A METHODOLOGY FOR CLOUD MASKING UNCALIBRATED LIDAR SIGNALS

Ioannis Binietoglou[1,2]*, Giuseppe D'Amico[3], Holger Baars[4], Livio Belegante[1], Eleni Marinou[2]

[1]*National Institute for R&D for Optoelectronics, Romania, *ioannis@inoe.ro*
[2]*National Observatory of Athens, Greece*
[3]*Consiglio Nazionale delle Ricerche - Istituto di Metodologie per l'Analisi Ambientale, Italy*
[4] *Leibniz Institute for Tropospheric Research (TROPOS), Germany*

ABSTRACT

Most lidar processing algorithms, such as those included in EARLINET's Single Calculus Chain, can be applied only to cloud-free atmospheric scenes. In this paper, we present a methodology for masking clouds in uncalibrated lidar signals. First, we construct a reference dataset based on manual inspection and then train a classifier to separate clouds and cloud-free regions. Here we present details of this approach together with an example cloud masks from an EARLINET station.

1 INTRODUCTION

Lidar systems have evolved into key observing tools of atmospheric aerosols, providing measurements of aerosol optical properties with high spatial and temporal resolution. In Europe, the European aerosol research lidar network (EARLINET) is allowing the study of aerosol vertical structure at a continental scale. Two major concerns, however, of such a distributed lidar networks is the homogeneity of the processed results and the ability to provide measurements in real-time. To tackle both issues, EARLINET is developing the Single Calculus Chain (SCC), a tool for automatic processing signals from heterogenous lidar systems [1].

Most aerosol processing algorithms in the SCC can be only applied in cloud-free conditions. This is a crucial step, as even a small misinterpretation of clouds as aerosols, could have strong effect in the retrieved lidar products and hinder any further use of the data. In EARLINET, the process of identifying cloud-free regions is done separately at each lidar station before submitting data for processing in the SCC, but this procedure is manpower-intensive and unfit for real-time analysis. Therefore, an accurate, automatic cloud-masking procedure is a necessary first step for constructing a fully automatic processing chain.

In this paper, we outline a methodology for cloud-masking based on uncalibrated lidar systems. We follow a supervised learning approach, trying to find the optimal discrimination parameters based on a reference cloud-masked dataset. The cloud masking procedure is developed for the SCC using data from several EARLINET lidar systems. In this paper, we present an example using the Polly^{XT} lidar systems of the National Observatory of Athens [2]. In section 2 we outline a general supervised learning methodology for cloud-masking and give an example application that is being tested for the SCC, while in section 3 we present specific results of this procedure.

2 METHODOLOGY

The aim of the methodology is to assign a cloud label to each lidar bin. The algorithm takes as input uncalibrated pre-processed lidar signals. We treat each lidar scene separately, assuming no knowledge of the instrument constant of the measuring system. This is a realistic assumption as the algorithm is designed for research lidar systems, that their parameters can change often. Being based on uncalibrated lidar signals, the results of the algorithm cannot be completely objective, and so they are not designed to be used for studying clouds. The aim instead is just to identify cloud-free regions suitable for the retrieval of aerosol properties.

2.1 Supervised learning

Supervised learning is separated in two phases (Fig. 1) [3]. In the *training* phase, a flexible classification algorithm is tuned to the specific problem of cloud masking based on a pre-classified reference dataset. The aim of this step is to find the optimal parameters to separate cloudy and cloud-free bins. In the second phase, the trained classifier is used to classify new data i.e. create a cloud mask for new lidar observations.

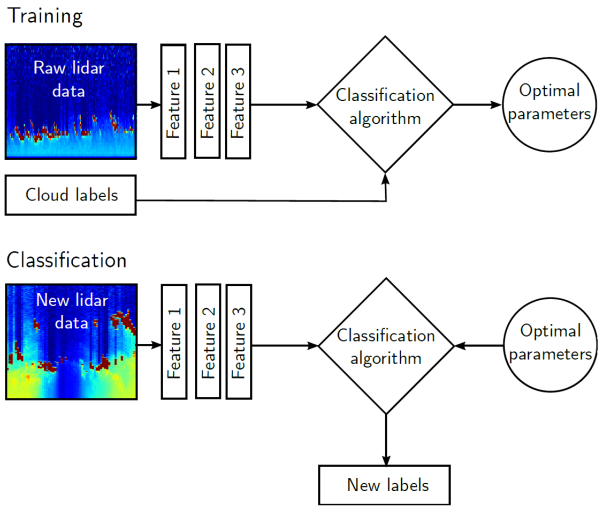

Figure 1. Overview of the two phases of the classification procedure.

Specifically, the first step of the training phase is *feature extraction*: instead of performing the classification using actual lidar data, each lidar bin is described by a set of parameter (features) that are selected to maximize the difference of aerosol and cloud regions. Such features could be, for example, the value of the backscattered signal, the slope in adjacent pixels, the variability of the signal in the region around the bin, etc. Second, we construct a reference dataset that contains a set of lidar scenes and a known cloud mask. The reference cloud mask can be created e.g. through manual inspection of lidar signals, or using ancillary collocated instruments like a cloud radar. Third, we use the reference dataset to tune (train) a classification algorithm to assign to each possible combination of features to the cloudy or non-cloudy categories. Seen in another way, the aim of this step is to separate the feature space in two sub-spaces, one assigned as cloudy and one as cloud-free. The surface that separates the two spaces is called the decision boundary of the classification. There are several well documented algorithms to find optimal decisions boundaries, depending on the number of features and the required stability [4]. Some of these algorithms can also give a probability that each new point belongs to one of the two classes. Finally, in practice the reference dataset is split in three parts: The first part, called the training dataset, is used to find the optimal decision boundary; the second part, called the validation dataset is used to evaluate the performance of each classification methods. Finally, a last part of the references dataset is used to quantify the performance of the chosen classifier. The end results of this process is a trained classifier i.e. a fixed procedure to assign each new set of features to one of the two classes, cloudy or non-cloudy.

In the classification phase (Fig. 2, bottom), the algorithm is used to classify lidar data not involved in the training procedure. The feature extraction procedure is performed in an identical way as the training phase. The output of this phase is the new cloud mask and, depending on the used classification algorithm, the probability that each bin belongs to the cloud or non-cloud categories.

2.2 Reference dataset

We constructed a reference dataset using 8 atmospheric scenes measured by Polly^{XT} lidars. Polly^{XT}s are autonomous, portable, multi-wavelength lidar systems developed by the TROPOS institute in Leipzig, Germany. They are typically operated with a temporal resolution of 30s and a vertical resolution of 7.5m. For constructing the reference dataset, we use measurements from Leipzig (Germany), Athens (Greece), and Finokalia (Greece). The 6-hour-long scenes were selected to include a wide range of cloud types, from low-level water clouds to optically thin cirrus, and aerosol burdens, including desert dust intrusions. In this way, the classifier can be trained in a wide range of atmospheric scenarios, covering most situations encounter in European measurements sites. For each scene, the reference cloud mask was constructed using manually selected thresholds in the values of range-corrected signal and edge detection value.

2.3 Features extraction

The goal of feature extraction is to convert the uncalibrated, pre-processed, lidar signals to a set of features, appropriately selected to assist the classification procedure. To make the mask as less restrictive as possible, we used only the 1064nm elastic channel. Regions with low signal-to-noise ratio are excluded from further analysis. The range-corrected signals are first normalized, aiming to homogenize the data from different lidar systems. For the normalization, we select data between full overlap range and 12km, and use their median as a normalization factor. To exclude extreme values

from this region, we apply an iterative 2-sigma clipping procedure, i.e. rejecting data that are more than 2 standard deviations away from the mean value [5]. In this way, we roughly exclude clouds from the normalization dataset and make the normalization procedure more robust.

For this demonstration, we perform the cloud masking using three features. First, we use the Sobel operator as a 2D edge detection filter. The Sobel operator is a very efficient filter for calculating the discrete gradient of the signal in both the time and vertical axis. To each bin, we assign the norm of the 2D gradient vector. Second, we estimate the standard deviation of the normalized signal for a box 5x5 bins centered around the bin, as cloudy regions are expected to have much larger variability than aerosol regions. Third, we estimate the ratio of minimum to maximum normalized value for a box 5x15 centered around the bin. For cloudy regions, this ratio will have values close to zero, while in aerosol regions the value will be closer to 1. These three features were found helpful to discriminate cloudy from non-cloudy regions, but other features are also evaluated to further improve the procedure.

2.4 Classification algorithm

For the classification, we use a simple logistic regression classifier, which outputs a linear decision boundary and provides the probability that each bin belongs to one of the two classes. To prevent overfitting, we apply a L_2 regularization constraint, i.e. we try to find the optimal decision boundary including an extra penalty factor for the square of the boundary coefficients [3]. The optimization problem is solved using stochastic gradient descent that is a very efficient algorithm for fitting linear classifiers [4].

As a post-processing step, profiles classified as cloud free and having no signals at the full-overlap altitude are marked as “fog/low clouds”, that is used as a general category to mark cases were the laser beam is completely extinguished and a consequence not properly observed.

3 RESULTS

In this section, we present an example of the cloud masking procedure, applied to a complex lidar scene obtained in Cyprus using the Polly^{XT}-NOA system. The top panel of Fig. 2 shows the logarithm of the range-corrected (RC) signal at 1064nm, while the bottom panel shows the constructed cloud mask. The scene covers the period from 6:00 to 12:00 UTC during which the planetary boundary layer (PBL) is rising and around 10:00 UTC clouds start forming on its top. A lofted aerosol layer is locater around 3km a.g.l, while several cloud layers are observed initially above 8km but then down to 4km.

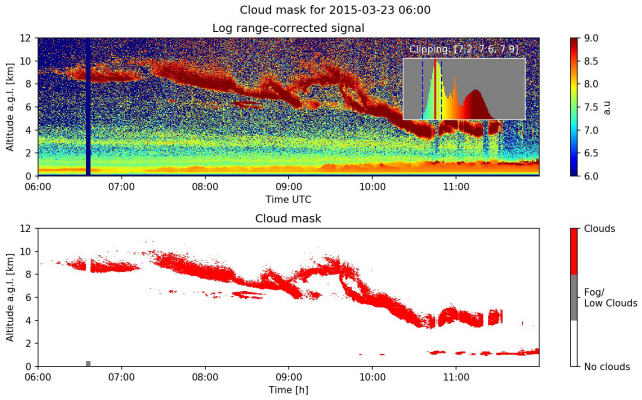

Figure 2. Example cloud mask applied to Polly^{XT}-NOA measurements on 23d of March, 2015. (top panel) Log RC signal at 1064nm. (inset figure) Histogram of log RC signal; the vertical lines shown the min, median, and max values of the clipping procedure. (bottom panel) Produced cloud mask.

The inset figure of Fig. 2 shows the histogram of log-RC values from 500m to 12km. The vertical dashed lines indicate the limits of the 2-sigma clipping procedure: only points between these values were used to calculate the median value used for normalization. Most cloud values have been excluded from the normalization making the procedure more robust.

As seen at the bottom panel of Fig.2, the cloud-masking algorithm can successfully detect most cloudy bins and separate the cloudy structures from the rest of the atmosphere. In the cases of low clouds, the regions marked as clouds are wider than the ones observed in the range-corrected signal. This is expected, as all features used in the classification take into account a region around each bin. Also, note that several bins at the edges of cirrus clouds are not classified consistently. This is also expected from a bin-based method, as such bins typically have characteristics similar to the ones observed within aerosol layers.

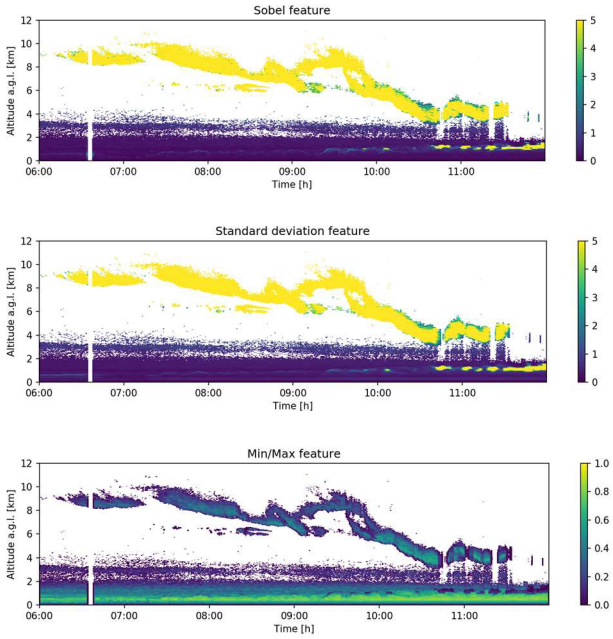

Figure 3. Time-height plots for the features used for the classification. (top panel): Sobel operator feature. (middle panel) standard deviation feature. (bottom panel): min/max feature.

Fig. 3 presents time-height plots of the three features uses in the example. In the specific case, both the Sobel operator and the standard deviation feature seems to clearly highlight cloudy and non-cloudy regions. The use of a combination of features, however, proves to make the classification more robust, and provide reliable cloud mask even in cases were simple thresholds in one parameter would fail.

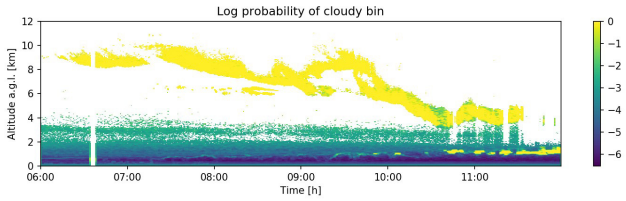

Figure 4. Log-probability that each bin is cloudy.

Finally, Fig. 4 presents the predicted log probability that a pixel is cloudy. Using this information, the strictness of the algorithm output can be easily tuned, by changing the probability threshold of what is considered cloud.

4 CONCLUSIONS

We have presented an approach for assigning a cloud mask to uncalibrated lidar signals using a supervised training algorithm. We have also presented a set of features that seem to be effective in the classification procedure. However, several challenges remain to be solved. First, we need to study the effect of different spatial and temporal resolutions of the input lidar data in the quality of the cloud-mask. Seconds, we need to expand the reference dataset to include a broader set of atmospheric scenes observed by different lidar systems. Finally, we need to extend the approach to other wavelengths, making the use of the developed algorithm as broad as possible.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation program under grant agreements No. 654169 (ACTRIS) and No. 602014 (ECARS).

References

[1] D'Amico, G., Amodeo, A., Baars, H., Binietoglou, I., Freudenthaler, V., Mattis, I., Wandinger, U., and Pappalardo, G., 2015: EARLINET Single Calculus Chain – overview on methodology and strategy, *Atmos. Meas. Tech.*, 8, 4891-4916.

[2] Engelmann, R., Kanitz, T., Baars, H., Heese, B., Althausen, D., Skupin, A., Wandinger, U., Komppula, M., Stachlewska, I. S., Amiridis, V., Marinou, E., Mattis, I., Linné, H., and Ansmann, A., 2016: The automated multiwavelength Raman polarization and water-vapor lidar PollyXT: the neXT generation, *Atmos. Meas. Tech.*, 9, 1767-1784.

[3] Duda, R.O., Hart, P.E. and Stork, D.G., 2012. *Pattern classification*. John Wiley & Sons.

[4] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12(Oct), 2825-2830.

[5] Pergola, N., Pietrapertosa, C., Laçava, T. and Tramutoli, V., 2001: Robust satellite techniques for monitoring volcanic eruptions. *Ann. Geophys.*, 44(2), 167-177.