is optically too thick, the CALIOP cloud layer will describe only the uppermost part of the cloud, and the mid-layer value here would then still be representative for the AVHRR-detected (radiatively efficient) cloud top, in our opinion.

Taking these aspects into account, we have applied the following criteria for evaluating the cloud top height:

- The uppermost cloud layer (or layers) in the CALIPSO dataset is disregarded if the cloud optical thickness (summed if more than one layer) does not exceed the minimum cloud optical thickness (τ_{min}).
- Cloud top height is interpreted as the mid-level of the uppermost CALIOP cloud layer assumed to be detected in CLARA-A1, i.e. the mean of the cloud base and the cloud top altitude for that layer.

3.5 The collocated NOAA-AVHRR and CALIPSO-CALIOP dataset

We have adopted the following strategy for collecting the collocated NOAA-AVHRR and CALIPSO-CALIOP cloud observations to be inter-compared:

- Select the best complete collocations or matches, i.e. entire global orbits with minimum observation time differences between NOAA-18 and A-Train/CALIPSO for every month where we have CALIPSO data available (in practice from October 2006 until December 2009).

Observe that the choice of NOAA-18 is explained by the fact that this satellite is placed in almost the same orbital plane as the Aqua-Train satellites with approximately the same Equator-crossing time. Thus, if choosing matches where the orbital tracks cross simultaneously (denoted Simultaneous Nadir Observations – SNOs) – in this case limited to within only 12 s – we can get measurements matched in near-nadir observation conditions for an entire global orbit and with a maximum time difference between observations of less than approximately 2 min for positions farthest away from the SNO point. Using this criterion we may theoretically get close to 3 such optimal matches each month. However, due to some losses of data (i.e. cases where we could not find both 1 km and 5 km CALIOP data) we ended up with a total of 99 global orbits evenly distributed over the period (see total coverage in Fig. 1). The geographical coverage is good, but we can see that for some regions (e.g. over South America, the North Atlantic Ocean, Africa and parts of the Pacific Ocean) the orbit coverage is less frequent than over other regions due to some loss of data. An example of one of the resulting orbits is shown in Fig. 2. The corresponding plot of CALIOP-observed cloud layers (green) and CLARA-A1/PPS cloud top height results (blue) is given in Fig. 3. Only small deviations (less than 10 degrees) from the nadir view are achieved for the matched AVHRR observations during such an orbit.

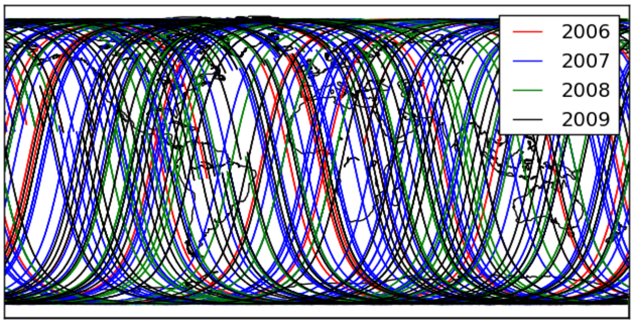

Fig. 1. Total global coverage of 99 matched CALIPSO-NOAA-18 orbits in the period October 2006 to December 2009. Different colours refer to different years.

The 99 collocated orbits resulted in a total of 725 900 matched AVHRR/CALIOP observations within a 2 min observation time difference (valid within each complete orbit) for the calculation of statistics and scores.

4 Results

4.1 Cloud screening efficiency

4.1.1 Overall results based on all collocations

A way to estimate the cloud detection efficiency is to plot and analyse various statistical scores as a function of the CALIOP-filtered cloud optical thickness. For clarity, we repeat that the filtering process means that whenever CALIOP-derived total cloud optical thickness in the column/FOV falls

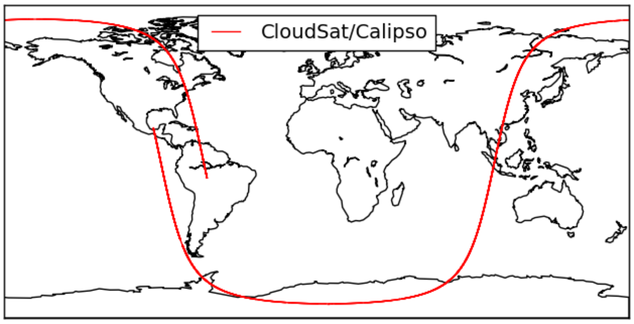

Fig. 2. Trajectory for one selected matched CALIPSO-NOAA-18 orbit from 6 October 2006 with first matched observation at 18:00 UTC (over South America).

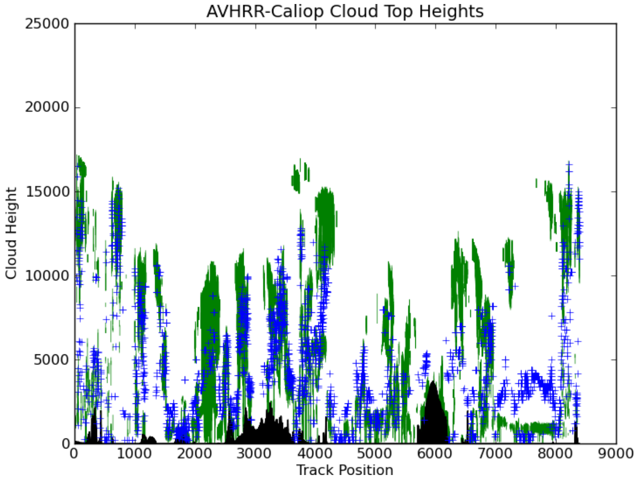

Fig. 3. Matched CALIPSO-CALIOP cloud mask (green) and NOAA-AVHRR cloud top height values (blue, in metres) from CLARA-A1/PPS for the same global orbit as shown in Fig. 2. Track position is given in number of AVHRR GAC pixels (to be multiplied by 4 to get roughly the distance in km). Significant topographic features are seen in black at track positions 6000 (Antarctica) and 3000 (Russia/China).

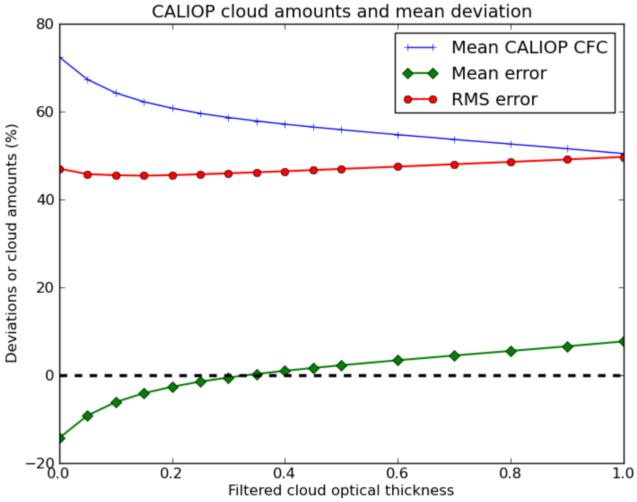

Fig. 4. Mean CALIOP cloud occurrences (CFC), mean error and RMS error as a function of filtered cloud optical thickness (explained in text) for CLARA-A1/PPS cloud masks calculated from 99 global matches of NOAA-18 with CALIPSO between October 2006 and December 2009.

below a specific cloud optical thickness threshold we will treat the observation as being cloud free when calculating the statistics. Figures 4–6 show the results for all statistical parameters described in Sect. 3.3 based on all collocated orbits. The use of a variety of statistical scores will help us later in deducing an appropriate way of defining the cloud detection limit (outlined in Sect. 4.1.2).

The basic mean error and RMS error quantities are shown together with the resulting total cloud fraction (i.e. percentage cloudy segments of all segments) in the CALIPSO-CALIOP dataset in Fig. 4. We notice that after filtering clouds having optical thicknesses up to 1.0, the total cloud fraction for CALIOP reduces from approximately 73 % to 50 %. At the same time the mean error changes from −14 % to +8 % and the RMS changes from 47 % to 50 %. Based on mean error results alone one might conclude that the optimal agreement is reached after filtering all cloudy columns with optical thickness values below 0.35. The fact that mean errors become positive for higher filtered optical thickness thresholds only means that some cloudy CALIOP columns are now treated as being cloud free even if they were detected successfully by CLARA-A1/PPS, thus giving a positive mean error. When comparing with Fig. 6 showing hit rates and Kuiper's skill scores, we see that the skill now peaks at slightly lower values of the filtered cloud optical thickness threshold, namely at about 0.2 for hit rate and 0.1 for Kuiper's skill score. This shows that from these different statistical measures it is not easy to come to a very clear conclusion about cloud detection limits.

However, results of POD and FAR in Fig. 5 also reveal some further features of CLARA-A1/PPS results which are not evident in Fig. 4 or Fig. 6 and which are not directly related to how thin or thick clouds are. We first note that the FAR quantity for clear segments initially reduces rapidly with increasing values of the filtered cloud optical thickness. This is what we should expect if very thin cloud layers are not detected by CLARA-A1/PPS, i.e. scores would improve if these CALIOP observations were also treated as being cloud free. Similarly, POD for cloudy conditions improves with increasing values of filtered optical thickness. However, more serious is the observation that the FAR quantity for cloudy conditions amounts to 8 % initially for unfiltered CLARA-A1/PPS results. Thus, we seem to have a significant misclassification of clear segments labelled as cloudy, which also explains why POD results for clear conditions are relatively far away from 100 % in the unfiltered mode. This shows that the cloud detection efficiency cannot be judged

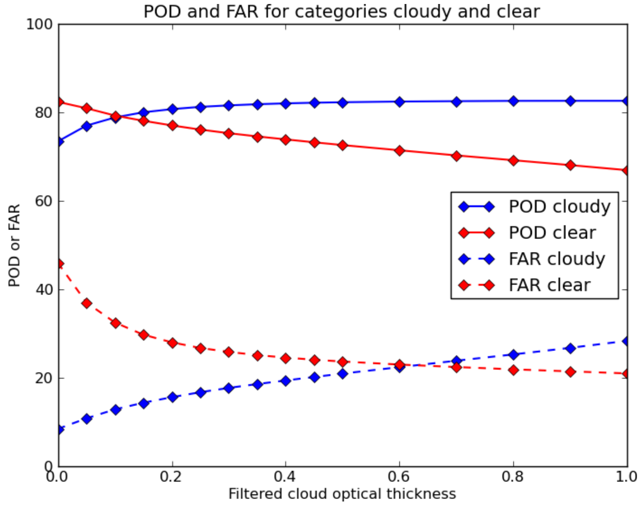

Fig. 5. Same visualisation as in Fig. 4 but for probability of detection (POD) and false alarm rates (FAR) for cloudy and clear categories.

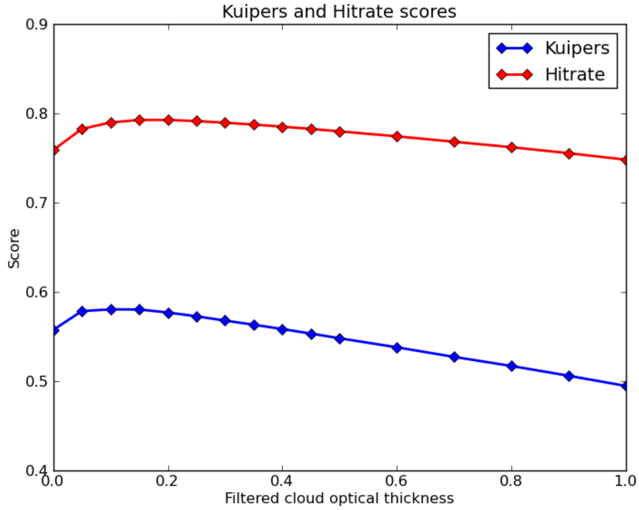

Fig. 6. Same visualisation as in Fig. 4 but for the hit rate and Kuiper's skill scores.

solely from studies of how thin or thick clouds are. It is clear that there are also Earth surfaces that have appearances that resemble those of clouds regardless of whether clouds are thin or thick. The most obvious example is the case when interpreting a cold ground surface at night as being a cloud when using an inappropriate value of the assumed ground surface temperature (i.e. being too warm). Another case is when a bright land surface (e.g. desert) is mistaken for a cloud because of using inappropriate (i.e. too dark) surface reflectance thresholds. We conclude that some measures must be taken to try to remove the influence from this latter type of misclassifications which could be interpreted as a constant bias in our results not related to the thickness of the clouds.

4.1.2 Results after excluding misclassified cloud-free surfaces

The most obvious way of trying to isolate the results depending mainly on the cloud optical thickness value of clouds would be to remove or ignore all cases being misclassified as cloudy in the completely unfiltered mode. In other words, let us restore the cloudy CLARA-A1/PPS pixels in evidently cloud-free CALIOP segments to become clear. Thus, these 8 % of the cases in the FAR category for cloudy pixels in the unfiltered mode are now being correctly classified as clear. Ideally, we should also try to exclude or ignore the oppositely misclassified cases, i.e. when clouds are misclassified as clear regardless of their optical thickness (i.e. for non-separability reasons meaning that cloud-free surfaces cannot easily be separated from a cloud). However, these cases are not as easily identified as the cases of misclassified clear pixels. More clearly, they can occur at any cloud optical thickness, meaning that these cases are inherently mixed with all the cases we actually aim at, namely those cases when cloud detection will clearly depend on the cloud optical thickness value. In that sense these misclassifications exist as an almost constant bias in our results. They are best identified in Fig. 5 as explaining why the $\mathrm{FAR}_{\mathrm{clear}}$ value is still high (20 %) even at the maximum filtered cloud optical thickness of 1.0. This means that in 20 % of all cases in which CLARA-A1/PPS gives a cloud-free result there are actually clouds in reality and they have cloud optical thickness values higher than 1.0. Further details on when these misclassifications occur will be revealed in forthcoming Sects. 4.1.3 and 4.1.4.

The revised results for the statistical scores (after ignoring misclassified clear cases labelled as cloudy) are now shown in Figs. 7–9. We notice in Fig. 7 that now the mean error quantity does not reach the zero level until a filtered cloud optical thickness of 0.7. This is a high value and indicates that the CLARA-A1/PPS cloud screening method is generally rather cloud conservative. But it doesn't necessarily mean that the cloud detection limit is best described by this value of optical thickness. Rather we should use a quantity which is more uniquely decided by and dependent on the filtered cloud optical thickness. The two quantities that best fit this description seem to be POD for cloudy conditions and FAR for clear conditions. The first quantity improves with increasing filtered optical thicknesses until "all" clouds are detected. The fact that the $\mathrm{POD}_{\mathrm{cloudy}}$ saturation level does not reach 100 % means that the difference with respect to the 100 % level represents all those cases where clouds remain undetected regardless of their cloud optical thickness. Similarly, the FAR for clear conditions behaves in the same way where the apparent convergence level defines the same misclassified cases (i.e. that portion of the CLARA-A1/PPS clear cases that actually are undetected clouds even for higher optical depths).

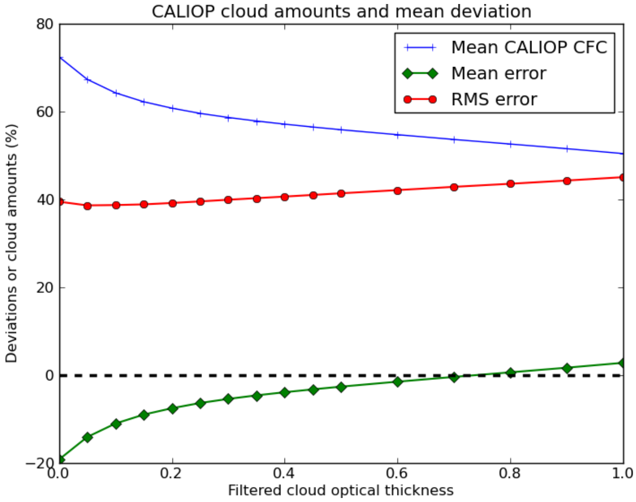

Fig. 7. Same as Fig. 4 but after treating CLARA-A1/PPS misclassified clear pixels as truly clear and not as cloudy (see motivation in text).

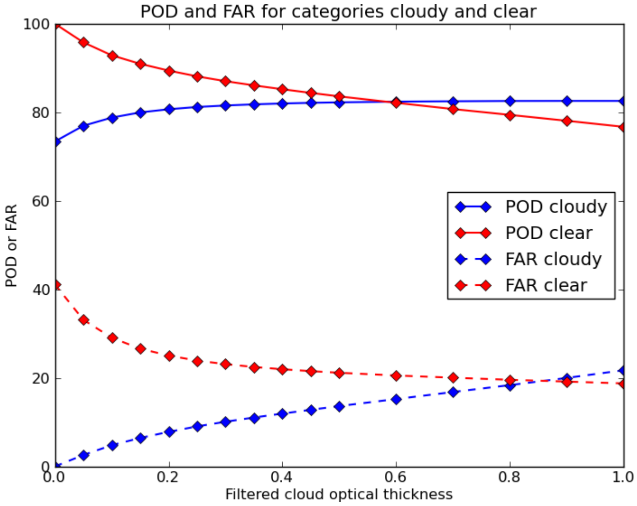

Fig. 8. Same as Fig. 5 but after treating CLARA-A1/PPS misclassified clear pixels as truly clear and not as cloudy (see motivation in text).

The significant increase at optical thicknesses lower than 0.7 of the POD quantity for cloudy conditions, and the corresponding decrease of the FAR quantity for clear conditions, in Fig. 8 shows that much thinner clouds than what the mean error quantity indicates are indeed detected. A better value of the minimum optical thickness detected could then be suggested to be derived from the rate of change of the mentioned POD and FAR quantities for cloudy and clear conditions, respectively. The minimum optical thickness to be determined would then be the value found when the improvement of these two quantities has slowed down or "saturated" (i.e. approaching constant or almost constant values). The interpretation of this value would be that at this cloud optical thickness all clouds are detected, unless other problems not related to how thick clouds are exists. For lower cloud optical thicknesses some clouds will be detected, but for very low optical thicknesses no clouds at all will be detected. Referring to the lacking clear recommendations on how to define the cloud detection limit, we suggest the following definition for finding this cloud optical thickness limit based on the principles outlined above:

$$\left(\frac{\delta \mathrm{POD}_{\mathrm{cloudy}}}{\delta \tau} + \frac{\delta \mathrm{FAR}_{\mathrm{clear}}}{\delta \tau}\right) < 1\,\%. \qquad (8)$$

This means that we will interpret the cloud detection limit as the first (i.e. lowest) cloud optical thickness value where this inequality is fulfilled while checking for higher and higher filtered cloud optical thickness values. To repeat, the ambition is to construct a robust measure that tells us where the rate of change of the POD and FAR quantities have decreased to very small values. The value 1 % is perhaps rather arbitrarily chosen, but it was considered reasonable as a value for representing the case when the two quantities had reached almost constant values. Consequently, when applying this definition we get the overall cloud detection limit at optical thickness 0.35.

Sensitivity studies testing the effect of using threshold values in Eq. (8) of 0.5 % and 2 %, respectively, showed that the deduced limit then varies between 0.25 and 0.5. However, both these options were found less useful. For the higher value (2 %) the rate of change is still quite high, indicating that we are still far from any saturation of the values. As a contrast, the 0.5 % threshold gives a result (optical thickness 0.5) which is quite far away from the point where the visual inspection of the figures (e.g. Fig. 8) would suggest that we have reached a decent saturation. Consequently, we believe that the 1 % threshold is a reasonable value.

An additional sensitivity test was performed to check the importance of the cloud fraction limit in tests 2a and 2b as previously described in Sect. 3.2. This cloud fraction limit based on the 1 km columns in a 5 km segment was here set to 50 % in deciding whether the 5 km segment should be considered cloudy or cloud free. We tested using a slightly lower value (30 % = at least 2 cloudy columns) and a slightly higher value (70 % = at least 4 cloudy columns). Conclusions from these tests were that the method is robust in that more or less the same results were achieved regardless of these applied thresholds. In practice, this means that individual values of curves in the figures are raised or lowered when changing this threshold but that the actual shapes of the curves (in particular the rate of changes) remains practically the same. Thus, we consider our results as quite robust.

When comparing with Fig. 9 we see that the optical thickness limit of 0.35 is also relatively close to where the maximum of the hit rate score occurs (although peaking at slightly lower optical thickness values). However, the Kuiper's score does not really help us here. Remembering that this score

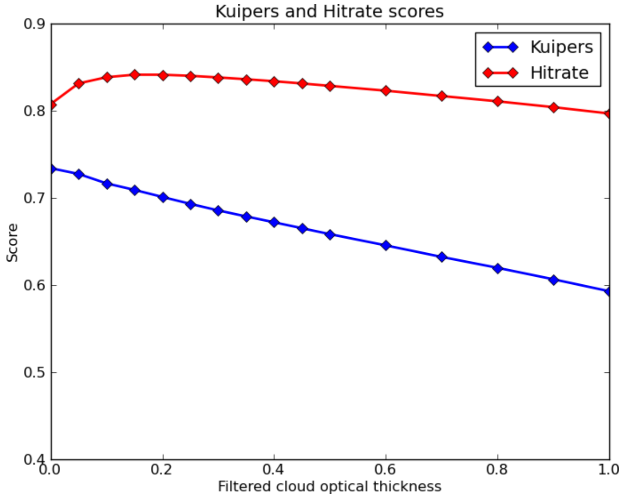

Fig. 9. Same as Fig. 6 but after treating CLARA-A1/PPS misclassified clear pixels as truly clear and not as cloudy (see motivation in text).

is a measure of how well cloudy and cloud-free situations are separated, it is clear that this will now occur in the unfiltered case (after having removed all obviously misclassified cloudy cases).

4.1.3 Results subdivided into day and night portions

Since the overall results include results from both illuminated and dark conditions, an interesting aspect is to study what happens if we look at both conditions separately – basically, looking at the impact of having access to visible spectral channels (i.e. information on reflected sunlight) or not. Figures 10–12 show corresponding results for all statistical scores in the day and at night (as defined in Sect. 3.3). All figures show convincingly how cloud detection efficiency degrades for nighttime conditions. For example, in Fig. 10 we see that, while the mean error reaches the zero level already at cloud optical thickness 0.2 during the day, it never reaches this level at night (i.e. remains negative). It is clear that a large fraction of all clouds are not detected at night, even at large cloud optical thicknesses. This is also well illustrated in Fig. 11 with decreasing $\mathrm{POD}_{\mathrm{cloudy}}$ and increasing $\mathrm{FAR}_{\mathrm{clear}}$ at night (i.e. $\mathrm{FAR}_{\mathrm{clear}}$ at filtered cloud optical thickness of 1.0 increases from about 10 % during the day to about 25 % during the night). Skill scores in Fig. 12 also show significantly better results during the day compared to during the night. Thus, the availability of information in the visible and shortwave infrared AVHRR channels appears to be quite important for the success of cloud detection.

Somewhat surprisingly, the derived value of the minimum cloud detection limit (according to Eq. 9) is found at cloud optical thickness 0.3 for both day and night conditions. Thus, the sensitivity to the filtered cloud optical thickness is relatively unchanged even if much fewer clouds are detected at night. We conclude that this must be explained by the increase in frequency of cases when clouds are completely missed at night (as indicated by the high $\mathrm{FAR}_{\mathrm{clear}}$ value at night). Thus, we are facing more general non-separability conditions of clouds and Earth surfaces at night, and this has nothing to do with how thick clouds are. The fact that the overall cloud detection limit was estimated to be at cloud optical thickness 0.35 in Sect. 4.1.2 (i.e. higher than the derived value for either day or night) indicates that conditions must be especially problematic at twilight conditions. Hence, the cloud detection limit is found to lie at a cloud optical thickness of 0.45 for twilight conditions. From Fig. 13, showing the POD and FAR quantities at twilight conditions, we conclude that this is explained by the rather slow increase in $\mathrm{POD}_{\mathrm{cloudy}}$ and the corresponding slow decrease of $\mathrm{FAR}_{\mathrm{clear}}$ for increasing filtered cloud optical thicknesses. Thus, at twilight we still miss a substantial fraction of optically thick clouds (more or less the same fraction as at night), but now we also face increasing difficulties in detecting thinner clouds.

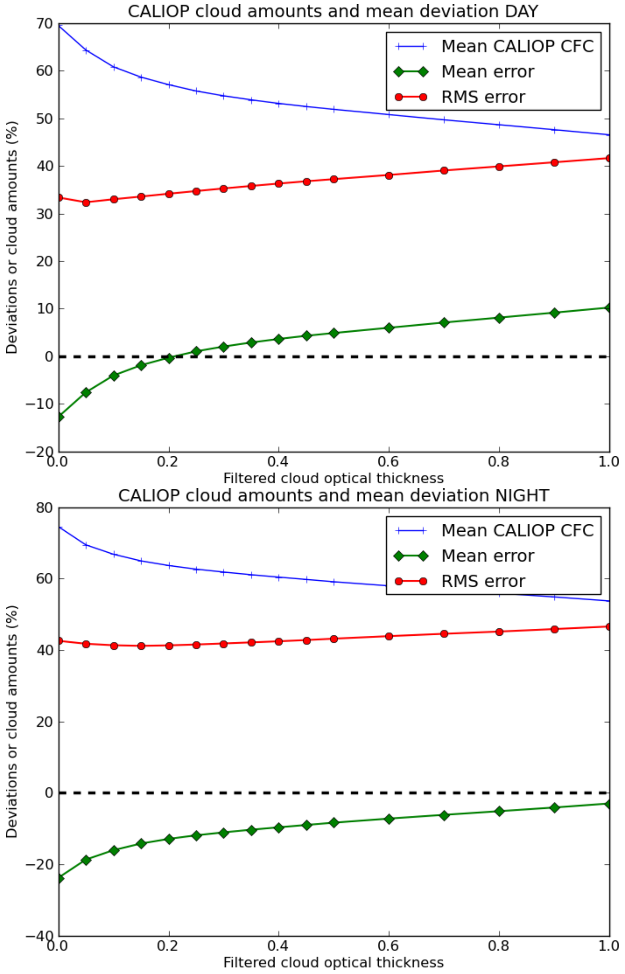

Fig. 10. Same visualisation as in Fig. 7 but for categories day (top) and night (bottom).

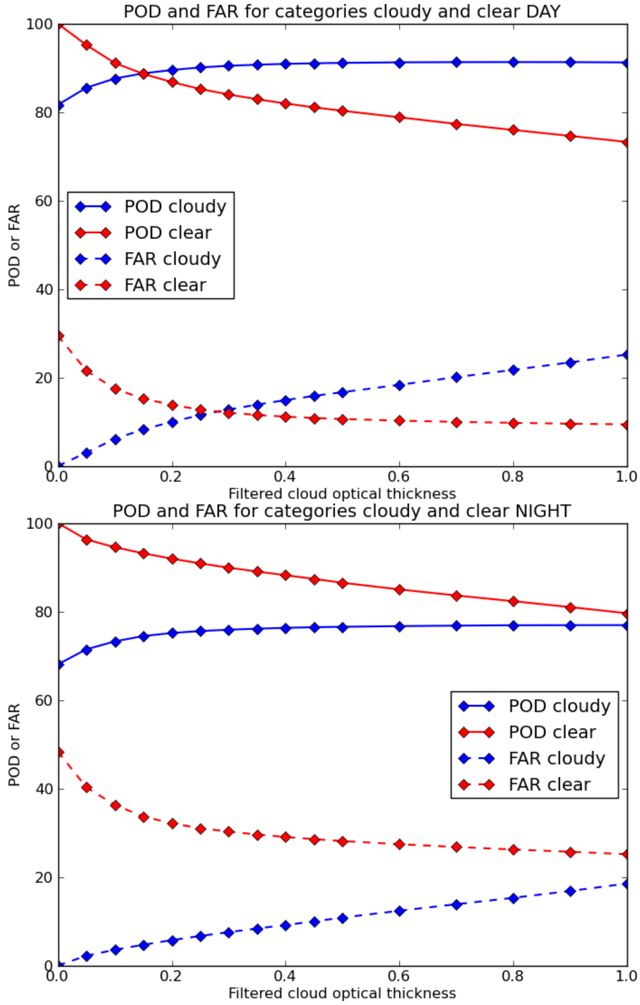

Fig. 11. Same visualisation as in Fig. 8 but for categories day (top) and night (bottom).

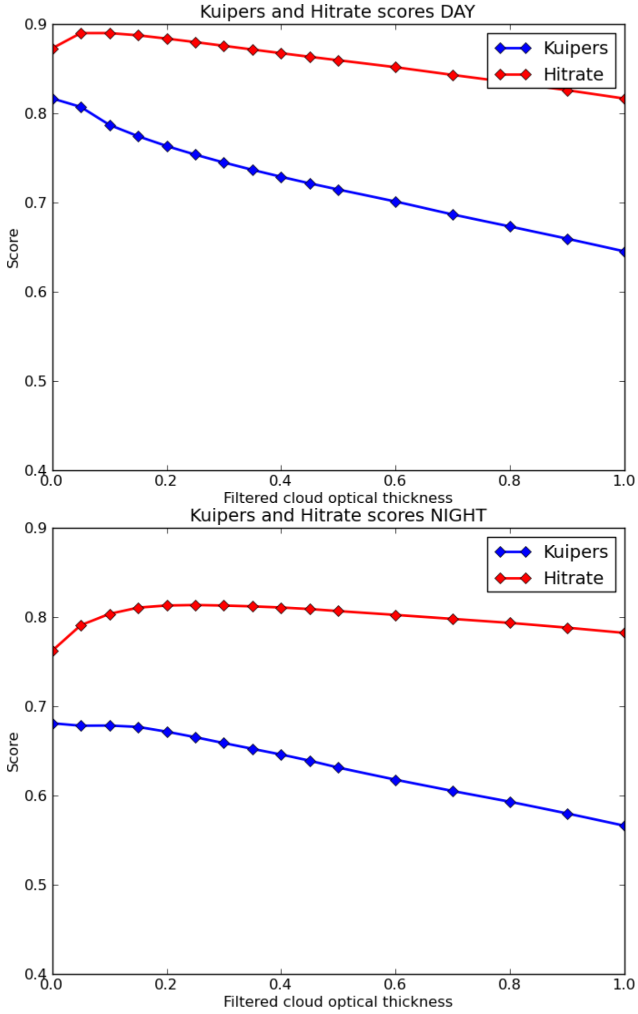

Fig. 12. Same visualisation as in Fig. 9 but for categories day (top) and night (bottom).

4.1.4 Global results subdivided into regions

Let us now look at the geographical variations of the validation results. This might shed some further light on where we encounter the problems of misclassified clear or misclassified cloudy conditions, i.e. those misclassifications that do not depend on existing clouds' optical thickness. First considering the unfiltered CALIPSO results, we will investigate if there are specific regions where misclassifications of cloud-free areas occur (i.e. explaining the 8 % of CLARA-A1/PPS misclassified clear cases mentioned in Sect. 4.1.2). These results are summarised for the mean error quantity in Table 3 for latitudinal bands defined in Table 2, for day, twilight and night categories and for land and ocean surfaces. We restrict the description to the mean error quantity since by this detailed subdivision of results the number of samples per category is sometimes too small to enable a proper estimation of all the statistical scores. For example, no samples for the twilight category could be found for tropical and mid-latitude regions.

As expected, for most categories in Table 3 we have a substantial underestimation of cloudiness explained by the inability to detect very thin cloud layers. However, one of the categories actually showing some overestimation (+6.2 %) is the category mid-latitude land. Near-zero results are also presented for the tropical land category. Further visual inspection of results revealed that misclassifications of clear conditions mainly occur over semi-arid land areas in the subtropical region, i.e. in the zone where desert regions change from being pure desert to being partly vegetation-covered. Thus, misclassifications do not occur over pure desert areas but where we have a seasonal transition from near-desert conditions to tropical vegetated conditions.

Table 4 shows results where we treat all CALIOP-detected clouds with cloud optical thicknesses lower than 0.35 as non-existent (i.e. as cloud-free cases). We notice that for the day category we now get dominantly positive values, i.e. we normally detect some clouds that are thinner than optical

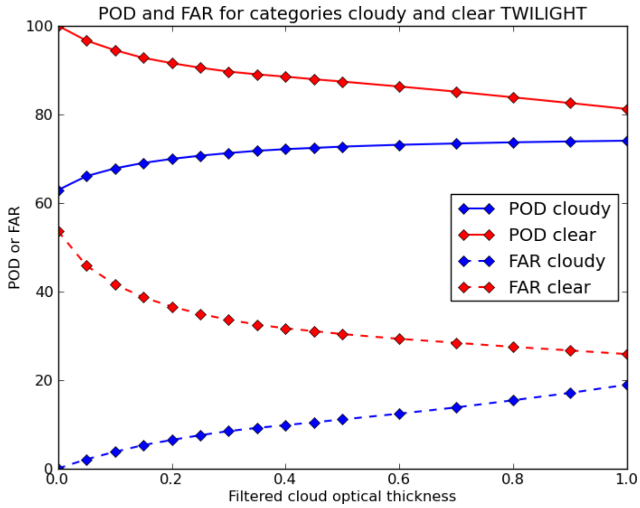

Fig. 13. Same visualisation as in Fig. 11 but for category twilight.

thickness 0.35. However, the overestimation is now quite excessive for categories mid-latitude land and tropical land, which even further emphasises the misclassification problems encountered here. Some positive values are also seen at night over mid-latitude and tropical categories, but otherwise we have dominantly negative results for night and twilight categories which are in line with the results discussed previously in Sect. 4.1.3. For these categories, we obviously do not detect a substantial fraction of all clouds regardless of their optical thickness. This occurs mainly in the polar regions but also during dark and snow-covered periods over high-latitude regions.

4.2 Cloud top height results

Results from the evaluation of cloud top height retrievals (following the method described in Sect. 3.4) are presented in Table 5. Results are compared with cases where we did not apply any filtering of very thin cloud layers and also where we compared with the cloud top boundary for the uppermost CALIOP cloud layer instead of the mid-layer value.

It is obvious from Table 5 that the chosen validation methodology has a tremendous impact on the achieved results. When including all thin cloud layers and when comparing with uppermost cloud boundary, a substantial underestimation of cloud top heights is found (on average more than 2.5 km). When instead taking into account the cloud detection limit at optical thickness value 0.35 and trying to represent clouds with a more radiatively relevant height, results improve drastically. The overall bias is now −274 m and the RMS error decreases by almost a factor 2. Even when filtering with the optical thickness value of 0.5, the bias almost disappears. However, we notice that the small total bias is largely a result of the sum of a large underestimation of high-level cloud tops (−1769 m) and a large overestimation of low-level cloud tops (+1137 m). Thus, there seems to be different behaviour of high-level clouds and low-level clouds. The low-level boundary layer cloud problem is the same as reported previously for MODIS cloud top products (Menzel et al., 2008). For CLARA-A1/PPS it can be explained as a problem with the reference atmospheric temperature profile taken from NWP analyses (here, ERA-Interim). For boundary layer clouds trapped in a temperature inversion, the reference profile is not detailed enough (i.e. too weak inversion, which is partly due to the mismatch between pixel and NWP grid resolution), leading to an overestimation of the cloud top height. A typical example of when this occurs can be seen in Fig. 3 between track positions 7000 and 8000.

The underestimation of high-level clouds reflects the problem of how to define the radiatively efficient cloud top height for thin and multiple cloud layers for an infrared channel of a passive imager. CALIOP measurements have also revealed the frequent existence of surprisingly thick (geometrically) single cloud layers which are optically very thin. Good examples of this are found in Fig. 3 at track positions 2400, 4000 and 6500. The use of a mid-layer representation of

Table 3. Mean error (%) of cloud detection separated according to latitude bands and illumination categories (defined in the text) and surface conditions (land or ocean). Statistics computed from 99 full globally matched NOAA-18 and CALIPSO orbits with a total of 725 900 individual pixel matches.

	DAY	TWILIGHT	NIGHT
TROPICAL Ocean	−10.0	–	−18.1
TROPICAL Land	−0.8	–	−22.4
MID-LATITUDE Ocean	−6.3	–	−14.9
MID-LATITUDE Land	6.2	–	−13.7
HIGH-LATITUDE Ocean	−4.7	−18.7	−18.4
HIGH-LATITUDE Snow-free Land	−6.6	−29.4	−27.0
HIGH-LATITUDE Snow-cover Land	−16.7	−36.6	−33.7
POLAR Ice-free Ocean	−6.0	−25.1	−39.2
POLAR Ice-cover Ocean	−13.4	−11.7	−37.5
POLAR Snow-cover Land	−16.5	−35.3	−25.0
POLAR Snow-free Land	−21.3	−38.9	−32.7

Table 4. Same as Table 3 but now after filtering results with cloud optical thickness threshold 0.35 (i.e. all CALIOP-detected clouds with smaller optical thickness are neglected and treated as a cloud-free observation).

	DAY	TWILIGHT	NIGHT
TROPICAL Ocean	11.2	–	4.8
TROPICAL Land	22.8	–	−0.6
MID-LATITUDE Ocean	9.9	–	−2.9
MID-LATITUDE Land	22.0	–	0.3
HIGH-LATITUDE Ocean	5.0	−9.0	−12.1
HIGH-LATITUDE Snow-free Land	9.9	−12.4	−11.3
HIGH-LATITUDE Snow-cover Land	0.1	−17.7	−17.5
POLAR Ice-free Ocean	1.9	−12.0	−32.2
POLAR Ice-cover Ocean	1.0	8.3	−21.1
POLAR Snow-cover Land	1.4	−14.5	−4.4
POLAR Snow-free Land	−6.0	−13.9	−21.6

Table 5. Cloud top height (CTO) results from CLARA-A1/PPS evaluated using unfiltered and filtered CALIOP results. Mean errors (Bias) and RMS errors are given for unfiltered (column 1) conditions and for two filtered conditions (columns 2 and 3) with two different cloud optical thickness thresholds. Mean errors are also given for the three cloud layer groups of low-level, medium-level and high-level clouds (explained in text).

	CTO results Total dataset Unfiltered	CTO results COT threshold 0.35	CTO results COT threshold 0.5
Samples	281 180	254 130	248 398
Bias (m)	−2762 (+593 Low −781 Medium −5339 High)	−274 (+1097 Low +199 Medium −2028 High	−78 (+1137 Low +280 Medium −1769 High
RMS (m)	4879	2511	2361

such a cloud layer is apparently still inadequate. Currently, CLARA-A1/PPS retrievals underestimate the height for all high-level clouds substantially even if the method of applying cloud filtering of the thinnest clouds have reduced the difference. It is clear that there are remaining ambiguities in the determination of an appropriate radiatively efficient height. A possible further improvement of the validation methodology could be to better try to estimate how deep (in the optical sense) into cloud layers we need to go to find this efficient height. The corresponding integrated cloud optical thickness should obviously be larger than the estimated detection limit of 0.35. But even if we cannot determine this value exactly, the systematic use of a stipulated value (e.g. optical thickness 1.0) could be valuable in the evaluation of different and upgraded cloud height retrieval methods in the future.

As a final remark, one must state that the AVHRR instrument is poorly equipped for detection of very thin cirrus clouds in comparison to other more advanced sensors (e.g. MODIS). The lack of spectral channels sensitive to conditions in the upper troposhere (e.g. CO_2 bands) means that the prospect for making accurate cloud height retrievals for thin cirrus clouds remains very challenging and without much potential to improve.

5 Conclusions

This study investigated the optimal validation methodology to be used when evaluating cloud retrievals from passive imagers to take full advantage of the measurements provided by the active cloud lidar CALIOP carried by the CALIPSO satellite. Some problems for adequately using the current CALIOP datasets for the validation purpose were identified and a method for mitigating the influence of those was proposed. The method was applied to evaluate a subset (covering the years 2006–2009) of the CMSAF CLARA-A1 dataset derived from historical global AVHRR data. It was demonstrated how the CALIOP-provided information of cloud presence and cloud optical thickness can be used to delineate the current cloud detection limitations of the methods used to compile the CLARA-A1 dataset. Although the cloud detection capability does vary with time of day and with the geographical environment, an overall cloud detection limit was estimated at a cloud optical thickness of 0.35. It means that at this cloud optical thickness most cloud layers are detected. Thinner clouds are detected but at decreasing efficiency with smaller cloud optical thickness. The diurnal variation showed that the detection limit is close to 0.3 for both day and night, while conditions deteriorate considerably at twilight conditions when the cloud detection limit is estimated at 0.45.

The study also revealed that there is a substantial fraction of cases where cloud detection results are not dependent at all on the thickness of existing clouds. In other words, there are cases where clouds are either completely missed or falsely identified. This explains why the probability of detecting clouds is limited to about 90 % during day but as low as 75–80 % during night and twilight conditions. Daytime misclassifications of semi-arid sub-tropical and tropical land surfaces as clouds were identified, as well as a substantial amount of missed clouds in the polar regions during the polar winter. Both deficiencies are well understood and reflect major challenges for most cloud retrieval schemes using data from passive imagery. The daytime problem is linked to the fundamental difference in the cloud-free spectral appearance of desert surfaces and tropical forested surfaces. While cloud screening seems to work well over both mentioned surfaces, problems arise in the transition zone between them where the appearance also changes seasonally. The current methodology has an inappropriate description of this transition zone and the associated temporal changes of its surface appearance (i.e. a static climatology is used). Thus, an improved methodology must address this limitation in the future.

The underestimation of cloudiness at high latitudes and especially during winter conditions is linked to another well-known problem for all cloud screening methods applied to passive imagery. It occurs when there is no distinct temperature difference between clouds and the underlying surface. The situation becomes even worse if the temperature difference is also reversed (i.e. if clouds are warmer than the surface), which is a frequent feature in the polar winter. Also, when ground temperatures become extremely cold (like over the Antarctic plateau in the polar winter) the radiometric accuracy of the AVHRR measurement is no longer accurate enough for estimating the brightness temperature difference between infrared channels – a quantity that is heavily used by many cloud screening methods. Altogether, this leads CLARA-A1 to substantially underestimate the cloudiness over polar regions in the polar winter and also during night and twilight conditions at high latitudes. Notice that even if this problem is common to most cloud screening methods applied in the polar region, the achieved results may differ depending on the actual method. For some methods like CLARA-A1/PPS the problems are manifested as missed clouds, while for other methods it could as well lead to overestimated cloudiness (misclassified cold cloud-free surfaces).

Complementary to the study on cloud detection efficiency, cloud top height assignments have been evaluated. The information on cloud detection limitations was taken into account, either by discarding too thin single-layer clouds or too thin uppermost cloud layers. Results were shown to differ substantially depending on whether the cloud top boundary was defined as the uppermost CALIOP-derived cloud layer boundary or as the mid-level (i.e. the mean of cloud base and cloud top) of the corresponding CALIOP-observed cloud layer. The latter definition gives a height that is closer to the radiatively efficient level of the cloud, which better resembles the level that is normally retrieved from passive imagery.

When using the latter approach a relatively small total cloud top height bias of −274 m was found. This can be compared to the cloud top height bias of −2762 m for the default method based on the uppermost cloud boundary and including all thin clouds. However, even after using the more realistic radiatively efficient level approximation it is clear that large underestimations of high-level cloud top heights and overestimations of low-level cloud top heights exist, which have to be addressed in a future reprocessing of the dataset.

In conclusion, we have demonstrated how CALIPSO-CALIOP results can be used to carry out a very detailed examination of cloud retrieval results from passive imagers. Results presented here are not entirely surprising or unexpected, but they are given with unprecedented detail. Although the current CALIOP datasets, as defined in a particular horizontal resolution, are not always directly applicable in comparisons with corresponding cloud datasets with the same resolution from passive sensors, we have shown how CALIOP datasets from different resolutions can be combined to construct a more reasonable validation reference. As such, we believe that its value is unprecedented and that it can be used as an invaluable reference for the evaluation of any cloud retrieval scheme based on data from passive imagers. In particular, we believe that even beyond the lifetime of the CALIPSO satellite, the extracted subset of collocated NOAA-18 and CALIPSO-CALIOP observations might serve as a benchmarking dataset for the testing of various AVHRR-based cloud retrieval methods. For the planned future upgrades of the CLARA-A1 dataset, the idea is to use the currently collected CALIPSO dataset in exactly this way. There is a limitation in that it is based exclusively on afternoon-orbit NOAA-18 data, but we believe that it can be complemented with a limited set of morning-orbit data from satellites carrying the modified AVHRR instrument with the additional 1.6 µm channel. For the latter the matched datasets are limited to latitudes near ± 70 degrees due to orbital considerations; i.e. this is the only latitude where simultaneous overpasses with CALIPSO occur for morning orbit satellites.

The next CLARA release (CLARA-A2) is scheduled for 2016, and we will utilise the current validation tool heavily in the work of upgrading and evaluating the methodology. But even concerning the current CLARA-A1 results, our findings should be very important for potential users. One particularly good example is the provision of essential background information for the construction of a CLARA-A1 simulator tool to be used for evaluation of cloud properties simulated by climate models.

Regarding the prospect of applying this methodology to data from other sensors than AVHRR, it is clear (or even trivial) that the method is directly applicable to data from the MODIS sensor (already available on the A-Train platform). The method is also directly applicable to data from the new Visible Infrared Imager Radiometer Suite (VIIRS) sensor on the Suomi-NPP satellite, also placed in an afternoon orbit very similar to the orbit of NOAA-18. As for the aforementioned morning-orbit NOAA and Metop satellites, the method should also be applicable at high latitudes for sensors like the Advanced Along-Track Scanning Radiometer (AATSR) and the Medium Resolution Imaging Spectrometer (MERIS) onboard the ENVISAT satellite.

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 authors want to thank Dave Winker in the CALIPSO Science Team and Ralph Kuehn at the Space Science and Engineering Center (SSEC) at the University of Madison, USA, for valuable discussions.

Edited by: A. Kokhanovsky

References

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., and John, V. O.: COSP: Satellite simulation software for model assessment, B. Am. Meteorol. Soc., 92, 1023–1043m doi:10.1175/2011BAMS2856.1, 2011.

Cesana, G., Kay, J. E., Chepfer, H., English, J. M., and de Boer, G.: Ubiquitous low-level liquid-containing Arctic clouds: New observations and climate model constraints from CALIPSO-GOCCP, Geophys. Res. Lett., 39, L20804, doi:10.1029/2012GL053385, 2012.

Chepfer, H., Bony, S., Winker, D., Cesana, G., Dufresne, J. L., Minnis, P., Stubenrauch, C. J., and Zeng, S.: The GCM-Oriented CALIPSO Cloud Product (CALIPSO-GOCCP), J. Geophys. Res., 115, D00H16, doi:10.1029/2009JD012251, 2010.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally,

A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteorol. Soc., 137, 553—597, doi:10.1002/qj.828, 2011.

Delanoë, J., Hogan, R. J., Forbes, R. M., Bodas-Salcedo, A., and Stein, T. H. M.: Evaluation of ice cloud representation in the ECMWF and UK Met Office models using CloudSat and CALIPSO data, Q. J. Roy. Meteorol. Soc., 137, 2064–2078, doi:10.1002/qj.882, 2011.

Devasthale, A. and Thomas, M. A.: A global survey of aerosol-liquid water cloud overlap based on four years of CALIPSO-CALIOP data, Atmos. Chem. Phys., 11, 1143–1154, doi:10.5194/acp-11-1143-2011, 2011.

Dybbroe, A., Thoss, A., and Karlsson, K.-G.: NWCSAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modelling – Part I: Algorithm description, J. Appl. Meteorol., 44, 39–54, 2005.

Haladay, T. and Stephens, G. : Characteristics of tropical thin cirrus clouds deduced from joint CloudSat and CALIPSO observations, J. Geophys. Res., 114, D00A25, doi:10.1029/2008JD010675, 2009.

Heidinger, A. K. and Pavolonis, M. J.: Gazing at Cirrus Clouds for 25 Years through a Split Window. Part I: Methodology, J. Appl. Meteorol. Clim., 48, 1100–1116, doi:10.1175/2008JAMC1882.1, 2009.

Heidinger, A. K., Evan, A. T., Foster, M. J., and Walther, A.: A Naive Bayesian Cloud-Detection Scheme Derived from CALIPSO and Applied within PATMOS-x, J. Appl. Meteorol. Clim., 51, 1129–1144, doi:10.1175/JAMC-D-11-02.1, 2012.

Holz, R. E., Ackerman, S. A., Nagle, F. W., Frey, R., Dutcher, S., Kuehn, R. E., Vaughan, M. A., and Baum, B.: Global Moderate Resolution Imaging Spectroradiometer (MODIS) cloud detection and height evaluation using CALIOP, J. Geophys. Res., 113, D00A19, doi:10.1029/2008JD009837, 2008.

Karlsson, K.-G. and Dybbroe, A.: Evaluation of Arctic cloud products from the EUMETSAT Climate Monitoring Satellite Application Facility based on CALIPSO-CALIOP observations, Atmos. Chem. Phys., 10, 1789–1807, doi:10.5194/acp-10-1789-2010, 2010.

Karlsson, K.-G., Riihelä, A., Müller, R., Meirink, J. F., Sedlar, J., Stengel, M., Lockhoff, M., Trentmann, J., Kaspar, F., Hollmann, R., and Wolters, E.: CLARA-A1: the CM SAF cloud, albedo and radiation dataset from 28 yr of global AVHRR data, Atmos. Chem. Phys. Discuss., 13, 935–982, doi:10.5194/acpd-13-935-2013, 2013.

Korpela, A., Dybbroe, A., and Thoss, A.: Retrieveing Cloud Top Temperature and Height in Semi-transparent and Fractional Cloudiness using AVHRR, Nowcasting SAF Visiting Scientist report, SMHI Reports Meteorology, 100, 48 pp., 2001.

Liu, Y., Key, J. R., Ackerman, S. A., Mace, G., and Quiqing, Z.: Arctic cloud macrophysical characteristics from CloudSat and CALIPSO, Remote Sens. Environ., 124, 159–173, doi:10.1016/j.rse.2012.05.006, 2012.

Menzel, W. P., Frey, R. A., Zhang, H., Wylie, D. P., Moeller, C. C., Holz, R. E., Maddux, B., Baum, B. A., Strabala, K. I., and Gumley, L. E.: MODIS Global Cloud-Top Pressure and Amount Estimation: Algorithm Description and Results, J. Appl. Meteorol. Clim., 47, 1175–1198, doi:10.1175/2007JAMC1705.1, 2008.

Minnis, P., Yost, C. R., Sun-Mack, S., and Chen, Y.: Estimating the physical top altitude of optically thick ice clouds from thermal infrared satellite observations using CALIPSO data, Geophys. Res. Lett., 35, L12801, doi:10.1029/2008GL033947, 2008.

Reuter, M., Thomas, W., Albert, P., Lockhoff, M., Weber, R., Karlsson, K.-G., and Fischer, J.: The CM-SAF and FUB Cloud Detection Schemes for SEVIRI: Validation with Synoptic Data and Initial Comparison with MODIS and CALIPSO, J. Appl. Meteorol. Clim., 48, 301–316, doi:10.1175/2008JAMC1982.1, 2009.

Schulz, J., Albert, P., Behr, H.-D., Caprion, D., Deneke, H., Dewitte, S., Dürr, B., Fuchs, P., Gratzki, A., Hechler, P., Hollmann, R., Johnston, S., Karlsson, K.-G., Manninen, T., Müller, R., Reuter, M., Riihelä, A., Roebeling, R., Selbach, N., Tetzlaff, A., Thomas, W., Werscheck, M., Wolters, E., and Zelenka, A.: Operational climate monitoring from space: the EUMETSAT Satellite Application Facility on Climate Monitoring (CM-SAF), Atmos. Chem. Phys., 9, 1687–1709, doi:10.5194/acp-9-1687-2009, 2009.

Stephens, G. L., Tsay, S.-C., Stackhouse, P. W., and Flatau, P. J.: The relevance of microphysical and radiative properties of cirrus clouds to climate and climatic feedback, J. Atmos. Sci., 47, 1742–1753, 1990.

Stephens, G. L., Vane, D. G., Boain, R. J., Mace, G. G., Sassen, K., Wang, Z., Illingworth, A. J., O'Connor, E. J., Rossow, W. B., Durden, S. L., Miller, S. D., Austin, R. T., Benedetti, A., Mitrescu, C., and the CloudSat Science Team: The CloudSat mission and the A-Train, B. Am. Meteorol. Soc., 83, 1771–1790, doi:10.1175/BAMS-83-12-1771, 2002.

Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., Di Girolamo, L., Getzewich, B., Guignard, A., Heidinger, A., Maddux, B. C., Menzel, W. P., Minnis, P., Pearl, C., Platnick, S., Poulsen, C., Riedi, J., Sun-Mack, S., Walther, A., Winker, D., Zeng, S., and Zhao, G.: Assessment of global cloud datasets from satellites: Project and Database initiated by the GEWEX Radiation Panel, B. Am. Meteorol. Soc., online first, doi:10.1175/BAMS-D-12-00117, 2013.

Vaughan, M., Powell, K., Kuehn, R., Young, S., Winker, D., Hostetler, C., Hunt, W., Liu, Z., McGill, M., and Getzewich, B.: Fully Automated Detection of Cloud and Aerosol Layers in the CALIPSO Lidar Measurements, J. Atmos. Ocean. Tech., 26, 2034–2050, doi:10.1175/2009JTECHA1228.1, 2009.

Virts, K. S., Wallace, J. M., Fu, Q., and Ackerman, T. P.: Tropical Tropopause Transition Layer Cirrus as Represented by CALIPSO Lidar Observations, J. Atmos. Sci., 67, 3113–3129, 2010.

Winker, D. M., Vaughan, M. A., Omar, A., Hu, Y., Powell, K. A., Liu, Z., Hunt, W. H., and Young, S. A.: Overview of the CALIPSO mission and CALIOP data processing algorithms, J. Atmos. Ocean. Tech., 26, 2310–2323, doi:10.1175/2009JTECHA1281.1, 2009.