Atmos. Chem. Phys., 10, 1789–1807, 2010
www.atmos-chem-phys.net/10/1789/2010/

Atmospheric Chemistry and Physics

Evaluation of Arctic cloud products from the EUMETSAT Climate Monitoring Satellite Application Facility based on CALIPSO-CALIOP observations

K.-G. Karlsson and A. Dybbroe

Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

Received: 14 May 2009 – Published in Atmos. Chem. Phys. Discuss.: 7 August 2009
Revised: 8 December 2009 – Accepted: 5 January 2010 – Published: 16 February 2010

Abstract. The performance of the three cloud products cloud fractional cover, cloud type and cloud top height, derived from NOAA AVHRR data and produced by the EUMETSAT Climate Monitoring Satellite Application Facility, has been evaluated in detail over the Arctic region for four months in 2007 using CALIPSO-CALIOP observations. The evaluation was based on 142 selected NOAA/Metop overpasses allowing almost 400 000 individual matchups between AVHRR pixels and CALIOP measurements distributed approximately equally over the studied months (June, July, August and December 2007). Results suggest that estimations of cloud amounts are very accurate during the polar summer season while a substantial loss of detected clouds occurs in the polar winter. Evaluation results for cloud type and cloud top products point at specific problems related to the existence of near isothermal conditions in the lower troposphere in the polar summer and the use of reference vertical temperature profiles from Numerical Weather Prediction model analyses. The latter are currently not detailed enough in describing true conditions relevant on the pixel scale. This concerns especially the description of near-surface temperature inversions which are often too weak leading to large errors in interpreted cloud top heights.

1 Introduction

Satellite-based monitoring of global cloud amounts and the associated various properties of clouds (e.g., cloud optical depths, cloud thermodynamic phase, cloud top heights, effective droplet/crystal sizes, droplet/crystal concentrations, liquid and ice water paths, etc) is an important task for understanding the role of clouds in climate, especially in terms of assessing the total effect of clouds in the current climate system and the role of clouds in various climate feedback processes (Stephens, 2005 and Bony et al., 2006). Several satellite-derived cloud datasets have been compiled during the last few decades based on data from different satellite sensors. A few examples are datasets based on utilization of visible and infrared passive imagery (Rossow and Schiffer, 1999), datasets based on advanced multispectral passive imagery (Platnick et al., 2003) and datasets based on satellite sounding data (Stubenrauch et al., 2006). Recently, attempts to compile larger joint sets of climate-relevant parameters (including cloud products) from a multitude of satellite sensors and satellite platforms have been initiated within the framework of the EUMETSAT Satellite Application Facility on Climate Monitoring (CMSAF, see Schulz et al., 2009). Common for all these datasets is the need to perform a proper validation of the derived cloud products to assess their reliability and credibility. This paper addresses some recent progress in this validation work related to the last mentioned dataset above.

Validation of satellite-derived datasets have for many years relied on comparisons to surface observations and/or on inter-comparisons to other satellite-datasets from passive sensors. The former validation reference is associated with problems due to completely different viewing conditions from ground compared to the satellite view as well as problems in enabling a proper temporal and spatial matching of the datasets. The latter kind of validation reference partly removes this problem but introduces new problems related to different sensor characteristics, new differences in temporal and spatial sampling characteristics and to the risk of finding the same kind of retrieval deficiencies in the reference dataset

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

(at least if using rather similar kind of sensors). The introduction of the A-train series of satellites (i.e., several satellites with different instrumentation flying close together and in the same orbit as the Aqua satellite) and especially its two satellites with active sensors onboard (CloudSat, described by Stephens et al., 2002 and CALIPSO – Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation, described by Winker et al., 2006) have drastically improved the conditions for validating satellite-derived cloud datasets. Now it is possible to align datasets almost simultaneously in space and time taking advantage of the overlapping parts of satellite orbits. Equally important is the possibility to obtain the same satellite view from the validation reference and the fact that active measurements assures that the measured signal comes exclusively from cloud and/or aerosol particles in the atmosphere and that it is not mixed up with contributions from the atmosphere or the surface (at least true for atmospheric layers not too close to the surface). Many studies taking advantage of this new situation have been initiated and have recently been reported (e.g. Ackerman et al., 2008; Holz et al., 2008; Weisz et al., 2007).

A previous paper by Reuter et al. (2009) made an attempt to evaluate CMSAF cloud top products from the SEVIRI instrument on the Meteosat-8 satellite based on CALIPSO measurements. This paper will describe the first thorough examination of CMSAF cloud products based on data from the NOAA Advanced High Resolution Radiometer (AVHRR). Furthermore, the validation area used is the Arctic region; an area where satellite-derived products have been very scarcely evaluated previously due to the very limited access to reference observations on ground. The validation effort is connected to the fact that the CMSAF is now expanding its product area coverage to include the Arctic region. This is mainly motivated by the fact that most climate model scenarios indicate that the largest climate change effects due to an increased greenhouse warming is expected to occur in the Polar Regions (IPCC4, 2007). Thus, noteworthy is that this study is not only presenting a new method of validation but it also addresses the general topic of cloud product generation in the Arctic environment; a task that is generally recognized as a very challenging task (especially in the polar winter) due to the lack of visible information during large parts of the year, the poor temperature contrast between clouds and Earth surfaces and the necessity to effectively discriminate between clouds and snow-covered surfaces.

Notice also that the study includes a direct comparison between the cloud mask results provided by the CMSAF method and the official cloud mask results from the Moderate Resolution Imaging Spectrometer (MODIS, Collection 5 cloud mask from the MODIS Science Team).

Section 2 describes the evaluated cloud products and the CALIPSO datasets used followed by a more thorough description of the applied validation method in Sect. 3. Main results are then presented and discussed in Sect. 4 followed by a summary and concluding remarks in Sect. 5.

2 Cloud products and validation datasets

2.1 CMSAF cloud products from AVHRR

The following three CMSAF cloud products have been evaluated in this study:

- Cloud fractional cover (CFC)
- Cloud type (CTY)
- Cloud-top height (CTH)

The processing of the Metop/NOAA AVHRR cloud products for a single overpass was performed by the PPS (Polar Platform System) cloud software package Version 2008. PPS is developed by the EUMETSAT Satellite Application Facility in support of Nowcasting and Very Short range Forecasting applications (NWCSAF, see http://nwcsaf.inm.es/). The computation of cloud products is sequential, i.e. the cloud fractional cover (or cloud mask) is derived first and is used as input to the cloud type and the cloud top height parameters retrieval.

A short description of all cloud products is given in the following sub-sections.

2.1.1 CLOUD FRACTIONAL COVER – CFC

The cloud mask retrieval algorithm is based on a multispectral thresholding technique where thresholds are scene-dependent and dynamically adjusted (Dybbroe et al., 2005a, b). However, PPS Version 2008 has been upgraded with improved threshold schemes adapted to Arctic polar night conditions (for further details, see Eliasson et al., 2007). The thresholds are based on pre-calculated radiative transfer simulations stored in look-up tables. Essential further input parameters are actual geographical data (e.g. land use, topography, etc.) and Numerical Weather Prediction (NWP) analyses. The latter are taken from the Deutscher Wetterdienst (DWD) GME model (see Majewski et al., 2002) with a temporal resolution of 3 h and a spatial resolution of about 40 km. The model has 40 atmospheric layers between ground and the topmost layer at 0.1 hPa. The CMSAF cloud fractional cover product (CFC) is calculated directly from the cloud mask by dividing the amount of cloudy pixels (including also what is denoted as cloud-contaminated pixels) with the total amount of valid (cloud-free or cloudy) pixels in coarse resolution grid squares.

2.1.2 CLOUD TYPE – CTY

The main objective of the NWCSAF cloud type product (CTY) is to provide a detailed cloud analysis. The original NWCSAF product distinguishes between 15 cloud classes while the CMSAF version of the product is less detailed and clouds are grouped as follows for all the pixels identified as cloudy in a scene:

- Low clouds (including fog)
- Medium clouds
- High Opaque clouds
- High Semi-transparent clouds
- Fractional clouds (i.e., cloud-contaminated pixels or pixels with sub-pixel scale cloud elements)

The CTY algorithm (briefly outlined in Dybbroe et al., 2005a) is a sequential threshold algorithm applied to pixels. It uses the pre-computed cloud mask and spectral and textural features which are derived from the multispectral satellite images and scene-dependent (dynamic) thresholds. What is especially important here is that the basic subdivision of opaque low-, medium- and high-level clouds is done by utilising temperature information at NWP-analysed pressure levels of 700 hPa and 500 hPa.

CMSAF CTY products are calculated in a similar way as the CFC product but now focusing on each CTY category, i.e., describing contributions to the CFC from respective CTY categories. Results can be described as either absolute or relative contributions. Although current operational CMSAF products give relative contributions we will investigate both options in this study.

2.1.3 CLOUD TOP HEIGHT – CTH

The CMSAF AVHRR-derived CTH product contains information on the cloud top altitude relative to the local topography for all pixels identified as cloudy in the satellite scene. The CTH product is derived using two different algorithms; one for opaque and one for fractional and semitransparent clouds, and it is applied to all cloudy pixels as given by the CTY product. The separation into opaque and semitransparent cloud groups is based on examination of 11 µm and 12 µm brightness temperature differences (BTD).

The algorithm for opaque clouds uses radiances and brightness temperatures for AVHRR channel 4 at 11 µm. Cloudy and cloud-free radiances are simulated applying the RTTOV radiative transfer model (Chevallier and Tjemkes, 2001) and using temperature and humidity profiles taken from NWP analyses. The overcast simulation results are available for each pressure level given by RTTOV and are derived using an emissivity of 1.0 ("black" clouds). The radiance simulations are made on a coarse horizontal resolution (on segments of high-resolution pixels). For the CMSAF implementation, a segment size of 32×32 pixels has been chosen.

The CTH opaque retrieval depends on the cloud type:

– For all pixels classified into one of the opaque cloud types the cloud top pressure is derived from the best fit between the simulated and the measured brightness temperatures. The temperature search goes from the surface and upwards with priority to choose the lowest solution in case of multiple solutions. The corresponding simulated cloud layer temperature from the segment closest in space to the given pixel is chosen as the associated cloud top temperature. The corresponding cloud altitude is calculated from geopotential relations (i.e., hydrostatic balance).

The algorithm for semi-transparent clouds uses a histogram technique based on the work of Derrien et al. (1988) and Inoue (1985) and detailed by Korpela et al. (2001). Two-dimensional histograms using AVHRR channel 4 and 5 brightness temperatures composed over larger image segments (e.g., 32×32 pixels) are constructed. By an iterative procedure a polynomial curve (simulating the arc shape) is fitted to the histogram values and the cloud top temperature and pressure (taken from NWP profiles) are retrieved. In this procedure, first guess values of surface temperatures are taken from NWP analyses as an external constraint.

2.2 Validation datasets from Aqua train satellites

This validation study focuses primarily on using data from the CALIPSO satellite launched in April 2006 together with the CloudSat satellite. CloudSat and CALIPSO fly in close formation. CALIPSO observes the same point on Earth only about 15 s later than CloudSat and about 1 min and 15 s later than the MODIS instrument onboard the Aqua satellite. The satellite carries the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) which 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 polarized components of the backscattered signal at 532 nm. The horizontal resolution of each single field of view (FOV) is 333 m and the vertical resolution is 30–60 m. The layer altitudes are given as the height above mean sea level. Lidar backscatter signals are directly linked to the optical thickness of clouds and aerosols at the two wavelengths. This means that the attenuation of the lidar pulse is strong for optically thick clouds. In practice this means that the instrument can only probe the full geometrical depth of a cloud if the total optical thickness is not larger than a certain threshold (assumed to be somewhere in the range 6-10). For optically thicker clouds only the upper portion of the cloud will be 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 2.01) and the associated information from the Lidar Level 2 Vertical Feature Mask product. These products define up to 10 cloud layers and each layer is classified into one of 10 cloud types according to Table 1. To be noticed here is that the International Satellite Cloud Classification Project (ISCCP) cloud typing convention has been used in the sense that the vertical separation of Low (categories

Table 1. Cloud categories according to the CALIOP Vertical Feature Mask product.

Category	Description
Category 0	Low, overcast, thin (transparent St, Sc and fog)
Category 1	Low, overcast, thick (opaque St, Sc and fog)
Category 2	Transition Stratocumulus
Category 3	Low, broken (trade Cu and shallow Cu)
Category 4	Altocumulus (transparent)
Category 5	Altocumulus (opaque, As, Ns, Ac)
Category 6	Cirrus (transparent)
Category 7	Deep convective (opaque As, Cb, Ns)

0–3), Medium (categories 4–5) and High (categories 6–7) clouds is defined by vertical pressure levels of 680 hPa and 440 hPa. However, the separation of thin and thick clouds is made using the information on whether the surface or lower layers below the current layer can be seen or not by CALIOP.

The CALIOP products are defined in five different versions with respect to the along-track resolution ranging from 333 m (individual FOVs), 1 km, 5 km, 20 km and 80 km. The four 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. We have used the 1 km resolution dataset since this resolution is closest to the nominal AVHRR image resolution. Consequently, this dataset might have somewhat smaller amounts of thin Cirrus cloud layers compared to what could be present in coarser resolution datasets (e.g., the 5 km dataset).

Important points to be noted here are also (as expressed in Data Quality Statements at the data retrieval website):

- Daytime measurements are less accurate than night time measurements since reflected solar radiation increases noise levels.
- The efficiency in the Cloud-Aerosol discrimination is currently estimated to 90% or better. It means in practice that
  - cases with heavy aerosol loadings in the troposphere are occasionally mis-classified as clouds
  - cases of very thin ice clouds are frequently misclassified as aerosols (especially in the Arctic region)

Apart from the use of the Vertical Feature Mask product we have also taken advantage of using some attached interesting auxiliary information related to the state of surface conditions when the measurements took place. This information concerns land cover characterization taken from the International Geosphere Biosphere Programme (IGBP) and ice and snow cover information taken from the National Snow and Ice Data Center (NSIDC).

The original concept for this validation study included plans to also compare with data from the cloud radar datasets from the CloudSat satellite. However, since CloudSat datasets do not permit studies of cloud conditions in the lowest kilometer of the atmosphere due to radar contamination from ground clutter it was decided to exclude this part of the study. This is motivated by the fact that near-surface clouds are frequent in the Arctic region and this limitation of the CloudSat dataset would therefore be detrimental for the analysis of the true cloud situation. Nevertheless, we have still taken advantage of the CloudSat datasets (i.e., the 2B-GEOPROF dataset, version 11, Release 04 based on the 2B-GEOPROF algorithm version 5.3) in that it allowed us to do a direct comparison with the MODIS cloud mask (Collection 5, described by Li et al., 2003 and Frey et al., 2008) along the CloudSat track. The close alignment of CloudSat and CALIPSO orbits permitted a re-navigation of the MODIS cloud mask to the CALIPSO track where it subsequently could be compared to both the CALIOP cloud mask and the CMSAF cloud mask. We will also visualize some CloudSat results for achieving a deeper understanding of the results (i.e., evaluating the CloudSat-CALIPSO agreement for medium-level and high-level cloud layers).

2.3 Chosen study area and validation period

The validation study was performed in the Arctic region as depicted in Fig. 1. This area covers all points north of approximately latitude 64° N.

Three months in the polar summer of 2007 were selected: June, July and August. The main motivation for choosing this period was for evaluating the cloud detection efficiency in a particularly critical period of ice melting in the Arctic (as reported by Kay et al., 2008). Efficient cloud screening is a prerequisite for successful estimation of other interesting surface properties (for example, sea ice albedo).

In addition to the chosen polar summer months, one month from the polar winter was chosen (December 2007).

CMSAF results from all available Arctic overpasses of the NOAA-17, NOAA-18 and Metop-A satellites were matched to and compared with CALIPSO and MODIS cloud datasets. AVHRR scenes were retrieved either using the AVHRR-extended EUMETSAT ATOVS Retransmission Service (EARS-AVHRR) or the global Metop-A 1-km resolution AVHRR scenes archived at the EUMETSAT Data Centre.

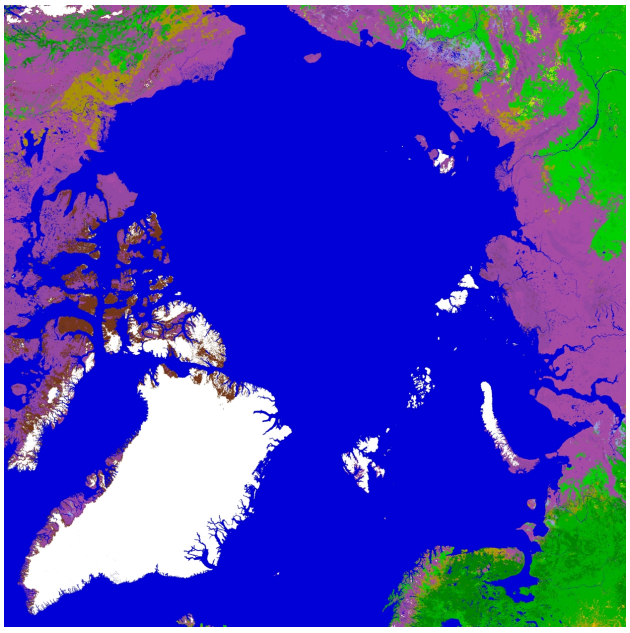

Fig. 1. Chosen region for CMSAF Arctic cloud products. Ocean surfaces are shown in blue and different land use categories (according to classification by US Geological Survey) are shown in different colours.

3 Validation method

3.1 Matching NOAA/Metop AVHRR and CALIPSO-CALIOP observations

The A-Train satellites overfly the NOAA and Metop AVHRR swaths regularly. At a latitude close to 70 degrees (on both hemispheres) the A-Train track frequently crosses the tracks of the NOAA and Metop satellites which means that the two satellites observe exactly the same point on the surface from the nadir view at approximately the same time. We denote these circumstances Simultaneous Nadir Observations (SNOs). For the selection of useful CloudSat/CALIPSO datasets we have set a very strict criterion on the maximum SNO time differences to 2 min in order to achieve as close as possible to simultaneous observations from both datasets. These cases occur roughly each second or third day which means that for each NOAA or Metop satellite we can compare to approximately 10–15 CloudSat/CALIPSO orbits per month in the Arctic region. In other words, about 30–45 comparisons per month are theoretically possible for the total set of the three satellites NOAA-17, NOAA-18 and Metop. Table 2 shows the final achieved frequency of useful cases. The variation between the months is explained by loss of some data (either CloudSat/CALIPSO overpasses or NOAA/Metop overpasses).

The character of the matched scenes (or rather tracks) depends on the relation between NOAA and Metop orbits and the A-Train orbit. SNOs can occur either at a very small angle between orbits (i.e., orbits are very much aligned along-track) or at a large angle between orbits (across-track). The best condition occurs for NOAA-18 tracks. When SNOs occur for this satellite, the NOAA-18 and CALIPSO orbital tracks are almost identical and this gives an opportunity to compare almost simultaneous observations at near nadir conditions for very long distances (e.g. more than 5000 km long tracks in the selected area of Fig. 1). For NOAA-17 and Metop-A the SNOs occur across-track and the length of the colocated observation section is then primarily limited by the AVHRR swath width. A drawback of across-track colocations is that the AVHRR viewing angle changes continuously along the matched track, i.e., from initial large viewing angles into nadir conditions and then back to large viewing angles). This also means that the time difference between AVHRR and CALIPSO observations will vary more (+/− 3–5 min) compared to the case of along track collocation (always close to or within 2 min). Notice here that even if the SNOs are limited to occur within 2 min (and only occurring at one specific pixel), the time difference along the entire matched observation track might exceed 2 min.

Table 2. Number of matched CloudSat/CALIPSO orbits (scenes) and total number of matched AVHRR/CALIOP pixels for all studied months.

Month (2007)	Matched scenes	Matched pixels
June	33	92 116
July	43	115 606
August	38	116 332
December	28	73 785

Figure 2 visualizes the final coverage of all matchup tracks over the Arctic region for all four months. We notice the very different patterns that emerge for the three different satellites. Best coverage over the whole area is given by the NOAA-18 satellite (in red). Also Metop-A matchups are well distributed over the region even if some loss of data can be seen for December 2007 in the North European part of the Arctic. Matchup tracks for NOAA-17 describe the most limited coverage since the selected tracks tend to line up at the same time of the day (meaning also at the same geographical position) for every occasion. Notice also that with the current orbit constellation of the A-Train satellites it is not possible to cover the area closest to the pole since we are here comparing with data from nadir-looking instruments.

3.2 Validation methods and validation scores

By utilizing the associated geo-location information in the AVHRR and the CloudSat/CALIPSO datasets it is possible

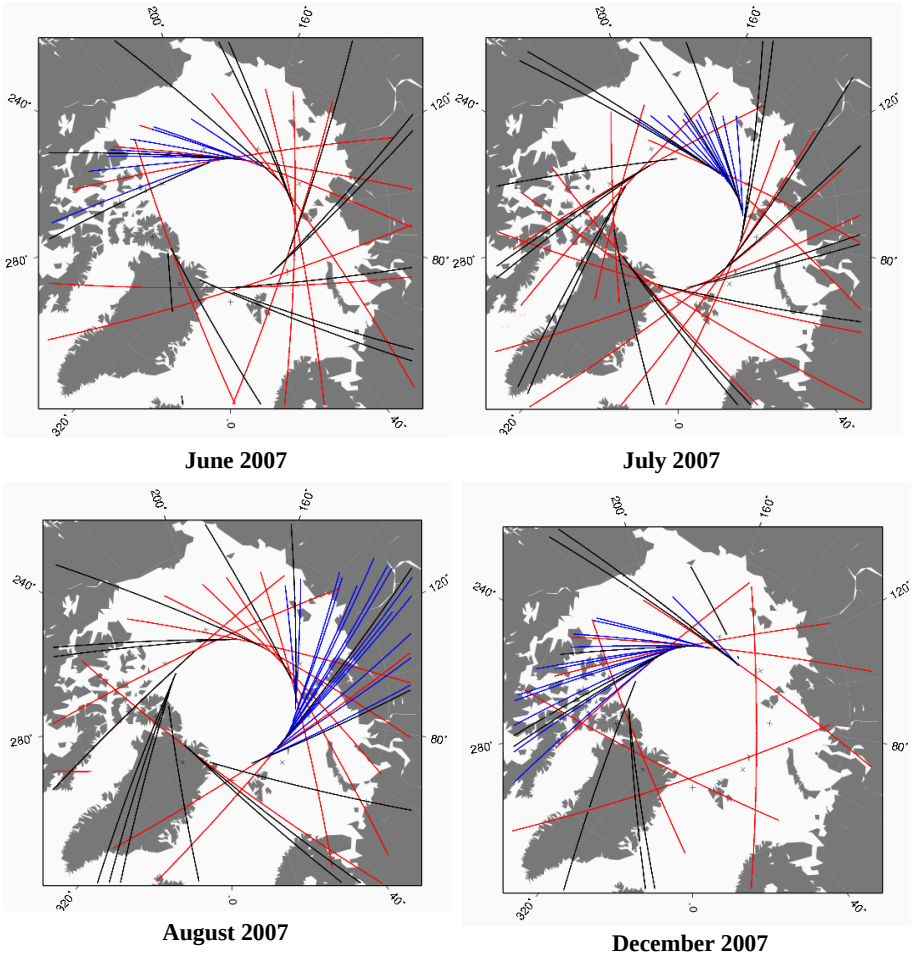

Fig. 2. Finally realised CloudSat/CALIPSO matchup tracks for METOP-02 (Black), NOAA-17 (Blue) and NOAA-18 (Red).

to plot all datasets in the same figure as demonstrated in Fig. 3 (PPS results are here represented by CTH values in blue crosses). Notice that the track starts (position 0 km in Fig. 3) in Russia (northern Siberia), overpasses the Arctic Ocean and finally passes Greenland. Notice also how the CloudSat observation misses near-surface clouds close to the position 2000 km in Fig. 3 as a contrast to the CALIPSO observation from CALIOP.

The calculation of various validation scores is now easily accomplished by comparing results for each individual matchup pixel along the matchup track. For all matchup tracks, each consisting of at maximum about 3000 pixels for NOAA-17 and Metop-A and up to 5500 pixels for NOAA-18, we calculated the following statistical validation scores (some of them only applicable to CFC and CTY products):

1. Mean-error (Bias)
2. Bias-corrected Root Mean Square Error (bc-RMS = standard deviation of Bias)
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)
6. Kuiper's skill score (KSS)

In the case of cloud occurrence (CFC) and cloud type occurrence (CTY), we have simply 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. Subsequently, we have then accumulated results over the full matchup track to get a mean CFC (according to Eq. (1) below) or CTY value and the associated Bias and bc-RMS values. As a final step, all matchup results for each individual month are then accumulated and averaged. Observe again that for CFC results we have also computed the

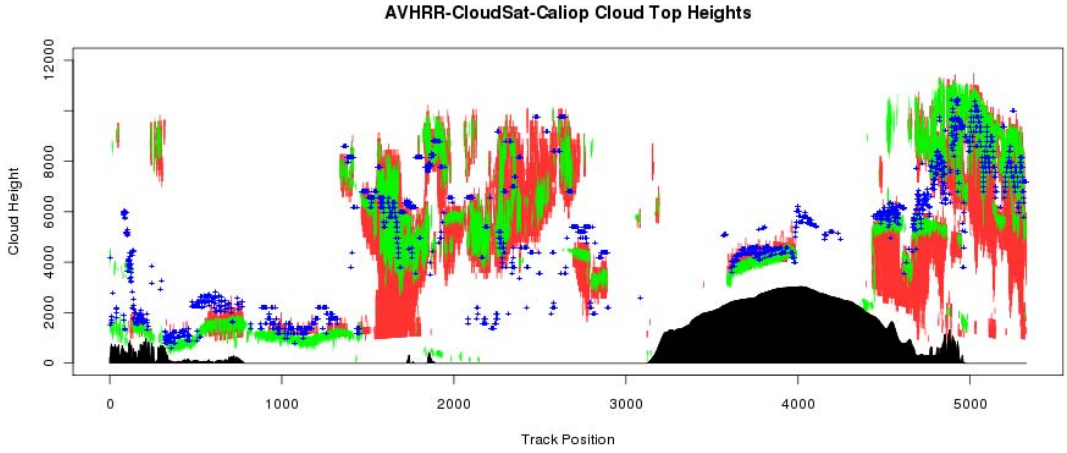

Fig. 3. Cross section plot of matched PPS (NOAA-18, 27 July 2007 at 06:12 UTC) and CloudSat/CALIPSO results as a function of the length along the matchup track (in km). Colour description: Red = CloudSat cloud mask, Green = CALIPSO cloud mask, Blue crosses = PPS cloud top heights. Topography along the matchup track is shown in black (e.g. parts of Siberia are visible in the leftmost part and parts of Greenland are visible in the right part between 3000 to 5000 km).

Table 3. Contingency matrix for the two different satellite observations.

Scenario	PPS Cloud-free	PPS Cloudy
CALIPSO Cloud-free	*a*	*b*
CALIPSO Cloudy	*c*	*d*

corresponding results from the MODIS cloud mask as an additional validation reference.

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

For the remaining four quantities we have used the following definitions (referring to notations in the contingency matrix in Table 3):

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

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

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

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

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

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

The POD and FAR quantities estimate how efficient 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 minimized. The hit rate HR (sometimes also denoted *Accuracy*) 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 would be missed by PPS the HR score will still be high. A more reasonable measure to use in such a condition is the KSS score that at least to some extent punishes misclassifications even if they are in a small minority of all the studied cases. The KSS score tries to answer the question how well the estimation separated the cloudy events from the cloud-free events. A value of 1.0 is in this respect describing the situation of a perfect discrimination while the value −1.0 describes a complete discrimination failure.

3.3 Further specifications of validation conditions for cloud fractional cover CFC and accompanying sensitivity tests

All of the six defined validation scores in the previous subsection were calculated on the total matchup dataset for the evaluation of the CFC parameter. However, in order to compensate for possible mismatches due to small navigation errors and SNO time differences, a post-processing of the PPS cloud mask was applied based on a majority voting procedure using three adjacent pixels (i.e., the centre pixel and the two adjacent pixels along the track).

In order to more easily interpret the results, the study included two sensitivity tests which are described in the following two sub-sections.

3.3.1 CFC sensitivity test 1 – importance of very thin cloud layers

According to the description of the CALIPSO-CALIOP dataset in Sect. 2.2 there is still some remaining uncertainty in the CALIOP-interpreted thin clouds part of the full cloud dataset. Some aerosol cases have apparently been noticed to be mis-classified into very thin clouds. To investigate the influence of this potential error and also for taking into account that there is definitely also a limit of methods based on passive imagery for detecting very thin Cirrus clouds (e.g. as accounted for by Karlsson et al., 2008) we have defined a sub-set of the dataset (denoted *Dominant clouds)* where the thinnest clouds have been removed. The *Dominant Cloud dataset* can in some sense be seen as the radiatively significant portion of all clouds.

Since we did not have access to more elaborated CALIOP parameters like the cloud optical thickness (at least not in the 1-km product) for making this definition of the dominant clouds, we can try to estimate the apparent top of atmosphere (TOA) cloud emissivity E_c according to the following formula (Heidinger, 2008, personal communication)

$$E_c = \frac{I - I_{\text{clear}}}{B(T_c) - I_{\text{clear}}} \tag{8}$$

where

I = *Total measured 11 micron TOA radiance (AVHRR channel 4)*

I_{clear} = *Simulated cloud-free 11 micron TOA radiance*

T_c = *Cloud layer temperature*

$B(T_c)$ = *Planck radiance for cloud layer temperature*

To estimate I_{clear} we can assume a surface emissivity of 1.0 (which should be valid in the Arctic environment) and then use the Planck radiance for the surface temperature in PPS auxiliary datasets (i.e., GME-analysed surface temperatures). We will then also neglect the influence of e.g. emissions from atmospheric water vapor which is relatively small in the dry Arctic atmosphere. The cloud layer temperature is available in the CALIOP dataset (parameter *mid- layer temperature)*. However, notice that we have not applied this method to cloud layers below an altitude of 2 km in order to take into account uncertainties in the GME-analysed surface temperature.

After some initial tests it was decided to use the threshold E_c=0.2 to separate the thin cloud part and the *Dominant Cloud dataset*.

3.3.2 CFC sensitivity test 2 – importance of underlying surfaces

As mentioned briefly in Sect. 2.2 it is possible to sub-divide CALIOP results into different categories related to the existing surface conditions under which the measurements took place. By doing this we might get some more clues as to what is causing particular problems for the cloud detection process. In addition, it can also provide some more details for understanding the differences between PPS and MODIS results.

Table 4. Definition of five specific Earth surface categories along the CloudSat/CALIPSO track.

Name of category	Description
ICE_COVER_OCEAN	*NSIDC ice cover >10%* *IGBP land cover =17 (water bodies)*
ICE_FREE_OCEAN	*NSIDC ice cover <10%* *IGBP land cover =17 (water bodies)*
SNOW_COVER_LAND	*NSIDC ice cover = 101 (Permanent ice) or 104 (snow)* *IGBP land cover ≠17 (all surfaces except water bodies)*
SNOW_FREE_LAND	*NSIDC ice cover = 0* *IGBP land cover ≠17 (all surfaces except water bodies)*
COASTAL_ZONE	*NSIDC ice cover = 255 (mixed pixels at coastlines where it is not possible for microwave-based algorithm to correctly separate ice/snow from ice-/snow-free surfaces)*

Thus, by combining the NSIDC ice product and the IGBP land use classification it is possible to isolate the study to focus on one of five categories according to Table 4. Regarding the last category here (COASTAL_ZONE) it should be mentioned that the original NSIDC grid has a horizontal resolution of 25 km which then determines the character of the used coastal zone (i.e., here a rather wide zone along coasts).

3.4 Further specifications of validation conditions for cloud type CTY

If comparing cloud type definitions according to the PPS in Sect. 2.1.2 and CALIPSO-CALIOP in Sect. 2.2 we notice that a direct comparison is not possible without rearrangement of some categories. Thus, we have applied the cloud type matching given in Table 5. Notice that the fifth PPS CTY category of Fractional clouds cannot be matched to any specific CALIOP Vertical Feature Mask category. Thus, account must also be taken separately to occurrences in this category when analyzing the results.

The binary approach for calculation of validation scores has also been applied to the study of the CTY parameter where now the binary cloud mask was modified as follows (see also Table 6):

Table 5. Evaluated groupings of cloud categories (left column) related to original CALIOP-VFM categories (middle column) and CM-SAF categories (right column).

Cloud Type category	Corresponding CALIOP-VFM categories	Corresponding CM-SAF Cloud Type categories
Low-level clouds	0 = Low, overcast, thin 1 = Low, overcast, thick 2 = Transition stratocumulus 3 = Low, broken	Opaque Low-level clouds
Medium-level clouds	4 = Altocumulus 5 = Altostratus	Opaque Medium-level clouds
High-level clouds	6 = Cirrus 7 = Deep convective	Opaque High-level clouds + Semi-transparent Cirrus clouds

Table 6. Contingency matrix concerning cloud type categories.

Scenario	PPS Cloud-free or other category than chosen	PPS cloud type category
CALIPSO Cloud-free or other category than chosen	*a*	*b*
CALIPSO cloud type category	*c*	*d*

– Category cloudy is replaced by the specific cloud type category (one of three)

– Category cloud-free is replaced by all other possible CTY realizations (i.e., one of the other two cloud type categories or the case of cloud-free conditions)

3.5 Further specifications of validation conditions for cloud top height CTH

We have compared PPS cloud top height products (CTH) with the corresponding measured maximum cloud top (i.e., the upper boundary of the uppermost cloud layer) as interpreted from the CALIPSO-CALIOP vertical cloud mask information. This can be done on a pixel-by-pixel basis, just as for previously reported CFC and CTY evaluations.

When analysing the CALIOP-retrieved cloud top information we have also used the information in the Vertical Feature Mask product to separate results into sub-groups Low-level clouds, Medium-level clouds and High-level clouds. This could help us to identify if there is any height dependence in the quality of the cloud top product. For this stratification of the results we have used the same sub-division of CALIOP VFM categories as previously outlined in Sect. 3.4 and Table 5. The statistical evaluation of results will naturally be restricted to describe results for the common cloud dataset, i.e. when both PPS and CALIOP report a cloud.

In the CTH case, we will restrict the number of validation scores to just two, namely *Bias* (mean error) and *bc-RMS* (bias-corrected root-mean-squared differences).

4 Results and discussion

4.1 Results for CFC

Table 7 summarizes all matchup results for individual months for the mean CFC, bias and bc-RMS validation scores. Concerning the bias it is clear that PPS performs very good in the polar summer months where biases are only a few percentage and indicating some underestimation of CFC compared to CALIOP results. A trend of an increasing underestimation towards the end of the polar summer period (i.e., August) is seen. This can be attributed to a gradually increasing frequency of less favourable observation conditions (i.e., twilight and night conditions). Overall results are quite comparable to the MODIS results, although the latter seem to give slightly higher CFC values (even higher than CALIOP). The situation is much worse in December when the PPS bias of −30% indicates that PPS leaves a large fraction (indeed close to 50%) of all clouds undetected. In comparison, MODIS CFC results are here much better (bias of −7%).

The bc-RMS values for MODIS results are higher than PPS for December which indicates a somewhat lower precision of MODIS estimations despite the better accuracy in terms of lower bias value. This fact together with the indication of a small overestimation of MODIS cloud cover in

Table 7. Summary of PPS CFC mean, bias and bc-RMS results over all matchup cases per month. All results are given in cloud cover units (%). Corresponding results based on the MODIS cloud mask are given in brackets.

Month (2007)	Mean CFC CALIOP (%)	Mean CFC PPS (%)	Bias (%) (MODIS)	bc-RMS (%) (MODIS)
June	67.54	66.24	−1.30	40.32
			(1.38)	(40.41)
July	75.26	73.63	−1.63	38.66
			(1.96)	(38.82)
August	79.59	72.59	−7.00	37.92
			(−0.69)	(38.44)
December	62.44	32.11	−30.33	54.67
			(−6.90)	(59.48)

Table 8. Accumulated results for POD, FAR, HR and KSS validation scores for each month. Corresponding values are given for the MODIS CFC in brackets.

Month (2007)	POD Cloudy (%)	POD Clear (%)	FAR Cloudy (%)	FAR Clear (%)	Hit Rate	KSS
June	87.00	76.92	11.31	26.03	0.84	0.64
	(90.16)	(75.27)	(11.65)	(21.39)	(0.85)	(0.65)
July	89.00	73.05	9.06	31.48	0.85	0.62
	(91.67)	(66.74)	(10.65)	(27.52)	(0.86)	(0.58)
August	86.26	80.72	5.42	39.90	0.85	0.67
	(91.83)	(71.49)	(7.37)	(30.84)	(0.88)	(0.63)
December	44.41	88.34	13.63	51.13	0.61	0.33
	(63.92)	(74.17)	(29.25)	(32.22)	(0.69)	(0.38)

the polar summer months and further combined with systematically lower FAR Clear values compared to PPS point at a general tendency for the MODIS cloud mask to be more clear conservative (i.e., a tendency to rather create some artificial clouds than to miss some clouds) than PPS. This behaviour could be understood and motivated for the sake of improving the estimation of surface properties (i.e., minimizing the risk of mistakenly interpreting a cloudy pixel as cloud-free) but it also means that cloud climatologies could be slightly biased. Nevertheless, it is clear that the overall performance of the MODIS cloud mask is significantly better than PPS during the polar winter.

Results for the remaining validation scores (i.e., POD, FAR, HR and KSS) are summarized for each month in Table 8. Again, it is clear that PPS and MODIS CFC estimations appear to be of very similar quality for the polar summer months. Especially, HR and KSS scores are practically identical for these months. Polar winter results for MODIS in December are significantly better than for PPS (especially for parameters like FAR Clear). However, the rather low HR and KSS values prove that the polar winter is still a major challenge for both sensors.

In Table 9 results are shown in the same form as in Table 7 but for the *Dominant Cloud dataset* (defined in the first subsection of Sect. 3.3.1) which results after removing contributions from all columns with the topmost cloud layer having estimated cloud emissivities less than 0.2.

For June, July and August the amount of removed cases with thin clouds was only around 3–5% (3077, 5008 and 6015 samples, respectively) and therefore the validation scores are much the same as in Table 7 for the polar summer months. However, for December the change is considerable. Here almost 25% (17 914 samples) of the CALIOP-detected cloud profiles was characterized as having thin clouds at the top according to the used separation method.

Overall, results improve for PPS (especially in December) indicating that almost half of the CFC deficit seen in Table 7 for December is related to the treatment of the thinnest clouds. The remaining half of the deficit is likely to be explained either by problems in detecting cloud layers below 2 km altitude or by remaining problems for the correct identification of thick high and medium cloud layers.

A noticeable feature in Table 9 is that the MODIS results show an overall overestimation of the amount of *Dominant Clouds* for all months and, in particular, in December. This means that it appears as the MODIS cloud mask algorithm has a higher tendency than PPS to create artificial clouds in areas described by CALIOP as being cloud-free (also

Table 9. Summary of PPS CFC mean, accuracy (bias) and precision (bc-RMS) results over all matchup cases per month but restricted to the *Dominant Cloud dataset* (all clouds with Ec>0.2). All results are given in cloud cover units (%). Corresponding results for MODIS are given in brackets.

Month (2007)	Mean CFC CALIOP (%)	Mean CFC PPS (%)	Bias (%) (MODIS)	bc-RMS (%) (MODIS)
June	66.46	67.18	0.72	38.43
			(2.89)	(38.49)
July	74.17	73.96	−0.21	37.57
			(2.92)	(37.70)
August	78.51	73.08	−5.43	36.64
			(0.22)	(37.06)
December	50.47	32.58	−17.89	51.23
			(8.30)	(57.54)

Table 10. Bias (mean error) results per month for each Earth surface category. Corresponding values for MODIS are shown in brackets. Notice that the number of samples for the SNOW_FREE_LAND category in December is very small.

	Bias (%) (MODIS)			
Name of category	June	July	August	December
ICE_COVER_OCEAN	1.8	0.9	−0.9	−31.0
	(3.1)	(2.2)	(2.4)	(−16.0)
ICE_FREE_OCEAN	1.6	5.1	−5.6	−7.9
	(7.3)	(12.9)	(3.5)	(−3.8)
SNOW_COVER_LAND	−9.2	−11.6	−19.5	−25.6
	(−8.0)	(−4.0)	(−5.5)	(−0.5)
SNOW_FREE_LAND	−7.1	−6.4	−10.5	−29.1
	(−2.1)	(−4.7)	(−8.8)	(1.31)
COASTAL_ZONE	−9.4	−1.9	−5.6	−35.9
	(0.9)	(1.6)	(2.3)	(−17.6)

supported by the high MODIS false alarm rates for cloudy conditions in Table 8). These results confirm the suggestion expressed earlier that the MODIS cloud mask is more clear conservative than PPS.

Concerning the study of the influence of the underlying surface, we limit ourselves here to only showing the results for the Bias parameter in Table 10. The reason is mainly because the subdivision into several sub-categories results in less significant statistical samples (e.g., for categories like SNOW_FREE_LAND in December 2007). Results allow us at least to get some indication under which surface conditions the PPS cloud masking method works best. This occurs clearly under conditions of ICE_FREE_OCEAN where especially results in December are much better than for other categories. Remarkable and encouraging is the good performance for ICE_COVERED_OCEAN during the three polar summer months. In contrast, we notice large problems for category SNOW_COVER_LAND where large CFC underestimations occur for all months. This means that PPS cloud detection show some problems over the Greenland ice cap also during the polar summer months. As a contrast, we see an overall PPS underestimation of CFC in December for all surface categories but especially for the COASTAL_ZONE category. A typical case when PPS misses a large part of thin polar winter clouds is shown in Fig. 4. We also see here a PPS tendency to give too high cloud tops for the Low-level clouds (to be discussed further below).

MODIS results in Table 10 show some similar features as PPS but also several deviations from the PPS behavior. For example, MODIS CFC results over ocean surfaces (especially over ice free ocean) in the polar summer indicate some overestimation of CFC as a contrast to the small underestimation of CFC that is seen for PPS. However, the most remarkable feature is the results over the SNOW_COVER_LAND category in December where MODIS actually has a negligible bias (−0.5%) in comparison to the large negative bias (−25.6%) seen for PPS. Thus, in spite of the fact that results degrade also for MODIS in the polar winter (also shown by Holz et al., 2008), it is clear that the access to more spectral channels (including sounding channels) improves cloud masking capabilities in the MODIS case. A remaining question here is why this improvement is not as prominent over ice covered ocean as over land.

A final remark in this context is that the study of the dependence on underlying Earth surfaces for cloud masking results should be made with much larger datasets in the future to improve the statistical significance.

4.2 Results for cloud type CTY

Table 11 shows overall results for the Low-level cloud category for Bias and bc-RMS parameters and for all studied months. The subsequent Tables 12 and 13 show similar results for the Medium-level and High-level cloud categories. The contribution from the missing category Fractional clouds for CMSAF PPS is given in Table 14. Notice that the contributions from the Fractional category must be added to the sum of the bias for all three cloud categories in order to be consistent with the total bias for the CFC parameter in Table 7.

Table 11. Relative and absolute contributions from Low-level clouds together with acquired bias and bc-RMS values (%) calculated for absolute contributions.

Month (2007)	LOW_{rel} CALIOP (%)	LOW_{rel} PPS (%)	LOW_{abs} CALIOP (%)	LOW_{abs} PPS (%)	Bias (%)	bc-RMS (%)
June	49.8	29.8	33.6	19.7	−13.9	49.7
July	38.4	22.6	28.9	16.6	−12.3	46.3
August	42.1	24.4	33.5	17.7	−15.8	48.5
December	55.7	25.2	34.8	8.1	−26.7	47.7

Table 12. Relative and absolute contributions from Medium-level clouds together with acquired bias and bc-RMS values (%) calculated for absolute contributions.

Month (2007)	$MEDIUM_{rel}$ CALIOP (%)	$MEDIUM_{rel}$ PPS (%)	$MEDIUM_{abs}$ CALIOP (%)	$MEDIUM_{abs}$ PPS (%)	Bias (%)	bc-RMS (%)
June	26.6	28.6	18.0	18.9	1.0	50.6
July	32.2	27.3	24.2	20.1	−4.1	49.5
August	29.6	22.7	23.6	16.4	−7.1	46.3
December	15.2	20.1	9.5	6.4	−3.0	44.7

It is clear from these tables that there is currently an imbalance between the three vertical cloud groups in the PPS results. More clearly, the contribution from the Low-level cloud category is too low (generally −10–15%, in December even down to −27%) in comparison to the corresponding contribution given by CALIOP. However, for the Medium-level and High-level groups differences are much less. Concerning the latter two categories results are very good for the High-level group while some underestimation is also seen for the Medium-level category. From the results we conclude that even if CFC results agree very well with CALIOP-retrieved CFC for the polar summer months (according to Table 7) there is a clear indication that PPS produces a too small contribution to the cloud category Low-level clouds. Even if a large part of the missing contribution apparently goes into the category Fractional clouds (see Table 14) or possibly to higher cloud categories it is clear that there is still a fraction missing for the Low-level cloud category. Thus, some of the Low-level clouds remain undetected in the polar summer and largely explain the small underestimation of total cloud amounts that is seen in Table 7. This loss of Low-level clouds is drastically increased in December. Here, we miss a contribution of 26.7% to the absolute total cloud cover from all Low-level clouds. Since the contribution from the Fractional cloud group only contains 2.9% of Low-level clouds and higher cloud categories also show some underestimation it is clear that the majority of these clouds have not been detected at all. Consequently, we have strong evidence that the majority of the missing clouds in December are predominantly Low-level clouds.

Additional statistical validation scores are given in Tables 15 and 16. From these tables we get some additional clues for the interpretation of the results. We recall that the goal is to get the absolute contributions to each vertical cloud group as correct as possible *AND* that for each vertical cloud group the cloud type labeling shall be as correct as possible. For example, we could erroneously label Low-level clouds as Medium-level clouds but still have a correct absolute fraction of Low-level clouds if also Medium-level clouds are misclassified in the same proportion as Low-level clouds. Thus, we want as well the quantity probability of detection (POD) to be maximized and the quantity false alarm rate (FAR) to be minimized. This translates further into the desire to have as high values of the hit rate (HR) and Kuipers skill score (KSS) as possible (where the latter also have to be positive).

A closer look at Tables 15 and 16 suggests that there are special problems related to the correct labeling of the Medium-level cloud category. Relatively low POD values combined with high FAR values indicate that the fairly acceptable bias values seen for Medium-level clouds in Table 12 are to some extent explained by compensating errors. For example, in June almost 72% of all pixels classified

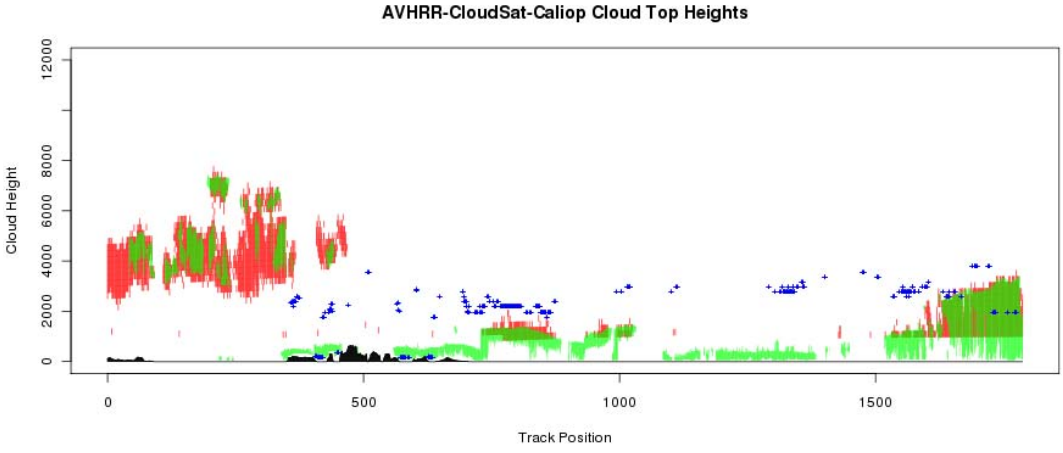

Fig. 4. Cross section plot of matched PPS (NOAA-17, 13 December 2007 at 18:51 UTC) and CloudSat/CALIPSO results as a function of the length along the matchup track (in km). Colour description: Red = CloudSat cloud mask, Green = CALIPSO cloud mask, Blue crosses = PPS cloud top heights. Topography along the matchup track is shown in black (the portion seen here at position 0–50 km and 300–700 km is from islands in northern Canada).

Table 13. Relative and absolute contributions from High-level clouds together with acquired bias and bc-RMS values (%) calculated for absolute contributions.

Month (2007)	$HIGH_{rel}$ CALIOP (%)	$HIGH_{rel}$ PPS (%)	$HIGH_{abs}$ CALIOP (%)	$HIGH_{abs}$ PPS (%)	Bias (%)	bc-RMS (%)
June	23.6	28.2	15.9	18.7	2.8	42.4
July	29.4	32.6	22.1	24.0	1.8	45.7
August	28.2	35.1	22.5	25.5	3.0	45.2
December	29.1	45.6	18.2	14.6	−3.5	41.9

as Medium-level clouds are actually from other cloud categories or they are cloud-free (Table 15). Since we have previously identified an overall lack of Low-level clouds we suspect that a large fraction of those 72% could be misclassified Low-level clouds. Indeed, when studying the composition of these misclassifications in June a bit closer we found that 61% are Low-level clouds, 32% are High-level clouds and 7% were cloud-free. The same kind of failures in the classification of Medium-level clouds is seen also for other months with more or less similar profiles in the misclassified categories. Noteworthy is also the relatively large portion of mis-classified High-level clouds erroneously labeled as Medium-level clouds. We also realize that a large portion of the true Medium-level clouds must then be misclassified as High-level clouds. Otherwise the bias values in Table 12 for Medium-level clouds should have become largely positive. Indeed, FAR values for High-level clouds are also relatively high and if studying the composition of these pixels more closely we found that 70–80% of all misclassified pixels come from the Medium-level category for the polar summer months. However, in December some misclassified High-level clouds also come from the Low-level category.

These mis-classifications are especially reflected in Table 16 in the KSS score which punishes misclassifications harder than other validation scores, even if mis-classifications occur relatively seldom. This leads to rather low KSS values for the Medium-level and High-level cloud categories. Thus, we conclude that the most correct cloud type labeling is actually made for Low-level clouds and that results for both Medium-level and High-level clouds are worse. However, since the true absolute (CALIOP) amount of Low-level clouds is larger than for the other two cloud categories (e.g. as seen in Tables 11–13) we still cannot ignore problems for the Low-level cloud category. This concerns especially the low POD values for December 2007 (Table 15).

Table 14. Absolute contributions (%) from the PPS Fractional cloud category and the corresponding distribution of PPS Fractional clouds among the three CALIOP cloud categories (given in brackets).

	June 2007 *(Low,Medium,High)*	July 2007 *(Low,Medium,High)*	August 2007 *(Low,Medium,High)*	December 2007 *(Low,Medium,High)*
Absolute contribution (%)	8.8	12.9	13.0	2.9
Fractional clouds	*(4.4, 2.6, 1.8)*	*(6.6, 3.5, 2.8)*	*(8.4, 2.6, 2.0)*	*(2.2, 0.6, 0.1)*

Table 15. POD and FAR scores for the three different cloud categories and for all studied months.

Month (2007)	POD Low (%)	POD Medium (%)	POD High (%)	FAR Low (%)	FAR Medium (%)	FAR High (%)
June	52.1	32.11	54.7	32.9	71.6	55.5
July	52.2	40.3	59.6	31.3	53.3	47.4
August	50.4	39.0	65.3	24.1	49.2	48.4
December	39.2	16.7	86.0	25.8	89.5	51.3

Finally, a closer look at the cases of complete misclassifications of cloudy situations (i.e., cases when clouds remain undetected) revealed that for December misclassifications actually concern all cloud categories but with highest contributions from Low-level clouds (generally more than 50%). For the polar summer months, PPS mainly misses Low-level clouds. For the cases of false detected clouds we can see that for polar summer months these are mainly Low-level clouds while for December false detection concerns all cloud categories. Overall, the fraction of false detected clouds is anyhow relatively low. Furthermore, the number of truly false-detected clouds is very likely to be even smaller since the effect of small navigation errors for the collocation of NOAA/Metop AVHRR pixels with CALIPSO-CALIOP measurements will give rise to an equal portion of falsely detected clouds and undetected clouds (at least when aggregating results from a large number of matched orbits).

4.3 Results for cloud top height CTH

Table 17 summarizes overall results for comparisons with CALIPSO-CALIOP observations. We notice that PPS is obviously underestimating CTH but that the magnitude of this error does not appear alarming at first sight. However, considering that bc-RMS values are as high as about 2000 m it is clear that the precision in CTH estimations is not good. A further illustration of this problem is given in Table 18 showing the results sub-divided into vertical cloud categories. Here, it is clear that there is a large amount of compensating errors for categories Low-level clouds and High-level clouds. For the former we notice an overall overestimation of CTH of about 1000 m while for the latter we have an underestimation of about 2500 m. We conclude that PPS has a systematic overestimation of cloud tops for Low-level clouds and a systematic underestimation of cloud tops for High-level clouds when applied in the Arctic region. The latter circumstance was already noted in Fig. 3 (most clearly seen along sections 1900–2700 km and 4800–5300 km) while indications of the latter can be clearly noticed in Fig. 4. These results appear to be robust and valid for both polar summer and polar winter months.

Conceptionally, it is understandable to find that High-level cloud tops are underestimated since we are here comparing directly with the uppermost detected cloud boundary from high-sensitive CALIOP measurements. Since high clouds (with ice crystals at top levels) are often diffuse or thin in their uppermost portions a satellite measurement will tend to be based on an average radiance contribution from the upper portion of the cloud rather than just the uppermost cloud boundary. Thus, when matching effective radiances (and associated brightness temperatures) to reference profiles the selected cloud tops will generally be lower than the uppermost cloud boundary.

As a contrast, the underestimation of Low-level clouds is more surprising and less obvious to understand intuitively. The reason is that low clouds most often have relatively small water droplets at high concentrations at the cloud top which leads to clouds with much higher optical thicknesses in upper portions of the cloud compared to high-level ice clouds. Thus, the measured brightness temperatures should

Table 16. HR and KSS scores for the three different cloud categories and for all studied months.

Month (2007)	HR Low	HR Medium	HR High	KSS Low	KSS Medium	KSS High
June	0.84	0.73	0.85	0.46	0.13	0.35
July	0.86	0.81	0.82	0.50	0.29	0.40
August	0.84	0.77	0.82	0.52	0.34	0.40
December	0.89	0.84	0.93	0.52	−0.03	0.38

Table 17. Summary of PPS CTH Bias and bc-RMS results over all matchup cases per month compared to CALIPSO-CALIOP observations.

Month (2007)	Bias (m)	bc-RMS (m)
June	−50	2063
July	−381	2026
August	−492	2007
December	−228	2521

be close to true cloud top temperatures and the matching to reference temperature profiles should therefore be more straight-forward and safe. Nevertheless, an investigation of one of the most extreme cases of overestimating Low-level cloud tops convincingly reveals the basic nature of this problem. Figure 5 shows this case matching NOAA-18 and CALIPSO observations from 2 June 2007 at 20:58 UTC. Remarkably high differences (at some places almost 3 km) between PPS-derived and CALIOP-derived cloud tops for the lowest clouds are seen for sections 0–1400 km and 1800–2600 km along the matched tracks. Two points along the matched track (denoted a and b in Fig. 5) were selected for which the corresponding reference temperature profiles are plotted in Fig. 6. The motivation for choosing exactly these two points for the investigation was that it was considered important to understand why cloud top estimations could be of so different quality within such a short distance.

Figure 6 shows that the reason for the strange "jump" in PPS-interpreted CTH values in a region with only slowly varying cloud top heights (according to CALIOP measurements) is that the NWP-analysed reference profile is not capable of reproducing the low-level temperature inversion accurately enough. Only in a few positions (like in point a) it is possible to match the temperature correctly as the first occurring (searching from surface and upwards) match with the simulated temperature profile (i.e., profile after having corrected for atmospheric moisture effects). In most other positions the first match does not occur until after reaching quite high in the troposphere, e.g., in point b in Fig. 6 at about 3000 m or at the 700 hPa level. In reality (i.e., as indicated by the measured brightness temperatures), the temperature inversion is apparently stronger than the profile provided by the GME analysis. This lack of detail in the reference temperature profile obviously leads to tremendous problems for satellite-based methods trying to estimate cloud top heights.

5 Summary and conclusions

We have been able to examine in detail, and by co-ordinated use of one and the same reference dataset (CALIPSO-CALIOP observations), the performance of the following three CMSAF cloud products: cloud fractional cover (CFC), cloud type (CTY) and cloud top height (CTH). In addition, the evaluation was made over the Arctic region, a region where cloud retrievals from passive imagery are known to be very problematic, mainly because of the very poor contrast in satellite imagery between cloudy and cloud-free surfaces throughout the year. A very strong feature of this study was the possibility to match the NOAA/Metop AVHRR and CALIPSO-CALIOP observations very close in time (mainly less than 2 min time difference for NOAA-18 and less than 5 min time difference for NOAA-17 and METOP) and in space (matched within a few km).

The evaluation was based on 142 selected NOAA/Metop overpasses allowing almost 400 000 individual matchups between AVHRR pixels and CALIOP measurements distributed approximately equally over the four studied months (June, July, August and December 2007).

CALIOP results suggest that CMSAF CFC estimations are very accurate during the polar summer season (June–August 2007) when PPS CFC values differ only a few percent in absolute values from CALIOP results. This was also supported by results in a separate study comparing monthly mean CMSAF results to corresponding MODIS Level 3 products

Table 18. Summary of PPS CTH bias and bc-RMS results over all matchup cases per month compared with CALIPSO-CALIOP observations used as reference. Results are sub-divided into results for groups Low-level, Medium-level and High-level clouds according to CALIOP Vertical Feature Mask classification.

Month (2007)	Bias Low-level (m)	Bias Medium-level (m)	Bias High-level (m)	bc-RMS Low-level (m)	bc-RMS Medium-level (m)	bc-RMS High-level (m)
June	1203	−359	−2632	1631	1428	3177
July	894	−44	−2489	1417	1288	3093
August	566	21	−2558	1214	1234	3223
December	938	−807	−2886	2124	2059	3486

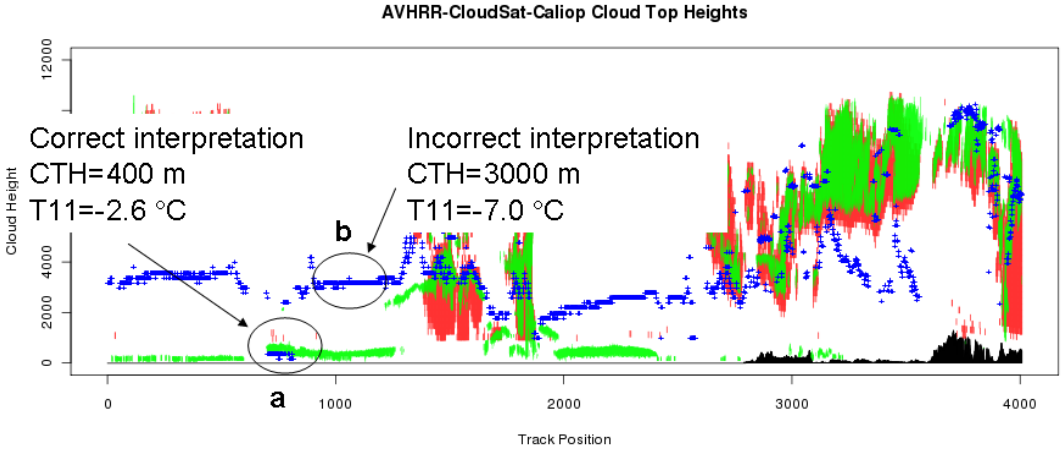

Fig. 5. Cross section plot of matched PPS (NOAA-18, 2 June 2007 at 20:58 UTC) and and CloudSat/CALIPSO results as a function of the length along the matchup track (in km). Colour description: Red = CloudSat cloud mask, Green = CALIPSO cloud mask, Blue crosses = PPS cloud top heights. Two specific points along the cross section (points a and b) were selected for deeper investigations of the results (see text).

(Karlsson, 2008). However, AVHRR-based cloud detection during the dark and cold polar winter period is very challenging and, as expected, results during the studied polar winter month were not as good as for the polar summer months. A very large part of CALIOP-observed clouds were left undetected leading to an underestimation of CFC of 30% in absolute cloud amount units. This is almost 50% of all encountered clouds since the overall CALIOP-retrieved CFC was about 62%. Sensitivity tests with removal of the thinnest clouds improved results showing that the failing cloud detection is to a large extent connected to the identification of thin cloud layers. Sensitivity tests also showed that cloud detection in the Arctic region works very efficiently over ice-covered ocean in the polar summer but that serious problems occur over all very cold snow- and ice-covered surfaces in the polar winter. Since the latter conditions prevail during the entire year over the Greenland ice plateau we also find some problems here in the polar summer. The persisting low surface temperatures here make identification of thin cirrus clouds more problematic compared to e.g. over the warmer ice covered ocean surfaces. Because strongly reflecting clouds with cloud top temperatures close to surface temperatures will still be identified, the net effect is that the polar summer cloud detection performance over Greenland will be better than the corresponding performance during the polar winter but worse than over all other surfaces in the Arctic region.

Comparisons with corresponding results from MODIS showed very good agreement during the polar summer although somewhat higher MODIS CFC values indicate a more clear conservative behaviour of the MODIS cloud mask (especially over ice free ocean). Results from December 2007 show significantly better results for MODIS (overall CFC underestimation limited to 7% compared to 30% for PPS), especially over the land portions of the Arctic. Despite this, the low hit rates and Kuipers Skill scores indicate remaining severe problems in the polar winter for both schemes.

A final remark on the CALIOP-based study of CFC is that the overall underestimation of cloud amounts in December

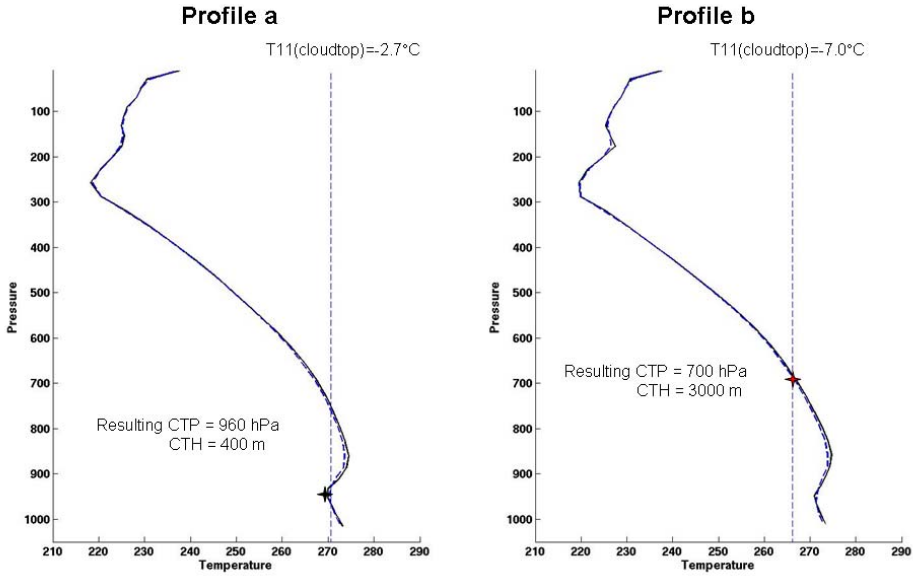

Fig. 6. Temperature profiles from GME analyses for positions closest to the two selected points **(a** and **b)** in Fig. 5. The dashed curve is the resulting top of atmosphere brightness temperature if accounting for atmospheric moisture contributions. Vertical line indicates the measured brightness temperatures (T11) in the two points. Interpreted cloud top pressure (CTP) and cloud top heights (CTH) is also indicated in the figure.

2007 might have been smaller if a better distribution of matchup tracks had been accomplished. If comparing with Fig. 2 it is seen that most observations are (thanks to the occurrence of NOAA-17 matchups) taken from the Alaskan-Canadian part of the Arctic while the North European part is rather sparsely covered. Surface temperatures were therefore biased towards the colder side which clearly made the AVHRR cloud detection task more problematic.

Concerning the cloud type product (CTY) we have found an underestimation of the contribution from the Low-level cloud category by 12–16% in absolute units during the polar summer. Contributions from Medium-level and High-level categories were at the same time relatively close to CALIOP-derived contributions (generally within +/−5% in absolute units). Since total cloud amounts have been found to be rather accurate this means that a large part of the lacking Low-level clouds have been captured in the category Fractional clouds. The remaining part of the mis-classified Low-level clouds went into the Medium-Level category. However, the study also revealed that since false alarm rates for the detection of individual categories were rather high for Medium-level and High-level cloud categories it was clear that also a significant portion of Medium-level clouds were mis-classified as High-level clouds and vice versa.

For the month of December the lack of Low-level clouds increased to almost 27% in absolute units. Here, it was clear that the majority of these missing clouds were also missed in the initial cloud detection process. However, it was also clear that many of the detected Low-level clouds were falsely labeled as Medium-level clouds leading to a high false alarm rate and very low Kuipers skill scores for Medium-level clouds.

Evaluation of cloud top height results (CTH) showed quite fair overall results. i.e., an underestimation of less than 500 m. However, further studies revealed that large deviations with opposite signs occurred for cloud tops of Low-level clouds (being overestimated) and High-level clouds (being underestimated) which then explain the good overall results. For High-level clouds the underestimation is very large and for all studied months between 2500 and 3000 m. Results improved if removing cases where the highest cloud tops could be judged to be very thin and we conclude that the large deviation is to a large extent explained by the existence of clouds with diffuse upper portions.

Currently used cloud top height retrieval methods based on passive satellite imagery in infrared window regions are generally not able to retrieve the corresponding true height of the uppermost boundary of the clouds. Satellite measurements tend to be based on an average radiance contribution from the upper portion of the cloud rather than just the uppermost cloud boundary. Consequently, one could claim that for diffuse clouds we should be more interested in this radiatively efficient cloud height representation than in the true upper boundary of a cloud. In that sense the current validation method is punishing the results unreasonably strong for the high cloud group. Attempts to compensate for this effect could be considered but the lack of consensus on how to do it appropriately made us to stick here to the simple comparison to the uppermost CALIOP-observed cloud boundary.

The discovered overestimation of the Low-level cloud tops (especially during polar summer) is a more serious problem since the majority of these clouds are water clouds which are optically thick. It was found that the problem originates from the fact that the used reference temperature profile is not capable of re-producing strong enough temperature inversion in the lowest layers. This fact and the circumstance that temperature variations often are very small from the surface up to about 3 km or 700 hPa in the polar summer explains why Arctic CTH retrievals are very problematic. We conclude that for cloud top estimation methods like the current PPS method, relying on the matching of brightness temperatures to simulated profiles of brightness temperatures from opaque ("black") clouds inserted at various levels in the troposphere, results for near-surface clouds in the Arctic region will be poor. Improvements here will only be realized (but not guaranteed) if NWP data assimilation methods are improved to better describe true temperature profiles in the Arctic environment throughout the year. However, even if this will be realized it is clear that problems will still remain since typical Arctic conditions in the lower troposphere in the polar summer season are very close to being completely isothermal in the lowest 1–3 km of the troposphere.

It should also be noted that the different problems seen for cloud top information retrieved for Low-level and High-level clouds, respectively, explain to a large extent the previously reported mis-classification problems for the Cloud Type product. The reason is naturally that Cloud Type discrimination relies very much on retrieved cloud altitude information.

As a final remark we want to stress that the access to CALIPSO-CALIOP measurements has proven to drastically improve the possibilities to evaluate cloud products from traditional satellite data sources. The current validation experiment is by far the most detailed evaluation of CMSAF cloud products that has been carried out so far. We will continue to exploit this new observation dataset for evaluating CMSAF products from both polar and geostationary satellite platforms and for other locations on the globe.

Acknowledgements. We would like to thank Andrew Heidinger for valuable contributions concerning the actual method for matching the NOAA/Metop and A-train datasets. Valuable comments on the content of the manuscript were also given by Abhay Devasthale, Vincenzo Levizzani and Jürgen Fischer.

This work was co-sponsored by SMHI, EUMETSAT and the Swedish National Space Board (contract 106/08:1).

Edited by: J. Quaas

References

Ackerman, S. A., Holz, R. E., Frey, R., Eloranta, E. W., Maddux, B., and McGill, M. J.: Cloud detection with MODIS: Part II Validation, J. Atmos. Ocean. Tech., 25, 1073–1086, 2008.

Bony, S., Colman, R., Kattsov, V. M., et al.: How well do we understand and evaluate climate change feedback processes?, J. Climate, 19, 3445–3482, 2006.

Chevallier, M. M. F. and Tjemkes, S.: An improved general fast radiative transfer model for the assimilation of radiances observations, Technical Memorandum 345, European Center for Medium Range Weather Forecasting (ECMWF), 2001.

Derrien, M., Lavanant, L., and Gleau, H. L.: Retrieval of the cloud top temperature of semi-transparent clouds with AVHRR, In: Proceedings of IRS'88, Lille, France, 199–202, 1988.

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

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

Eliasson, S., Tetzlaff, A., and Karlsson, K.-G.: Prototyping an improved PPS cloud detection for the Arctic polar night, SMHI Reports Meteorology, available from: SMHI, Folkborgsvägen 1, 601 76 Norrköping, Sweden), No. 128, 37 pp., 2007.

Frey, R. A.; Ackerman, S. A., Liu, Y., Strabala, K. I., Zhang, H., Key, J. R., and Wang, X.: Cloud detection with MODIS: Part I Improvements in the MODIS Cloud Mask for Collection 5, J. Atmos. Ocean. Tech., 25, 1057–1072, 2008.

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

Inoue, T.: On the temperature and effective emissivity determination of semi-transparent cirrus clouds by bi-spectral measurements in the 10 µm window region, J. Meteorol. Soc. Jpn., 63(1), 88–98, 1985.

IPCC4: 4th Assessment report Intergovernmental Panel on Climate Change: Climate Change 2007 – Synthesis report, Core Writing Team, edited by: Pachauri, R. K. and Reisinger, A., IPCC, Geneva, Switzerland, 104 pp., 2009.

Karlsson, K.-G.: CMSAF Arctic cloud studies during the 2007 polar summer extreme sea ice anomaly event, in: Proceedings of 2008 EUMETSAT Satellite Conference, Darmstadt, Germany, 8–12 September 2008, 7 pp., 2008.

Karlsson, K.-G., Willén, U., Jones, C., and Wyser, K.: Evaluation of regional cloud climate simulations over Scandinavia using a 10-year NOAA Advanced Very High Resolution Radiometer cloud climatology, J. Geophys. Res., 113, D01203, doi:10.1029/2007JD008658, 2008.

Kay, J. E., L'Ecuyer, T., Gettelman, A., Stephens, G., and O'Dell, C.: The contribution of cloud and radiation anomalies to the 2007 Arctic sea ice extent minimum, Geoph. Res. Lett., 35, L08503, doi:10.1029/2008GL033451, 2008.

Korpela, A., Dybbroe, A., and Thoss, A.: Retrieving Cloud Top Temperature and Height in Semi-transparent and fractional cloudiness using AVHRR, SMHI Reports Meteorology, available from SMHI, Folkborgsvägen 1, 60 176 Norrköping, Swe-

den, 100, NWCSAF Visiting Scientist Report. 2001.

Li, J., Menzel, P., Yang, Z., Frey, R., and Ackerman, S.: High-spatial-resolution surface and cloud-type classification from MODIS multispectral band measurements, J. Appl. Meteor., 42, 204–226, 2003.

Majewski D., Liermann, D., Prohl, P., Ritter, B., Buchhold, M., Hanisch, T., Paul, G., Wergen, W. and Baumgardner, J.: The operational global icosahedral-hexagonal grid point model GME: Description and high resolution tests, Mon. Weather. Rev., 130, 319–338, 2002.

Platnick, S., King, M. D., Ackerman, S. A., Menzel, W. P., Baum, B. A. Riedi J., and Frey, R. A.: The MODIS cloud products: Algorithms and examples from Terra, IEEE T. Geosci. Remote, 41, 1–15, 2003.

Reuter, M., Thomas, W., Albert, P., Lockhoff, M., Weber, R., Karlsson, K.-G., and Fischer, J.: The CMSAF 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, 2009.

Rossow, W. B. and Schiffer, R. A.: Advances in understanding clouds from ISCCP, B. Am. Meteorol. Soc., 80, 2261–2288, 1999.

Schulz, J., Albert, P., Behr, H.-D., et al.: Operational climate monitoring from space: The EUMETSAT Satellite Application Facility on Climate Monitoring (CMSAF), J. Atmos. Chem. Phys., 9, 1687–1709, 2009, http://www.atmos-chem-phys.net/9/1687/2009/.

Stephens, G. L.: Cloud feedbacks in the climate system: A critical review, J. Climate, 18, 237–273, 2005.

Stephens, G. L., Vane, D. G., Boain, R. J., et al., and CloudSat Science team: The CLOUDSAT mission and the A-Train, B. Am. Meteorol. Soc., 83, 1771–1790, 2002.

Weisz, E., Li, J., Menzel, W. P., Heidinger, A. K., Kahn, B. H., and Liu, C. Y.: Comparison of AIRS, MODIS, CloudSat and CALIPS cloud top height retrievals. Geoph. Res. Lett., 34, L17811, doi:10.1029/2007GL030676, 2007.

Stubenrauch, C. J., Chédin, A., Rädel, G., Scott, N. A., and Serrar, S.: Cloud properties and their seasonal and diurnal variability from TOVS Path-B, J. Climate, 19, 5531–5553, 2006.

Winker, D., Vaughan, M., and Hunt, B.: The CALIPSO mission and initial results from CALIOP, Proc. SPIE, 6409, doi:10.1117/12.698003, 2006.