New Limb Cloud Detection Algorithm Theoretical
Basis Document

Patricia Liebing

January 22, 2016

Contents

1 Introduction 3

2 Theoretical Basis 3
  2.1 Motivation ........................................... 3
  2.2 SCIATRAN Studies .................................. 4

3 Data Analysis 6
  3.1 Statistical Distributions .............................. 6
  3.2 Threshold Determination .............................. 8
  3.3 Fits to the histogram distributions .................. 11

4 Determination of Probabilities 12

5 Application to the Data 15
  5.1 Data Containers ...................................... 16
  5.2 Data Preprocessing ................................... 17
  5.3 Determine LUT Cell ................................... 19
  5.4 Determine PSC Gradients ............................. 20
  5.5 Retrieve Probabilities from LUTs and Apply Apriori
      Weights. .............................................. 20
  5.6 Perform Low Cloud Check ............................ 22
  5.7 Determine PSC Flag and PSC Top TH ................. 23
  5.8 Determine Cloud Flag and Cloud Top TH ............ 24
  5.9 Determine Tropospheric Cloud Flag and Tropospheric
      Cloud Top TH ......................................... 25

6 Results 26
  6.1 Cloud Identification ................................. 26
  6.2 Cloud Statistics ...................................... 27
  6.3 Preliminary Investigation of PSCs ................... 28
7 Algorithm Description Summary

7.1 Preprocessing: Reflectances and Color Ratios, Apriori Weights 31
7.2 Gradients, LUT Bins, Probabilities and Statistical Cloud Weight 31
7.3 Low Cloud Check 32
7.4 PSC Flags and Parameters 32
7.5 General Cloud Flags and Parameters 33
7.6 Tropospheric Cloud Flags and Parameters 33
7.7 Summary of Initialization Parameters 34
7.8 Recommended Output 36
1 Introduction

The new limb cloud detection algorithm is outlined here, along with the analysis that led to the determination of the probabilities used in the algorithm. The goal is to assign to each limb measurement a probability that it is affected by a cloud, and then to determine the cloud top height or, rather, the highest tangent height at which a cloud is likely present in a limb profile. The motivation behind this new scheme is discussed in section 2, together with model studies. The strategy to arrive at the probabilities from the measurements is presented in section 3 and the algorithm for the application of the probabilities and determination of the cloud flags will be described in section 5.

2 Theoretical Basis

2.1 Motivation

The current L1-2 algorithm for limb cloud flagging is based on SCODA developed by K.U. Eichmann [L1-2 ATBD (2015), Eichmann et al.(2015)]. It uses the ratio of measured intensities at 750 and 1090 nm, $C_{750}$, and determines an index $S_{750}$ by taking the ratio of $C_{750}$ from the next higher tangent height (TH) to that from this TH, i.e.,

$$S_{750} = \frac{C_{750}(i_{TH+1})}{C_{750}(i_{TH})} = \frac{I_{750}/I_{1090}(i_{TH+1})}{I_{750}/I_{1090}(i_{TH})}.$$  (1)

If this ratio is larger than a given fixed threshold, this TH is flagged, and the highest TH at which this occurs is defined to be the cloud top height (CTH).

An additional cloud flag is determined using the same principle, but with the ratio $C_{1550} = I_{1550}/I_{1670}$ of intensities at roughly 1550 and 1670 nm, and different thresholds. This ratio is supposed to be have an enhanced sensitivity to ice clouds, therefore, a cloud flagged with this flag is likely an ice cloud. For both flags, the algorithm relies heavily on clouds having a steep gradient in density, such that the ratio of 2 subsequent THs results in a large value. On the other hand, the color ratio $C_{750}$ itself is very sensitive to both aerosols and clouds, with the ratio depending on both the density and size parameters. That means, if large gradients in aerosol density occur as it was, for instance, the case during the period after the Sarychev eruption in June 2009, those aerosol layers will be misidentified as clouds. The sensitivity of the $C_{1550}$ ratio to aerosols is much smaller due to the close proximity of the two wavelengths. However, the absorption cross sections for liquid water and ice are different between the two wavelengths, resulting in a different ratio for liquid water clouds, ice clouds and aerosols. A drawback of this ratio is that for the low signals expected in some limb geometries (higher altitudes, clear atmospheres or backward scattering angles), the intensity at 1670 nm is extremely susceptible to noise and offset calibration errors.
In SCODA, the thresholds for both cloud flags are fixed, though, in order to achieve roughly similar detection efficiencies for all scattering geometries and conditions, they should vary with latitude or scattering angle and time. It can therefore be expected that the purity of the “cloud free” sample selected by SCODA depends systematically on the scattering geometry and time. The threshold for the “ice cloud” ratio of 1.28 means that a lot of optically thin or geometrically small clouds will be missed. The use of the gradients is in principle convenient, as it cancels multiplicative errors and reduces sensitivity to albedo effects, but it is also problematic w.r.t. the zig-zag scan pattern of SCIAMACHY, where horizontally inhomogeneous features may appear in one TH step but not in the one before or after.

It is therefore worthwhile to improve the algorithm such that aerosols and clouds can be distinguished clearly while increasing the sensitivity to thin clouds and reducing errors from calibration or horizontal sampling. The new algorithm is building on SCODA in the sense that it makes use of similar color ratios, though in a very different way.

2.2 SCIATRAN Studies

Two sets of SCIATRAN V3.1 scenarios [Rozanov et al.(2013)] are compared. The first set contains scenarios without clouds, but varying albedo, tropospheric and stratospheric aerosol profiles and types. The other set contains clouds of varying optical depth and type at different altitudes on top of several sets of aerosol background and albedo. Not all possible cloud scenarios have been simulated, meaning that conclusions drawn will have limited validity for some cases. The behavior of the color ratio $C_{750}$ vs. $C_{1550}$ has been investigated with the goal of developing a strategy that uses both ratios to optimally identify clouds and distinguish them from high aerosol. Figure 1 depicts the general behavior of pure aerosol and cloudy cases.

The $C_{1550}$ of the pure aerosol cases is always above 1 for all values of $C_{750}$, while for many of the cloudy cases it is the lower the lower the value of $C_{750}$ is. For optical depths (OD) larger than about 0.1, $C_{750} \approx 1$ and $C_{1550}$ is distinctly lower for ice than for water clouds. Nevertheless, on a case by case basis water and ice clouds cannot clearly be distinguished unless other parameters are known. As is shown on the right side of the figure, optically thin clouds become more similar to the pure aerosol case, very thin clouds (OD $\leq 0.001$) can hardly be distinguished from pure aerosol cases. For fixed $C_{750}$, $C_{1550}$ depends on the microphysical properties of the cloud or aerosol particles and scattering angle (not shown). However, for aerosols the dependence is fairly weak. An additional complication arises due to the reflection from lower clouds, as is demonstrated by the magenta point on the right side. These clouds are entirely below the field of view (FoV), though, for optically thick clouds, their $C_{1550}$ can reach values typical for thin clouds inside the FoV. This problem is to a large part avoided with the gradient approach in SCODA, but from just the values at a single TH as shown here it is impossible to distinguish the two situations.
Figure 1: The color ratios $C_{750}$ vs. $C_{1550}$ (the axis labels are wrong, $x$-axis is $C_{1550}$, $y$-axis is $C_{750}$) for SCIATRAN simulations without clouds (black and gray points) and with clouds (other colors). On the left, water (green) and ice (blue) clouds with a CTH of 15 km with a wide range of optical depths are shown. On the right, optically thin clouds ($\tau < 0.01$, light blue) are shown as well as the effect of optically thick clouds with a CTH $\leq 13$ km (magenta) are shown in addition.

In the left and middle panels of figure 2 the distribution of aerosols and clouds, respectively, is plotted with the common logarithm of the corresponding optical depth (OD) in color code. In both cases, the $C_{750}$ decreases monotonically as the OD does up to values of $C_{750} \approx 1$, when it is saturated. The value of $C_{1550}$ also decreases monotonically with the (A)OD, but with a very different slope for aerosols and clouds. In the right panel of figure 2, the color code represents the AOD of the background aerosol used for the cloudy scenes, showing that the limits for thin clouds in both color ratios are determined by the aerosol load. This implies that

Figure 2: The color ratios $C_{750}$ vs. $C_{1550}$ (the axis labels are wrong, $x$-axis is $C_{1550}$, $y$-axis is $C_{750}$) for SCIATRAN simulations without clouds (left) and with clouds (CTH=15 km, middle and right). On the left and right, the color code is the common logarithm of the stratospheric AOD, in the middle it is the common logarithm of the cloud optical depth (OD).
it is in general not possible to find a fixed threshold in \((C_{750}, C_{1550})\) which would allow to distinguish thin clouds from aerosols with a sufficient amount of efficiency. In particular, the upper limiting value of \(C_{750}\) depends strongly on scattering angle, aerosol type and density. Note that this also applies for the case of gradients, as the value of \(C_{750}\) depends on the aerosol density of both this TH and the one above, and the aerosol distribution is of course time dependent. The cloud detection efficiency, as well as the purity of the cloud free data sample therefore depends strongly on season, latitude, TH, and variability of aerosols. The lower limit in \(C_{750}\) is typically saturated around 1 for high aerosol density (stratospheric AOD > 0.02) or cloud optical depth > 0.1. with the value of \(C_{1550}\) typically below 1 for clouds and above 1 for pure aerosol.

### 3 Data Analysis

From the discussion in the previous section the following conclusions can be derived:

- The color ratio \(C_{1550}\) is sensitive to the type of scatterers.

- Its sensitivity is enhanced, if the density of scatterers, or a proxy for it, is known.

- The color ratio \(C_{750}\) can serve as a proxy for the density of scatterers.

- Cloud aerosol separation can then be achieved by investigating statistical distribution of the real data in \(C_{1550}\) as a function of \(C_{750}\), time, latitude, TH and surface conditions (land or sea surface).

The strategy is therefore to first analyze the data distributions of \(C_{750}\) vs. \(C_{1550}\) in optimized slices in time, latitude and TH, and then to determine threshold values that separate clouds from aerosol for each of these slices. In order to optimize the sensitivity to thin clouds a hard threshold should be avoided, rather it is preferable to investigate probability distributions.

#### 3.1 Statistical Distributions

The two-dimensional histograms of \(C_{750}\) vs. \(C_{1550}\) are obtained in time slices of one month, 30 degree latitude bins and for each TH step until about 30 km. The current analysis is based on seasonal variation only, that is, for each month of the year the statistics of all years are integrated. The integration optimizes the statistics per bin while at the same time increasing the information content regarding the aerosol properties. On the other hand, it inhibits a time dependent self calibration, i.e., a potential time dependent radiometric calibration error cannot be regarded. Eventually, it was found by comparison that a seasonal variation analysis gives the more physical results than a truly time dependent analysis.
The L1 V8 data set was used, along with radiometric calibration using Version 9.02 key data and M-Factors, for both irradiance and radiances. No polarization correction was applied. The average of the four limb dark measurements was subtracted from the limb radiance. Data quality cuts were applied w.r.t. bad dates or orbits, e.g., after decontaminations. If any of the limb or limb dark measurements were inside the SAA region, they were rejected. The pixel signals were averaged over a wavelength window around the central values (see below). For channel 6+, a special bad pixel mask was applied, which was derived from the dark limb measurements (State ID 32) of about one orbit per day, requiring that the pixel was stable at least 99% of the time.

In addition, the reflectances have to be above a minimum threshold:

\[
\begin{align*}
R_0 &\equiv R(750\text{nm}) > 10^{-3} \\
R_1 &\equiv R(1090\text{nm}) > 10^{-3} \\
R_2 &\equiv R(1670\text{nm}) > 3 \cdot 10^{-3}.
\end{align*}
\]  

The reason for this is that reflectance values below these thresholds lead to extreme noise in the color ratio, mainly in \(C_{1550}\), but on occasion also in \(C_{750}\). On the other hand, neither from model nor data it is expected that data points with such low reflectances contain clouds. Note that the expected noise levels from the dark currents and gain values are typically an order of magnitude smaller.

The TH steps are “equivalent” TH steps in the sense, that a reference TH of about 35 km is defined to be step 12. Depending on the limb scan profile, the actual TH steps are corrected w.r.t. to this reference step. This ensures that even when the THs in each step change, as after the orbit change in 2010, the data will still be collected into the bins corresponding to the equivalent TH from the stable period between 2004 and 2010.

Figure 3 shows typical distributions for a volcanically quiescent period (August 2004). Except for the lowest tangent height, it is fairly easy to identify the cloudy data points as lying below the upper right peak in both the \(C_{750}\) and \(C_{1550}\) directions. The upper right is more or less pronounced, depending on the relative abundance of cloud free measurements. At 15km and 22km, a pure aerosol peak can be clearly identified.

After the eruption of Sarychev in June 2009, the aerosol load increased in the Northern hemisphere. Figure 4 demonstrates how the aerosol distribution changes at 15 km under these conditions, mainly in the \(C_{750}\) ratio. It can be directly compared to the corresponding plot in the lower left of figure 3. In this case it is obviously not trivial to identify an aerosol peak nor separate cloudy from cloud free measurements clearly. A strategy to derive a threshold for each histogram, depending on \(C_{750}\), that works sufficiently well for all conditions is described in the following.
3.2 Threshold Determination

The cloud–aerosol threshold is determined for each month of the year, latitude and equivalent TH index larger than 2, and each surface (land, ocean and combined data). First, it is necessary to determine the position of the upper right “aerosol peak”, \( (C_{\text{max}, 1550}, C_{\text{max}, 750}) \) and its standard deviation along the x-axis \( \sigma_{1550} \). This is achieved in the following procedure:

1. All data which are lying above a minimum value of \( C_{1550} = 1.1 \) are projected onto the y-axis.

2. This projection is used to determine the minimum \( (y_{\text{min}}) \) and maximum \( (y_{\text{max}}) \) values of of \( C_{750} \) which are filled with at least 5 entries.

3. The next step is to consider only data which are lying between \( (y_{\text{min}} + y_{\text{max}})/2 \)
Figure 4: 2D Histogram of $C_{750}$ vs. $C_{1550}$ for August 2009 between 30°N and 60°N, at a TH of 15.

(if $\delta y = y_{\text{max}} - y_{\text{min}} > 1$) or $y_{\text{max}} - 0.5$ (if $\delta y < 1$) and the lower value of $y_{\text{max}}$ and 3.75. The maximum bin content $n_{y_{\text{max}}}$ of a histogram projected along $y$ within this range is determined.

4. Starting from one $y$-bin below $y_{\text{max}}$, the data of this bin plus the one below and above are projected onto the x-axis. If the total content of this histogram exceeds $n_{y_{\text{max}}}$, the histogram is smoothed. If not, the projection onto the x-axis is repeated one $y$-bin lower, until the histogram content is larger than $n_{y_{\text{max}}}$.

5. Estimates of $C_{1550}^{\text{max}}$ and $\sigma_{1550}$ are obtained from the position of the maximum in this histogram (if the equivalent TH step is lower than 5) or the mean of the histogram (for equivalent TH steps of 5 and above) and the histogram RMS, respectively. This distinction proved necessary due to the specific properties of some of the distributions at the lower TH steps. The distribution often has long tails toward small values of $C_{1550}$ such that the mean will be considerably lower than the position of the maximum. At higher TH, the difference is typically not very large.

This procedure is only performed if the global maximum of $C_{750}$ is below 3.75. Otherwise, the histogram mean and RMS along $C_{1550}$ are taken directly, and $C_{750}^{\text{max}}$ is set to be the histogram mean along the $y$-axis.

If, for whatever reasons, no valid threshold values could be found, the results from the last successful lower TH step are copied to this one.

From the model, the slope of $C_{1550}^{\text{RTM}}$ w.r.t. $C_{750}$ can be determined by calculating

$$b_{\text{RTM}}^{\text{RTM}} = \frac{C_{1550}^{\text{RTM}}(C_{750}^1) - C_{1550}^{\text{RTM}}(C_{750}^0)}{C_{750}^1 - C_{750}^0}. \quad (3)$$
are TH dependent initialization parameters. \( C^{RTM}_{1550} \) is a value determined from the SCIATRAN pure aerosol simulations. It is the minimum of the distribution of \( C_{1550} \) for each \( C_{750} \) bin and depends on the scattering angle.

An example of the set of parameters derived for a TH of 15 km is shown in figure 5.

The cloud–aerosol threshold for all values \( C_{750} < 3.75 \) is calculated as follows:

\[
\begin{align*}
C^{Cut}_{1550} &= C^{max}_{1550} - N_{RMS}(\cos(\theta), TH)\sigma_{1550}, & \text{if } C_{750} > \max(C^{max}_{750}, 3.0) \\
C^{Cut}_{1550} &= C^{max}_{1550} - N_{RMS}(\cos(\theta), TH)\sigma_{1550} + p^{RTM}(C_{750} - C^{max}_{750}), & \text{if } C_{750} \leq \max(C^{max}_{750}, 3.0) \\
C^{Cut}_{1550} &= \max(C^{RTM}_{1550} (C_{750} = 1.0), C^{Cut}_{1550}).
\end{align*}
\]

The parameter \( N_{RMS} \) is a scattering angle dependent parameter and indicates the number of standard deviations of the estimated peak. It is activated only if the equivalent TH step is above a lower limit \( i^{Max}_{TH} \) given in table 1.
### Table 1: Indices of maximum equivalent TH steps $i_{TH}^{Max}$ above which $N_{RMS}$ can be different from 1.

<table>
<thead>
<tr>
<th>index</th>
<th>Latitude Range</th>
<th>$i_{TH}^{Max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$-90^\circ - 60^\circ$</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>$-60^\circ - 30^\circ$</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>$-30^\circ - 0^\circ$</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>$0^\circ - 30^\circ$</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>$30^\circ - 60^\circ$</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>$60^\circ - 90^\circ$</td>
<td>5</td>
</tr>
</tbody>
</table>

The scattering angle dependence is given in five bins of $\cos(\theta)$, as shown in table 2. For TH steps up to the values in table 1, $N_{RMS} = 1$. If $C_{750} > 3.75$, $N_{RMS} = 2$.

![Table 2: Number of standard deviations of the estimated peak (see eq. 4).](image)

<table>
<thead>
<tr>
<th>$\cos(\theta)$</th>
<th>-1.0 to -0.6</th>
<th>-0.6 to -0.2</th>
<th>-0.2 to 0.2</th>
<th>0.2 to 0.6</th>
<th>0.6 to 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{RMS}$</td>
<td>1.0</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
<td>2.0</td>
</tr>
</tbody>
</table>

The possible values of $\sigma_{1550}$ are further limited from below by 0.03, and from above by 0.08 (for equivalent TH steps smaller than a latitude dependent maximum), only if $C_{750} < 3.75$.

For TH steps below 3, the model values ($C_{1550}^{RTM}(C_{750})$) for the lowest $TH \approx 2$ km are used directly. In order to derive the probability maps later on, missing values for TH steps above 2 are interpolated in time.

Figure 6 shows an example for the derived thresholds as a magenta graph on top of the 2D histograms for THs of 9 and 15 km in August.

### 3.3 Fits to the histogram distributions

The data are now fitted to a series of Gaussian peaks in $C_{1550}$, maximum 3 to the left and 2 to the right of the cloud–aerosol threshold. The maximum likelihood method of TMinuit, implemented in ROOT, is used for this. Exactly how many peaks are fitted depends on the range of the data in $C_{1550}$, it is roughly one every 0.1 on the left, and one every 0.05 on the right. To get the distribution, the bins in $C_{750}$ are integrated until either a minimum of data points, a minimum of the mean value per filled bin or a minimum of the maximum bin content is reached. This implies that the range of $C_{750}$ for which a fit is valid is variable. The initial parameters and limits are set through an analysis of histogram range, the rightmost left peak is limited to $C_{1550}^{Cut}$ from below, the leftmost right peak to the same value from above. In a second round, the fit is initialized with the fit parameters from the first fit, but this
Figure 6: Example for cloud–aerosol thresholds, derived for $TH \approx 9\text{km}$ (left) and $TH \approx 15\text{km}$ (right) for August, for the latitude bin from $30^\circ N$ to $60^\circ N$. The magenta curve indicates the derived aerosol-cloud discrimination threshold.

time the fit merely requires that all peaks after the first have a position higher than the previous peak. Otherwise, the limits are released. In many cases, the result is very similar to the original fit, and in many other cases the $\chi^2$ is better after the second fit. The reduced $\chi^2$ is on average 0.6, with high values larger than 4 only for a small fraction of all fits (less than a per cent). Note that in a maximum likelihood fit, all bins of a histogram are taken into account, also the empty ones, thus it is expected that the reduced $\chi^2$ is smaller than 1 if not all bins are filled. Figure 7 shows examples for the fit results for different slices of $C_{750}$ in August 2009, 15 km and $30^\circ N-60^\circ N$.

4 Determination of Probabilities

The conditional probability that a cloud is present in the FoV, given a measurement of $C_{1550}$ can be expressed as:

$$P(\text{Cloud}|C_{1550}) \equiv P_{\text{Cloud}}(t, TH, lat, a) = \frac{P(C_{1550}|\text{Cloud})P(\text{Cloud})}{P(C_{1550})}$$

according to Bayes’ Theorem. $P(\text{Cloud})$ is the total (apriori) probability that there is a cloud at this bin in time ($t$), latitude ($lat$), TH, surface type ($a$) and for this slice in $C_{750}$. The probability of a measurement giving $C_{1550}$ is simply

$$P(C_{1550}) = \frac{N(C_{1550})}{N} = \frac{\sum_{i=0}^{n_p} f_i(C_{1550})}{\sum_{i=0}^{n_p} \int dx f_i(x)}$$

where the sum runs over all non-zero peaks. From the Gaussian fit analysis we can identify each peak as belonging to a cloud or pure aerosol distribution. For this,
Figure 7: Example for fits to the data at $TH \approx 15\text{km}$ August 2009 (right), for the latitude bin from $30^\circ N$ to $60^\circ N$. Each panel depicts a slice in $C_{750}$, increasing clockwise from the top left. The vertical line indicates the cloud–aerosol threshold with $N_{RMS} = 2$. Note that this is an example with data for a single month, while the final analysis uses the data integrated over the whole mission for each month.

Each peak has to be checked again as to whether it lies on the cloud or aerosol side of $C_{1550}^{Cut}$. The formalism to calculate $C_{1550}^{Cut}$ is analogous to Eq. 4.

Then

$$P(C_{1550}|\text{Cloud}) = \frac{\sum_{i=0}^{n_c} f_i(C_{1550})}{\int dx f_i(x)},$$

where this time the sum runs over all peaks identified as “cloud” peaks. The apriori probability for $P(\text{Cloud})$ is not known, but can in principle be inferred from the same analysis:

$$P(\text{Cloud}) = \frac{\sum_{i=0}^{n_p} \int dx f_i(x)}{\sum_{i=0}^{n_p} \int dx f_i(x)}.$$  

Equation 5 then reduces to

$$P_{\text{Cloud}} = \frac{\sum_{i=0}^{n_c} f_i(C_{1550})}{\sum_{i=0}^{n_p} f_i(C_{1550})} = \frac{N_{\text{Cloud}}}{N_{\text{Total}}},$$
i.e., the simple ratio of the sum of the values of all cloud peaks at a measured $C_{1550}$ to the sum of the values of all peaks at this $C_{1550}$. The right part of this equation defines the expected number of total ($N_{Total}$) and cloud events ($N_{Cloud}$) in bin of $C_{1550}$ with a given width. The values of $P_{Cloud}$ have been mapped as look-up-tables (LUTs) as a function of $(C_{1550}, C_{750})$ for each bin in $(t, TH, lat, a)$.

A few complications with this LUT approach need to be resolved, though. First of all, for some ranges in $C_{750}$ no valid fit results may exist because, e.g., the $\chi^2$ is too large. For these bins, the probabilities from the adjacent bins in $C_{750}$ are linearly interpolated. In some cases (particularly over land), some time bins are not filled due to low statistics. In these cases a linear interpolation is performed over the adjacent monthly bins.

A second issue is more fundamental in nature and requires a deviation from the pure probability approach. The values of the fit functions are in principle defined for all values of $C_{1550}$, although in reality the values should lie within some physical boundaries. On the other hand, the fits cannot be expected to exactly reproduce tails in the data which originate from either rare, exceptional meteorological conditions or which are simply outliers due to (electronic) noise, in particular in channel 6+. This means that both “outliers” in the data as well as the tail regions of the fitted distributions are not described appropriately and need some fudging. The procedure described in the following is based on empirical studies with the goal of producing as little artifacts of possible and reproducing a somewhat naively expected behavior in which the cloud probability decreases more or less monotonically as $C_{1550}$ increases.

- Cloud peaks are cut off at $2.5 \sigma$ above their mean if $C_{1550} > C_{Cut}^{1550}$.
- Aerosol peaks are cut off at $2.5 \sigma$ below their mean if $C_{1550} < C_{Cut}^{1550}$.
- If $C_{1550} > C_{Cut}^{1550}$ and $P_{Cloud} > P_{prev}_{Cloud}$, the previous value $P_{prev}_{Cloud}$, which was obtained for a previous value $C_{Prev}^{1550}$ with $C_{Prev}^{1550} < C_{Cut}^{1550} < C_{1550}$ is assigned to $P_{Cloud}$.
- If $N_{Total} < 10^{-6}$, assign $P_{Cloud} \equiv 10^{-6}$ if $C_{1550} \geq C_{Cut}^{1550}$.
- If $N_{Total} < 10^{-6}$ and $N_{Cloud} > 0$, assign $P_{Cloud} \equiv 1$ if $C_{1550} < C_{Cut}^{1550}$.
- If $N_{Total} < 10^{-6}$ and $N_{Cloud} = 0$, assign $P_{Cloud} \equiv 0.5$ if $C_{1550} < C_{Cut}^{1550}$.
- If $P_{Cloud} < 10^{-6}$, assign $P_{Cloud} \equiv 10^{-6}$. This is only necessary to distinguish bins for which information on $P_{Cloud}$ exists from those for which none is available and which have a value of 0.

In a final step, the two bins adjacent to the last filled bin in the $C_{750}$ direction are repeated with the last filled values. This is needed to ensure that subsequent smoothing does not affect the originally filled bins.
In figure 8 two examples for the cloud probability maps corresponding to the data in figure 6 are shown. There is still considerable jitter between the $C_{750}$ slices which is mainly due to fit errors and subsequent misidentification of peaks. This jitter can in principle be reduced by smoothing the maps. For comparison, figure 9 shows the same distributions as in figure 8, but with smoothed probability maps. In general it seems that there is not much loss in information due to smoothing. After all, the slicing in bins of $C_{750}$ is arbitrary and there is good reason to assume that conditions change smoothly with $C_{750}$.

5 Application to the Data

This section describes the application of the resulting cloud probabilities to the data. It presumes that the probabilities are available as ASCII-data (or converted format) for each month, 30°-latitude bin, TH step and surface type. The surface type can be determined with the land-sea mask delivered together with the probability LUTs. The generation of the probabilities also used a particular bad pixel mask for channel 6 which should be applied for the processing as well, and which is being delivered with the LUTs. The probabilities will be calculated for $n_{TH} = 11$ equivalent TH steps (see eq. 20 below). In this procedure it is also necessary to perform a PSC detection, for which an algorithm very much like that of the current SCODA algorithm is applied. This is done for $n_{TH}^{PSC} = 12$ equivalent TH steps. The reason for this synchronized approach is that the intended output should distinguish between “normal”, i.e., tropospheric, clouds and PSCs where applicable.
Figure 9: \( P_{\text{Cloud}} \) for the same cases as Fig. 8, but with the probabilities smoothed. The contour plots are the data distributions for August 2004 in orange and for August 2009 in turquoise.

5.1 Data Containers

The following arrays need to be created and initialized for each limb profile:

1. An integer array \( c_{\text{below}} \) with \( n_{\text{TH}} \) elements, initialized to 0. It holds the TH step where a low cloud influencing the color ratio at this TH step is suspected (see section 5.6).

2. A double array of the reflectances \( R_{0,1,2,3} \) each with as many elements as there are limb scans, initialized to 0.

3. A double array of the errors of reflectances \( R_{2,3}, \delta R_{2,3} \), each with as many elements as there are limb scans, initialized to 0.

4. A double array of the limb dark currents \( D \) each with at least \( n_{\lambda} = 4 \) elements, initialized to 0.

5. A double array of the limb dark current errors \( \delta D \) each with at least \( n_{\lambda} = 4 \) elements, initialized to 0.

6. A double array of the color ratio \( C_{1550} \) with at least \( n_{\text{TH}} + 1 \) elements, initialized to 0.

7. A double array of the errors of the color ratio \( C_{1550}, \delta C_{1550} \), with at least \( n_{\text{TH}} + 1 \) elements, initialized to 0.

8. A double array of the gradient of the color ratio \( C_{1550} \) w.r.t. the next TH step, \( \Delta C_{1550} \), with \( n_{\text{TH}} \) elements, initialized to 0.
9. A double array of the color ratio $C_{750}$ with at least $n_{TH}^{PSC} + 1$ elements, initialized to 0.

10. A double array of the (relative) gradient of the color ratio $C_{750}$ w.r.t. the next TH step, $\Delta C_{750}$, with at least $n_{TH}^{PSC}$ elements, initialized to 0.

11. A double array of the probabilities $P_{\text{Cloud}}$ with $n_{TH}$ elements, initialized to -1.

12. A double array of the cloud apriori weights $w_{\text{Cloud}}$ with $n_{TH}$ elements, initialized to 1.

13. An integer or boolean array of the all cloud flag $f_{\text{Cloud}}^{0}$ with $n_{TH}$ elements, initialized to 0 (or possibly -1 if integer). This particular array is not actually necessary, but may be stored for convenience.

14. An integer or boolean array of the tropospheric cloud flag $f_{\text{Cloud}}^{1}$ with $n_{TH}$ elements, initialized to 0 (or possibly -1 if integer). This particular array is not actually necessary, but may be stored for convenience.

15. An integer or boolean array of the PSC flag $f_{PSC}$ with $n_{TH}^{PSC}$ elements, initialized to 0 (or possibly -1 if integer). This particular array is not actually necessary, but may be stored for convenience.

16. A double scalar for the highest TH with a cloud, one each for normal and tropospheric clouds and PSCs, i.e., 3 values.

17. A double scalar for the value of $P_{\text{Cloud}}$ (resp. $\Delta C_{750}$) at the highest TH of each normal and tropospheric clouds and PSCs, i.e., 3 values.

5.2 Data Preprocessing

Four wavelength windows are needed, their center and width is given in table 3. The individual detector pixels are to be offset and memory effect (?)/nonlinearity corrected, but not corrected for the limb dark signal (the correction is done later in the process, see below), and also not for the GADS dark current. Radiometric calibration and degradation correction (V9.02 key data and M-factors) has to be applied, and radiance is divided by the corresponding irradiance.

$$R^n = \pi \frac{I^n}{I_0}$$ (10)

for each pixel $n$ inside the window. The limb dark signal $D^n$ is computed in the same way for the TH step where the limb dark measurement is executed. The average
The average limb dark reflectance for each window is computed in the same way, but in addition a simple averaging over all limb readouts is performed:

\[
\langle D_i \rangle = \frac{1}{N_{\text{Readouts}}} \sum_{j=0}^{N_{\text{Readouts}}} \langle D_i \rangle_j
\]

\[
\delta \langle D_i \rangle = \frac{1}{N_{\text{Readouts}}} \sqrt{\sum_{j=0}^{N_{\text{Readouts}}} \delta^2 \langle D_i \rangle_j}.
\]

<table>
<thead>
<tr>
<th>index</th>
<th>(\lambda_{\text{Center}} ) [nm]</th>
<th>(\Delta \lambda ) [nm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>750</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1090</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1550</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1670</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 3: Wavelength windows for the cloud detection algorithm. All pixels with wavelengths \(|\lambda_{\text{Center}} - \lambda^n| \leq \Delta \lambda\) are to be integrated over.

The color ratios can then be computed from the average reflectances:

\[
C_{750} = \frac{\langle R_0 \rangle - \langle D_0 \rangle}{\langle R_1 \rangle - \langle D_1 \rangle}
\]

\[
C_{1550} = \frac{\langle R_2 \rangle - \langle D_2 \rangle}{\langle R_3 \rangle - \langle D_3 \rangle}.
\]

The error of the color ratio \(C_{1550}\) can be computed from the error of the involved reflectances and dark signals:

\[
\delta C_{1550} = |C_{1550}| \sqrt{\left(\frac{\delta^2 \langle R_2 \rangle + \delta^2 \langle D_2 \rangle}{\langle R_2 \rangle - \langle D_2 \rangle} \right)^2 + \left(\frac{\delta^2 \langle R_3 \rangle + \delta^2 \langle D_3 \rangle}{\langle R_3 \rangle - \langle D_3 \rangle} \right)^2}.
\]
For the purpose of the low cloud check described in section 5.6 below it is recommended to fill the arrays of color ratios, reflectances, the errors of the average reflectances $\langle R_2 \rangle$ and $\langle R_3 \rangle$ and the values of the average limb dark signal and its errors, $\langle \delta \rangle (D_2)$ and $\langle \delta \rangle (D_3)$ for at least $n_{TH} - 1 \cdot (n_{PSC}^{TH} + 1)$ TH steps before the loop that determines the probabilities.

### 5.3 Determine LUT Cell

The steps described in this section as well as in section 5.5 can be performed individually for each TH step, starting from the lowest TH. It is assumed here that the L1 data have been correctly sorted into one of the four limb profiles of each state. First, the equivalent TH step $i_{TH^*}$ corresponding to the current TH step $i_{TH}$ has to be determined. For this, the step with the TH closest to 35 km, $i_{35}$,  

$$i_{TH^*} = i_{TH} + 12 - i_{35}, \quad (20)$$

where $i_{35}$ is the TH index in the scan which is closest to 35 km. The loop stops if $i_{TH^*} = n_{TH}$. If the equivalent step is smaller than 2, the LUT values for $i_{TH^*} = 2$ should be used. All indices are counted from 0.

The surface index is fixed to 2, meaning that regardless of the actual surface type, the LUTs for land and sea surfaces combined are being used. This is because on the one hand overall there was not a lot of difference in the distributions between land and ocean, on the other hand quite a number of the land LUTs lacked sufficient statistics.

The time bin is simply  

$$i_{T} = (m - 1). \quad (21)$$

The month $m$ is here simply the month of the year, i.e., running between 1 and 12.

The latitude bin $i_L$ is determined from the latitude of the 10th TH step:  

$$i_{L} = \text{floor}\left(\frac{\text{lat}_{10} + 90}{30}\right). \quad (22)$$

The probability LUTs are given in cells with a width of (0.01, 0.1) in $(C_{1550}, C_{750})$ with the corner points (0.5,0.5) and (1.5,4.5). Values of $(C_{1550}, C_{750})$ exceeding these limits should be fudged to the closest limit. If $C_{750} < 0$ or $C_{750} > 5$ the data point should be rejected because of bad data quality.

$$i_{750} = \text{floor}\left(\frac{C_{750} - 0.5}{40}\right), \quad (23)$$

$$i_{1550} = \text{floor}\left(\frac{C_{1550} - 0.5}{100}\right). \quad (24)$$

Typically only rows in $C_{750}$ which are above a certain minimum and below a certain maximum are actually filled. Unfilled values in the LUTs have a value of exactly 0,
while the minimum filled value is $10^{-6}$. So if the LUT value of the bin $(i_{1550}, i_{750}) = 0$, $i_{750}$ should be replaced with the first or last filled bin, and the LUT value taken from this cell.

5.4 Determine PSC Gradients

The PSC gradient ratio for this TH step $i_{TH}$ is defined as:

$$
\Delta C_{750} = \frac{C_{750}(i_{TH} + 1)}{C_{750}(i_{TH})}.
$$

(25)

Note: Even though this gradient ratio is only used for PSC identification, it may still be valuable for diagnosing the final data set.

5.5 Retrieve Probabilities from LUTs and Apply Apriori Weights.

The value of $P_{Cloud}$ is the value of the cell $(i_{1550}, i_{750})$ for the time bin $i_T$, equivalent TH bin $i_{TH}$, latitude bin $i_L$, and surface type $i_S = 2$. The apriori weight needs to be applied for situations where there is no cloud with a high (apriori) probability at the given latitude and TH, but the data are not reliable enough to determine the cloud probability with this method. The weight $w_{Cloud}$ should be set to 1 by default. If, on the other hand:

$$
\langle R_0 \rangle - \langle D_0 \rangle < R_0^{min} \text{ or } \\
\langle R_1 \rangle - \langle D_1 \rangle < R_1^{min} \text{ or } \\
\langle R_3 \rangle - \langle D_3 \rangle < R_3^{min}
$$

(26)

then $w_{Cloud} = 0$. The minimum values of the reflectances, $R_i^{min}$, are set as initialization parameters in table 4.

To further evaluate the apriori weight (if $w_{Cloud} = 1$), the gradient of $C_{1550}$ w.r.t. to the next TH step has to be computed as well as its error. The gradient is defined as

$$
\Delta C_{1550} = C_{1550}(i_{TH} + 1) - C_{1550}(i_{TH}).
$$

(27)
Its error can be computed as follows:

\[
\left( \frac{\partial C_i}{\partial D^2_i} \right)^2 \equiv C_i^2 \left[ \frac{\delta^2 R^2_i}{(R^2_i - D^2_i)} + \frac{\delta^2 R^3_i}{(R^3_i - D^3_i)} \right]
\]

for \( i = i_{TH}, i_{TH} + 1 \) \( (28) \)

\[
\left( \frac{\partial \Delta C_{1550}}{\partial D^2_2} \right)^2 = \frac{1}{(R^2_{TH} + 1 - D^2_3)^2} + \frac{1}{(R^3_{TH} - D^3_3)^2} \left[ 2 \frac{(R^2_{TH} + 1 - D^2_3)(R^3_{TH} - D^3_3)}{(R^2_{TH} + 1 - D^2_3)^2} \right] \]

\[
\left( \frac{\partial \Delta C_{1550}}{\partial D^3_3} \right)^2 = \frac{C_{1550}^2(i_{TH} + 1)}{(R^2_{TH} + 1 - D^2_3)^2} + \frac{C_{1550}^2(i_{TH})}{(R^3_{TH} - D^3_3)^2} \left[ 2 \frac{C_{1550}^2(i_{TH} + 1)C_{1550}(i_{TH})}{(R^2_{TH} + 1 - D^2_3)^2} \right] \]

\[
\delta \Delta C_{1550} = \left[ C_{1550}^2(i_{TH} + 1)(\partial C_{i_{TH} + 1})^2 + C_{1550}^2(i_{TH})(\partial C_{i_{TH}})^2 \right]^{1/2} \]

\[
(\partial C_i)^2 \equiv C_i^2 \left[ \frac{\delta^2 R^2_i}{(R^2_i - D^2_i)} + \frac{\delta^2 R^3_i}{(R^3_i - D^3_i)} \right]
\]

Given the gradient and its statistical error, the Cumulative Distribution Function (CDF) of Student’s t-distribution can be used to estimate the probability that both \( C_{1550}(i_{TH} + 1) \) and \( C_{1550}(i_{TH}) \) are from the same sample. This is done by assuming the null hypothesis for a sample with one degree of freedom (i.e., both data points come from a sample with the same mean and variance) and computing the probability that this is the case. The weight is then multiplied by the 1 minus this probability:

\[
t \equiv -\frac{\Delta C_{1550}}{\delta \Delta C_{1550}} \]

\[
P_{\text{Student}} = CDF_{\text{Student}}(t, 1) \]

\[
w_{\text{Cloud}} = w_{\text{Cloud}}(1 - P_{\text{Student}}) \]

The reason for taking the negative value of the gradient in the argument of the CDF is that later we are interested in the probability of \( t \) being smaller than a certain value (see section 5.6). Note that the CDF for one degree of freedom can be analytically computed (see https://en.wikipedia.org/wiki/Student%27s_t-distribution):

\[
CDF_{\text{Student}}(t, 1) = \frac{1}{2} + \frac{1}{\pi} \arctan(t)
\]

In this way, values of \( C_{1550} \) which are within errors close to the value at the next TH step are assigned a low weight, meaning that it is highly improbable that this TH is affected by a cloud. If it were, one would expect a significant gradient. On
the other hand, if the cloud is below the FoV but affects this TH due to its albedo, no significant gradient is expected. This method avoids having to compare with absolute thresholds which is problematic in two ways:

1. A fixed threshold implies a fixed minimum cloud signal, while for the thinnest clouds of course the signal can be arbitrarily small.

2. With one or both $C_{1550}$ values involved in the gradient unaffected by clouds, any of the values can be quite meaningless due to noise. Comparing to a threshold without regard to the statistical errors can therefore lead to more or less random selections.

Thus, this strategy makes the most out of the information available from the measurement while maximizing the sensitivity. Of course this also implies that there is a detection limit and implicitly a (situation dependent) threshold.

### 5.6 Perform Low Cloud Check

The low cloud check is performed to identify “false” clouds which could occur for two main reasons:

- Due to tails in the $C_{1550}$ distribution which are either true tails in the aerosol distribution or appear because of calibration uncertainties and noise, a peak could be wrongly identified as a cloud peak. This can be sometimes just an insignificant peak affecting only a handful of data points. Since the low cloud check is based mainly on gradients, it is highly likely, though not guaranteed that all misidentified events will be rejected. Gradients due to changes in aerosol are in general expected to be smaller than those due to clouds, but they still may be significant, for instance when volcanic aerosol layers are present or in the lower troposphere.

- The other reason is a genuine physical effect due to the albedo of a cloud below the field of view (see Fig. 1). A thick low ice cloud can affect $C_{1550}$ up to high altitudes.

In both cases the effect can be mitigated by checking whether a thick cloud is present at lower altitudes. The low-cloud check should be performed for all events where

$$
\begin{align*}
i_{TH}^* & > 3 \quad \text{and} \\
P_{\text{Cloud}} & > 10^{-6} \quad \text{and} \\
w_{\text{Cloud}} & > 0.9 \quad \text{and} \\
\Delta C_{1550} & < 0.2 \quad \text{(37)}
\end{align*}
$$

$$
\Delta C_{1550} < 0.2 \quad \text{(38)}
$$
If this data point is selected, the following loop has to be performed, that starts one step below this one \((i_{iTH})\) and stops if either the equivalent step \((i_{TH}^{below})\) < 3 or the criteria in Eq. 39 are met. To start with, set \(i_{TH}^{below} = i_{iTH} - 1\). Then for each step in the loop, while \(i_{TH}^{below} \geq 3\), check if

\[
P_{\text{Cloud}}(i_{TH}^{below}) > 0.5 \text{ and } w_{\text{Cloud}}(i_{TH}^{below}) > 0.9 \text{ and } C_{1550}(i_{TH}^{below}) < 0.95 \text{ and } \Delta C^* \equiv \Delta C_{1550}(i_{iTH}) - \Delta C_{1550}(i_{TH}^{below}) < -2.5 \delta \Delta C^*. \tag{39}\]

The error of the difference in gradients \(\Delta C^*\) can be approximated from the errors of the individual color ratios as follows:

\[
\delta \Delta C^* = \sqrt{\delta^2 C_{1550}(i_{iTH}) + 1} + N \delta^2 C_{1550}(i_{TH}^{this}) + N \delta^2 C_{1550}(i_{TH}^{below} + 1) + \delta^2 C_{1550}(i_{TH}) \tag{40}\]

with

\[
N = \begin{cases} 
1 & \text{if } i_{iTH} > i_{TH}^{below} + 1 \\
2 & \text{if } i_{iTH} = i_{TH}^{below} + 1. 
\end{cases} \tag{41}\]

While the loop runs, decrease \(i_{TH}^{below}\) by 1 in each iteration. If these conditions are met, the loop is stopped and

\[
c_{\text{below}}(i_{TH}) = i_{TH}^{below}. \tag{42}\]

If, after the low cloud check, \(c_{\text{below}}(i_{TH}) > 0\) it is likely that the cloud is \(c_{\text{below}}(i_{TH})\) steps below the current one. The current TH step can be considered cloud free.

### 5.7 Determine PSC Flag and PSC Top TH

The PSC parameters to be determined are the PSC flag \(f_{\text{PSC}}\) for each TH step, the PSC Top TH, \(TH_{\text{PSC}}\), and the value of \(\Delta C_{750}\) at \(TH_{\text{PSC}}\), \(\Delta C_{\text{PSC}}\). PSC flagging should be only enabled for the seasons and latitudes when and where they are to be expected. Set the PSC flag \(f_{\text{PSC}}\) to 1 if:

\[
l_{10} < -l^{\text{PSC}} \quad \text{and} \quad m_{\text{Start}[0]}^{\text{PSC}} \leq m \leq m_{\text{End}[0]}^{\text{PSC}} \text{ or } m_{\text{Start}[1]}^{\text{PSC}} \leq m \leq m_{\text{End}[1]}^{\text{PSC}}\tag{43}\]

However, the PSC flag should be vetoed, i.e., set back to 0 if:

\[
\langle R_0 \rangle - \langle D_0 \rangle < R_0^{\text{PSC,min}} \text{ or } \langle R_1 \rangle - \langle D_1 \rangle < R_1^{\text{PSC,min}} \text{ or } \langle R_3 \rangle - \langle D_3 \rangle < R_3^{\text{PSC,min}} \text{ or } TH < TH_{\text{PSC,min}}. \tag{44}\]
The initialization parameters for potential PSC latitudes \((\text{lat}^{PSC})\) and seasons, i.e., \(m_{\text{Start,End}}^{PSC}\), for both hemispheres and the lower thresholds for the measured reflectances \(R_{0,1,3}\) and TH \((TH_{\text{min}}^{PSC})\) are given in table 4.

If the conditions of Eq. 43 are met, a loop over all \(n_{TH}^{PSC}\) steps should be performed, starting from the equivalent TH step \(n_{TH}^{PSC} - 1\) and stopping (latest) when \(TH\text{km} < 14\text{km}\).

Initialize \(TH^{PSC}\) to a large negative value and \(\Delta C^{PSC}\) to 0. Both values have to be updated in the loop if \(\Delta C^{PSC}\) is larger than its previously determined value at a higher TH, i.e., if

\[
f_{PSC} > 0 \text{ and } \Delta C_{750} > \Delta C_{PSC} \text{ and } C_{1550} < C_{1550}^{PSC}(\text{lat}, i_{TH^*}),
\]

then

\[
\begin{align*}
TH^{PSC} &\equiv TH \\
C^{PSC} &\equiv \Delta C_{750}.
\end{align*}
\]  

The parameter \(C_{1550}^{PSC}\) indicates a hemisphere and TH dependent value of the maximum of \(C_{1550}\) allowed for a PSC, it is also given in table 5. If at any iteration \(\Delta C_{PSC} > \Delta C_{PSC}^{\text{thresh}}\), the loop is stopped. Else, at each iteration where this is not the case, set \(f_{PSC}\) to 0. The parameter \(C_{1550}\) indicates a hemisphere and TH dependent value of the maximum of \(C_{1550}\) allowed for a PSC, it is given in table 5. The threshold \(\Delta C_{PSC}^{\text{thresh}}\) is calculated from hemisphere and TH dependent values given in table 5 as follows:

\[
\Delta C_{PSC}^{\text{thresh}}(\text{lat}, i_{TH^*}) = C_{750}^{0}(\text{lat}, i_{TH^*}) + p_{750}(\text{lat}, i_{TH^*})(C_{1550} - 0.9) * k_{1550}
\]

where \(k_{1550} \equiv \begin{cases} 1 & \text{if } C_{1550} \geq 0.9 \\ 0 & \text{if } C_{1550} < 0.9. \end{cases} \) (48)

This PSC flagging algorithm constitutes a refinement of the previous algorithm inasmuch as that the thresholds are adapted to TH and latitude specific situations, and that the minimum reflectance requirements avoids artifacts from measurements with very low signals and relatively high noise. The thresholds are derived from the statistical relationship between \(\Delta C_{750}\) and \(C_{1550}\) and take into account both the higher sensitivity of \(C_{750}\) and the at times higher aerosol load in the Northern high latitude regions.

### 5.8 Determine Cloud Flag and Cloud Top TH

To eventually arrive at a cloud flag, another loop has to be performed, starting from the highest TH step corresponding to \(i_{TH^*} = n_{TH} - 1\), until whichever of \(i_{TH}\) or \(i_{TH^*}\) is 0 first. To mark an entire limb profile, and determine the highest TH of the cloud, two more variables are needed:
1. $P_{\text{Cut}}^0$, which is either the highest probability of the profile or the probability at the highest TH which is larger than a given threshold $P_{\text{Thresh}}$ (see table 4). The former is relevant only if there is no TH step in the profile for which $P_{\text{Cloud}} > P_{\text{Thresh}}$.

2. $TH_{\text{Cut}}^0$, the tangent height at which $P_{\text{Cut}}^0$ is set.

These two values should be first initialized to 0 and a large negative default value, respectively, and updated in the loop if

$$P_{\text{Cloud}} > P_{\text{Cut}}^0 \text{ and } w_{\text{Cloud}} > w_{\text{min}} \text{ and } c_{\text{below}}(i\text{TH}) = 0,$$

then

$$TH_{\text{Cut}}^0 \equiv TH$$
$$P_{\text{Cut}}^0 \equiv P_{\text{Cloud}}.$$  

If the conditions in Eq. 49 are met and $P_{\text{Cloud}} > P_{\text{Thresh}}$, then the loop should be stopped to ensure that the highest TH where a cloud is detected is marked. The minimum cloud weight $w_{\text{min}}$ is given in table 4. To determine these values this way opens up the possibility to investigate profiles with a nonzero probability, even if the highest probability is lower than $P_{\text{Thresh}}$.

The simplest cloud flag is the one which marks a data point at a given TH step at or below the cloud TH as affected by a cloud:

$$f_{\text{Cloud}}^0(i\text{TH}) = 1(\text{true}) \text{ if } P_{\text{Cloud}} > P_{\text{Thresh}} \text{ and } TH \leq TH_{\text{Cut}}^0 \text{ and } c_{\text{below}}(i\text{TH}) = 0.$$  

5.9 Determine Tropospheric Cloud Flag and Tropospheric Cloud Top TH

The cloud parameters determined by the algorithm described in section 5.8 are set regardless of whether the cloud is likely to be a PSC or a “normal” cloud. It may be necessary to make a distinction between the two types of clouds. One reason for this may be that people may be interested in PSCs yet still be able to exclude tropospheric clouds. Another reason is that even though a good part of the PSCs can be detected, some will be identified as aerosol (see section 6.3). In other words, the flag based on $C_{1550}$ is a potential, but not a reliable indicator for PSCs. It may therefore be worth the while to introduce a “normal” cloud flag, i.e., a flag that is set when the threshold conditions are met, but the cloud is most like not a PSC. For this flag, the PSC flag determined in section 5.7 is used to veto the “normal” cloud flagging. Two additional variables are needed:

1. $P_{\text{Cut}}^1$, which is the probability at the highest TH which is larger than a given threshold $P_{\text{Thresh}}$. 

2. $TH_{Cut}^1$, the tangent height at which $P_{Cut}^1$ is set.

These two values should be first initialized to 0 and a large negative default value, respectively, and updated in the loop if

$$P_{\text{Cloud}} > P_{\text{Thresh}} \text{ and } w_{\text{Cloud}} > w_{\min} \text{ and } c_{\text{below}}(i_{TH}) = 0 \text{ and } f_{PSC} = 0,$$

then

$$TH_{Cut}^1 \equiv TH$$

$$P_{Cut}^1 \equiv P_{\text{Cloud}}.$$

If the conditions in Eq. 53 are met, then the loop should be stopped to ensure that the highest TH where a cloud is detected is marked. The minimum cloud weight $w_{\min}$ is given in Table 4.

The “normal” cloud flag $f_{\text{Cloud}}^1$ is the one which marks a data point at a given TH step as affected by a “normal” cloud:

$$f_{\text{Cloud}}^1(i_{TH}) = 1(\text{true}) \text{ if }$$

$$P_{\text{Cloud}} > P_{\text{Thresh}} \text{ and } TH \leq TH_{Cut}^1 \text{ and } c_{\text{below}}(i_{TH}) = 0 \text{ and } f_{PSC} = 0.$$

This strategy implies the assumption that during PSC season and at the affected latitudes, tropospheric clouds do not affect THs above roughly 14 km, and, vice versa, PSCs have only a negligible effect below 14 km. When, at the end of the PSC season, their location is fairly low, there could be some mingling effects. However, these cannot be come by without a detailed analysis of profile shapes etc., if at all. Outside PSC season and latitudes, the tropospheric general cloud flags should be the same.

6 Results

6.1 Cloud Identification

Figure 10 shows the quantity $P_{\text{Cloud}}w_{\text{Cloud}}$ applied to the data, again for August 2004 and 9 at a TH of about 15 km, between 30°N and 60°N. Note the different scales in $C_{1550}$ and $C_{750}$ in both plots! The probability maps used here were not smoothed, so the $C_{750}$ bins are clearly visible. The low cloud check removes a good portion of the data which seem ambiguous by eye already, about 16% of the data in 2004 and 4% in 2009. Overall it seems that the high misidentification fraction for the high aerosol conditions in 2009 is substantially reduced.
6.2 Cloud Statistics

Figure 11 shows the fraction of data points at each TH and latitude affected by clouds, i.e., the ratio of all data points with $P_{\text{Cloud}} \cdot w_{\text{Cloud}} > 0.5$ and $c_{\text{below}}(i_{\text{TH}}) = 0$, again for August 2004 and 2009. Both plots look similar, though it seems that the average cloud probability at $TH > 9$ km and North of $40^\circ$ N is reduced compared to 2004.

Figure 12 shows the occurrence rate of clouds with the apparent top at the corresponding TH, again for August 2004 and 2009. Again, both plots look very similar, though it indeed seems as if the cloud rate at higher altitudes and high Norther latitudes is diminished compared to 2004.

Finally, figures 13 (14) show the occurrence rate of clouds with the apparent top at each of six measured THs, for February (August) 2004, vs. latitude and longitude...
6.3 Preliminary Investigation of PSCs

The data quality of $C_{1550}$ is degrading with lower signals, such that it is difficult to analyze extremely thin clouds with non-typical properties like PSCs, in particular for fixed TH steps.

Of course, these statistical distribution eventually have to be validated thoroughly, but the distribution look roughly as expected.

Figure 11: Fraction of data points affected by clouds, for August 2004 (left) and August 2009 (right).

Figure 12: Occurrence rate for clouds with cloud TH at the measured TH, for August 2004 (left) and August 2009 (right).
Figure 13: Latitude–longitude distribution of the occurrence rate for clouds with cloud TH at the measured TH, for February 2004. Clockwise from upper left: 2 km, 5 km, 9 km, 12 km, 15 km and 18 km.

at the large scattering angles as they occur in Southern hemispheric winter. Nevertheless a good fraction of PSCs can be identified with the probability method, as shown in Fig. 15. Here one more the $C_{750}$ vs. $C_{1550}$ distribution is shown, for Northern and Southern PSC seasons (February and August 2011). Here, all data are plotted (the clouds below the FoV are excluded, though), meaning that the PSCs can also be above 18 km. However, there are higher density features which have a low cloud probability because their $C_{1550}$ resembles that of aerosol. Of course there are different types of PSCs, both in composition and particle size (7), such that it is possible that certain types of PSCs appear as aerosol. In this case, this analysis in fact opens up possibilities for a closer study of PSC properties. It may be recommendable, though, that for the identification of PSCs the old gradient approach in $C_{750}$ should be retained, after thresholds are confirmed or updated. This would also avoid issues with the large scatter of the $C_{1550}$ ratio as it can be seen in the top of the figures.
Figure 14: Latitude–longitude distribution of the occurrence rate for clouds with cloud TH at the measured TH, for August 2004. Clockwise from upper left: 2km, 5 km, 9 km, 12km, 15 km and 18 km.

Figure 15: Examples for PSCs identified with the probability method at about 18 km, for February 2011 between 60°N and 90°N (left) and for August 2011 between 60°S and 90°S (right). The axes and color code are analogous to Fig. 10.
7 Algorithm Description Summary

This section gives a summary of the proposed algorithm, in the order at which the relevant subprocesses should be implemented.

7.1 Preprocessing: Reflectances and Color Ratios, Apriori Weights

1. Generate and initialize the arrays introduced in section 5.1.

2. Determine the limb profiles of average reflectances \( R_i \) for the wavelength windows given in table 3, according to Eq. 10 and Eqs. 11 to 13 in section 5.2.

3. Determine the limb profiles of the errors of the average reflectances \( \delta R_{2,3} \) in windows 2 and 3 using Eqs. 12 and 14 in section 5.2.

4. Determine the limb dark signal (Eq. 15) and its error (Eq. 16) from the reflectances and their errors at the last TH step for each of the four wavelengths in table 3. Store values in the arrays \( D \) and \( \delta D \). See section 5.2.

5. Determine the profile of color ratios \( C_{750} \) (Eq. 17) and \( C_{1550} \) (Eq. 18) as described in section 5.2.

6. Determine the profile of error of \( C_{1550}, \delta C_{1550} \), (Eq. 19) as described in section 5.2.

7. For the purpose of determining the LUT bins and assigning apriori cloud weights, retrieve the geolocation parameters (\( \text{lat}_{10}, \text{lon}_{10} \)) from the 10th TH step for each limb profile. Also, determine the month \( m \) from the measurement start time of the state.

8. Determine the apriori value of the cloud weight \( w_{\text{Cloud}} \) using Eq. 26 in section 5.5.

9. Set a boolean variable \( kPSC \) to true if the statement in Eq. 43 in section 5.7 is true. This variable is valid for the entire limb profile. Then, set the apriori value of the PSC cloud flag \( f_{PSC} \) for each of the \( n_{TH}^{PSC} \) steps to one or 0 using Eq. 44 in section 5.7.

7.2 Gradients, LUT Bins, Probabilities and Statistical Cloud Weight

In the next loop, the color ratio gradients and their errors can be determined:
1. Determine the PSC gradients $\Delta C_{750}$ for each limb profile from Eq. 25 (section 5.4).

2. Determine the gradient of $C_{1550}$, $\Delta C_{1550}$, using Eq. 27 in section 5.5.

3. Determine the error of the gradient of $C_{1550}$, $\delta \Delta C_{1550}$, using Eqs. 28 to 31 in section 5.5.

4. Update the cloud weight as described in section 5.5 (Eqs. 33 to 35).

5. Determine the LUT bin as described in section 5.3, using Eqs. 20 to 24. To determine the LUT cell in $(i_{1550}, i_{750})$, take care of data quality and fudge to the closest filled values in $(C_{1550}, C_{750})$.

6. Determine the cloud probability $P_{\text{Cloud}}$ for the LUT$(i_{T}, i_{TH^*}, i_{L}, i_{S} = 2)$ and the cell $(i_{1550}, i_{750})$.

### 7.3 Low Cloud Check

After the second loop has been performed, candidates for the low cloud check have to be selected and checked.

1. In another loop, starting from the lowest TH step, determine whether this measurement is a candidate for the low cloud check using the criteria in Eq. 37 in section 5.6.
   
   - If this measurement is a candidate for the low cloud check, follow the instructions in section 5.6 to perform a loop starting from the TH step below this one going down in order find a potential low cloud that meets the conditions in Eq. 39.

2. If such a cloud has been found, set the low cloud veto $c_{\text{below}}(i_{TH})$ to the corresponding TH step.

3. Set the cloud flag $f_{\text{Cloud}}^{\theta}$ according to Eq. 52 in section 5.8 for all $n_{TH}$ steps.

### 7.4 PSC Flags and Parameters

See section 5.7. This needs only to be performed if $k_{PSC} = \text{true}$.

1. Initialize $TH_{PSC}$ to a large negative value and $\Delta C_{PSC}$ to 0.

2. Perform a loop, starting from the equivalent TH step $n_{TH}^{PSC} - 1$ going down and check for the conditions of Eq. 45 and update $TH_{PSC}$ and $\Delta C_{PSC}$ according to Eqs. 46 and 47. Set $f_{PSC} = 0$ if $\Delta C_{PSC} \leq \Delta C_{PSC}^{\text{Thresh}}$.

3. Stop the loop if $\Delta C_{PSC} > \Delta C_{PSC}^{\text{Thresh}}$ or $TH < 14 \text{ km}$. 
7.5 General Cloud Flags and Parameters

See section 5.8.

1. Initialize $TH_{Cut}^0$ to a large negative value and $P_{Cut}^0$ to 0.

2. Perform a loop, starting from the equivalent $TH$ step $i_{TH^*} = n_{TH} - 1$ going down and check for the conditions in Eq. 49. If they are met, update $TH_{Cut}^0$ and $P_{Cut}^0$ (Eqs. 50 and 51).

3. Stop the loop if $P_{Cut}^0 > P^{Thresh}$ or, equivalently, $f_{Cloud}^0 = 1$.

7.6 Tropospheric Cloud Flags and Parameters

See section 5.9.

1. Initialize $TH_{Cut}^1$ to a large negative value and $P_{Cut}^1$ to 0.

2. Perform a loop from the lowest $TH$ until $i_{TH^*} = n_{TH} - 1$ and determine $f_{Cloud}^1$ using Eq. 56.

3. Perform a loop, starting from the equivalent $TH$ step $i_{TH^*} = n_{TH} - 1$ going down and check for the conditions in Eq. 53. If they are met, update $TH_{Cut}^1$ and $P_{Cut}^1$ (Eqs. 54 and 55) and stop the loop.
7.7 Summary of Initialization Parameters

<table>
<thead>
<tr>
<th>name</th>
<th>n_{val}</th>
<th>value(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_{TH}</td>
<td>1</td>
<td>11</td>
<td>Number of equivalent TH steps for which normal cloud flags are determined</td>
</tr>
<tr>
<td>n_{TH}^{PSC}</td>
<td>1</td>
<td>12</td>
<td>Number of equivalent TH steps for which PSC flags are determined</td>
</tr>
<tr>
<td>lat_{PSC}</td>
<td>1</td>
<td>60</td>
<td>Absolute value of minimum latitude for PSC detection</td>
</tr>
<tr>
<td>m_{Start}^{PSC}</td>
<td>2</td>
<td>(6,11)</td>
<td>Start of the PSC season (SH, NH), i.e. (lat &lt; 0, lat &gt; 0)</td>
</tr>
<tr>
<td>m_{End}^{PSC}</td>
<td>2</td>
<td>(10,3)</td>
<td>End of the PSC season (SH, NH), i.e. (lat &lt; 0, lat &gt; 0)</td>
</tr>
<tr>
<td>TH_{min}^{PSC}</td>
<td>1</td>
<td>14</td>
<td>Minimum TH in km of PSCs, or maximum TH for tropospheric clouds during PSC season</td>
</tr>
<tr>
<td>R_{0,min}</td>
<td>1</td>
<td>0.001</td>
<td>Minimum value of ( \langle R_0 \rangle - \langle D_0 \rangle ) to activate normal cloud weight</td>
</tr>
<tr>
<td>R_{1,min}</td>
<td>1</td>
<td>0.001</td>
<td>Minimum value of ( \langle R_1 \rangle - \langle D_1 \rangle ) to activate normal cloud weight</td>
</tr>
<tr>
<td>R_{3,min}</td>
<td>1</td>
<td>0.003</td>
<td>Minimum value of ( \langle R_3 \rangle - \langle D_3 \rangle ) to activate normal cloud weight</td>
</tr>
<tr>
<td>R_{0,PSC,min}</td>
<td>1</td>
<td>0.006</td>
<td>Minimum value of ( \langle R_0 \rangle - \langle D_0 \rangle ) to activate PSC weight</td>
</tr>
<tr>
<td>R_{1,PSC,min}</td>
<td>1</td>
<td>0.002</td>
<td>Minimum value of ( \langle R_1 \rangle - \langle D_1 \rangle ) to activate PSC weight</td>
</tr>
<tr>
<td>w_{min}</td>
<td>1</td>
<td>0.9</td>
<td>Minimum cloud weight for the highest cloud TH</td>
</tr>
<tr>
<td>P_{Thresh}</td>
<td>1</td>
<td>0.5</td>
<td>Minimum cloud probability above which a data point can be flagged as cloudy</td>
</tr>
<tr>
<td>\Delta C_{Thresh}^{PSC}</td>
<td>2 × 6</td>
<td>see table 5</td>
<td>Thresholds values for the gradient ratio of ( C_{750} ) for PSC detection (SH, NH)×(TH)</td>
</tr>
</tbody>
</table>

Table 4: Global initialization parameters for cloud detection.
Table 5: Threshold values for PSCs for Southern (SH) and Northern (NH) hemispheres and seasons.

<table>
<thead>
<tr>
<th>$i_{TH^*}$</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{750}^n$ (SH)</td>
<td>1.35</td>
<td>1.25</td>
<td>1.25</td>
<td>1.2</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>$C_{750}^o$ (NH)</td>
<td>1.3</td>
<td>1.1</td>
<td>1.2</td>
<td>1.25</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>$p_{750}$ (SH)</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$p_{750}$ (NH)</td>
<td>0.0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>$C_{1550}^{PS}$ (SH)</td>
<td>1.3</td>
<td>1.3</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>$C_{1550}^{PS}$ (NH)</td>
<td>1.2</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
<td>1.3</td>
</tr>
</tbody>
</table>
### 7.8 Recommended Output

<table>
<thead>
<tr>
<th>name</th>
<th>n_val</th>
<th>type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{1550}$</td>
<td>$n_{Scans}$</td>
<td>double</td>
<td>limb profile of $C_{1550}$</td>
</tr>
<tr>
<td>$\delta C_{1550}$</td>
<td>$n_{Scans}$</td>
<td>double</td>
<td>limb profile of errors of $C_{1550}$</td>
</tr>
<tr>
<td>$C_{750}$</td>
<td>$n_{Scans}$</td>
<td>double</td>
<td>limb profile of $C_{750}$</td>
</tr>
<tr>
<td>$\Delta C_{1550}$</td>
<td>$n_{Scans}$</td>
<td>double</td>
<td>limb profile of gradients of $C_{1550}$</td>
</tr>
<tr>
<td>$\Delta C_{750}$</td>
<td>$n_{Scans}$</td>
<td>double</td>
<td>limb profile of gradient ratios of $C_{750}$</td>
</tr>
<tr>
<td>$f_{Cloud}$</td>
<td>$n_{TH}$</td>
<td>boolean</td>
<td>Profile of limb cloud flags</td>
</tr>
<tr>
<td>$w_{Cloud}$</td>
<td>$n_{TH}$</td>
<td>double</td>
<td>Profile of limb cloud weights</td>
</tr>
<tr>
<td>$P_{Cloud}$</td>
<td>$n_{TH}$</td>
<td>double</td>
<td>Profile of limb cloud probabilities</td>
</tr>
<tr>
<td>$f_{PSC}$</td>
<td>$n_{PSC}$</td>
<td>boolean</td>
<td>Profile of limb PSC flags</td>
</tr>
<tr>
<td>$TH_{0}^{Cut}$</td>
<td>1</td>
<td>double</td>
<td>highest TH at which any cloud detected above threshold probability</td>
</tr>
<tr>
<td>$TH_{1}^{Cut}$</td>
<td>1</td>
<td>double</td>
<td>highest TH at which a tropospheric detected above threshold probability</td>
</tr>
<tr>
<td>$TH_{PSC}$</td>
<td>1</td>
<td>double</td>
<td>highest TH at which a PSC detected above threshold</td>
</tr>
<tr>
<td>$P_{0}^{Cut}$</td>
<td>1</td>
<td>double</td>
<td>Probability at $TH_{0}^{Cut}$</td>
</tr>
<tr>
<td>$P_{1}^{Cut}$</td>
<td>1</td>
<td>double</td>
<td>Probability at $TH_{1}^{Cut}$</td>
</tr>
<tr>
<td>$\Delta C_{PSC}$</td>
<td>1</td>
<td>double</td>
<td>$C_{750}$ gradient ratio at $TH_{PSC}$</td>
</tr>
</tbody>
</table>

Table 6: Global initialization parameters for cloud detection.
References

