ATLAS

ATLID Algorithms and Level 2 System Aspects

Algorithm Theoretical Basis Document (ATBD) for A-FeatureMask

Code	**:**	EC-TN-KNMI-ATL-010
Issue	**:**	2.2
Date	**:**	26/05/2011
Authors	**:**	G-J van Zadelhoff D.P. Donovan S. Berthier

This page intentionally left blank

Issue	Revision	Date	Reason for Change	Author
1.0	Baseline	2008-12-01	Version delivered aat CASPER	GJvZ [KNMI]
1.1	Update	2010-06-01	Updated draft ATLAS	GJvZ, DD & SB [KNMI]
2.0	Update	2010-08-19	Final version MTR ATLAS	GJvZ, DD [KNMI]
2.1	Update	2011-05-10	Updated version for FM ATLAS (adapt to other ATBD formats)	GJvZ
2.2	Update	2011-05-26	Updated document based on comments from final review	GJvZ|

Table of Contents

1. Purpose and Scope 5
2. Applicable and Reference Documents 6
2.1. Applicable documents 6
2.2. Reference & Related documents 6
2.3. Scientific References 6
3. Scientific Background of the algorithm 7
3.1. Algorithm history 7
3.2. Algorithm introduction 7
3.3. Physical/mathematical Background 8
3.4. Summary of changes to the ATLAS algorithm 10
4. Justification for the selection of the algorithm 10
5. Mathematical algorithm Description 11
5.1. Input parameters 11
5.2. Configuration parameters 12
5.3. Output parameters 13
5.3.1. Operational output parameters 13
5.3.2. Extra evaluation parameters (set by the configuration parameters) 14
5.4. Algorithm flow chart 15
5.5. Algorithm Definition 16
5.5.1. Signal probability 16
5.5.2. Median Hybrid method 17
5.5.3. Hybrid Median technique for Rayleigh and Mie signals 18
5.5.4. Maximum entropy method concept 19
5.5.5. Calculation of the optimum image 20
5.5.6. Robustness of the maximum entropy retrieval 23
5.5.7. Separating noise and signals 24
5.5.8. Gaussian fit to the noise 25
5.5.9. Integrating convolution and Hybrid median results 27
6. Algorithm performance, sensitivity studies, limitations 27
6.1. ECSIM scenes 27
6.1.1. ECSIM standard scene. 27
6.1.2. Retrievals for different solar angles. 29
6.1.3. ATLID-Featuremask & CALIOP data 31
6.1.4. CALIOP night orbits 32
6.1.5. CALIOP day-time orbits 34
7. Validation status 36
Annex A: Technical implementation 37
External models 37

1. Purpose and Scope

This document describes the Lidar only Featuremask algorithm developed within the ATLAS project. This algorithm outputs the probability of particle return detection at native resolution (horizontal approx 0.2-km and vertical approx 100 m) The relationship between this algorithm and other algorithms developed within ATLAS is shown in Figure 1. This document presents theoretical background of the algorithm (Section 3) as well as describing practical implementation aspects such as inputs (Section 5.1), outputs (Section 5.3) and algorithm structure (Section 5.3.2). Examples applications of the algorithm are given in Section 6 and an overview of the status of the algorithm is given in Section 7.

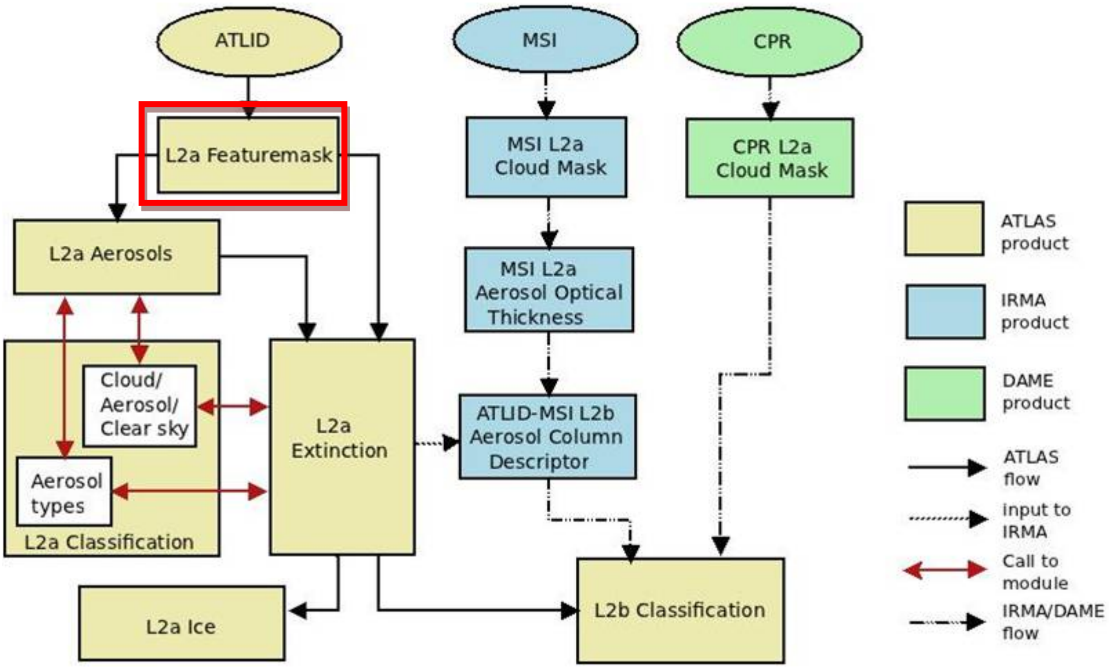

Figure 1: Schematic relationship of the algorithm described in this ATBD (red-box) with respect to other lidar-only (L1a) algorithms as well as relevant MSI and CPR synergetic (L2b) algorithms.

2. Applicable and Reference Documents

2.1. *Applicable documents*

Reference	*Code*	*Title*	*Issue*	*Date*
[MRD]	EC-RS-ESA-SY-012	EarthCARE System Requirements Document	5	Nov 2 2006

2.2. *Reference & Related documents*

Reference	*Code*	*Title*	*Issue*	*Date*
[CASPER-PARD]	CASPER-DMS-PARD-001	CASPER Products and Algorithms Requirement Document (PARD)	2.0	30 Oct 2008
[CASPER-FINAL]	CASPER-DMS-FR-01	CASPER Final Report	1.1	30/01/2009
[A-FM-PPD]	EC-TN-KNMI-ATL-011	ATLAS Featuremask PDD	0.1	
[ATL-PARD]	EC-TN-KNMI-ATL-005	ATLAS Products and Algorithms Requirements Document (PARD)	1.1	10-03-10
[EarthCARE]	EC-ICD-ESA-SYS-0314	EarthCARE product Table	1.3	15/06/2010

2.3. *Scientific References*

Keyword	**Reference**
[VPW05]	Mark A. Vaughan Kathleen A. Powell, David M. Winker, 2005, CALIOP Algorithm Theoretical Basis Document; Part 2: Feature Detection and Layer Properties Algorithms, PC-SCI-202 Part 2 [http://www-calipso.larc.nasa.gov/resources/pdfs/].
[V09]	Vaughan, Mark A., and Coauthors, 2009: Fully Automated Detection of Cloud and Aerosol Layers in the CALIPSO Lidar Measurements. J. Atmos. Oceanic Technol., 26, 2034-2050.
[FFTW]	Matteo Frigo and Steven G. Johnson, MIT, http://www.fftw.org/.
[HK05]	Hogan, R. J., and S. F. Kew, 2005: A 3D stochastic cloud model for investigating the radiative properties of inhomogeneous cirrus clouds. Q. J. R. Meteorol. Soc., 131, 2585-2608.

Keyword	Reference
[HJR07]	Hansen, P.C., Jensen, T.K., Rodriguez, G, 2007: An adaptive pruning algorithm for the discrete L-curve criterion, J. of Computational and Applied Mathematics Volume 198, Issue 2, Pages 483-492.
[HP93]	Christain Hansen and Dianne Prost O'Leary, 1993, The use of the L-curve in the regularization of discrete ill-posed problems, SIAM J. Sci. Comput., Vol 14, No6, pp 1487-1503.
[R07]	John C. Russ, 2007, The Image Processing Handbook, 5th Ed., Published by CRC Press, ISBN 0849372542, 9780849372544, chapter 4, correcting imaging defects. 214-224.
[RL08]	de Roode, S. R., and A. Los, 2008: The effect of temperature and humidity fluctuations on the liquid water path of non-precipitating closed cell stratocumulus clouds. Accepted for publication in the Quart. J. Roy. Meteor. Soc.
[SG85]	John Skilling and S.F. Gull, 1985, Algorithms and Applications; Eds. C. Ray Smith and W.T. Grandy, Jr. Maximum-Entropy and Bayesian Methods in Inverse Problems, 83-132, 1985 by D. Reidel Publishing Company .

3. Scientific Background of the algorithm

3.1. Algorithm history

The presented algorithm has built upon the featuremask algorithm developed within the CASPER project.

3.2. Algorithm introduction

The algorithm finds the feature mask based on the correlation of the data without focussing on a number of hard coded or input dependent thresholds. The main reason for this is the relatively large number of noise counts present in the ATLID signals. As the signal strength of aerosol or very optically thin ice clouds on the single shot grid can be comparable to the noise levels it was chosen to rely on image reconstruction techniques and not on signal to noise ratios and thresholds. The main reason why an image reconstruction technique can be so effective for the ATLID data is that in principle the Mie signals contain only particle backscatter and noise due to the Mie-Rayleigh cross-talk. No molecular backscatter, e.g. no variable background signal, should be present in these channels.

Two methods were employed in the CASPER algorithm to retrieve the feature mask: the median-hybrid method [R07] and the maximum entropy [SG85] method, both using the detection probability. Based on these two methods, coherent structures can be defined. The maximum entropy method however does not always converge and focuses on the stronger features in the noise and can miss some of the more tenuous widespread aerosol layers. To make sure that the algorithm is robust enough for the usage of space-based data, the algorithm has been simplified and now uses 3 or 4 convolved images instead of attempting to retrieve the rigorous maximum entropy defined image.

The features with a low signal to noise ratio within the images can be distinguished from the noise only signals by performing a Gaussian fit to the signal probability distribution. Both the high and low SNR routines implicitly use horizontal and vertical information from neighbouring pixels to define structures.

3.3. Physical/mathematical Background

The lidar used by EarthCARE (ATLID) is a high-spectral-resolution lidar. That is, the contribution to the return signal from the thermally broadened Rayleigh return and spectrally narrow elastic backscatter return (Mie) is separated by means of a spectral filter. Thus, in principle, the extinction profile at the lidar wavelength along with the corresponding backscatter profile may be independently derived. Before this can be performed a feature mask needs to be created as an input to the extinction retrieval algorithm. The Mie signals are separated by a Fabry-Perot etalon. Due to this configuration, there is a large noise component in both the Mie and Rayleigh channels, referred to as cross-talk. Specifically, cross-talk is Mie signal which ends up in the Rayleigh channel and vice versa. The cross-talk combined with the more standard noise sources (background, dark count, element efficiencies), hampers more general methods purely based on signal to noise ratios. It is still possible to apply such methods for the ATLID signals but will only enable the masking of very high signals and therefore very optically thick ice clouds and water clouds. An example of the expected ATLID Mie Co-Polar signals is presented (Figure 2) for a standard case consisting of an ice cloud a liquid water cloud and a background aerosol layer. Note that the individual noise values, e.g. between along track 0 and 3 km and above 2 km in height, has similar values as the aerosol field (between 0 and 3 km horizontally and below 2 km). The main difference is that the aerosol field shows a more coherent field compared to the high variability in the molecular regions.

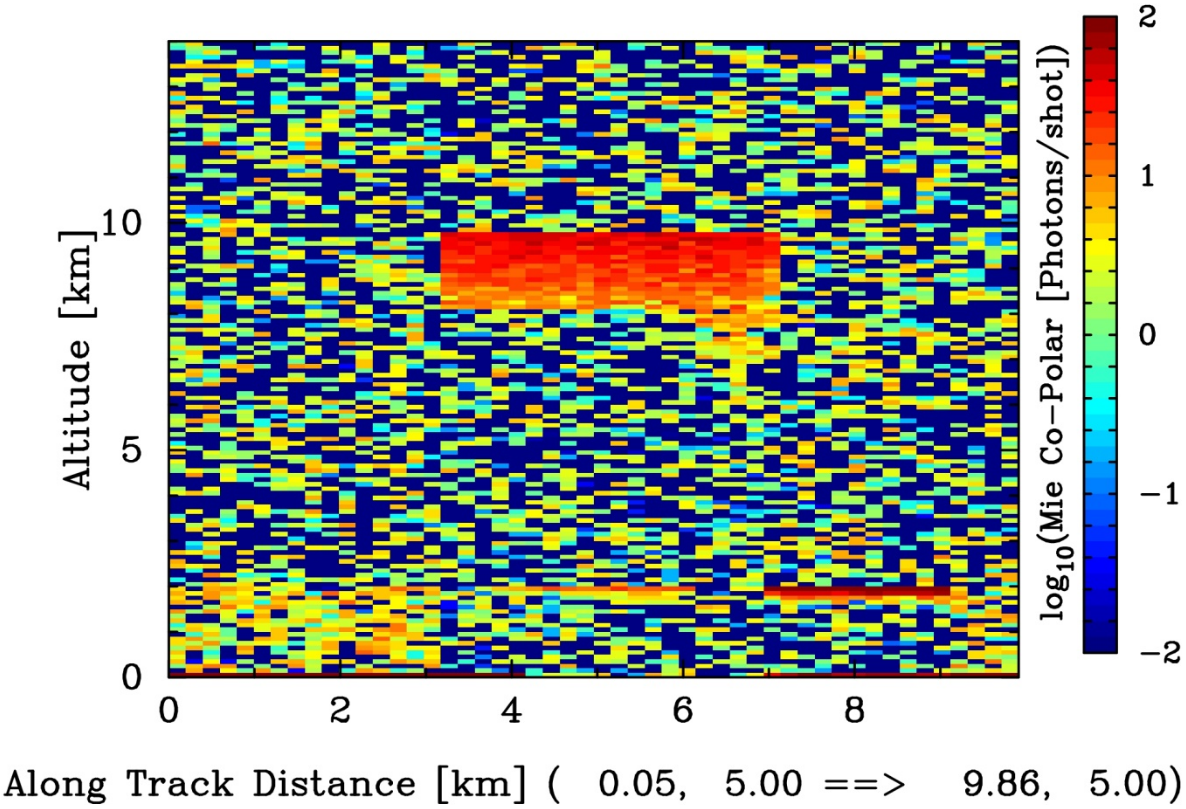

Figure 2: Mie Co-polar channel signal for a standard scene consisting of an ice cloud, a liquid water cloud and a background aerosol field.

In most cases it will not be possible to mask aerosol layers on an isolated shot-by-shot basis. If the cross-talk noise is a random process, e.g. if there is no preferred height or positional relation for the cross-talk, we can treat it as true noise. On the other hand the true signals in the Mie channel only arise from the backscatter from cloud or aerosol particles and not from air molecules, which all reside in the Rayleigh channel. Also, particle features in general are not single pixel events, but in a lot of cases extended in both the vertical and horizontal. The combination of these two features (pseudo random noise and pure extended particular signal) point to the use of image reconstruction techniques. These techniques can implicitly take into account information from surrounding pixels and correlations. The choice of the methods given in Section 5.5 is based on the specific benefits each of them have. The median-hybrid method is particularly good at finding coherent features while keeping edges constant (no smoothing effects beyond the features). The maximum entropy method and the simplified form in which it is finally convolves the data iteratively until a good balance is found between smoothing of noisy features while still remaining a good comparison to the original data.
Note that as mentioned, both methods assume that particle features are not single point events. Events of this nature would be missed by this algorithm, except when the signal probability is very close to 100%. Future datasets should help indicating if this would lead to a large set of missed events.

3.4. Summary of changes to the ATLAS algorithm

The following list are the main changes within the algorithm compared to the CASPER algorithm version.

- Convolution takes place in Fourier space (faster).
- New (faster & better) Gaussian fit routine.
- Maximum entropy algorithm was re-created to function automatically, but has been changed into a more robust version in which the user can assign different convolved images to be checked for features.
- Updated featuremask values assignment. Gives a more even spread between 0 and 10. CASPER version used [0,3,4,5,6,7,8,9,10].
- Added -2 as an extra assignment (surface or below surface).
- Reads in full CALIPSO/EarthCARE orbit. Automatically split orbit on horizontal blocks which can be run parallel.
- The featuremask is now compatible with the CloudSAT/CPR data (for L2a radar featuremask needed by L2b algorithm).
- Updated the noise fitting procedure to enable processing of the low SNR day time CALIPSO data.

4. Justification for the selection of the algorithm

This algorithm generates a feature mask based on the correlation present in the data itself without critically relying on a-priori input thresholds. At the single-shot scale the signal strength associated with aerosol or very optically thin ice clouds can be comparable to the noise levels, therefore it was chosen to rely on image reconstruction techniques and not on a-priori signal to noise ratios and thresholds. The main reason why an image reconstruction technique can be so effective for the ATLID data is that, in principle, the Mie signals contain only particle backscatter and noise due to the Mie-Rayleigh cross-talk. No molecular backscatter, e.g. no variable background signal, should be present in these channels

5. Mathematical algorithm Description

5.1. Input parameters

Variable	Description	Unit	Source	Dim	Type
Time	UTC time	S	ATLID-L1b	Time	Real*8
Height	Height of each radar/lidar gate above mean sea level	Km	ATLID-L1b	Time	Real
Tot_perp	Cross-Talk corrected background-subtracted Total Perpendicular return.	Photon Counts per Shot	ATLID-L1b	time, height	Real
Tot_perp_err	Standard deviation Cross-Talk corrected background-subtracted Total Perpendicular return.	Photon Counts per Shot	ATLID-L1b	time, height	Real
Mie_para	Cross-Talk corrected background-subtracted Mie parameter return.	Photon Counts per Shot	ATLID-L1b	time, height	Real
Mie_para_err	Standard deviation of the Cross-Talk corrected background-subtracted Mie parallel return.	Photon Counts per Shot	ATLID-L1b	time, height	Real
Ray_para	Cross-Talk corrected background-subtracted Rayleigh parallel return.	Photon Counts per Shot	ATLID-L1b	time, height	Real
Ray_para_err	Standard deviation of the Cross-Talk corrected background-subtracted Rayleigh parallel return.	Photon Counts per Shot	ATLID-L1b	time, height	Real

5.2. *Configuration parameters*

Variable	Symbol	Description	Unit	Dim	Type
always_feature	P_{ft}	Minimum probability for which the pixel is always assigned as feature (close to 1)"	None	1	Real
prob_min_val	P_{mh}	Minimum probability value accepted by hybrid median masking (between 0 and 100). Calipso (70)/ECSIM(30)	None	1	Real
med_hyb_size	N_{mh}	Number of pixels (AxA) used in the hybrid median routine (use 5,7,9, etc)	None	1	Integer
gauss_ratio	G_{rat}	Ratio for which signal has to exceed the gaussian fit to be qualified as feature	None	1	Real
Convols	N_{con}	Array with the image-numbers (times of iterative convolution) for which the SNR data will be checked. First value defines the convolution used for the main settings (20)	None	3 or 4	Integer
nx_size	N_x	Horizontal block size for which the featuremask will be calculated (4000)	None	1	Integer
dx_size	D_x	Overlap between the different blocks (100)	None	1	Integer

Testing	-	Add extra output, including all the histograms and Gaussian fits (0=no, 1=yes)"	None	1	Integer

5.3. Output parameters

5.3.1. Operational output parameters

Variable	Description	Unit	Destination	Dim	Type
Time	UTC time	S	A-AER, A-EBD	Time	Real*8
Latitude	Latitude of the ATLID footprints	Deg.	A-AER, A-EBD	Time	Real
Longitude	Longitude of the ATLID footprints a	Deg.	A-AER, A-EBD	Time	Real
Height	Height above mean sea level	M	A-AER, A-EBD	Height	Real
Surface_altitude	Height of surface above mean sea level	M	A-AER, A-EBD	Time	Real
Mask_Pa	-2 = below ground surface -1 = totally extinguished data 0 = most likely molecular, 1-5 = likely molecular but increasing chance of being a feature 6-9 = likely feature, decreasing chance of being molecular 10 = most likely feature detection	None	A-AER, A-EBD	Time, Height	Integer
Block_start_end	Boundaries of the data blocks for which the featuremask is derived	None	A-AER, A-EBD	nblocks	Integer

5.3.2. *Extra evaluation parameters (set by the configuration parameters)*

Variable	Description	Unit	Dim	Type
Signal_low_res	Signal probability of the data after subtraction of the features found through the median hybrid routine	None	Time, Height	Real
Histo_x	Signal probabilities array for which the histogram is calculated	None	Nblocks, Nhisto	Real
Histo_y	Normalized number count for the bins defined in histo_x, in each data block	None	Nblocks, Nhisto, Ncon	Real
Gauss	Gaussian fit which was used within the featuremask determination, in each data block	None	Nblocks, Nhisto, Ncon	Real

5.4. *Algorithm flow chart*

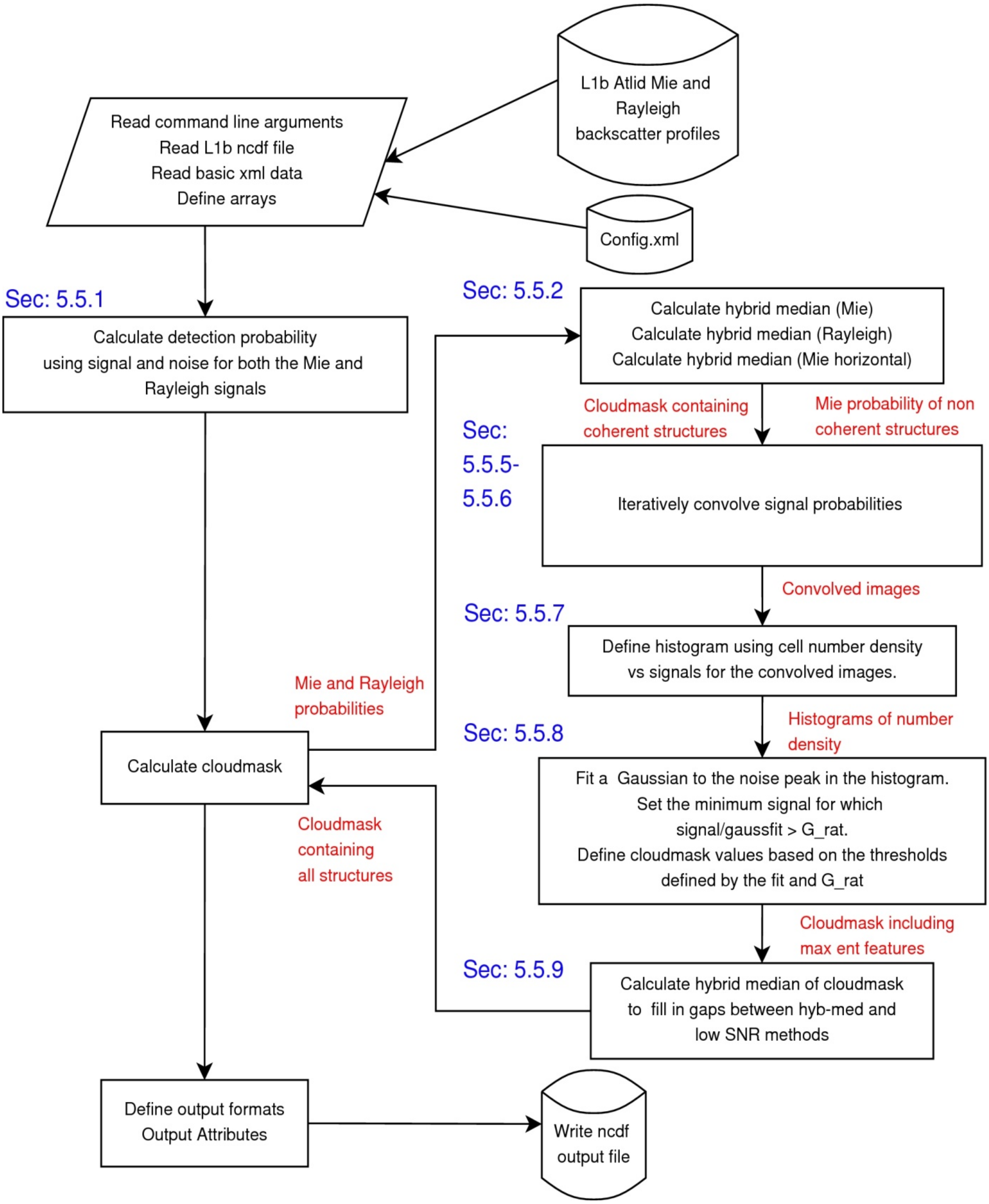
L1b Atlid Mie and Rayleigh backscatter profiles
Read command line arguments
Read L1b ncdf file
Read basic xml data
Define arrays
Config.xml
Sec: 5.5.1
Calculate detection probability using signal and noise for both the Mie and Rayleigh signals
Sec: 5.5.2
Calculate hybrid median (Mie)
Calculate hybrid median (Rayleigh)
Calculate hybrid median (Mie horizontal)
Cloudmask containing coherent structures
Mie probability of non coherent structures
Sec: 5.5.5-5.5.6
Iteratively convolve signal probabilities
Convolved images
Sec: 5.5.7
Define histogram using cell number density vs signals for the convolved images.
Mie and Rayleigh probabilities
Sec: 5.5.8
Histograms of number density
Calculate cloudmask
Fit a Gaussian to the noise peak in the histogram.
Set the minimum signal for which signal/gaussfit > G_rat.
Define cloudmask values based on the thresholds defined by the fit and G_rat
Cloudmask containing all structures
Cloudmask including max ent features
Sec: 5.5.9
Calculate hybrid median of cloudmask to fill in gaps between hyb-med and low SNR methods
Define output formats
Output Attributes
Write ncdf output file

Figure 3: Atlid Featuremask flow diagram. The red annotations along the arrows describe the relevant data flow, the blue text the Sections in which more information can be found

5.5. Algorithm Definition

The feature mask algorithm can be defined in a six steps procedure where each of the steps can be followed in the flow diagram in Section 5.4 (Figure 3). In this section the steps will be dealt each in turn trying to keep the logical flow. The basic idea behind the algorithm is to first extract the high signal to noise features. This is followed by a method to check for low signal to noise features by smoothing the image to an appropriate degree. The six steps can be summarized by:

1. Calculating the signal probabilities (Sec. 5.5.1).
2. Applying the hybrid median edge preserving technique to retrieve the coherent high signal to noise regions (Sec 5.5.2).
3. Iteratively convolve the remaining low SNR signals with a 2D Gaussian smoothing kernel (Sec 5.5.5 & 5.5.6).
4. Calculate the probability distribution histogram for the appropriate convolved images (Sec. 5.5.7).
5. Separate the noise from the signals by fitting a Gaussian noise peak within the histograms (Sec 5.5.8).
6. Apply the hybrid median technique to combine the results from the high SNR and low SNR results (Sec 5.5.9).

5.5.1. Signal probability

We follow the description and rationale used by the CALIOP team [VPW05 & V09] in assuming that Gaussian statistics are a reasonable approximation for the detected ATLID signals. If this is not the case after the ATLID configuration is finalized and the information is available, this step can be easily updated to the correct statistical approach. If we assume that the L1b data contains both the signal and noise levels of the three ATLID channels we take the noise levels to be the standard deviation of the signal. In that case the probability of detection can be written as:

$$P_d = 1 - \frac{1}{\sqrt{2\pi\sigma_s}} \int_{\sigma_s}^{\infty} e^{-s^2/2\sigma_s{}^2}\, ds, \qquad \textit{(1)}$$

where S is the signal, σ_s is the standard deviation of the signal and P_d the detection probability. This integral can be re-written using the error function (erfc):

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2}\, dt \qquad \textit{(2)}$$

Equation 1 in this case can be converted into the following description for the detection probability of signal S:

$$P_s = 1 - \frac{1}{2} erfc\left(\frac{S - \sigma_s}{\sqrt{2}\sigma_s}\right) \qquad (3)$$

The first step is to define the detection probabilities for both the Rayleigh and Mie channels. In case of the Rayleigh channel the available signal (S_{ray}) and standard deviation (σ_{ray}) are directly used from the L1b data. The results from this step are the two probabilities, P_{ray} and P_{mie}. For all $P_{mie} > P_{ft}$ (see table in Section 5.2) the feature mask is set to 10.

5.5.2. Median Hybrid method

The hybrid median filter checks the entire image pixel by pixel using an *n x n* box, where n is an odd integer of 5, 7 or 9. The centre pixel is calculated using the two diagonal and the horizontal and vertical rows within this box [R07; Figure 4]. For each of the rows the median value is calculated, after which the median value of these four median values is taken. As this latter median is from an even number we take the third value of the sorted value (not the mean of two values in the centre). The algorithm is very effective in removing single noise events and filling empty gaps. The median hybrid algorithm is run iteratively five times to ensure that the image has converged, e.g. there are no more changes in the image between this iteration and the next. As only median values are used, there are no smoothing edge effects, resulting in the need for only a few iterations in order to converge. This procedure is performed for both the Rayleigh signals and the Mie-co-polar signals

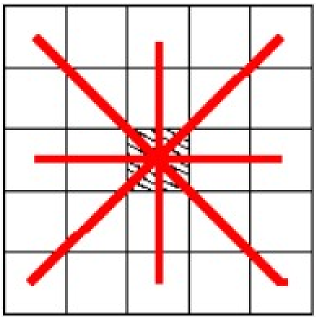

Figure 4: The hybrid median filter example uses a 5x5 box to determine the value of the center box (shaded). For each red line the median value is calculated. From the resulting 4 values, again the median is calculated. As this is the median of an even number, the value is determined by the third value of the sorted array.

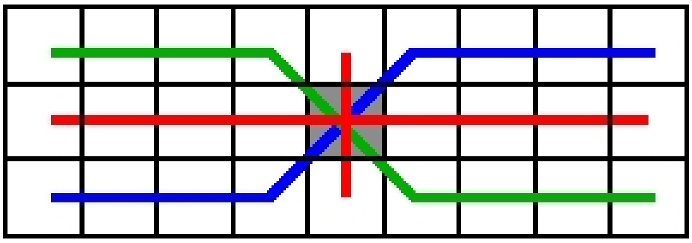

Figure 5: The hybrid median filter example uses a *9 x 3* box to determine the value of the center box (shaded). For each line (2x red, blue and green) the median value is calculated. From the resulting 4 values, the median is calculated. As this is the median of an even number, the value is determined by the third value of the sorted array.

Together with this filter a second filter is calculated for the Mie-co-polar signals, more oriented at finding horizontally oriented features. The procedure is similar to the one described above but uses an *n x 3* median hybrid filter. In Figure 5 the corresponding *9 x 3* filter is depicted.

From the three derived images, two different masks are created. First a Rayleigh mask which looks for any coherent features in the Rayleigh-image. Added to these are the lowest detections in the Mie-Image. The resulting Rayleigh mask helps identifying those regions for which the lidar signals are completely extinguished. It is needed for separating the molecular (0) and unknown (-1) in the final mask.

The task of the second mask is to separate the features and molecular regions based on the Mie signals using both the *n x n* and *n x 3* images.

5.5.3. Hybrid Median technique for Rayleigh and Mie signals

The first step in finding the feature mask is to let the P_{ray} and P_{mie} undergo a 2D filtering technique to reduce noise and delete individual pixels with a large difference compared to its direct surroundings, by both removing pixel values but also filling in missing pixel values. As we are interested in a feature mask it is of extreme importance that the algorithm correctly detects edges with no smoothing beyond the features or cutting corners. For this purpose the Hybrid Median filter algorithm is adopted [Sec 5.5.5]. The hybrid median size (n x n) is set through the user defined $\mathbf{N_{mh}}$ [Section 5.2]
Figure 6 depicts the iterative process of the median hybrid algorithm using the Mie signal probability for a test case consisting of an ice cloud, a water cloud and an aerosol layer. The image is constructed using data from the ECSIM forward and instrument modules after which its probability is calculated using Equation 3. The image shows the original data and the data after 1 and 5 iterations for n equal to 7. The noisy initial image loses most of the noise already in the first iteration with only part of the aerosol signal remaining. Note that the aerosol signal and noise signal have roughly the same detection probabilities.

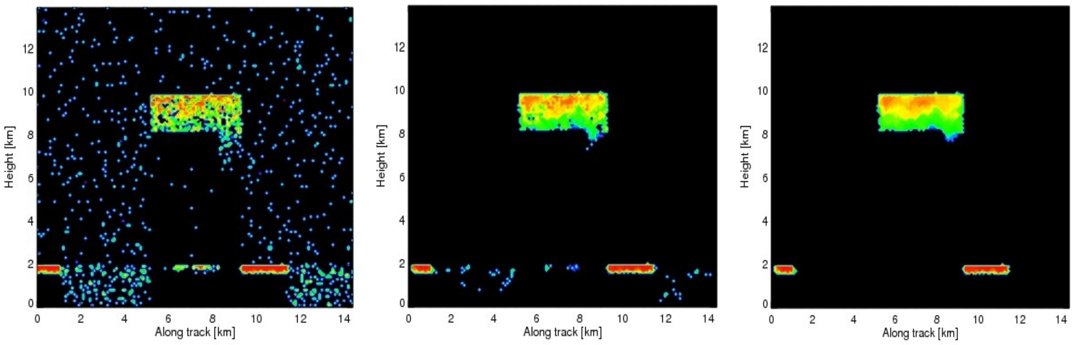

Figure 6: Three images showing the effects of iteratively adopting the hybrid median technique. The colors show the probability of detection (black =0%, red=100%). From left to right the original image, the image after 1 and 5 iterations is shown.

After 5 iterations only part of the ice cloud and the water clouds remain in the image with all gaps filled in. The image with 100 iterations, not shown here, is exactly the same as the one with only 5 iterations. The only coherent structures which will not be detected this way are structures with a vertical or horizontal width of 1 or 2 pixels.

Particularly horizontal stretched structures are at risk here as high optically thick water clouds (stratus or cumulus) may yield only 2 pixel thick clouds before the backscatter is completely reduced. To keep these important structures within the mask the hybrid median technique is used in a slightly altered version by using an *n x 3* box ensuring that also features of only two pixels thick, e.g. water clouds, are detected. The two masks are compared and only those additional features in the *n x 3* hybrid median results are added to the feature mask. The *n x n* version is considered to be the superior masking routine as it takes into account both vertical and horizontal coherence and is capable of filling in larger gaps. If only the *n x 3* version is used the features in Figure 6 are still detected but with a higher variability and 'noisy' behaviour within the features.

To ensure that only real features are found and not coherent noise the median-hybrid image is checked for a minimum value defined by the user defined threshold (**P_{mh}).** This threshold does not imply that parts of the data are not used. Any feature that is not recognised due to a too low signal probability will still be used in the maximum entropy retrieval.

The feature mask values at position (i,j) are given according to the following criteria:

$$Mask_{i,j} = Int(\frac{P(i,j)_{Med\_Hyb}}{0.2}) + 5 \tag{4}$$

This results in values between 5 and 9 which indicate that it is more likely to very likely a feature and not molecular, e.g. for pixels with a median-hybrid signal probability of 90% the equivalent mask value is 9. As the **P_{mh}** threshold is in general chosen around 60-70% the median hybrid value only results in high probable features based on the signals and the local coherence.

As the *n x n* mask might not detect horizontal features of 2 pixels high, the mask is checked for coherent signals in the *n x 3* image. For those pixels which show previously undetected features the mask is updated using the same criteria as given in Eq. 7. The previously retrieved Rayleigh-image is now used to set those regions which show no features in the Rayleigh and Mie signals by setting the Mask to -1. Finally the surface mask is added by using the L1b given surface height. The mask is set to -2 in those areas where the pixel height is equal or lower compared to the surface height.

5.5.4. Maximum entropy method concept

With the high SNR features retrieved using the hybrid median filter the remaining signals have to be checked for any features. In the original version of the featuremask algorithm this was performed by retrieving that image that best represents the truth using image reconstruction techniques. The concepts of dealing with the data in the current featuremask are based on this procedure and hence this explanation is added in the document, even though the current version does not calculate the maximum entropy anymore. In practice the maximum entropy method (MEM) selects that particular feasible image, from a large number of possible representations of the true image, which has the greatest entropy (E), taking into account the chi-square (χ^2) difference. This constrained maximum of E will be at an extreme of E-$\lambda\chi^2$ for a suitable Lagrange multiplier λ [SG85].

The entropy of a probability distribution is a measure of the information content and can be defined as:

$$E = -\sum [P_i \log(P_i)] \tag{5}$$

With P_i (>0) the probability in pixel i and Σ P_i=1. Next to the entropy there is also the misfit (χ^2- difference) for any retrieved image compared to the actual measured data. This 'misfit' is represented by a single number depending on the original input probability data (D_i) and the probability data retrieved after a mathematical procedure to the data (I_i):

$$\chi^2 = \sum \frac{(D_i - I_i)^2}{\sigma_i^2} \tag{6}$$

with the local noise given by the standard deviation σ_i calculated using the values within a 5x5 box around cell i. When less than 3 values are within this box, the σ_i is calculated using the mean of the available σ values within the image (σ_i=2$\underline{\sigma}$). The maximum entropy method looks for the best combination of the two statistical representations (E and χ^2) by taking a linear combination of the two and maximizing E-$\lambda\chi^2$, introducing the unknown regularization parameter λ. For each possible λ, an optimal image (j) can be retrieved which best represents both the original data and the maximum entropy criteria. The curve of optimum image number 'j', or χ^2(j)/E(j) vs. λ(j) follows a distinct L-shape (the L-curve), in which the solution 'j' changes in nature from being dominated by smoothing on one hand and being dominated by single error events on the other hand [HP93]. Hence, the corner of the L-curve corresponds to a good balance between too much and too little smoothing.

The optimum λ is estimated in the following way. An initial estimate is made by calculating the ratio of max(χ^2) and max(E), giving both the factors equal weight. With this value, the following array is set up:

$$\lambda = [\sum_{i=-8}^{8} 2^i] \frac{\max(E)}{\max(\chi^2)} \tag{7}$$

where a parameter space is set up ranging from 2^{-8} up to 2^8 around the initial estimate. For each λ the E-$\lambda\chi^2$ function is checked for the availability of a local maximum. The range of the lambda's is adjusted, so that the first and last lambda have no local maximum (entropy is always larger than χ^2 and vice versa). The remaining lambda's all have a local maximum. This ensures that the L-curve is optimally probed. Subsequently the optimum λ can be calculated by determining the corner in the L-curve, thereby determining the convolved image linked to this λ.

5.5.5. Calculation of the optimum image

With the main big features found in the Median Hybrid calculation the next step is to find structures within the noisy part of the image. For this we use the original Mie probability data (P_{Mie}) as was constructed in Section 5.5.1. All the probabilities where features were detected in section 5.5.3 are set to 0 (the resulting signals are from here on referred to as $P_{HM,Mie}$). As the most obvious features are already detected and only a noisy image remains a very simple version of the MEM (Section 5.5.4) is needed to check for more coherent features within the noise. Within the MEM method a number of different images are compared to the original input data ($P_{HM, Mie}$) looking for the optimum maximum entropy (following Eqs. 5-7). These different images are calculated by iteratively convolving the previous image with the following *5 x 3* convolution kernel, defined by 2-D Gaussian (Eq 9) with a maximum of 8 and standard deviations of $\sigma_{x=}$sqrt(4./log8) and $\sigma_{y=}$sqrt(1./log8).

$$K = \begin{pmatrix} 0.13 & 0.59 & 1.00 & 0.59 & 0.13 \\ 1.00 & 4.75 & 8.00 & 4.75 & 1.00 \\ 0.13 & 0.59 & 1.00 & 0.59 & 0.13 \end{pmatrix} \tag{8}$$

The kernel is normalized to ensure no signal loss when performing the convolution. For both the Kernel and the original image ($P_{HM,Mie}$) are converted to their FFT images using the fftw3 module [FFTW]. In Fourier space the matrices can be multiplied iteratively. For each image the entropy and χ^2 differences are calculated in real space, the original data resulting from the same test case as was shown in Figure 6 is used as an example.

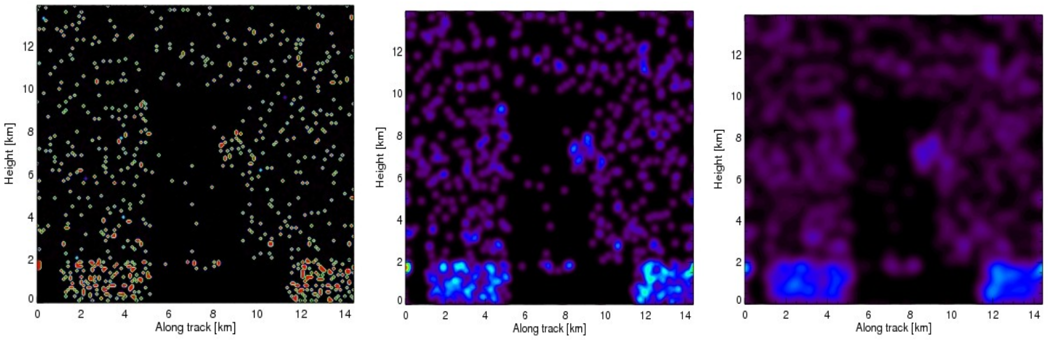

Figure 7: Three images showing the results of the convolution kernel. The colors show the probability of detection (black =0%, red=50%). From left to right the original image, the image after 5 convolutions and after 20 convolutions is shown.

The aerosol layers are nicely filled due to the convolution while the more uniform spread out noise is fading to very low detection probabilities. For each of the images the entropy and χ^2 can be calculated.

The L-curve is constructed by calculating the local maximum entropy depending on different λ's. In Figure 8 the determination of the maximum entropy is shown, based on CALIOP data, for 6 different combinations of the entropy and χ^2 by changing the λ value. For each curve the local maximum is calculated, each of which can be directly linked to a specific number of convolutions.

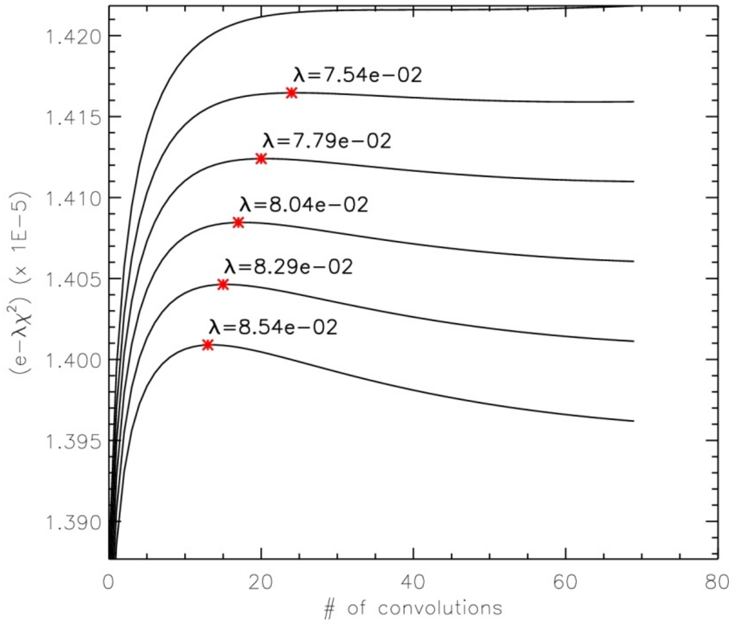

Figure 8: Curves of E-λχ^2 for six different λ values. The local maximum for each lambda is given by the red star. Each local maximum can be linked to a respective number of convolutions.

The local maximum links the number of convolutions to lambda and the entropy difference. In Figure 9 an L-curve is presented based on similar data. Shown is the logarithm of the entropy vs. the local maximum of the difference between the entropy and χ2. Since both the entropy as well as the max differences is increasing with respect to the number of convolutions the x-axis was inverted and scaled to have the same vertical extent of the y-axis. This transforms the graph to a standard form of the L-curve with a maximum visible curvature.

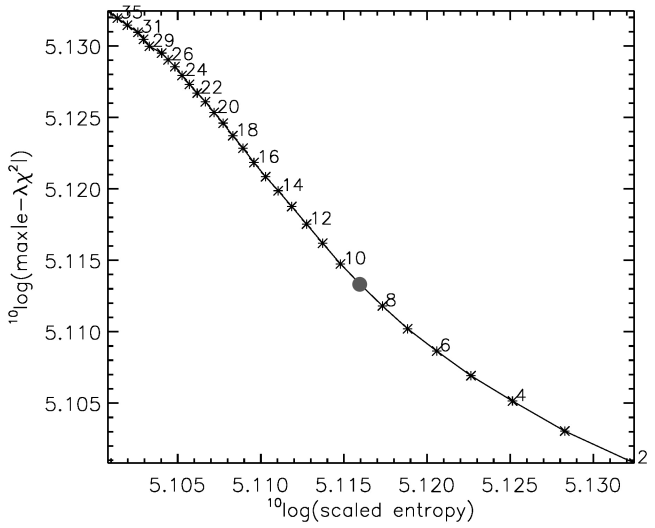

Figure 9: L-curve of the scaled entropy (see text) versus the local maximum of the entrop-chi2. Next to the points the number of convolutions are given, The adaptive pruning corner detection algorithm retrieved the optimum image to be calculated after 9 iterative convolutions.

The detection of the corner is performed using the adaptive pruning method (HJR07). When the optimum entropy probability image ($P_{MA, mie}$) is retrieved, using the L-curve criteria, showing the best comparison to the original data and maximum entropy it can be subsequently checked for its signal characteristics.

5.5.6. Robustness of the maximum entropy retrieval

The L-curve solution presented in the previous section shows that the implemented maximum entropy method can retrieve good convergence and there for feature masks from low SNR data. The question remaining is if this method is robust and if the retrieved image shows all potential features within the image. One of the potential issues with the L-curve as shown in Figure 9 is the shallow slope of the curve and the therefore hard to determine corner. This indicates that there is not a well defined optimum image (i) to be retrieved and that very similar results are reached when looking at image i-1 or i+2. In the testing phase using CALIOP data there were a number of regions where the L-curve showed a near linear line in the entropy vs. local maximum entropy parameter-space making it impossible to define the optimum image (e.g. Figure 10).

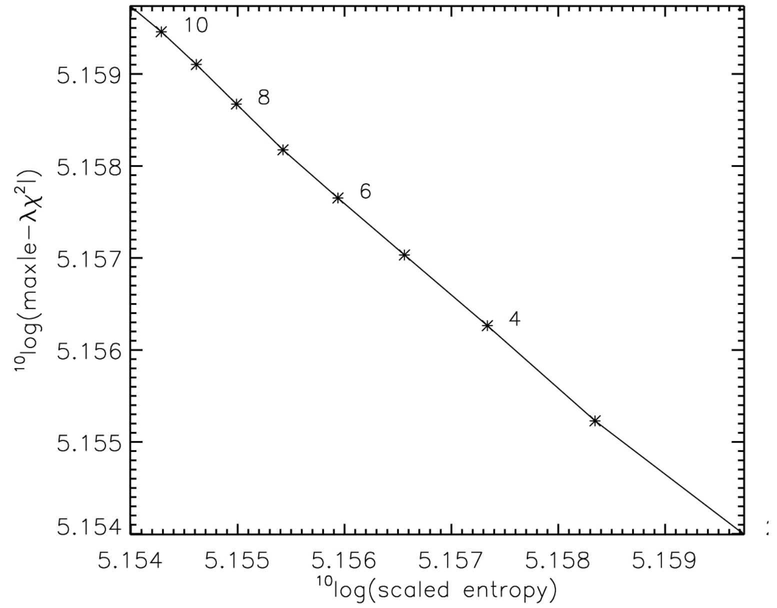

Figure 10: L-curve of the scaled entropy (see text) versus the local maximum of the e-□2 difference. Next to the points the number of convolutions is given. The adaptive pruning algorithm was not able to determine a corner in this case, nor is there any other way if selecting a best fit based on these data.

Based on six CALIOP 1064nm orbits, 168 blocks of data were examined resulting in a retrieval of the optimal image in between 6 and 22 iterations in most cases. In the case of the CALIPSO spacecraft velocity this is roughly equivalent to a binning of 2 and 7.3 km respectively. Many of the aerosols features detected in the VFM mask can take up to 80 km horizontal binning [VFM mask], which suggests that only looking up to 6 km would results in too few feature detections, especially of the very thin aerosol layers. After checking the features it was indeed concluded that the maximum entropy method was good in finding most but not all, especially faint, features. This can be solved by adding information from the 50th iteration (~16.7 km) and the 110th iteration (~36.6 km).

One final issue regarding the maximum entropy retrieval is that for each convoluted image a reverse Fast Fourier Transform is needed to calculate the entropy and χ^2 difference in real space. This is extremely CPU-time consuming when one regards the total amount of data involved per orbit.

Combining the argument that the 'corner' in the L-curve is very smooth and that it cannot be detected for all the regions, that additional lower resolution information is needed and that the procedure is very time consuming results in a more practical approach to this problem. For the results presented here the features are retrieved using the combination of 4 convolved images (10, 20, 50 and 110). This reduces the number of FFT transformations from 70 times to only 4 times. The lowest two number of convolutions are used to detect the features originally found by the maximum entropy routines while the latter two are needed for low SNR horizontally spread aerosol features.

5.5.7. Separating noise and signals

The smoothing of the image was performed to lower the intensity of noisy pixels more compared to pixels within a feature. As soon as the relative values of the features exceed the basic noise values an attempt to separate the signals from noise can be made. As the convolution kernel is Gaussian shaped and the uniform noise is assumed to be Gaussian in nature it is a feature which may be exploited. From the $P_{MA,Mie}$ image the detection probability histogram is calculated. This histogram depicts the number of pixels within a probability bin, with bin sizes of 0.005 ranging from 0.0 up to 0.8, normalized to the largest number count within a bin.

In Figure 11 the histogram after 16 iterations for the test case presented in Figure 6 and Figure 7 is presented. Up to a probability of 0.08 a noise peak is visible; the excess with respect to the Gaussian signal beyond 0.08 is due to coherent features. Note that due to the convolutions the probabilities are smaller than the original probabilities entering the maximum entropy algorithm. The 8% level does not imply that there was originally only 8% chance of being a correct detection.

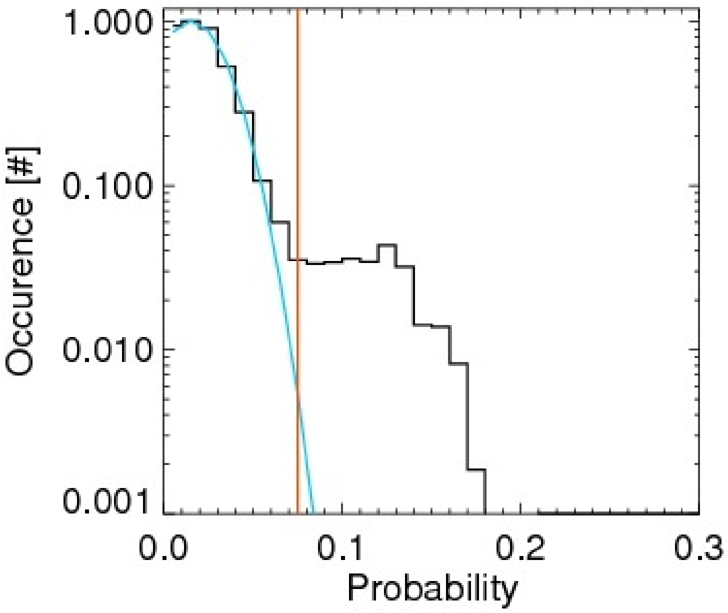

Figure 11: Histogram of the 16th convolved image (see Figure 5- for examples). The black line denotes the image histogram, normalized to its maximum value. The blue line is the Gaussian fit and the red line the user set threshold. This threshold is defined as the ratio of the Gaussian fit over the true data, for which in this case the value 4 was used

5.5.8. Gaussian fit to the noise

The histogram calculated in Sec 5.5.7 (Figure 11) shows two main components, a Gaussian peak which can be attributed to the noise to the left and an excess to the Gaussian for larger probabilities attributed to signals coming from clouds or aerosols. This noise and signals can be separated by fitting the Gaussian noise peak. Before the fit is calculated the position of the peak (i_{peak}) in the histogram is defined.

Performing a Gaussian fit to the data can cause a number of problems as the noise peak is not always symmetric around the maximum. Especially the left side can be steeper compared to the right side of the noise peak in case of extra background noise (daytime). In case of a non-symmetric peak there is an interest to have a very good fit to the right side and less interest in the slope on the left side. A second difficulty is that the signal probability values related to the features, the excess, can be very close to the maximum value of the noise peak. Therefore the Gaussian fit has to be performed close to the peak, limiting the number of usable points to fit the function and thereby increasing the error in the data.

The Gaussian fit is performed using an unconstrained linear least-squares calculation (lsq). However as a fit to a function with a large gradient, like a Gaussian:

$$n(i) = a_0 e^{(p(i)-a_1)^2/(2a_2^2)}, \tag{9}$$

gives rise to large uncertainties, where n(i) is the normalized number count and p(i) the probability bins, a trick has to be employed. To reduce the uncertainties a fit is performed to the logarithm of the function [Eq. 9]. This reduces the fit to a second order polynomial where the retrieved coefficients (A_0, A_1, A_2) can be directly related to the Gaussian coefficients (a_0, a_1, a_2):

$$\log(n(i)) = A_0 + A_1 p(i) + A_2 p(i)^2 \tag{10}$$

The fitting procedure is first performed for three fits adopting 20, 10 and 10 points, respectively fitting around i_{peak} , right side of i_{peak} but including this value and completely in the right wing of the noise peak. If all three result in the same fit, the result is clear. If there are noticeable differences, the right wing of the noise peak is sampled using 8 points starting at i_{peak}-4 and performing a fit starting 2 points further until the best fit is found.

The best fit is performed by comparing the total difference to the data on the 7 consecutive points starting at the bin i_{peak}+1, the retrieved width of the Gaussian and the maximum value at i_{peak}. The reason for using only these few points is related to the possible fast offset from the Gaussian due to signals from features.

From the Gaussian fit the noise is separated from the signals using the user set $\mathbf{G_{rat}}$ (Section 5.2). The width of the noise peak at which the histogram data exceeds the Gaussian fit by this ratio is called σ_{user}, assuming i_{peak} as the origin. In Figure 11 both the Gaussian fit and the $\mathbf{G_{rat}}$ value are over plotted. In this example the line is drawn for the first bin with an excess of more than 4 times the fit.

The feature mask is consequentially set using the standard deviation of the Gaussian fit [σ_{fit}=a_2] and the found probability at which the ratio is the user set value [σ_{user}]. As the features are derived within the noise it is chosen not to accept the feature mask to be extremely accurate on particle detections, e.g. its maximum value will be 9 (not 10). The spread of the feature mask settings is performed in two different ways.

The $\mathbf{N_{con}}$ array consists of three to four values which can be user defined. The first two values are related to values equivalent to what a maximum entropy algorithm would find, the last one or two are related to finding very faint wide spread features. The standard array used in this work is

$$N_{con} = [20,\ 10,\ 50,\ 120] \tag{11}$$

The first value in the $\mathbf{N_{con}}$ array will be sampled using a larger set of criteria compared to the other 3, and it was experienced that the convolutions between 15 and 20 were more suitable for this compared to the 6 to 15 number of convolutions.

The remaining three convolution values in the array are treated in a simpler way and are only intended to give additional information. The first convoluted image checked (in general around 20 or 25 times the convolutions) is treated in the following way:

- If the pixel was previously found by the hybrid median algorithm the value is not updated (change would be due to smoothing effects)
- If the detection mask indicates a -1 or -2, the feature mask remains -1 or -2 (changes would be due to smoothing effects)
- For signal bins $P_{20,\,mie}$ > a1+5*σ_{user} → Feature mask = 9

- For signal bins $a1+3*\sigma_{user} < P_{20,\,mie} < a1+5*\sigma_{user}$ → Feature mask = 8
- For signal bins $a1+2*\sigma_{user} < P_{20,\,mie} < a1+3*\sigma_{user}$ → Feature mask = 7
- For signal bins $a1+1*\sigma_{user} < P_{20,\,mie} < a1+2*\sigma_{user}$ → Feature mask = 5
- For signal bins $a_1+2*\sigma_{fit} < P_{20,\,mie} < a1+1*\sigma_{user}$ → Feature mask = 4,

with a1 the centre of the histogram Gaussian i_{peak},

The remaining three convolution numbers are only checked for high signal probability values:

- Where Featuremask < 7 and $P_{10,\,mie} > a1+3*\sigma_{user}$ → Feature mask = 7
- Where Featuremask < 6 and $P_{50,\,mie} > a1+2.5*\sigma_{user}$ → Feature mask = 6
- Where Featuremask < 6 and $P_{120,\,mie} > a1+2.5*\sigma_{user}$ → Feature mask = 6

5.5.9. Integrating convolution and Hybrid median results

At this point in the procedure as much 2D information has been obtained from the original observations as possible. There is only one task remaining, and that is to integrate the results from the Hybrid median and the maximum entropy algorithms. As all P_{mie} values going into the maximum entropy algorithm were set to 0 when a feature was detected there is the chance of miss classification or gaps where different features from the two algorithms are next to each other. This integration is performed by applying the hybrid median n x n routine iteratively on top of the retrieved feature mask (FM). The resulting mask (FM_{hm}) is compared to FM. For all cells which have a zero value in FM while it is a non-zero value in FM_{hm} results in an update of the feature mask to the FM $_{hm}$ value.

For all features which disappeared in FM_{hm} compared to FM the corresponding feature mask value is reduced by 1.

6. Algorithm performance, sensitivity studies, limitations

6.1. ECSIM scenes

6.1.1. ECSIM standard scene.

The first test case is presented in Figure 12. Intermediate results of this test case were presented Figure 6 to Figure 11. The scene itself consists of three particle regions. An ice cloud, with an optically thin and optically thick part, a water cloud and an aerosol layer up to 2 km. The results are presented in three plots: 1st the extinction slice through the scene as was created in the EarthCARE simulator; 2nd the detection probability between 0 (black) and 100 % (red) for the combined Mie channels and 3rd the feature mask as was derived from the Mie channels (black=-1, blues are from the max-entropy method, green→ orange from the hybrid median method). The ATLID calculations are performed using the ECSIM forward and instrument models and assuming an ocean surface layer. Both the lidar calculations and the retrieved featuremask are shown in Figure 13. This results shows that the expected results from ATLID will be noisy but with an algorithm, as described here, all features are identified.

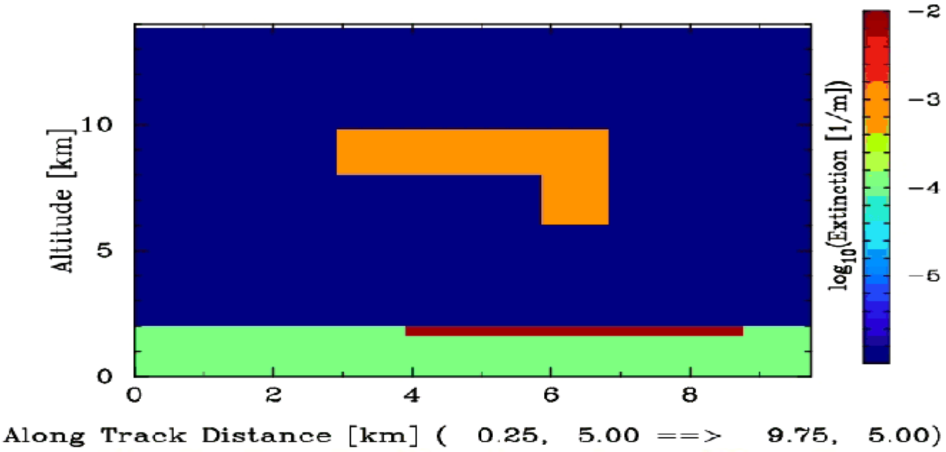

Figure 12: Slice of the extinction through the ECSIM simple scene as was used in Section 5. The scene consists of three different features. A blocky ice cloud which is thicker on the right side of the cloud, a liquid water layer below the ice cloud and a constant aerosol layer throughout the scene.

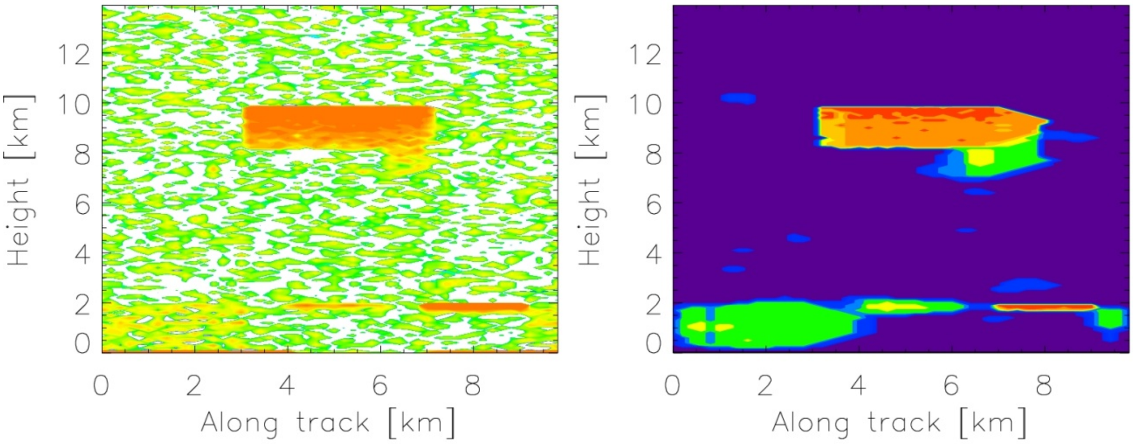

Figure 13: From left to right: Co-polar Mie channel L1b of the scene in Figure 6-1 over a snow surface with a solar elevation of 70 degrees and the retrieved featuremask. The dark blue represents values between 2 and 4, light blue: 5, green: 6, yellow 7 and the orange to red colors 8 to 10.

Note that the Rayleigh mask should have obtained the fully attenuated region as was described in Section 5.5.3. Due to the latest updates in the ECSIM lidar background signals the original calibration settings of this median hybrid mask, originating from the CASPER algorithm was not able to retrieve the fully attenuated regions. The new settings for this part of the mask will have to be studied in a future validation effort. This results in regions, e.g. below the thicker part of the ice cloud and below the liquid layers in the images in Figure 13 and Figure 14, to be purple instead of black.

6.1.2. Retrievals for different solar angles.

The standard scene presented in the previous section was calculated assuming an ocean surface and a solar elevation angle of 45 degrees. The diurnal cycle is very prominent in the CALIOP data and it is therefore of importance to check for the influence of the solar angle on the possibility of retrieving the features in the standard scene. In Figure 6-3 the standard scene is placed over a snow background and calculated at 0, 45, 70 and 90 degrees Solar elevation. The biggest differences in solar background are expected in the Rayleigh channel and not the Mie channels due to the narrow FP filter. Visually there are no major differences in both the Co-polar Mie channel and the retrieved Featuremask.

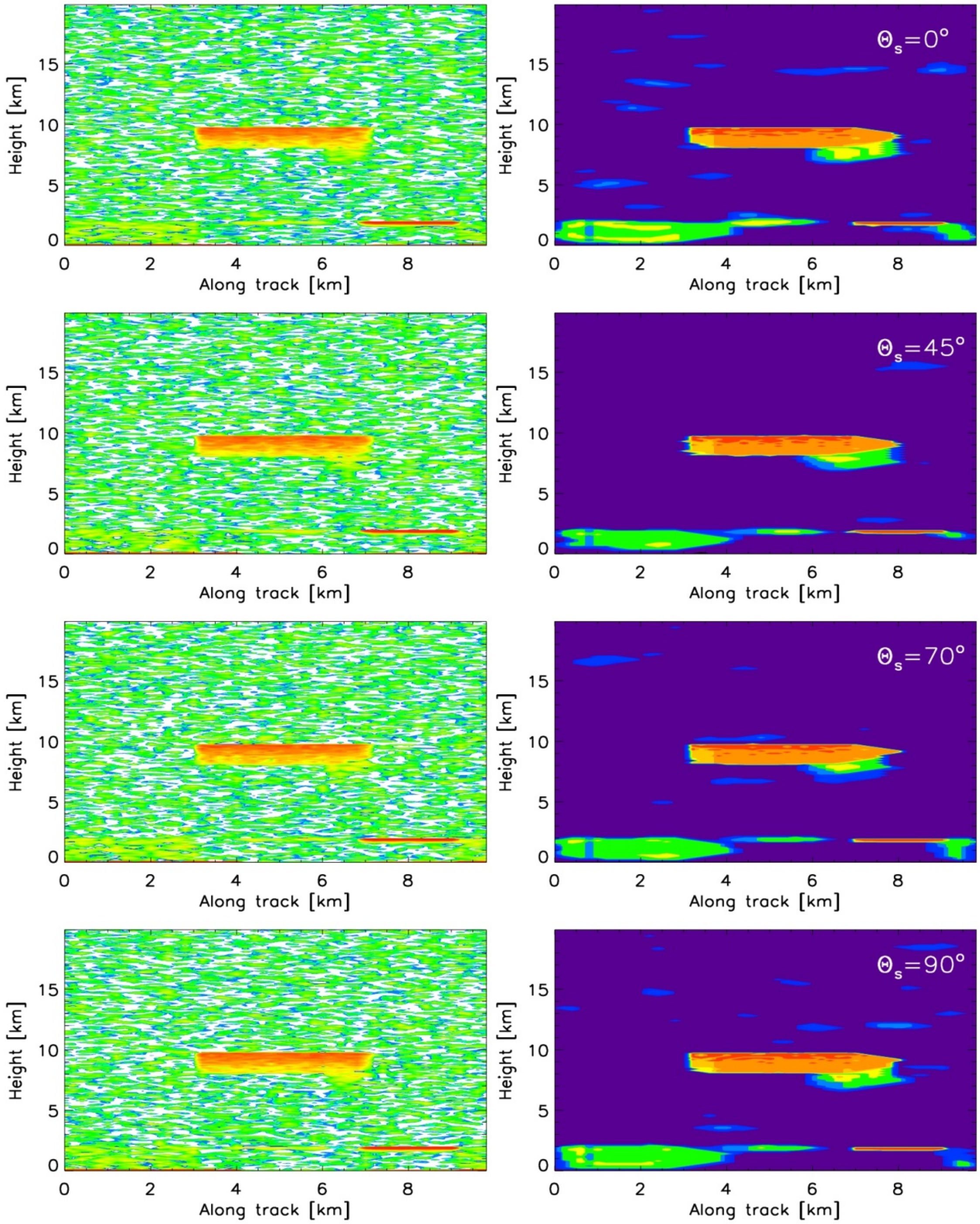

Figure 14: Comparison of the Co-polar Mie channel L1b and retrieved featuremask of the standard scene over a snow surface for four different solar elevation angles. No major differences are found for the different retrievals.

6.1.3. ATLID-Featuremask & CALIOP data

The Featuremask algorithm is intended for signals which show a combination of real backscatter signals from features with additional noise sources. It does not expect a background field due to molecular backscatter. This means that Co-polar 532nm channel of CALIOP is not suitable for calculating the featuremask. The 1064nm channel however hardly shows any molecular backscatter and is noisier in comparison to the 532nm channel. As the expected ATLID Co-polar Mie channel will have a relatively high noise level it is a very good data set for testing and validation of the featuremask.

The CALIPSO team has a similar data product, the VFM mask [V09]. This mask is based on the 532nm channel and is therefore more sensitive to the very small aerosol particles compared to results based on the 1064nm channel. There is a large conceptual difference between the ATLID-Featuremask and the VFM mask in the sense that the ATLID-Featuremask is used for processing the lidar signals before these are used and the VFM mask is a higher order algorithm. The VFM mask requires a high enough signal to noise ratio data to be able to assign cloud and aerosol types and requires a larger binning of the data. This requirement of high SNR will be reflected by the larger and blockier structures found in the VFM mask compared to the ATLID-Featuremask. As this is the only available lidar dataset from space which can be directly downloaded and is kept up to date it is the ideal dataset for evaluating the algorithm presented here.

The CALIOP data changes in vertical resolution at 8.2 km from 30 to 60 m. The change in resolution can potentially change the noise structure in the image. To exclude any influence of the change in resolution the featuremask is calculated in two separate runs. One focussing on the field below 8.2 km and one dealing with the data from 8.2 km up to 19 km. This is reflected by the horizontal line in the featuremask. It is also visible in the retrieved value of the features just above and below 8.2 km.

In all the figures the 1064nm raw data is shown in the top figure, the ATLID Featuremask in the center figure and the corresponding VFM mask in the bottom figure. The color scale of the Featuremask represents the chance of a pixel being molecular (0) or containing cloud or aerosol particles (10), while the VFM mask represents the classification of the pixel (1: clear sky, 2: clouds, 3:aerosols, 4:stratospheric features, 5:surface, 6: sub-surface and 7 fully attenuated). As the noise in the 1064nm channel hampers the detection of the attenuated regions the attenuated area's retrieved from the VFM mask have been added to the ATLID-Featuremask.

6.1.4. CALIOP night orbits

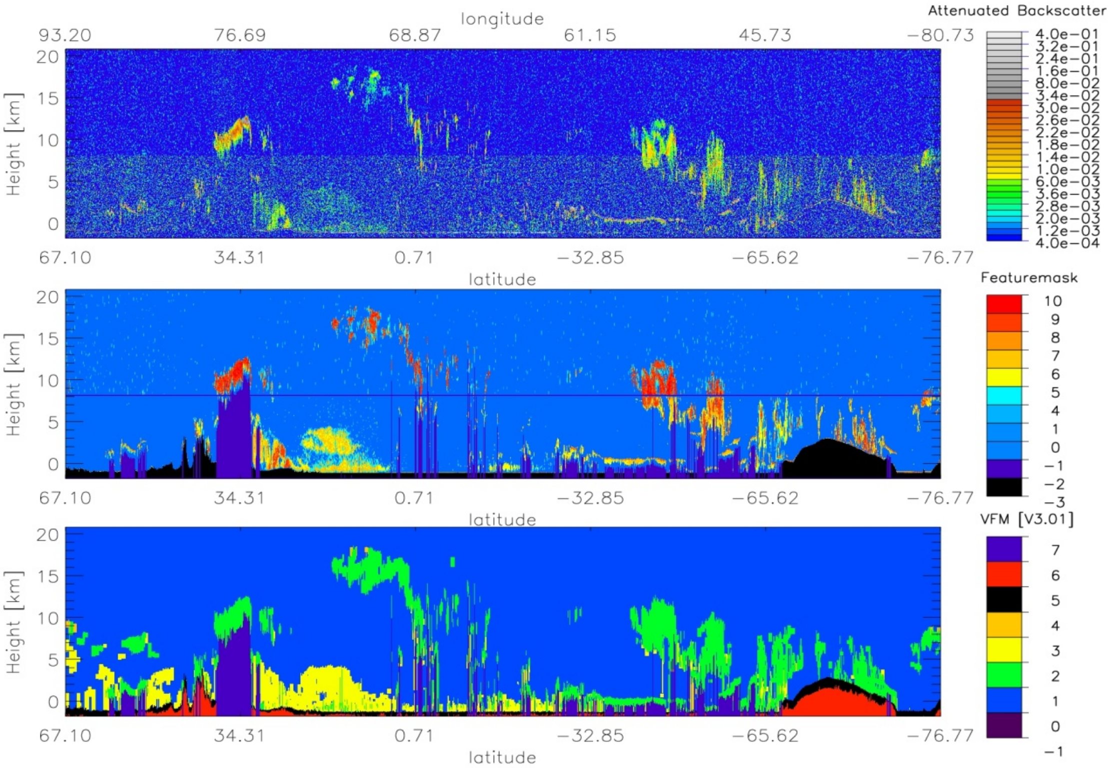

Figure 15: Full orbit of CALIOP night time data (CAL_LID_L1_ValStage-V3-01.2010-04-18T20-52-36ZN), the top Figure represents the raw 1064nm data, the center panel the featuremask and the bottom panel the VFM mask.

The Night time CALIOP data is relatively clean, even though one can see a clear difference between the data below and above 8.2 km. In Figure 15 to **Figure 17** the 1064nm Calipso data from 18 April 2010 is presented (CAL_LID_L1_ValStage-V3-01.2010-04-18T20-52-36ZN). In the first of these the entire orbit is presented, followed by two zooms into regions where a combination of aerosol, ice clouds and liquid clouds is presented. In **Figure 15** the full orbit is presented.

In general all the visible features in the raw data (top figure) are present in both masks. The VFM mask fills in a lot of gaps and has in general smoother and larger features compared to the ATLID-Featuremask. The Featuremask follows the raw data structures more closely however. In the left part of the image (between 67.10 and 34.31 longitude) the VFM mask finds many features which are not available in the Featuremask. Based on the data both algorithms have issues in this region and this should be looked at in the future.

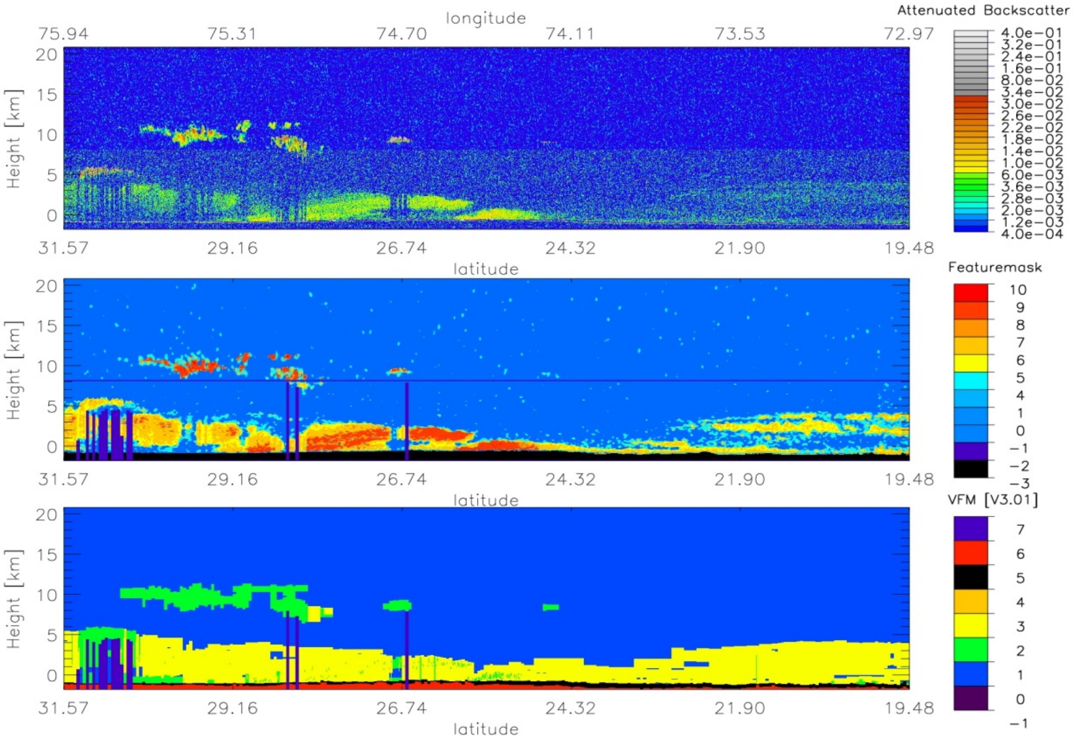

Figure 16: Zoom of an aerosol rich region from the orbit presented in Figure 6-4

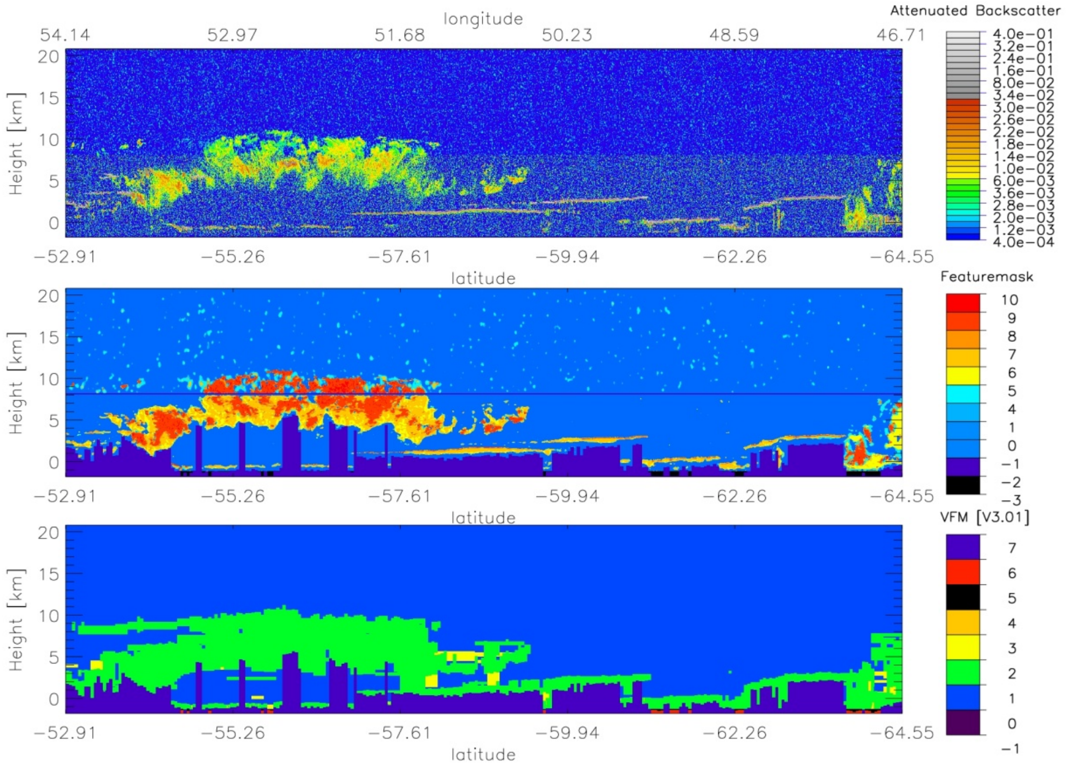

Figure 17: Zoom of a region from the orbit presented in Figure 6-4 where both ice clouds and liquid layers are present. There is an additional aerosol region in the bottom right.

In Figure 16 and Figure 17 two zooms are presented of this orbit to visualize both the agreement of the two algorithms to detect features in the data and show the differences between the two algorithms, which are due to their specific assumptions and needs. Figure 16 shows an aerosol rich region with a thin ice cloud and some scattered liquid cloud layers. The ATLID Featuremask follows the raw data very well and finds elevated aerosol layers at -21 degrees latitude. The VFM mask retrieves a continuous aerosol layer throughout this region. The difference is most likely caused by the VFM need of a high SNR and therefore a large horizontal binning. In Figure 17 a complex ice cloud structure is situated above a large number of liquid cloud layers. There is a potential aerosol layer on the right side of the figure. The median hybrid edge conserving method retrieves as much as possible the complex structure within the ice cloud. The retrieved liquid layers are thinner compared to the VFM results, also visible by the small separation seen often between the liquid layers and the fully attenuated regions taken from the VFM mask.

6.1.5. CALIOP day-time orbits

The day time CALIOP data is far noisier compared to the night-time data. This hampers the detection of low backscatter features like aerosol regions and thin ice clouds. In Figure 18 to Figure 20 the 1064nm CALIOP data from 30 April 2010 is presented (CAL_LID_L1_ValStage-V3-01.2010-04-30T15-29-14ZD).

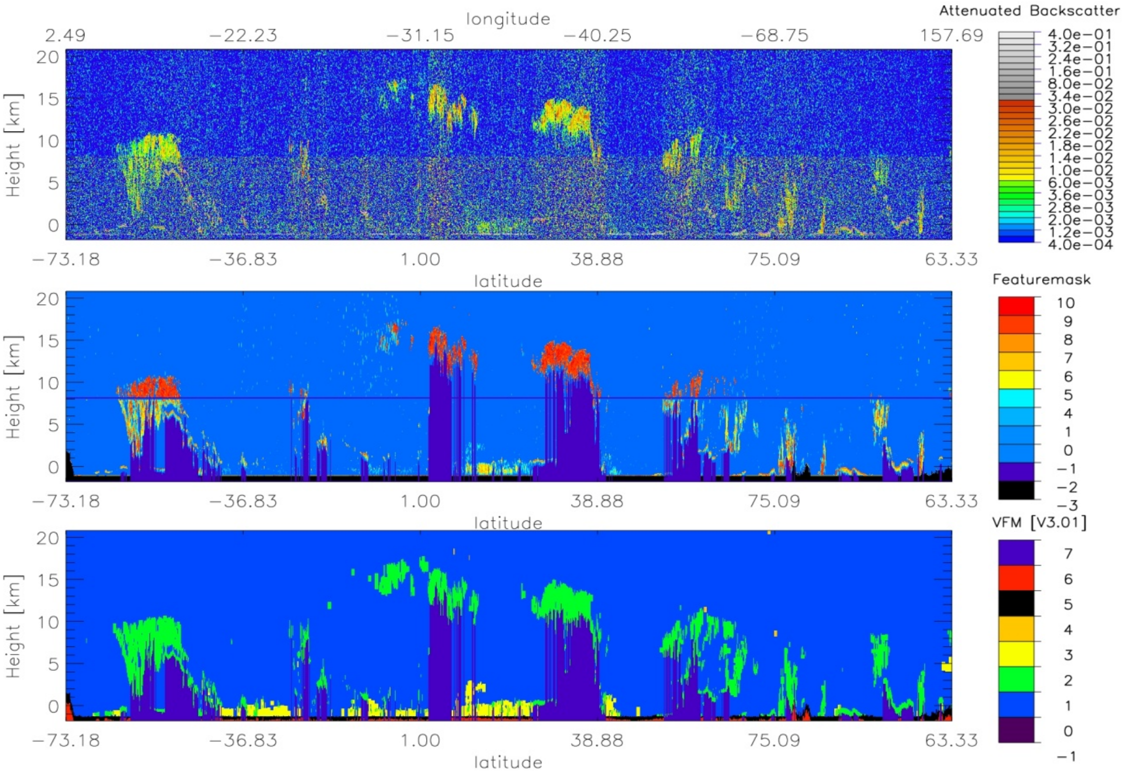

Figure 18: Full orbit of CALIOP day time data (CAL_LID_L1_ValStage-V3-01.2010-04-30T15-29-14ZD), the top Figure represents the raw 1064nm data, the center panel the featuremask and the bottom panel the VFM mask.

The most striking difference between the day time and night time the two masks is the lower amount of aerosol layers in the Featuremask. The current featuremask settings retrieve the missing regions with values in between 3 and 5. This suggests that the settings have to be optimized to detect these low SNR features in high background noise regions. In Figure 19 and Figure 20 two zoom regions are presented to look into the retrieval of ice cloud, liquid layers and aerosols in day time conditions.

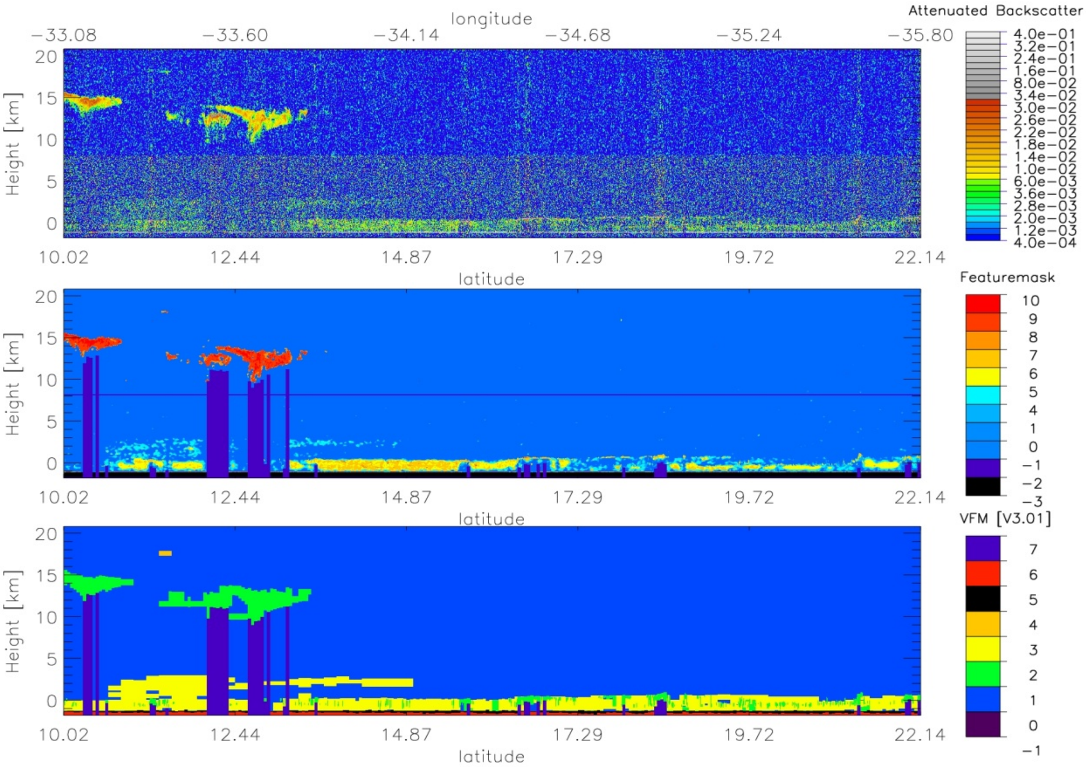

Figure 19: Zoom of a day-time region with ice clouds and thick aerosol regions.

In Figure 19 a zoom is shown with a few small (but optically thick) ice clouds. Close to the surface there is an extended aerosol layer with a few small liquid clouds. The featuremask retrieves the ice clouds and detects the edges very well. There is an issue with the retrieved aerosol layers. Especially the elevated layer between (10.5 and 14.87 degrees latitude) is only assigned with a five indicating that it is as well likely to be signals from aerosols as molecules. The same holds for parts of the extended surface layer. Since there are no hints of false positives in the entire region, the retrieval can still be improved by changing the settings within the retrieval algorithm. This will require additional calibration in case of day-time CALIOP data.

Finally in Figure 20 a combination of ice and liquid cloud layers is presented for the same orbit as is presented in Figure 18. The ice cloud complex shows a large variability, showing the characteristic shape of a frontal system and fall streaks. The VFM mask fills in most of the gaps within the system, even though the raw data shows clear gaps. There are bridges between the ice clouds and liquid layer at latitude >54^{o}S within the VFM mask which are not seen in the data. The Featuremask detects all ice clouds and most of the gaps. Also the liquid layers are fully separated from the ice clouds.

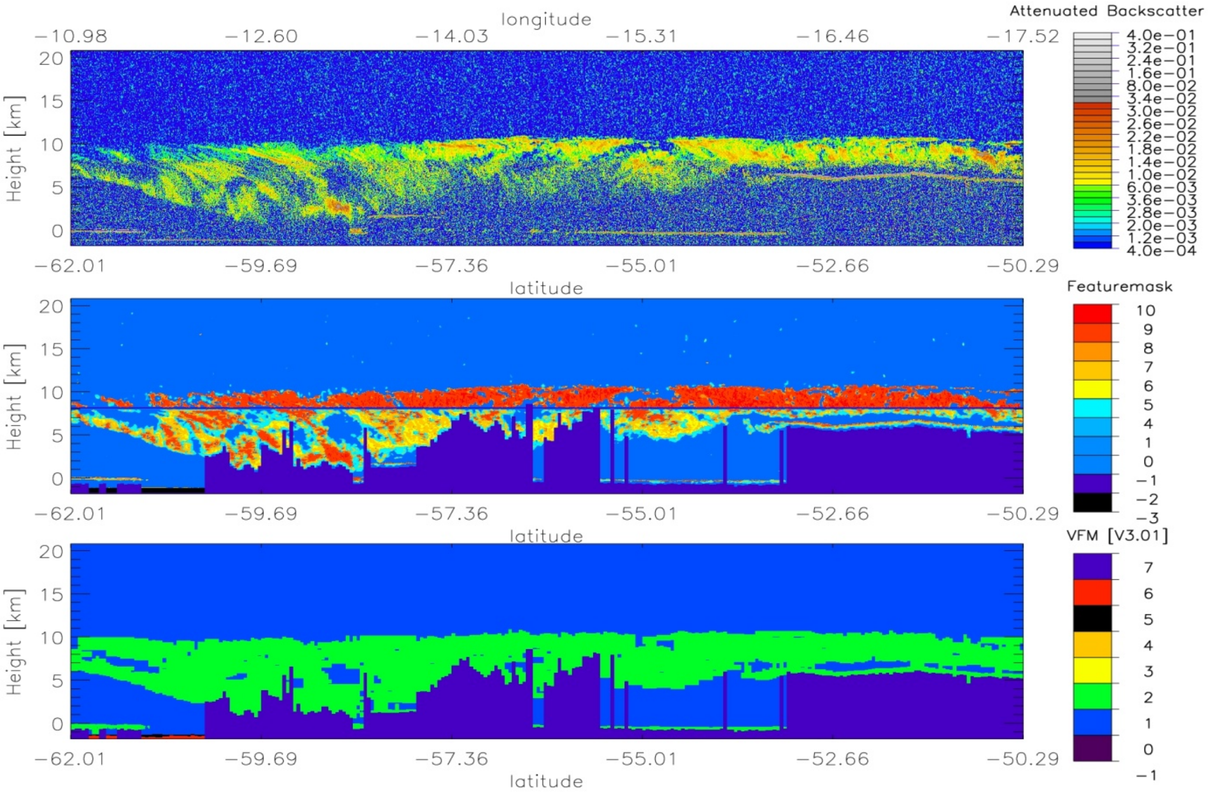

Figure 20: Combination of ice clouds and liquid layers in part of the orbit presented in Figure 18.

7. Validation status

The first version of the ATLID-Featuremask which was developed in the CASPER project solely focussed on results from the ECSIM lidar modules. At that time the background noise and cross-talk coefficients were underestimated making relatively easy to retrieve a good featuremask.

In the presented updated version of the Featuremask the ECSIM lidar module show more realistic signals and (background) noise behaviour. The algorithm has been extensively checked using 4 night-time and 3 day-time CALIPSO orbits. The CALIOP 1064nm data is currently the only available data-set from space for this type of validation. The night time orbits seem to represent the data very well. In case of the day time orbits the algorithm settings have to better optimised.

Future validation needs:

The most important validation still lacking for the current use of the algorithm is the determination of those regions which are fully attenuated. The current Rayleigh channel settings are based on an older version of the ECSIM lidar module and these are no longer valid for the current noise levels within the forward modelled data.

The algorithm has to be tested against more realistic ECSIM scenes, e.g. the ones created in the ICAROHS project. Secondly the optimisation for CALIOP data has to be performed after which a statistical comparison can be done using the raw data, the VFM mask and the Featuremask. Thirdly; campaign data from one of the future airborne instruments should be used for validation.
Finally all the settings of the algorithm have to be recalibrated in the commissioning phase of the EarthCARE satellite after launch. This will be a vital and delicate procedure for retrieving the best possible results from the ATLID instrument.

Annex A: Technical implementation

A full orbit calculation for Caliop data takes in the current mode and on a moderate (32bits) workstation 20 minutes. The code has not been optimized for speed and the total runtime can therefore be shortened, The biggest speed increase however can be very easily obtained by parallelizing the code. The configuration file has two parameters (**nx_size** and **dx_size**) which define blocks of data. The data in each of these data-blocks are independent to the other data blocks and these can therefore be separately retrieved on different nodes. This can lower the run-time to 1 minute in case of 20 to 21 computer-nodes.

External models

- The algorithm uses a module by Alan Miller taken from his website with his personal permission to use and distribute the routines (http://users.bigpond.net.au/amiller/)
  The algorithm used is: lsq.f90: Module for unconstrained linear least-squares calculations, used for the calculation of the Gaussian fit

- A second external module general_math_codes is distributed under the terms of the GNU General Public License. Within the module the error function and median of an array are calculated.

- A third external module is the fftw3 module [http://www.fftw.org/] used for the FFT transforms in the iterative convolution part.