Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

Detailed characterisation of AVHRR global cloud detection performance of the CM SAF CLARA-A2 climate data record based on CALIPSO-CALIOP cloud information

Karl-Göran Karlsson[1], Nina Håkansson[1]

[1]Swedish Meteorological and Hydrological Institute, Folkborgsvägen 17, 601 76 Norrköping, Sweden

Correspondence to: Karl-Göran Karlsson (Karl-Goran.Karlsson@smhi.se)

Abstract. The cloud detection performance of the cloud mask being used in the CM SAF cloud, albedo and surface radiation dataset from AVHRR data (CLARA-A2) cloud climate data record (CDR) has been evaluated in detail using cloud information from the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) onboard the CALIPSO satellite. Validation results, including their global distribution, have been calculated from collocations of AVHRR and CALIOP measurements over a ten-year period (2006-2015). The sensitivity of the results to the cloud optical thicknesses of CALIOP-observed clouds were studied leading to the conclusion that the global cloud detection sensitivity (defined as the minimum cloud optical thickness for which 50 % of clouds could be detected) was estimated to 0.225. After applying this optical thickness threshold to the CALIOP cloud mask, results were found to be basically unbiased over most of the globe except over the polar regions where a considerably underestimation of cloudiness could be seen during the polar winter. The probability of detecting clouds in the polar winter could be as low as 50 % over the highest and coldest portions of Greenland and Antarctica, showing that also a large fraction of optically thick clouds remains undetected here. The study included an in-depth analysis of the probability of detecting a cloud as a function of the vertically integrated cloud optical thickness as well as of the cloud's geographical position. Best results were achieved over oceanic surfaces at mid-to-high latitudes were at least 50 % of all clouds with an optical thickness down to a value of 0.075 were detected. Corresponding cloud detection sensitivities over land surfaces outside of the polar regions were generally larger than 0.2 with maximum values of approximately 0.5 over Sahara and the Arabian Peninsula. For polar land surfaces the values were close to 1 or higher with maximum values of 4.5 over the geographically highest parts of Greenland and Antarctica. The validation method is suggested to be applied also to other satellite-based CDRs and validation results are proposed to be used in Cloud Feedback Model Intercomparison Project (CFMIP) Observation Simulation Package (COSP) simulators for cloud detection characterisation of various cloud CDRs from passive imagery.

1 Introduction

Monitoring the global amount and distribution of clouds as well as assessing the optical properties of clouds appear increasingly important following the growing insight that cloud description and cloud-aerosol feedback processes stand out as key uncertainty factors in climate change analysis and in climate predictions from climate models (Stocker et al., 2013). However, encouraging in this aspect is the steadily increasing amount of observations from space from passive and active sensors (an excellent overview is available at https://www.wmo-sat.info/oscar/) and the continuous prolongation of the observational record for some initial satellite sensor families since the time of introducing reliable and sustainable satellite observation systems back in the 1970's. These early satellite observations, basically consisting of spectral radiance measurements, can be used to retrieve information on clouds and other relevant Earth-Atmosphere parameters. Most important is that they have now evolved into time series of observations with lengths approaching four decades which qualifies them for use as climate data records (CDRs) in combination with other Earth surface-based climate observations.

Examples of CDRs built upon such observations are described by Rossow and Schiffer (1999), Karlsson et al. (2017), Heidinger et al. (2014) and Stengel et al. (2017).

From the climate analysis perspective, the advantage of satellite-based observations is naturally the global view. A similar view is very difficult to achieve from surface-based observations alone because of the inhomogeneous and sometimes lacking coverage of the surface-based observational network. Similar to many other kinds of observations, this concerns also observations of cloudiness and the information on cloud properties, where large parts of the Earth (e.g. oceanic and polar regions) are still poorly covered. However, the different observation capabilities and conditions for space-based sensors and Earth surface-based observations lead to problems when trying to characterise the accuracy of space-based CDRs. Although the quality of observations may be estimated for selected Earth positions or for smaller regions with dense surface networks, it is very difficult to achieve a representative and homogenous view of the accuracy over the entire globe using surface observations. The importance of the CDR quality aspect reflects the fact that observations used for climate monitoring must be very accurate to be able to reliably estimate potential climate change signals (Ohring et al., 2004) which is a central aspect in the planning and definition of the global climate observing system (Dowell et al., 2013). Linked to this are also recent efforts for becoming more stringent in the description of the uncertainty of CDRs by following international metrological norms (Merchant et al., 2017).

One solution for achieving both the global coverage and a better prospect for quality description is to introduce and make use of high-quality reference measurements from space-borne platforms (Dowell et al., 2013). This has already been successfully demonstrated by utilizing data delivered by the A-Train (i.e., Aqua Train or sometimes referred to as the Afternoon Train) concept, a system of satellites operating in the same orbit configuration and with close to simultaneous observation times (Stephens et al., 2002). Particularly important for the cloud observation topic has been one of the satellites in the A-Train: the CALIPSO satellite with the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) onboard (Winker et al., 2009). The sensitivity of CALIOP to clouds in the atmosphere is much higher than for other space-based sensors and this makes it a natural reference for evaluating the cloud detection efficiency in data records compiled from passive sensor data (e.g., as demonstrated by Heidinger et al., 2016).

This paper presents a detailed CALIOP-based evaluation of the cloud detection efficiency and uncertainty of the cloudiness information provided by the CLARA-A2 (The CM SAF cLoud, Albedo and surface RAdiation dataset from AVHRR data" - second edition) CDR (Karlsson et al. 2017). This CDR was released in 2017 by the Climate Monitoring Satellite Application Facility (CM SAF); a project being a part of the satellite ground segment of the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT, Schulz et al., 2009). The evaluation of this CDR is based on an original validation method described by Karlsson and Johansson (2013) but now extended with several new features. The method has been updated to use the latest revision of the CALIPSO-CALIOP dataset and it is also taking advantage of the greatly extended CALIOP observation period (here covering almost 10 years) allowing the monitoring of globally averaged cloud conditions with fine details. A specific enhancement of the method is the estimation of the geographical distribution of cloud detection probability as a function of cloud layer optical thickness. Section 2 describes the CLARA-A2 and CALIPSO datasets, Section 3 outlines the extended validation method and is followed by results in Section 4. Finally, Section 5 summarizes the results and discusses potential applications.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

2 Data

2.1 The CLARA-A2 climate data record.

CLARA-A2 is based on historic measurements of the Advanced Very High Resolution Radiometer (AVHRR) operated onboard polar orbiting NOAA satellites as well as onboard the MetOp polar orbiters operated by EUMETSAT since 2006. AVHRR is measuring in five spectral channels (two visible and three infrared channels) with an original horizontal field of view (FOV) resolution at nadir of 1.1 km, although the data used in CLARA is a reduced resolution (5 km) global area coverage (GAC) resampled version of these measurements. Only GAC data is available globally (i.e., being archived) over the full period since the introduction of the AVHRR sensor in space. The resampling of original data into GAC representation means that four out of five original FOVs are selected for the first scan line while the next two scan lines are ignored. Radiances for these four selected FOVs are averaged and then used to represent the GAC FOV consisting of 15 original full resolution FOVs. Thus, only about 25 % of the nominal GAC FOV is actually observed.

This second CLARA edition is an improved and extended follow-up of the first version of the record (Karlsson et al., 2013) and is now covering a 34-year time period (1982-2015). Original visible radiances were inter-calibrated and homogenised, using MODIS data as a reference, before applying the various parameter retrievals. The inter-calibration was based on an original method introduced by Heidinger et al. (2010) which now has been updated (MODIS Collection 6) and extended (six years have been added). CLARA-A2 features a range of cloud products: cloud mask/cloud amount, cloud top temperature/pressure/height, cloud thermodynamic phase, and (for liquid and ice clouds separately) cloud optical thickness, particle effective radius and cloud water path. Cloud products are available as monthly and daily averages in a 0.25 by 0.25 degree latitude-longitude grid and also as daily resampled global products (Level 2b) in a 0.05 by 0.05 degree latitude-longitude grid. The daily resampled products are valid per satellite and orbit node (ascending or descending). The daily average product is an average of all daily resampled products and the monthly products are averages of all daily average products. Cloud parameter results are also presented as multi-parameter distributions (i.e., joint frequency histograms of cloud optical thickness, cloud top pressure and cloud phase) for daytime conditions. Besides cloud products CLARA-A2 also includes surface radiation budget and surface albedo products. Examples of CLARA-A2 products can be found in Karlsson et al. (2017).

In this study, we focus exclusively on the quality of the original AVHRR GAC cloud mask because of its central importance for the quality of all other CLARA-A2 products. Validation results for other CLARA-A2 products can be found in Karlsson et al. (2017) and in CM SAF 1 (2017). The method for generating the CLARA-A2 cloud mask originates from Dybbroe et al. (2005) but significant improvements and adaptations have been made for enabling reliable processing of the historic AVHRR GAC record (CM SAF 2, 2017).

2.2 The CALIPSO-CALIOP cloud information.

An extensive description of the existing CALIPSO-CALIOP cloud and aerosol datasets can be found in Vaughan et al. (2009). In short, the used Cloud Layer product from CALIOP (denoted CLAY) provides information of up to 10 individual cloud layers in the vertical. However, the detection of a cloud layer requires that all layers above a given layer are thin enough to allow the lidar signal to penetrate down to that particular layer. Thus, in reality the number of layers may be higher if clouds are optically thick. The CLAY products are provided in three different horizontal resolutions (along track): 333 meter (“single shot”), 1 km and 5 km. Coarser resolutions than 333 meters are constructed through averaging over several single shots. This is done to increase the signal to noise ratio for allowing detection of thinner clouds than what can be achieved at the original single shot resolution. Thus, CALIOP products at coarser resolution will be capable of including

more clouds than at finer resolutions and, in particular, studies of thin Cirrus clouds should preferably be based on products in the coarse resolution at 5 km. Notice that the nominal single shot resolution does not correspond to the true lidar FOV but rather to the along-track sampling distance. Consequently, with the true lidar FOV of 70 meters (Winker et al., 2007), less than 5 % of the nominal single shot FOV size is actually observed.

An estimation of the cloud optical thickness of each layer is also provided but only for a FOV resolution of 5 km. To bear in mind here is that, because of signal saturation in thick clouds, the cloud optical thickness values are only reliable for relatively thin clouds, i.e., with cloud optical thickness values below approximately 3 (Vaughan et al., 2009 and Sassen and Cho, 1992). In this study we have used the optical thickness interval 0-5, claiming that useful information also slightly above the suggested value of 3 seems to be available.

The CALIPSO satellite follows the A-Train track in a sun-synchronous orbit with an equator-crossing local time of 01:30. It means that observations from the NOAA satellites can be matched to CALIPSO-CALIOP data in near-nadir conditions for a full orbit if being in an orbit with the same or very close to the same equator-crossing time. For all other NOAA satellite orbits (and now also including the METOP satellites), matchups are only possible at high latitudes close to +/- 70 degrees. Since CALIPSO is operated in a slightly lower and faster orbit than the NOAA/METOP satellites, close matchups in time (i.e., with observation time differences less than 5 minutes) can only be found with an interval of 3-5 days.

In this study, we have used the fourth reprocessed version of the CALIOP CLAY datasets (version 4.10) which was released in 2016. The main features of this updated version are described at https://www-calipso.larc.nasa.gov/resources/calipso_users_guide/qs/cal_lid_l2_all_v4-10.php.

Regarding the basic CALIOP cloud mask, the most relevant changes are

- Revised and improved basic cloud-aerosol-discrimination method
- Removal of mis-classifications of aerosols and dust as clouds at certain locations at high latitudes (as discussed by Jin et al., 2014)
- Inclusion of information on single shot cloud detection in the 5 km dataset (implications to be discussed further in Section 3.2).

3 Validation and analysis methods

3.1 Some theoretical considerations about clouds

Cloudiness is not an absolute well-defined quantity like other cloud properties or most other geophysical parameters. Firstly, it depends on the scale of interest, i.e., you need to specify the aerial extension over which cloud cover has to be calculated. Secondly, and perhaps most important, you need to define what you mean by a cloud, e.g., how thin or thick should a cloud be to be called a cloud? The second aspect is in most cases not well-defined which has made the use of this quantity rather difficult. A good example is comparison studies between satellite-derived and manual surface-observed cloudiness (e.g., Sun et al., 2015). Results from such studies are difficult to interpret because of the different observation geometries for the compared datasets and the lack of an objective and clear definition of the clouds being observed in the surface reference dataset. Because of this ambiguity it has often been recommended to use other parameters than cloudiness or cloud cover (as mentioned in WMO 1, 2012) to instead describe the effect of clouds (e.g. “cloud albedo”, “effective cloud cover” or “joint histograms of cloud top pressure and cloud optical thickness”) in climate analysis and climate model evaluation studies. Nevertheless, the need to get the geographical distribution of modelled clouds correct is still a crucial requirement (as

pointed out in WMO 1, 2012), also bearing in mind that parameters describing the effect of clouds are still critically depending on how you define the underlying cloud or cloud mask. This calls for continued studies of cloud cover from both the observational and modelling perspective. We claim here that the access to high-quality reference cloud observations from CALIPSO-CALIOP may help us to take a significant step forward regarding this aspect. A very strict definition of the clouds we are observing can be made using CALIOP data. Thus, the ability to observe similar clouds in data records based on passive imagery can then be assessed which will augment the usefulness of these data. The following sub-sections will outline the way forward to enhance the value of results from such cloud validation studies.

3.2 Basic CALIOP matching method and adaptation to version 4 CLAY products

The underlying method for matching the two cloud datasets is described in detail by Karlsson and Johansson (2013). It uses a combination of the CALIOP CLAY products at 1 km and 5 km resolutions. The reason for combining the two CLAY datasets with different resolutions, rather than using exclusively the CLAY 5 km version with the same nominal resolution as AVHRR GAC data, was the observation that global cloudiness estimated for each individual orbit is not always increasing when switching from the 1 km dataset to the 5 km dataset. A non-negligible fraction (~ 3-5 %) of all investigated cases in a preparatory study actually showed lower cloud amounts for the 5 km resolution. Ideally, there should be an increase since CALIOP products with coarser resolution should contain more thin clouds which are not seen in the higher resolution products. This inconsistency comes as a side effect of the actual method used for creating the coarser resolution CALIOP datasets (David Winker, CALIPSO Science Team, 2017, pers. comm.). Prior to performing the horizontal averaging of the CALIOP scattering signal over several single shots, some single shot views are excluded from the analysis if containing strongly reflecting boundary layer clouds or aerosols. In the vast majority of cases the number of these removed single shots is less than 50 % of all single shot measurements within the 5 km FOV which would then still justify labelling of the 5 km FOV as cloud free if no other cloud layers are detected. However, in some areas the frequency of small-scale convective clouds may be high and for these cases this could lead to underestimated cloudiness in the 5 km products. Another important aspect is that strongly reflecting clouds on the sub-pixel scale of AVHRR GAC data may still be detectable because of non-linear radiance contributions (with similarities to the “hot spot” effect from fires) in the short-wave infrared channel at 3.7 µm (Saunders and Grey, 1985, and Saunders, 1986). Thus, to not include these clouds in the CALIOP datasets would possibly punish AVHRR-based methods in an unfortunate and undeserved way in the validation process. Karlsson and Johansson (2013) showed also that validation scores improved for AVHRR-based cloud products when adding clouds from the 1 km datasets if 3 or more of the 1 km FOVs within the 5 km FOV were cloudy in cases when the original 5 km products were deemed cloud-free. For these added clouds from 1 km data, the 5 km cloud optical thickness (not estimated in CLAY 1 km data) was set to 5, i.e., at the maximum upper end of realistically estimated cloud optical thicknesses. This should be reasonable since these clouds are by definition strongly reflecting and in most of the cases it would lead to effective cloud optical thicknesses close to or above 5.

An important preparatory step in this study was to check if the method used by Karlsson and Johansson (2013) would still be applicable to the new version 4 release of the CALIOP CLAY product in 2016 and if validation results changed in any systematic way. Of importance here is that the fundamental retrieval method for the CALIOP CLAY product has basically remained the same despite the implemented modifications mentioned at the end of section 2.2. Consequently, the above mentioned inconsistencies between fine and coarse resolution CALIOP datasets are likely to remain and would need a similar post-processing adjustment as for previous version 3 products. However, the new version of the 5 km CALIOP cloud product (i.e., in this study we have used the standard CLAY product version 4.10) has been expanded to include full information of which single shots that were removed during the averaging process. Thus, the previous use of 1 km data in the method by Karlsson and Johansson (2013) could in principle be abandoned and be replaced by the direct use of this

single shot removal information (the latter to be called "modified method" in the following). An additional improvement in the used version 4.10 dataset is that the removed single shot FOVs have also been labelled as being either cloudy or filled with thick aerosols. This separation was not available in version 3 where all removed single shot FOVs were assumed to be cloudy.

The modified method was compared against the old method for a limited test dataset of 80 NOAA-18 matched orbits between October and December 2006 and the results are presented in Figs. 1 and 2. Figure 1 shows validation results for the two different approaches based exclusively on CALIOP CLAY products version 4.10. We are here using the same visualisation of the results for two validation scores (Hitrate and Kuipers score, see also discussion and definition in section 3.3) as in Karlsson and Johansson (2013). Results of using the original CALIOP cloud mask is given by the leftmost value with a filtered cloud optical thickness of 0.0. The curves are constructed by validating against a successively reduced CALIOP cloud mask were clouds being optically thinner than the values at the x-axis have been transformed from cloudy to clear cases. In this way we can estimate for which CALIOP cloud mask (i.e., for which filtered cloud optical thickness) we get the highest scores. Figure 1 shows practically identical results for the two methods or actually slightly improved results for the method using the single shot information. The improvement may come from the improved cloud-aerosol labelling of removed single shots. Figure 2 shows the overall effect of introducing the new matching method and the new version 4 dataset compared to the results achieved using the former version 3 dataset and the previous matching method. We notice a small increase in the overall results (maximum scores) and a progression of the maximum values towards larger optical depths. We believe that the improvement in results reflects an improved CALIOP product and that the shifting of peak score values towards larger filtered cloud optical depths results from more realistic and larger optical depths in CALIOP version 4.10 data (as confirmed by David Winker, CALIPSO Science Team, 2017, pers. comm.). This is quite in line with expectations and we conclude that the modified method is an appropriate basis for further validation studies based on the updated CALIOP CLAY dataset.

3.3 Applied validation concept and validation scores

An important difference in this study compared to the conditions prevailing for the study by Karlsson and Johansson (2013) is the access to CALIOP data for a much longer validation period; almost 10 years (2006-2015). This means that not only overall mean conditions can be approximated but also the geographical distribution of validation results. More clearly, after ten years we have compiled a sufficiently large amount of AVHRR-matched nadir looking CALIOP observations to estimate cloud detection conditions over all locations. Thus, for the first time we can evaluate the quality of a cloud CDR in a (close to) homogeneous way over the entire globe except in the near vicinity to the poles were CALIOP measurements are not available. Consequently, we will here focus on presenting our results in global maps rather than as tables with global mean values. For the plotting of global maps we have rearranged and calculated the results in a global equal-area grid. We have used a Fibonacci grid with 28878 grid points evenly spread out around the Earth approximately 75 km apart. The resulting grid has almost equal area and almost equal shape of all grid cells. Fibonacci grids behave the same near the poles as at the equator, compared to traditional latitude-longitude grids which often behave in a strange way near the poles. For further details on Fibonacci grids, see González (2009) and Swinbank and Purser (2006).

We will estimate the same set of validation scores as those described and defined by Karlsson and Johansson (2013), namely

- Mean error (bias) of cloud amount (%)
- Bias-corrected Root Mean Square Error (RMS) of cloud amount (%)

- Probability of Detection ($0 \leq POD \leq 1$) for both cloudy and cloud-free conditions relative to all observed cloudy or clear cases
- False Alarm Rate ($0 \leq FAR) \leq 1$) for both cloudy and cloud-free conditions relative to all predicted cloudy and clear cases
- Hitrate: Frequency (value between 0 and 1) of correct cloudy and clear predictions relative to all cases
- Kuiper's skill score ($-1 \leq KSS \leq 1$)

Observe that we will treat both CLARA-A2 cloud masks and CALIOP cloud masks as binary values, i.e., each FOV is considered as either fully cloudy or cloud free. This approximation is acceptable for the estimation of cloud amount when we have a large number of matched observations as in this case. The Kuiper's skill score can be used for better identification of mis-classifications in cases when one of the categories is dominating. It 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.

According to, e.g. Merchant et al. (2017), a minimum requirement for describing the accuracy of a parameter is to estimate the mean error or bias (giving the systematic error) and the variance of the error (giving the random error or dispersion). However, for the identification of specific problems with cloud identification it is also useful to look at the other quantities mentioned above and especially in cases when one of the two categories "cloudy" or "clear" is dominating. The latter is motivated by the fact that any cloud contamination (even if it is just a few cases) can have serious implications for parameter retrievals further downstream in the processing. Thus, the use of a rather full toolbox of validation scores can be needed to identify all problematic and critical cases.

3.4 Extension of the original validation method: Enhanced analysis and introduction of cloud layer detection probability

The use of the CALIOP cloud mask for validation of cloud masking methods based on passive imagery is rewarding but also challenging. Especially, we know that a comparison of results with those from the original CALIOP cloud mask means that we compare results from a high-sensitive sensor to results from sensors with a lower sensitivity. The question is: How shall we handle this sensitivity difference and still get useful results?

Of importance here is that there are two major risks with comparing results to the original CALIOP cloud mask:

1. The CALIOP dataset will include sub-visible clouds (Martins et al., 2011) which are not possible to detect in passive imagery.
2. In areas where sub-visible clouds exist in abundance, a method may have been 'overtrained' or 'overfitted' (if trained with CALIOP data) to always predict clouds since this gives the best overall validation scores.

These two problems can be handled by zooming in on what happens for clouds having different vertically integrated optical thicknesses as provided by the CALIOP 5 km product. We recall that the use of successively reduced CALIOP cloud masks (as applied in Figs 1 and 2) means that we exclude the thinnest clouds from the analysis by transforming them to be interpreted as cloud-free FOVs. This also means that if we isolate clouds within finite cloud optical thickness intervals (i.e., resulting from subtracting two adjacent restricted CALIOP cloud masks with different filtered cloud optical thickness) we can calculate validation results exclusively for those clouds. If the cloud optical thickness interval is sufficiently small and

the number of samples within this particular interval is sufficiently high we may then estimate the method's efficiency in detecting a cloud (i.e., the cloud layer detection probability $POD_{cloudy}(\tau)$) with this particular cloud optical thickness given by the mean cloud optical thickness in this interval. We may then expect to see low detection scores for low optical thicknesses but scores that will increase with increasing cloud optical thickness values. We argue that a special situation occurs when this cloud layer detection probability for the first time exceeds 50 % for increasing cloud optical thicknesses. It marks an important performance point: At this cloud optical thickness we detect at least 50 % of all clouds. In the following we will denote this value of the filtered cloud optical thickness as the method's ***cloud detection sensitivity***. It is also clear that we should get a peak in the Hitrate parameter at exactly this point. For lower optical thicknesses, scores improve if we filter out thin clouds, while for higher optical thicknesses scores start to decrease since we then transform too many correctly detected clouds to the cloud-free case. We argue that the best way of evaluating a cloud masking method would be to estimate this cloud sensitivity parameter and to compute all validation scores after applying optical thickness filtering with exactly this value. This would describe a methods optimal performance when using CALIOP cloud masks as the reference. The cloud detection sensitivity parameter would define the method's cloud detection capability in terms of the thinnest cloud being detected with confidence and the validation scores computed at this particular value of the filtered optical thickness would define the method's optimal performance taking into account also false classifications. An important additional or complementary parameter in this context would be the false alarm rate in the unfiltered case ($FAR_{cloudy}(\tau=0)$) since this parameter is not depending on any filtering of thin clouds. This parameter could preferably be used to investigate the degree of overtraining of a method (according to second bullet above). In the following we will present results of the cloud detection sensitivity, a range of validation scores computed at the point of the cloud detection sensitivity and $FAR_{cloudy}(\tau=0)$. Most of the results will be presented as global maps.

4 Results

4.1 Data coverage

We have matched a total number of 5747 global afternoon orbits of the NOAA-18 and NOAA-19 satellites with corresponding CALIPSO-CALIOP data in the time period October 2006 and December 2015. The study does not include results from satellites in morning orbit since these can only be matched with CALIOP data at high latitudes (further discussed in Section 6). Due to increasing orbital drift of the NOAA-18 satellite after 2010 (with resulting deviation from the A-Train orbit and increasing off-nadir viewing angles for matchups), the matchup dataset is unfortunately not exclusively based on AVHRR near nadir observations even if they dominate. However, all NOAA-18 data were included here since we wanted a representative evaluation of the AVHRR CDR for the entire studied period. The observation time difference was limited to 3 minutes and the spatial matchup error was maximised to 2.5 km (as a consequence of using the nearest neighbouring technique and after assuming negligible navigation errors). This resulted in a total number of more than 23 million global matchups. The distribution of the matchups is visualized in Fig. 3 using a Fibonacci grid resolution of 75 km (which is also used in the following figures for the subsequent plotting of most of the results).

Figure 3 shows a quite varying degree of coverage as a function of latitude with a minimum of the number or matchups occurring at low latitudes and a maximum of matchups for the highest latitudes. Although the likelihood for a valid matchup to occur is the same everywhere on a particular matched orbit, the pattern of the matchup numbers is explained by the converging orbital tracks towards the poles. Furthermore, the large variation with some typical geographical features and variations also in the zonal direction shows that we have not been able to extract 100 % of all theoretically available matching cases (some periods with loss of data exist for both CALIOP and AVHRR). This means that the ambition of getting a homogeneous global coverage cannot be perfectly met but it is still the best effort in that direction that we can

make. Even at low latitudes the number of matches generally exceeds 300 for a grid resolution of 75 km. Some exceptions can be seen, particularly over the Pacific Ocean, but we still believe that the number of samples is sufficient for obtaining a fair estimation of the cloud screening performance.

4.2 Results based on one original and one restricted CALIOP cloud mask

Figure 4 shows the achieved global distribution of the Hitrate parameter when comparing to the original CALIOP cloud mask. Results indicate a fairly good performance over mid- to high latitudes (especially over oceans) but degraded results over most low latitudes and over the polar regions with the poorest results occurring over Greenland and Antarctica.
Further analysis of results is complicated by the fact that the original CALIOP cloud mask includes all CALIOP-detected clouds as explained in Section 3.4. In particular, we suspect that the rather poor results in Fig. 4 in the tropical region may be significantly influenced by the presence of sub-visible clouds.

If using all available matchups, we can calculate $POD_{cloudy}(\tau)$ and plot results for all values of τ (Fig. 5). Calculations have been based on optical thickness intervals of 0.05 in the range $0.0<\tau<0.5$, intervals of 0.1 in the range $0.5<\tau<1.0$ and intervals of 1.0 in the range $1.0<\tau<5.0$ (the latter results not shown in Fig. 5.). From Fig. 5 we deduce that the cloud detection sensitivity (i.e., where a probability of 50 % is reached) can be estimated to 0.225 for the investigated AVHRR-based results. Consequently, we will use this value to represent the optimal Hitrate results and we then get the global distribution of results as presented in Figure 6. As expected, results improve considerably for most places compared to Fig. 4, and especially over low latitudes. In most places, Hitrates above 80 % are achieved. The polar regions (at least the snow- and ice-covered parts) stand out as regions of poor quality with the worst conditions seen over central Greenland and Antarctica. Some degradation in the results is still seen over some regions at low-to-middle latitudes and this will be analysed further in the next section.

We claim that the results in Fig. 6 give a much better idea of the performance of the CLARA-A2 cloud mask than what was shown in Fig. 4, especially since it is now linked to a well-defined description of the involved clouds. We will use the same filtering approach for results to be shown in the next sub-section.

4.3 Complementary results based on a restricted CALIOP cloud mask

Figure 7 presents results for the systematic (bias) and random errors (bias-corrected RMS) of the CLARA-A2 cloud amounts. It is clear that the cloud detection problems over the polar regions, as indicated by the Hitrate parameter in Fig. 6, leads to a massive underestimation of cloud amounts, especially over the parts being normally covered with snow or ice. However, notice that this is an overall mean (close to an annual mean) and that results may be seasonally varying. For example, cloud detection in the polar summer season is considerably better than during the polar winter (as shown by Fig. 5 in Karlsson et al., 2017). The most unbiased results are found over mid-to-high latitudes while some overestimated cloud amounts are seen over lower latitudes, particularly over oceanic surfaces. RMS values are naturally also high in the polar regions but also over what can be described as oceanic sub-tropical high regions. This agrees also well with corresponding Hitrate results in Fig. 5. RMS values are also low over dry desert regions but mostly as a consequence of the general lack of cloudy situations here.

In a further search of the areas where we have significant misclassifications of cloudy and clear conditions we can study results of probability of detection of the two categories in Figure 8. For the cloudy category results are mostly already deduced from previous figures except possibly for the low probabilities of cloud detection over northern Africa and the Arabian Peninsula. It means that in this particular area, where cloudiness is generally low, we still find particular problems in detecting the few occurring cases of clouds. The reasons for this have to be investigated further but they are likely to be

linked to remaining uncertainties in the used surface emissivities over these dry and desert-like surfaces. The two maps in Figure 8 reveal another interesting feature: In areas where cloudiness is low (e.g., over sub-tropical ocean and land regions) POD_{cloudy} is low and where cloudiness is high (e.g., mid-latitude storm tracks and ITCZ near the equator) POD_{clear} is low. This contributes to give fairly low values of the Kuipers' score over these regions (Figure 9) leading to a slightly different distribution of results in comparison to the Hitrate (Fig. 6). However, we must remember that Hitrate is dominated by results for the dominating mode (cloudy or clear) while Kuipers punishes particularly the existence of misclassifications of the minority mode. Figures 8 and 9 reveal that even if the dominantly cloudy and clear regions are generally captured very well the few cases of the opposing mode have a high frequency of misclassifications. This is difficult to understand from the perspective of long-term experience of AVHRR cloud screening. More clearly, cloud screening is generally understood to work best over dark and warm ocean surfaces in good illumination. So, why are results not better here (e.g., over oceanic sub-tropical high regions)? We believe that this unexpected behaviour is a consequence of the limitations of both AVHRR GAC data and CALIPSO-CALIOP data when it comes to the sampling of the true conditions within the nominal 5 km FOV. It was already mentioned in Section 2 that only about 25 % of the nominal AVHRR GAC FOV of 5 km is actually observed and that the corresponding figure for CALIOP single shot nominal FOV of 330 meters is as low as 5 %. Notice that the latter means that CALIOP is only able to cover about 0.3 % of the nominal 5 km FOV. This has important consequences for all cases when we have cloud elements present which are smaller in size than the nominal 5 km FOV. We first conclude that only in the case of having cloud elements larger than the nominal 5 km FOV we can be confident in getting the same results from AVHRR and CALIOP observations. For all other cloud situations involving clouds being smaller in size than 5 km the two data sources will give different results since the sensors will probe different parts of the 5 km FOV. The situation is made even worse by the fact that the AVHRR scan lines are perpendicular to the CALIPSO track when matching the two datasets in the near-nadir mode. It means that the CALIOP sensor will consistently probe a different part of the nominal 5 km FOV than AVHRR. Theoretically, a maximum of 3 CALIOP single shot measurements out of totally 15 would actually be able to measure the same spot on Earth as the AVHRR GAC measurement within the FOV of 5 km. A consequence of this must be that in the case of dominating fractional cloudiness with cloud size modes below the 5 km scale the random errors and the false-alarm rates should increase even if the bias could remain small. This is exactly what is observed over the oceanic sub-tropical high regions (Figure 7 and Figure 10, upper panel) also explaining the degraded overall scores in this region (in particular the POD_{cloudy} score in Fig. 8). These regions have a reduced total cloud amount in the annual mean (e.g., see Fig. 6 in Karlsson et. al., 2017), mainly because of generally more stable conditions here with prevailing large-scale subsidence (poleward parts of the Hadley cell) suppressing cloudiness in mid- to high layers and basically only allowing convective and stratiform boundary layer cloudiness to form. This boundary layer cloudiness consists to a large degree of scattered small-scale cumulus and stratocumulus clouds, i.e., typically the kind of clouds for which we would expect enhanced disagreeing results for the AVHRR and CALIOP datasets following the above reasoning. It is interesting to notice that not only oceanic areas show this feature. Also some eastern parts of continents show similar results, e.g. easternmost part of South America and Africa. It could mean that scattered cumulus cloudiness is the dominant mode of cloudiness also here. Finally, notice also that we can see exactly the same effect for fractional clear areas, e.g. over northern and southern hemisphere stormtracks at mid- to high latitudes as shown by the large FAR_{clear} values here in Fig. 10. We conclude that, because of the problems to correctly representing cases of both small-scale cloudiness and small-scale holes in cloud decks in the two datasets, validation results are probably underestimated (i.e., giving too low scores) over these dominantly cloudy or dominantly clear regions of the globe.

4.3 Estimating the global variability of cloud detection limitations

The concept of presenting validation results after having removed all clouds with smaller optical depths than the cloud detection sensitivity parameter is undoubtedly a clear improvement compared to if only showing results based on the original

CALIOP cloud mask. However, we still have the problem that the currently used cloud detection sensitivity value is a global average meaning that we can still have large geographical variations in the performance. To investigate how serious this simplification is, we can plot the results of τ_{min}(POD>50) calculated exclusively for every Fibonacci grid point (Figure 11). To reduce the uncertainty in this calculation due to spuriously occurring low number of samples per grid point as indicated in Fig. 3 for low latitudes, we have here increased the radius of the Fibonacci grid from 75 km to 300 km. We notice a considerable variation in cloud detection sensitivity over the globe in Fig. 11. Especially we notice that the cloud detection sensitivity is generally considerably lower than the global average value of 0.225 over large parts of the oceanic areas as well as over tropical land areas. On the other hand, values are generally larger than 0.225 over dry and desert-like regions and over high-latitude and polar land areas. For the polar land areas the cloud detection sensitivity frequently exceeds 1 and for some grid points even reaches values close to 5. These values, when put in relation to the global average value of 0.225, tell us that more representative (and most likely higher) overall validation scores could have been achieved if re-calculating validation scores per Fibonacci grid point using these globally resolved cloud detection sensitivity values. However, we have not taken this step here because of the relatively low number of samples in some grid points (even at the 300 km scale).

We may also visualise the variable cloud detection sensitivity by plotting the same kind of cloud layer probability curves as in Fig. 5 for individual grid points in Fig. 11. Figure 12 shows such curves for the three locations marked out in Fig. 12. The three points describe three typical but also extreme situations. The blue curve in Fig. 12 shows cloud layer detection probabilities for a distant (from land) point in the North Atlantic Ocean. It marks a position where cloud detection clearly works the best in the global perspective. The cloud detection sensitivity value is as low as 0.075 here meaning that even very thin clouds are well detected here. Cloud detection performance is also reaching a maximum value of approximately 95 % already at $\tau = 0.5$. This is probably as high as can be reached because of remaining and unavoidable AVHRR-CALIOP mis-location and matching problems (both in time and space). As a contrast, a grid point located in the Sahel region (green curve in Fig. 12) shows less good results with a cloud detection sensitivity of 0.375 and it barely reaches maximum cloud detection performance at $\tau = 3.5$ and higher. However, even worse results are recorded for the point over central Greenland (red curve in Fig. 12). The cloud detection sensitivity is here as high as 1.5 and it is clear that not even at a maximum τ value of 4.5 we can come close to an optimal cloud detection performance. Thus, over a snow-covered and often extremely cold location we cannot even detect all optically thick clouds (which is in line with the low POD_{cloudy} results over Greenland and Antarctica in Fig. 8, upper panel).

The visualisation of the results in Fig. 12 reveals again that we are probably slightly undersampling the true conditions at some individual grid points. This is indicated by the unexpected decrease in POD at some points for increasing τ values. Theoretically, one would expect a steady increase in POD as a function of τ.

5 Discussion

We have shown that with the access to the cloud information provided by the high-sensitivity CALIPSO-CALIOP lidar, covering almost a full decade (2006-2015), it is possible to construct detailed global maps of cloud detection performance parameters of the cloud screening method used in the AVHRR-based CLARA-A2 cloud climate data record. A wide range of validation scores, several of them complementary to the essential scores describing systematic and random errors, have been used to get a very detailed picture of the cloud screening efficiency. Furthermore, by use of the CALIOP-derived information on cloud optical thickness, it has been possible to make a clear definition of what clouds that has been observed and for which the validation scores are valid. We believe this to be crucial for allowing further quantitative use of the results. The method as such is not specifically developed or valid for the CLARA-A2 cloud masking method but should be

applicable to any method utilizing CALIOP data as its reference. Because of this the proposed method is suggested to be used also in studies evaluating different methods in an objective way.

The necessity to specify the clouds being investigated is partly linked to the fact that the CALIOP sensor is capable of detecting clouds which could be considered as being fundamentally "sub-visible" for passive imaging sensors. Therefore, a globally estimated minimum cloud optical thickness value (denoted "Cloud detection sensitivity" and for which the majority of clouds would be detected) was estimated to 0.225 for the CLARA-A2 cloud masking method. This value was used to remove contributions to validation scores from thinner clouds than this minimum optical thickness, thus maximising the validation scores. For example, by utilising this definition of detectable clouds, resulting cloud amounts were found to be unbiased over most locations of the world except for a major underestimation over the polar regions. For the latter, a large part of all clouds still remain undetected during the polar night and this fraction can be as high as 50 % over the coldest and highest portions of Greenland and Antarctica. Under these conditions not even optically thick clouds may be detected due to the very similar thermal characteristics of clouds and Earth surfaces. Another observed deviation is a small overestimation of cloudiness over tropical ocean areas. Land-ocean differences were generally small with only results over Greenland and Antarctica standing out as clear exceptions.

The study revealed some small but noticeable degraded results over mainly sub-tropical ocean areas. Here, random errors were surprisingly high indicating decreased agreement between AVHRR and CALIOP observations despite otherwise very favourable cloud detection conditions (e.g., warm ocean temperatures and good illumination conditions).We argue that this is caused by insufficient and mutually different sampling conditions within the studied 5 km FOV of the AVHRR and the CALIOP sensors in cases when small-scale boundary layer cloudiness dominates the cloud situation. Because of this we suspect that the cloud detection performance over these areas is actually better in reality compared to what the presented results show.

An important novel feature of this study compared to many previous validation efforts based on CALIPSO-CALIOP data is the estimation of the probability of detecting an individual cloud as a function of its vertically integrated optical thickness and its geographical position on Earth. This was accomplished by isolating finite optical thickness intervals in the CALIOP cloud information and calculating validation scores for this subset of data in a coarse global grid. Results show a substantial variation compared to the previously mentioned global mean optical thickness value of 0.225 for the thinnest retained cloud in the CALIOP cloud mask to give optimal global performance of validation scores. The highest sensitivity to clouds in AVHRR data is generally found over mid-to-high latitude ocean surfaces. Here, clouds with cloud optical thicknesses as low as 0.075 can be detected efficiently. This can be compared to a value of approximately 0.2 over tropical oceans and generally larger than 0.2 over most land surfaces. The latter value reaches 0.5 at some dry and desert-like regions (e.g., Sahara and the Arabian Peninsula) and increases towards or beyond 1 over polar regions with a highest value of 4.5 found over Greenland and Antarctica. The latter indicates that not even optically thick clouds can be confidently identified over Greenland and Antarctica during the polar winter.

The presented validation method can be viewed upon as a step towards a more strict and universal validation method to be used consistently for cloud climate data records generated from passive imagery (as discussed in Wu et al., 2017). The more than decadal long CALIPSO-CALIOP cloud dataset should be used for benchmarking and for evaluation of current CDRs and future revisions of them. The method presented here could be seen as one candidate method. The possibility to derive globally distributed results makes it also easier to define and test global quality requirements on the CDRs. For example,

requirements could be formulated in terms of minimum global coverage within a certain quality threshold instead of today's often very generally formulated global requirement in one finite value or a value range (WMO 2, 2011).

A specific problem with the current method is the inability to assess the global quality of products from polar satellites in morning orbits (e.g., from the NOAA-17 and Metop satellites). Matchups with CALIPSO-CALIOP are here only possible at high latitudes leaving low-to-middle latitudes without reference CALIOP observations for AVHRR products. Comparisons have been made for morning satellites at high latitudes (CM SAF 1, 2017) showing good agreement with corresponding results from afternoon satellites. Thus, for cloud amount information (in contrast to some other cloud parameters, like cloud effective radius) there is no reason to suspect large differences between morning and afternoon results even if morning orbit data is partly using measurements in another spectral band (at 1.6 μm) in the short-wave infrared spectral region. However, this has to be further confirmed in the future, e.g., by use of reference data from the Cloud-Aerosol Transport System lidar (CATS, https://cats.gsfc.nasa.gov/) on the International Space Station or by use of data from the Earth Cloud Aerosol and Radiation Explorer (EarthCARE) mission (http://m.esa.int/Our_Activities/Observing_the_Earth/The_Living_Planet_Programme/Earth_Explorers/EarthCARE/ESA_s_cloud_aerosol_and_radiation_mission) for afternoon satellites with two coexisting short-wave infrared channels onboard.

One particular aim of this study was to provide a strict definition of the clouds being validated together with the main validation results. This has been accomplished by the use of a combination of the CALIOP-derived cloud mask and the CALIOP-estimated optical thickness of clouds. In this way we believe that results could become more quantitatively useful. One obvious application would be to incorporate this information about strengths and limitations of cloud detection capabilities into cloud dataset simulators of the Cloud Feedback Model Intercomparison Project (CFMIP) Observation Simulation Package (COSP, Bodas-Salcedo et al., 2011). Existing COSP simulators for cloud datasets generated from passive satellite imagery (e.g., ISCCP and MODIS) do not explicitly take into account potential inherent cloud detection problems. Instead, they concentrate on simulating some satellite-specific or retrieval-specific features (e.g., systematic underestimation of cloud top height of thin high clouds) leaving it to the user of the simulator to add existing knowledge on cloud detection efficiency in the final evaluation process. We are of the opinion that also aspects of cloud detection capabilities need to be explicitly taken into account in these simulators. A specific CLARA-A2 COSP simulator is therefore under development where the description of such quality aspects will be included based on the findings of this validation study.

Finally, we repeat our opinion of CALIPSO-CALIOP data as being a great asset for current and future evaluation of cloud CDRs based on passive satellite imagery. At the same time, we must express our concern about the current uncertainty regarding the long-term planning of possible replacements of both the A-Train satellites and the upcoming EarthCARE mission. Without follow-on missions it will be very difficult to assess the critical long-term stability of the CDRs which in turn means difficulties in assessing the reliability of possible climate trends deduced from these CDRs.

Author contributions

Karl-Göran Karlsson conducted the validation study and wrote the manuscript. Nina Håkansson prepared the calculation and visualisation of the global results.

Competing interest

The authors declare that they have no conflict of interests.

Acknowledgements

CALIPSO-CALIOP datasets were obtained from the NASA Langley Research Center Atmospheric Science Data Center. The authors want to thank Dr. David Winker in the CALIPSO Science Team for valuable advice regarding the use of CALIOP cloud information.

This work was funded by EUMETSAT in cooperation with the national meteorological institutes of Germany, Sweden, Finland, the Netherlands, Belgium, Switzerland and United Kingdom.

The CLARA-A2 data record is (as all CM SAF CDRs) freely available via the website https://www.cmsaf.eu.

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, Bull. Amer. Meteor. Soc., August 2011, 1023-1043, doi: 10.1175/2011BAMS2856.1 , 2011.

CM SAF 1: Validation Report - CM SAF Cloud, Albedo, Radiation data record, AVHRR-based, Edition 2 (CLARA-A2) – Cloud Products, SAF/CM/DWD/VAL/GAC/CLD version 2.3, Available at http://dx.doi.org/10.5676/EUM_SAF_CM/CLARA_AVHRR/V002, 2017.

CM SAF 2: Algorithm Theoretical Basis Document - CM SAF Cloud, Albedo, Radiation data record, AVHRR-based, Edition 2 (CLARA-A2) – Cloud Fraction, SAF/CM/DWD/ATBD/CMA_AVHRR version 2.0, Available at http://dx.doi.org/10.5676/EUM_SAF_CM/CLARA_AVHRR/V002, 2017.

Dowell, M., P. Lecomte, R. Husband, J. Schulz, T. Mohr, Y. Tahara, R. Eckman, E. Lindstrom, C. Wooldridge, S. Hilding, J. Bates, B. Ryan, J. Lafeuille, and S. Bojinski, 2013: Strategy Towards an Architecture for Climate Monitoring from Space. Pp. 39. This report is available from: www.ceos.org; www.wmo.int/sat; http://www.cgms-info.org/

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. Meteor, 44, 39-54, 2005.

González A.: Measurement of Areas on a Sphere Using Fibonacci and Latitude--Longitude Lattices. Mathematical Geosciences. 42 (1), 49-64. doi:10.1007/s11004-009-9257-x, 2009.

Heidinger, A., Foster, M., Botambekov, D., Hiley, M., Walther, A. and Li, Y.: Using the NASA EOS A-Train to Probe the Performance of the NOAA PATMOS-x Cloud Fraction CDR. Remote Sens., 8, 511, 2016.

Heidinger, A. K., Foster, M. J., Walther, A. & Zhao, Z.: The Pathfinder Atmospheres Extended (PATMOS-x) AVHRR climate data set. Bull. Am. Meteorol. Soc. 95, 909–922, 2014.

Heidinger, A.K., Straka, W.C., Molling, C.C., Sullivan, J.T. and Wu, X.Q.: Deriving an inter-sensor consistent calibration for the AVHRR solar reflectance data record. Int. J. Rem. Sens., 31(24), 6493-6517, doi: 10.1080/01431161.2010.496472, 2010.

Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Meirink, J.F., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., Benas, N., van Zadelhoff, G.-J., Schlundt, C., Stein, D., Finkensieper, S., Håkansson, N. and Hollmann, R.: CLARA-A2: The second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data. Atmos. Chem. Phys., 17, 5809–5828, doi: 10.5676/EUM_SAF_CM/CLARA-AVHRR/V002, 2017.

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: a cloud, albedo, and radiation dataset from 28 yr of global AVHRR data, Atmos. Chem. Phys., 13, 5351-5367, doi:10.5194/acp-13-5351-2013, 2013.

Jin, Y., Okamoto, J and Hagihara, Y.: Improvement of CALIOP cloud masking algorithms for better estimation of dust extinction profiles, J. Meteorol. Soc. of Japan, 92, 433-455, doi: 10.2151/jmsj.2014-502. 433-455, 2014.

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

Martins, E., Noel, V. and Chepfer, H.: Properties of cirrus and subvisible cirrus from nighttime Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP), related to atmospheric dynamics and water vapor, J. Geoph. Res., 116, D02208, doi: 10.1029/2010JD014519, 2011.

Merchant, C. J., Paul, F., Popp, T., Ablain, M., Bontemps, S., Defourny, P., Hollmann, R., Lavergne, T., Laeng, A., de Leeuw, G., Mittaz, J., Poulsen, C., Povey, A. C., Reuter, M., Sathyendranath, S., Sandven, S., Sofieva, V. F. and Wagner, W.: Uncertainty information in climate data records from Earth observation, Earth Syst. Sci. Data, 9, 511-527, doi: 10.5194/essd-9-511-2017, 2017.

Ohring, G., Wielicki, B., Spencer, R., Emery, B. and Datla, R.: Satellite instrument calibration for measuring global climate change. Bulletin of the American Meteorology Society, 86, 1303–1313, doi: 10.1175/BAMS-86-9-1303, 2005.

Rossow, W.B. and Schiffer, R.A.: Advances in understanding clouds from ISCCP, Bull. Am. Meteorol. Soc., 80, 2261-2287, doi:10.1175/1520-0477(1999)080%3C2261:AIUCFI%3E2.0.CO;2, 1999.

Sassen, K. and Cho, B. S.: Subvisual-Thin Cirrus Lidar Dataset for Satellite Verification and Climatological Research, J. Appl. Meteor., 31, 1275-1285. https://doi.org/10.1175/1520-0450(1992)031%3C1275:STCLDF%3E2.0.CO;2, 1992.

Saunders, R. W.: An automated scheme for the removal of cloud contamination from AVHRR radiances over western Europe, Int. J. Rem. Sens., 7, 867, 1986.

Saunders, R. W. and Grey, D. E.: Interesting cloud features seen by NOAA-6 3.7 μm images, Met. Mag., 114, 211, 1985.

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.

Stengel, M., Stapelberg, S., Sus, O., Schlundt, C., Poulsen, C., Thomas, G., Christensen, M., Carbajal Henken, C., Preusker, R., Fischer, J., Devasthale, A., Willén, U., Karlsson, K.-G., McGarragh, G. R., Proud, S., Povey, A. C., Grainger, D. G., Meirink, J. F., Feofilov, A., Bennartz, R., Bojanowski, J., and Hollmann, R.: Cloud property datasets retrieved from AVHRR, MODIS, AATSR and MERIS in the framework of the Cloud_cci project, Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2017-48, in review, 2017.

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. Bull. Amer. Meteor. Soc., 83, 1771–1790, doi: http://dx.doi.org/10.1175/BAMS-83-12-1771, 2002.

Stocker, T.F., D. Qin, G.-K. Plattner, L.V. Alexander, S.K. Allen, N.L. Bindoff, F.-M. Bréon, J.A. Church, U. Cubasch, S. Emori, P. Forster, P. Friedlingstein, N. Gillett, J.M. Gregory, D.L. Hartmann, E. Jansen, B. Kirtman, R. Knutti, K.

Krishna Kumar, P. Lemke, J. Marotzke, V. Masson-Delmotte, G.A. Meehl, I.I. Mokhov, S. Piao, V. Ramaswamy, D. Randall, M. Rhein, M. Rojas, C. Sabine, D. Shindell, L.D. Talley, D.G. Vaughan and S.-P. Xie: Technical Summary. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Sun, B., Free, M., Yoo, H. L., Foster, M. J., Heidinger, A., and Karlsson, K.-G.: Variability and trends in US cloud cover: ISCCP, PATMOS-x and CLARA-A1 compared to homogeneity-adjusted weather observations, Journal of Climate, 28, 4373–4389, doi: 10.1175/JCLI-D-14-00805.1, 2015.

Swinbank, R. and R. J. Purser.: Fibonacci grids: A novel approach to global modelling. Quarterly Journal of the Royal Meteorological Society, 132 (619), 1769–1793. doi:10.1256/qj.05.227, 2006.

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

Winker, D. M., Hunt, W. H. and McGill, M. J.: Initial performance assessment of CALIOP, Geoph. Res. Lett., 34, L19803, doi:10.1029/2007GL030135, 2007.

Winker, D. M., Vaughan, M. A., Omar, A., Hu, Y. and Powell, K. A.: Overview of the CALIPSO mission and CALIOP data processing algorithms, J. Atmos. Oceanic. Technol., 26, 2310-2323, doi: 10.1175/2009JTECHA1281.1, 2009.

WMO 1: Recommended methods for evaluating cloud and related parameters, WWRP 2012-1, Report of the WWRP/WGNE Joint Working Group on Forecast Verification Research (JWGFVR), available at https://www.wmo.int/pages/prog/arep/wwrp/new/documents/WWRP_2012_1_web.pdf, 2012.

WMO 2: Systematic observation requirements for satellite-based data products for climate - Supplement details to the satellite-based component of the "Implementation Plan for the Global Observing System for Climate Support of the UNFCCC (2011 update), GCOS-154, available at https://library.wmo.int/opac/doc_num.php?explnum_id=3710, 2011.

Wu, D. L., Baum, B. A., Choi, Y.-S., Foster, M., Karlsson, K.-G., Heidinger, A., Poulsen, C., Pavolonis, M., Riedi, J., Roebeling, R., Sherwood, S., Thoss, A. and Watts, P.: Towards Global Harmonization of Derived Cloud Products, Bull. of Amer. Meteor. Soc., February 2017, 49-52, doi: 10.1175/BAMS-D-16-0234.1, 2017.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

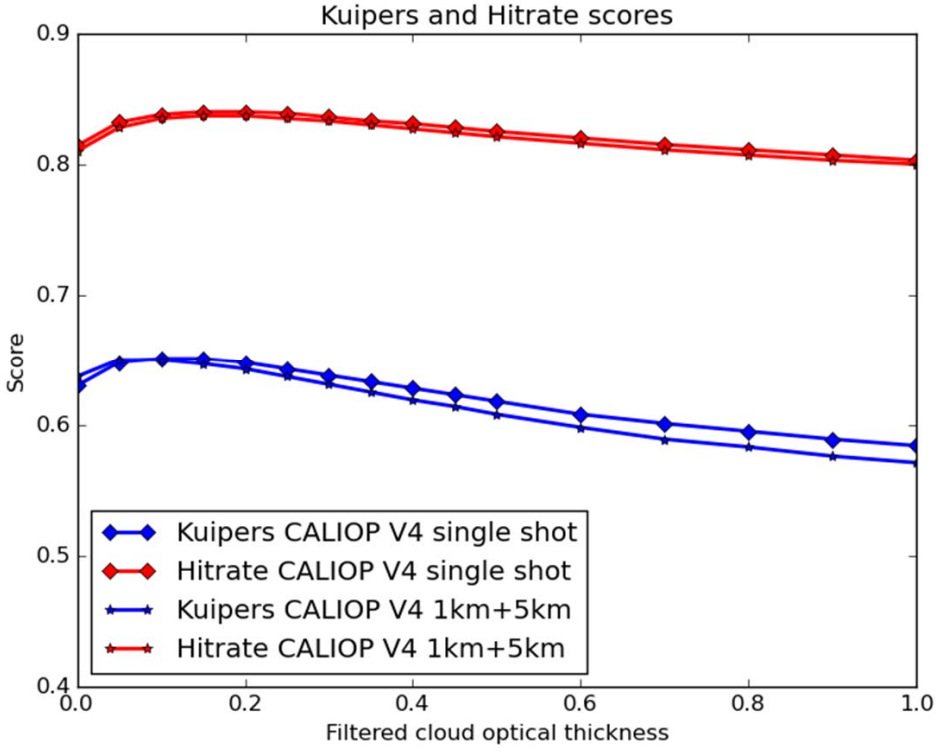


Figure 1: CALIOP-based validation scores (Hitrate and Kuipers) as a function of filtered cloud optical thickness (see text for explanation) for 80 matched NOAA-18 orbits between October and December 2006. Validation is based on CALIOP version 4.10 CLAY products and show results from two alternative validation methods (single shot or combined 1 km + 5 km, see text for explanation).

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

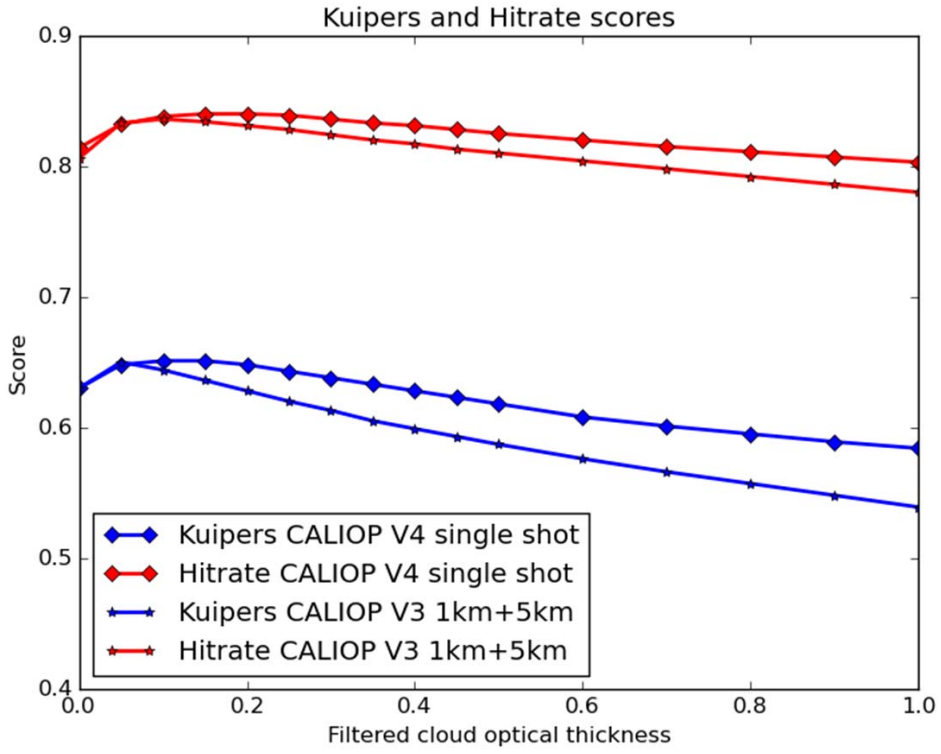


Figure 2: CALIOP-based validation scores (Hitrate and Kuipers) as a function of filtered cloud optical thickness (see text for explanation) for 80 matched NOAA-18 orbits between October and December 2006. The curves compare results based on CALIOP version 4.10 CLAY products computed with the new method based on single shot information (denoted "CALIOP V4 single shot") with results based on CALIOP version 3.01 CLAY products computed with the old method based on combined 1 km + 5 km data (denoted "CALIOP V3 1km+5km").

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

665

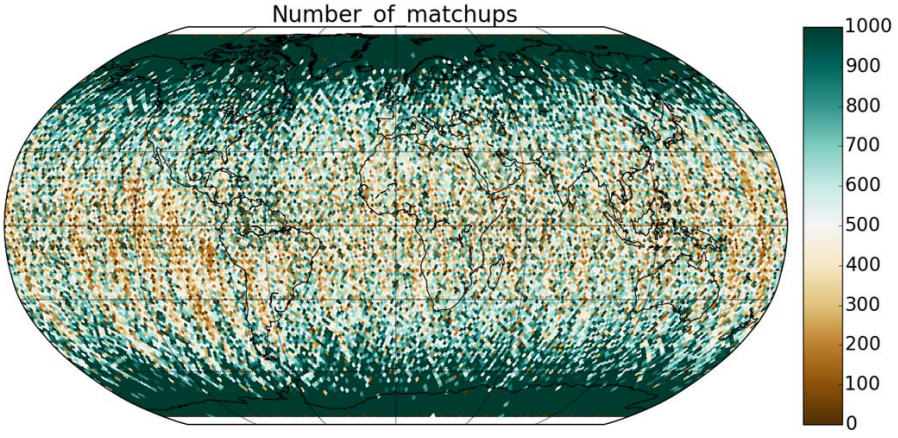


Figure 3: Total number of CALIPSO-CALIOP matchups with NOAA-18 and NOAA-19 AVHRR observations in the time period October 2006 to December 2015. Results presented in a Fibonacci grid with 75 km resolution.

670

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

675

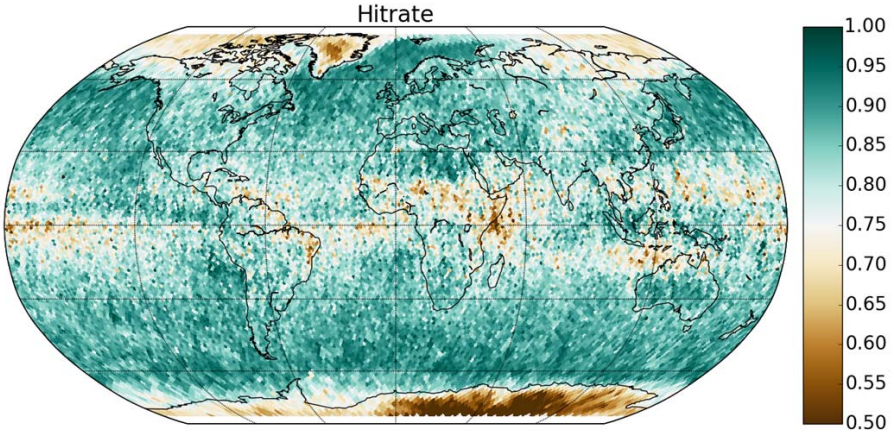


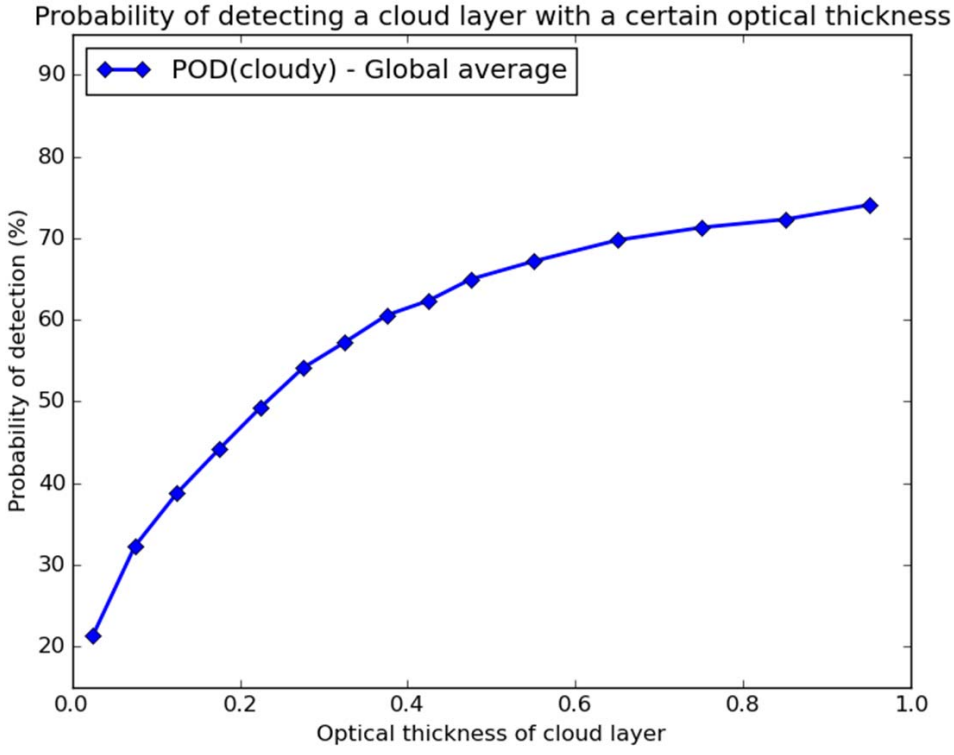
Figure 4: Global presentation of the CLARA-A2 cloud mask Hitrate parameter with a horizontal Fibonacci grid resolution of 75 km. Validation results are based on comparisons with the original CALIPSO-CALIOP cloud mask. Same underlying matchup dataset as in Fig. 3.

680

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

685



690

Figure 5: Global estimation of the probability of detecting a cloud with a certain cloud optical thickness. Calculations are based on all available AVHRR-CALIOP matchups over the time period October 2006 to December 2015.

695

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

700

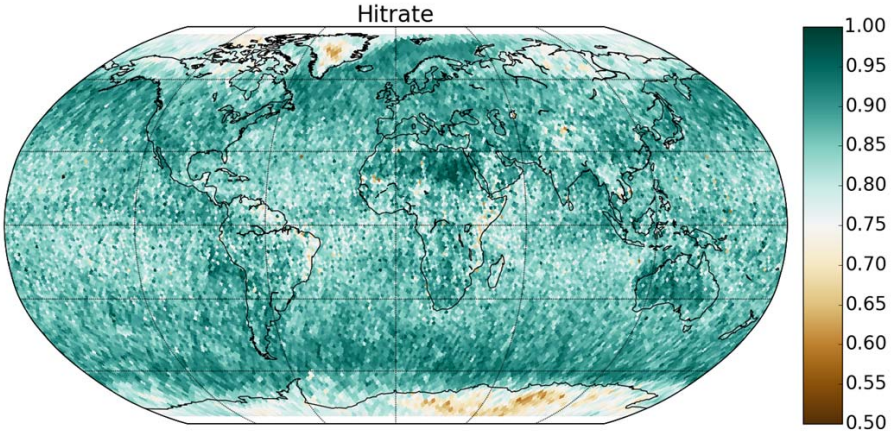


Figure 6: Peak Hitrate results for the CLARA-A2 cloud mask achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. 3.

705

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

710

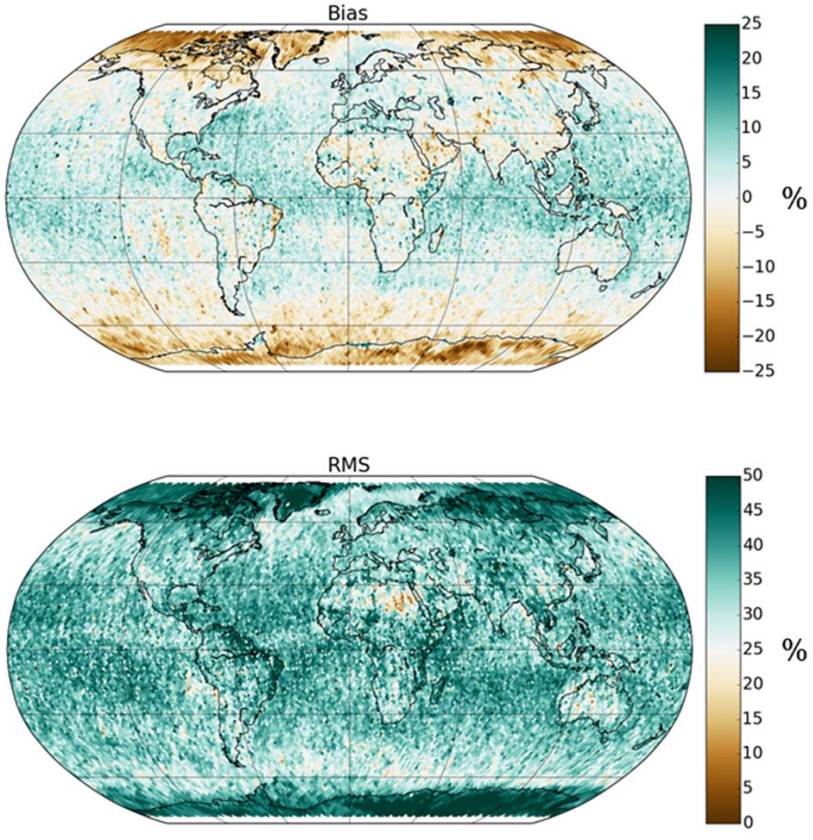


Figure 7: Mean Error (Bias) and bias-corrected Root Mean Squared Error (RMS) for the CLARA-A2 cloud amount achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. 3.

715

720

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

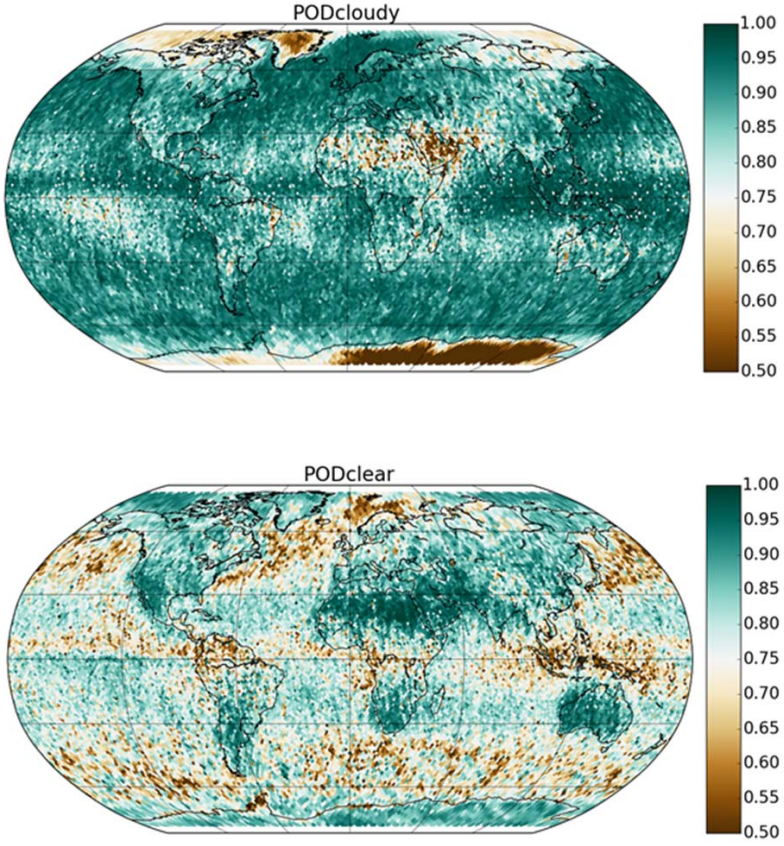


Figure 8: Probability of detection of cloudy (top) and clear (bottom) conditions for the CLARA-A2 cloud mask achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. 3.

725

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

730

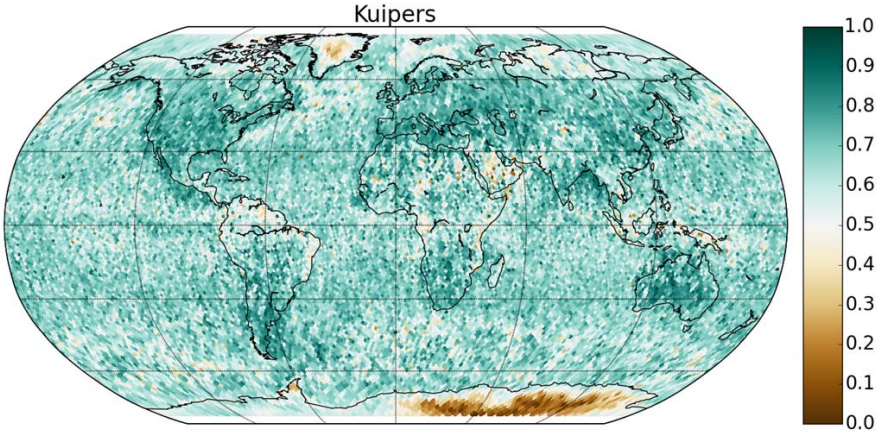


Figure 9: Kuipers score for the CLARA-A2 cloud mask achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. 3.

735

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

740

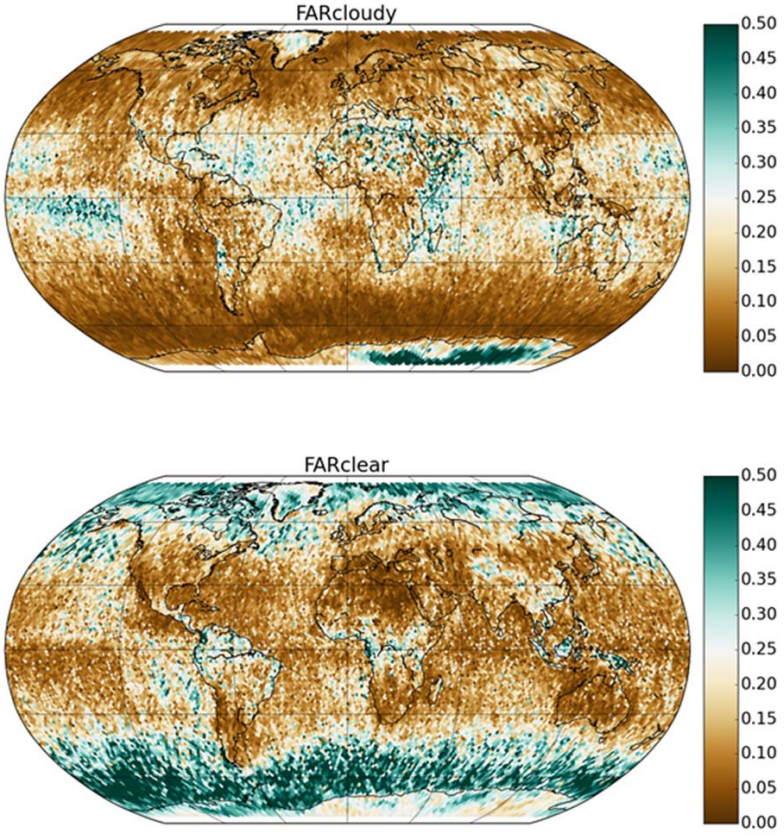


Figure 10: False alarm rates for cloudy (top) and clear (bottom) predictions for the CLARA-A2 cloud mask achieved after filtering the CALIOP cloud mask with the cloud optical thickness value of 0.225. Same underlying matchup dataset as in Fig. 3.

745

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

750

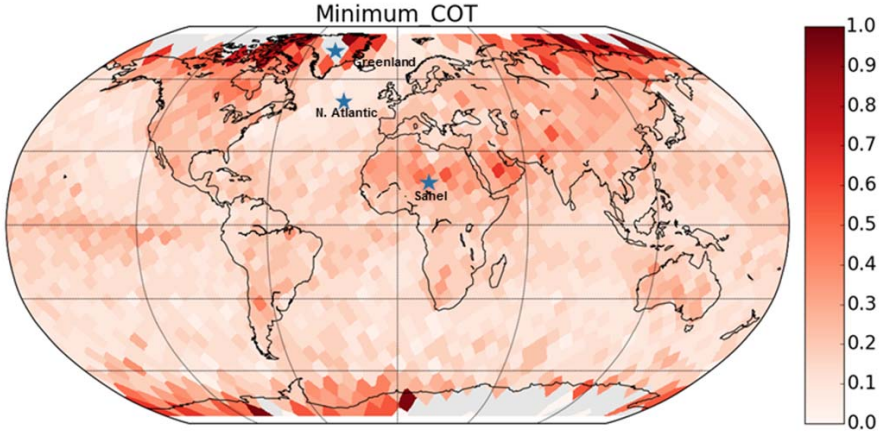


Figure 11: Global map of estimated cloud detection sensitivity of the cloud mask of CLARA-A2 (see text for explanation). Results are calculated from the same dataset as visualized in Figure 3 but in a coarser Fibonacci grid resolution of 300 km. Conditions in the three marked locations are analysed further in Fig. 12. Grey areas denote areas with values exceeding 1.0.

755

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2017-307
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 15 September 2017

Atmospheric Measurement Techniques Discussions

760

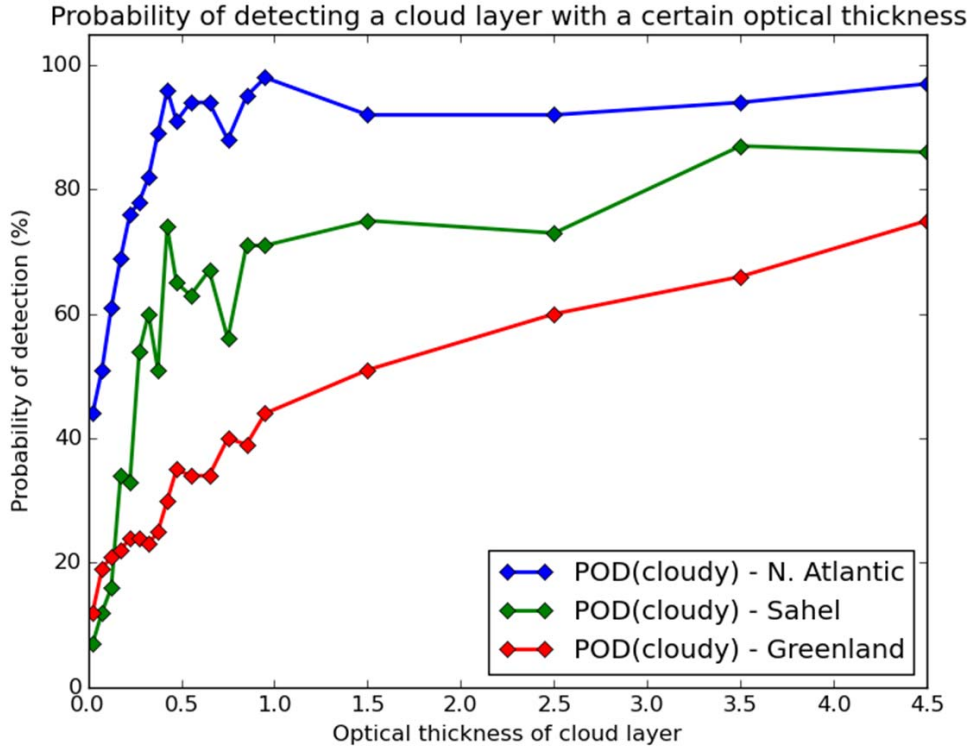


Figure 12: Same as Fig. 5 but for individual grid points marked out in Fig. 11.