SIBYL: a selective iterated boundary location algorithm for finding cloud and aerosol layers in CALIPSO lidar data

Mark A. Vaughan(a)*, David M. Winker(b), and Chris A. Hostetler(b)
(a) Science Applications International Corporation; (b) NASA Langley Research Center

ABSTRACT

This paper introduces SIBYL, a selective iterated boundary location algorithm designed to detect the base and top altitudes of clouds and aerosol layers in the backscatter lidar data that will be collected during NASA's upcoming CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations) mission[1]. SIBYL makes multiple passes through the data, constructing profiles at increasingly coarse spatial resolutions and repeatedly applying an adaptive threshold scheme to discriminate between clear air and "features" (i.e., clouds, aerosol layers, and/or surface returns). The sections below provide a brief overview of SIBYL's profile scanning engine and present specifics of SIBYL's iterated averaging scheme. Results are illustrated using data generated by the CALIPSO lidar simulator[2].

1. INTRODUCTION

During a planned three-year mission, the suite of active and passive sensors aboard the CALIPSO satellite will provide a unique data set that will address many of the current uncertainties about the role of clouds and aerosols in the Earth's climate. The primary instrument in the CALIPSO payload is a polarization-sensitive backscatter lidar that will operate continuously at both 532 and 1064 nm. Analysis of the lidar backscatter data will yield what is perhaps the most fundamental of all data products delivered by CALIPSO: the precise and accurate description of the vertical distribution of cloud and aerosol layers throughout the Earth's atmosphere. Given CALIPSO's large data volume (in excess of 2 gigabytes daily), timely analysis of the lidar data requires a fully automated scheme for detecting layer boundaries. Strong features such as stratocumulus clouds can be easily identified using simple techniques, as their backscatter intensity contrasts sharply with the much weaker scattering from the ambient molecular atmosphere. However, at the other end of the intensity spectrum aerosols and subvisible cirrus can be vexingly difficult to identify at high spatial resolutions, simply because the magnitude of scattering from these features is often indistinguishable from the local molecular background and its associated noise. To locate the boundaries of these weaker features requires enhancing the contrast between the scattering from the feature itself and the scattering in contiguous regions due solely to the Rayleigh atmosphere. Traditional feature-finding strategies accomplish this by averaging the profile data both horizontally and vertically prior to searching for weaker features. However, any such averaging scheme must be applied judiciously, else optically and/or meteorologically dissimilar features could be irretrievably commingled. Even within a single feature, preserving the natural spatial variability of the layer requires that averaging be limited to the minimum amount necessary. The SIBYL algorithm is designed specifically to address these averaging concerns. To properly identify all of the features within single segment of data, SIBYL scans a series of backscatter profiles constructed with successively greater horizontal averaging, corresponding to successively coarser spatial resolutions. Using a technique known as *feature clearing*, clouds and aerosols found at higher spatial resolutions (less averaging) are removed from consideration in successive scans. Doing so allows subsequent additional averaging to increase the visibility (i.e., contrast) of weaker features while simultaneously limiting the risk of spatial smearing.

2. SIBYL'S PROFILE SCANNING ENGINE

At the heart of SIBYL's retrieval scheme is a robust profile-scanning engine originally developed for application to ground-based observations[3] and later adapted for space-based analyses using the LITE data[4]. The core functionality of this profile scanner descends directly from the threshold algorithms long used to locate cloud boundaries in radar data[5]. In concept these algorithms are quite straightforward: a threshold level is established – either defined arbitrarily or

---

* m.a.vaughan@larc.nasa.gov; phone 1-757-864-5331; fax 1-757-864-7775; Mail Stop 435, NASA Langley Research Center, Hampton VA, USA 23681

chosen according to some heuristic – and the profile is scanned beginning in some segment that is known to be clear. Data points are examined sequentially, and features are identified as those regions where the profile data exceed the threshold values over any altitude range greater than some predetermined minimum feature thickness. Because the 532 nm channel is sensitive to a wider range of atmospheric targets than the 1064 nm channel, CALIPSO's feature-finding scheme is applied to the 532 nm total backscatter data (i.e., the sum of the contributions from both the parallel and perpendicular polarization detectors). The cost of this additional sensitivity is paid by the increased algorithmic complexity required for the practical implementation of an effective threshold scheme at 532 nm. In particular, due to the greater detection sensitivity and the nonlinear nature of the noise imposed on the backscatter signal, a "constant offset" threshold is generally inadequate for examining data over an extended altitude range. Consequently the CALIPSO threshold array is constructed as a function of altitude, with threshold values chosen to match the expected variation of the clear air (Rayleigh) backscatter signal as a function of altitude. Contributions to this variation fall into two distinct categories. The first category consists of noise that remains (on average) constant with respect to range from the lidar. Included in this segment of the noise budget are such things as detector dark current and the noise due to solar background contamination. The second category contains those sources for which the magnitude of the variation is function of range from the receiver. Within SIBYL this range-dependent component is assumed to be restricted to Poisson-distributed noise introduced by the photo-detectors. The threshold arrays used in SIBYL's profile scanner are built using a weighted combination of these two components. The "constant noise" contributions are measured on-board the satellite on a shot-by-shot basis and included in the downlinked data stream. The range dependent components are derived from site-specific molecular backscatter models computed using meteorological data. The combined result is a threshold scaled according to the expected noise envelope within any given backscatter profile.

The adaptive nature of this threshold construction scheme is demonstrated in Figure 1. Here thresholds constructed using a fixed set of weighting constants are applied to simulated backscatter data averaged horizontally over 5, 20, and 80 kilometers (Figure 1a, Figure 1b, and Figure 1c respectively). In each case the threshold line is seen to be riding just at the upper edge of the clear air noise envelope along the entire length of each profile. The 80-km average reveals a faint aerosol layer between 0 and 2-km that extends noticeably above the threshold line. Note too the discontinuities that appear in the threshold lines at 8-km and 20-km vertically. These adaptations account for the apparent step-function variations in signal-to-noise ratio (SNR) due to CALIPSO's non-uniform on-board data averaging scheme[1].

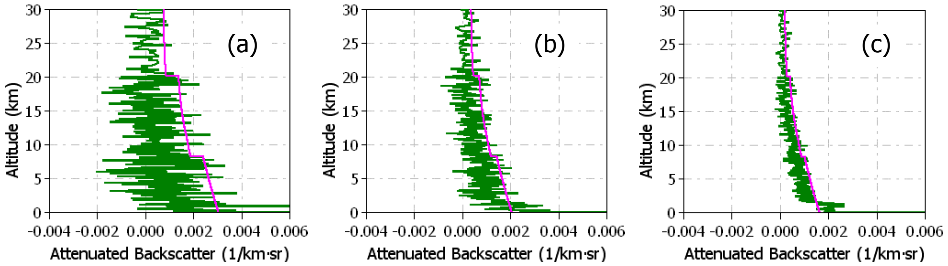

Figure 1: Threshold determination for (a) 5-km, (b) 20-km, and (c) 80-km horizontal averaging distances using simulated daytime backscatter data (surface albedo = 0.2)

3. THE ITERATED SEARCH FOR FEATURES AT MULTIPLE SPATIAL RESOLUTIONS

To identify spatial boundaries over the full range of backscatter intensities SIBYL relies primarily on a technique known as feature clearing. Feature clearing locates and *removes* all valid features within a single backscatter profile. Feature boundaries are located using the profile scanner described above. The actual feature clearing sequence is initiated at the top of the highest identified feature and progresses downward. As illustrated in Figure 2, feature removal is accomplished by (1) measuring the signal attenuation across each feature to estimate its two-way transmittance, (2) replacing the data within the feature boundaries with (a model of) "clear air", and finally (3) correcting all data beneath the excised feature for the attenuation ascribed to that feature. Ideally what remains following this procedure is a profile that represents the backscatter that *would have been measured had the feature(s) not been present*. The goal of the feature clearing process is to produce a series of contiguous profiles that can then be averaged further to increase the contrast between the ambient molecular scattering environment and any remaining weaker features (see Figure 2d). The feature clearing procedure can be repeated as necessary, using the more highly averaged profiles as new inputs to the scanning engine. The attenuation corrections inherent in the feature clearing technique ensure that profiles generated in subsequent averages are properly weighted throughout with respect to the varying optical depths of the (relatively) strong features removed in prior processing.

In principle the feature clearing technique can be applied over an arbitrarily large sequence of averaging intervals. In practice however the current SIBYL prototype iterates over a fixed sequence of three horizontal averaging intervals: a fundamental averaging distance of 5-km, an intermediate distance of 20-km, and a maximum horizontal averaging distance of 80-km. The 5-km interval was chosen in response to on-board data averaging considerations – the largest horizontal averaging distance in the data down-linked from the satellite is 5 km. Successive increments were selected based on the expected improvement in signal-to-noise ratio. Increasing the horizontal averaging distance by a factor of four each time will yield (on average) a factor of two increase in the profile SNR.

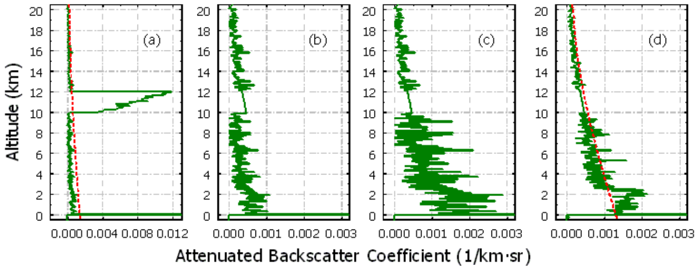

Figure 2: Feature Clearing. Figure 2(a) shows a 5-km averaged profile with a moderate cirrus layer overlying a weaker absorbing aerosol layer. Figure 2 (b) and Figure 2 (c) show, respectively, the removal of the cirrus layer and the correction of the underlying data for the estimated cirrus attenuation. Averaging four consecutive feature cleared 5-km profiles results in the 20-km horizontal average depicted in Figure 2 (d). This 20-km average reveals the previously obscured aerosol layer. The dotted lines in (a) and (d) are the threshold lines used by SIBYL's profile-scanning algorithm.

4. SELECTIVE FEATURE FINDING AND FALSE POSITIVE REJECTION

Each of SIBYL's averaging levels is implicitly focused on a specific (though not precisely defined) class of scatterers. For example, 5-km profile scans (i.e., SIBYL's first stage) seek to identify the boundaries of only the strongest features. And in fact, given the SNR requirements of our extinction retrieval algorithm[6], we specifically do not want to identify the fainter features detectable at a 5-km resolution. Doing so will result in inadequate SNR within these features and hence lead to sub-optimal retrievals of extinction and optical depth. Therefore, to ensure the delivery of robust, high SNR feature data to the extinction algorithms, SIBYL incorporates a *selective reporting scheme* that imposes a lower bound on the integrated attenuated backscatter (γ') of all valid features that can be identified during a profile scan at any particular averaging level. Potential features that fail to exceed this minimum γ' value are rejected, and thus are not removed in the feature clearing process. Genuine features rejected at a higher resolution will presumably be retrieved on subsequent profile scans at a lower spatial resolution where the γ' restrictions are necessarily less stringent.

The use of a selective, iterated averaging scheme has an additional benefit. Particularly when SNR is low (i.e., at higher spatial resolutions), random noise can generate false positives within the profile scanner. These false positives are essentially pseudo-features; that is, they represent noise excursions with a magnitude and extent sufficient to pass all of the scanning algorithm's feature acceptance tests. However, because SIBYL makes multiple passes through the data and levies strict acceptance criteria on the scanner results at each pass, we are in effect requiring that features exhibit a certain degree of horizontal persistence. Subsequent averaging will therefore improve the contrast between genuine features and the surrounding clear air. Simultaneously, the prominence of false positives will be decreased considerably by this same averaging process; in general rejected false positives will not persist in the more highly averaged profiles.

5. RESULTS

Typical results achieved by SIBYL are illustrated below. Figure 3a shows a segment of LITE data from orbit 22 that has been resampled using the CALIPSO lidar simulator[2]. The center section of this scene contains a tumultuous mix of cirrus, alto, and stratus clouds embedded in rising smoke layer (2 to 12 km vertically, ~300 to ~1300 km horizontally) and an extended lower layer (0 to 2 km vertically) of continental aerosol. The backscatter intensity of the features in this scene varies enormously; some of the alto-level clouds are completely opaque (e.g., at ~750 km horizontally, where the surface is totally obscured), whereas some wisps of smoke are almost invisible (e.g., at ~7.5 km vertically and ~1250 km horizontally). Figure 3b adds the feature boundaries determined by the current prototype of the SIBYL production code. Regions identified as features have been over-painted with color-coded solid rectangles. The dark red rectangles represent feature boundaries located during the 5-km scan, yellow rectangles denote features identified during the 20-km scan, and green rectangles indicate features found at the 80-km resolution. Note that within a single vertical column (e.g., at ~1100 km along the horizontal axis), features having vastly different backscatter intensities can be identified. Likewise, different feature types (smoke and cirrus) are correctly discriminated in the horizontal dimension (e.g., between 10 and 12 km on the vertical axis).

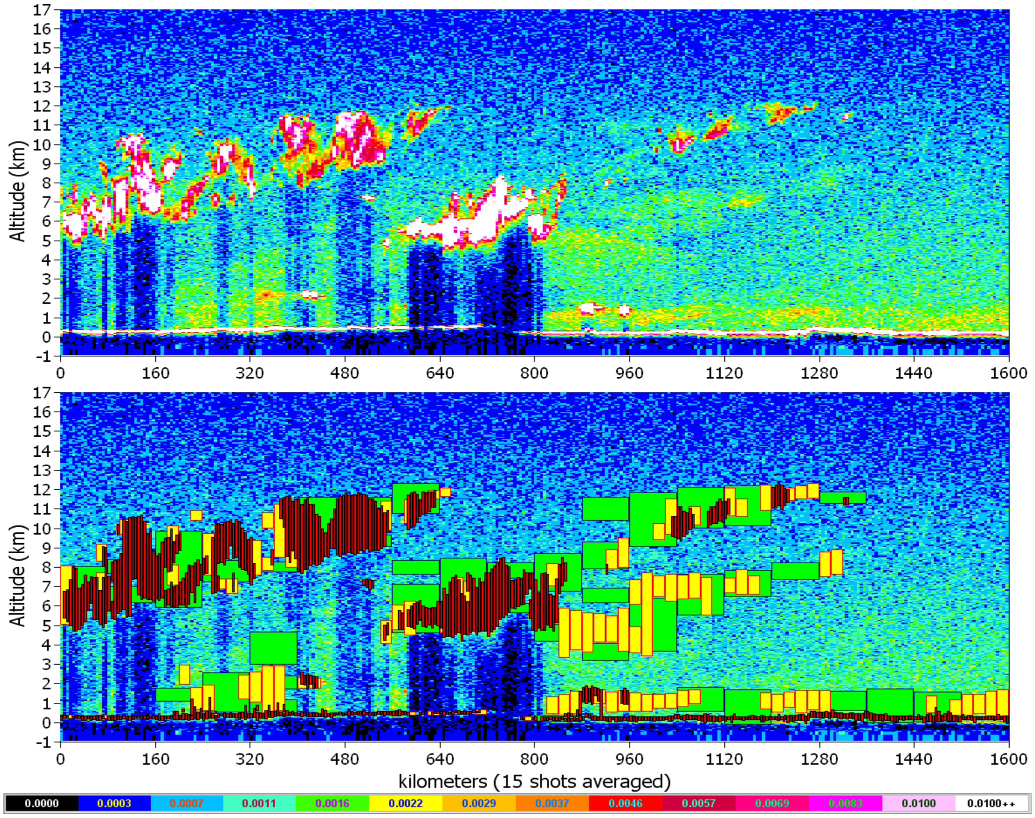

Figure 3: SIBYL boundary identification using simulated data derived from LITE

SIBYL is under active development, and numerous enhancements are planned. In particular, the production version of SIBYL will also be able to distinguish between and separate clouds and aerosols at single-shot resolution within the boundary layer.

REFERENCES

---

[1] Winker, D. M., “The CALIPSO Mission”, current proceedings (2002)

[2] Powell, K. A., “Simulations of CALIPSO Lidar Data”, current proceedings (2002)

[3] Winker, D. M. & M. A. Vaughan, “Vertical distribution of clouds over Hampton, Virginia observed by lidar under the ECLIPS and FIRE ETO programs”, Atmospheric Research, Vol. 34, pp. 117-133 (1994)

[4] Platt C. M. R., D. M. Winker, M. A. Vaughan, & S. D. Miller, “Backscatter-to-extinction ratios in the top layer of tropical mesoscale convective systems and in isolated cirrus from LITE observations”, Journal of Applied Meteorology, Vol. 38, No. 9, pp. 1330-1345 (1999)

[5] Uttal, T., L. I. Church, B. E. Martner, & J. S. Gibson, CLDSTATS: A Cloud Boundary Detection Algorithm for Vertically Pointing Radar Data, NOAA Technical Memorandum ERL WPL-233 (July 1993)

[6] Young, S. A., M. A. Vaughan, & D. M. Winker, “Adaptive methods for retrieving extinction profiles from space applied to CALIPSO lidar data”, current proceedings (2002)