Application of high-dimensional fuzzy $k$-means cluster analysis to CALIOP/CALIPSO version 4.1 cloud–aerosol discrimination

Shan Zeng$^{1,2}$, Mark Vaughan$^2$, Zhaoyan Liu$^2$, Charles Trepte$^2$, Jayanta Kar$^{1,2}$, Ali Omar$^2$, David Winker$^2$, Patricia Lucker$^{1,2}$, Yongxiang Hu$^2$, Brian Getzewich$^{1,2}$, and Melody Avery$^2$

$^1$Science Systems and Applications, Inc., Hampton, VA 23666, USA
$^2$NASA Langley Research Centre, Hampton, VA 23666, USA

Correspondence: Shan Zeng (shan.zeng@ssaihq.com)

Received: 22 May 2018 – Discussion started: 28 June 2018
Revised: 26 February 2019 – Accepted: 19 March 2019 – Published: 12 April 2019

Abstract. This study applies fuzzy $k$-means (FKM) cluster analyses to a subset of the parameters reported in the CALIPO lidar level 2 data products in order to classify the layers detected as either clouds or aerosols. The results obtained are used to assess the reliability of the cloud–aerosol discrimination (CAD) scores reported in the version 4.1 release of the CALIPO data products. FKM is an unsupervised learning algorithm, whereas the CALIPO operational CAD algorithm (COCA) takes a highly supervised approach. Despite these substantial computational and architectural differences, our statistical analyses show that the FKM classifications agree with the COCA classifications for more than 94% of the cases in the troposphere. This high degree of similarity is achieved because the lidar-measured signatures of the majority of the clouds and the aerosols are naturally distinct, and hence objective methods can independently and effectively separate the two classes in most cases. Classification differences most often occur in complex scenes (e.g., evaporating water cloud filaments embedded in dense aerosol) or when observing diffuse features that occur only intermittently (e.g., volcanic ash in the tropical tropopause layer). The two methods examined in this study establish overall classification correctness boundaries due to their differing algorithm uncertainties. In addition to comparing the outputs from the two algorithms, analysis of sampling, data training, performance measurements, fuzzy linear discriminants, defuzzification, error propagation, and key parameters in feature type discrimination with the FKM method are further discussed in order to better understand the utility and limits of the application of clustering algorithms to space lidar measurements. In general, we find that both FKM and COCA classification uncertainties are only minimally affected by noise in the CALIPO measurements, though both algorithms can be challenged by especially complex scenes containing mixtures of discrete layer types. Our analysis results show that attenuated backscatter and color ratio are the driving factors that separate water clouds from aerosols; backscatter intensity, depolarization, and mid-layer altitude are most useful in discriminating between aerosols and ice clouds; and the joint distribution of backscatter intensity and depolarization ratio is critically important for distinguishing ice clouds from water clouds.

1 Introduction

The Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) mission has been developed through a close and ongoing collaboration between NASA Langley Research Center (LaRC) and the French space agency, Centre National D’Études Spatiales (CNES) (Winker et al., 2010). This mission provides unique measurements to improve our understanding of global radiative effects of clouds and aerosols in the Earth’s climate system. The CALIPSO satellite was launched in April 2006, as a part of the A-Train constellation (Stephens et al., 2018). The availability of continuous, vertically resolved measurements of the Earth’s atmosphere at a global scale leads to great improvements in understanding both atmospheric observations and climate models (Konsta et al., 2013; Chepfer et al., 2008).
The Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP), aboard CALIPSO, is the first satellite-borne polarization-sensitive lidar that specifically measures the vertical distribution of clouds and aerosols along with their optical and geometrical properties. The level 1 CALIOP data products report vertically resolved total atmospheric backscatter intensity at both 532 and 1064 nm, and the component of the 532 nm backscatter that is polarized perpendicular to the laser polarization plane. The level 2 cloud and aerosol products are retrieved from the level 1 data and separately stored into two different file types: the cloud, aerosol, and merged layer product files (CLay, ALay, and MLay, respectively) and the cloud and aerosol profile product files (CPro and APro). The profile data are generated at 5 km horizontal resolution for both clouds and aerosols, with vertical resolutions of 60 m from 0.5 to 20.2 km, and 180 m from 20.2 to 30 km. The layer data are generated at 5 km horizontal resolution for aerosols and at three different horizontal resolutions for clouds (1/3, 1 and 5 km). The layer products consist of a sequence of column descriptors (e.g., latitude, longitude, time) that provide information about the vertical column of atmosphere being evaluated. Each set of column descriptors is associated with a variable number of layer descriptors that report the spatial and optical properties of each layer detected in the column.

The CALIOP level 2 processing system is composed of three modules, which have the general functions of detecting layers, classifying the layers, and performing extinction retrievals. These three modules are the selective iterated boundary locator (SIBYL), the scene classifier algorithm (SCA), and the hybrid extinction retrieval algorithm (HERA) (Winker et al., 2009). The level 2 lidar processing begins with the SIBYL module that operates on a sequence of scenes consisting of segments of level 1 data covering 80 km in along-track distance. The module averages these profiles to horizontal resolutions of 5, 20, and 80 km, respectively, and detects features at each of these resolutions. Those features detected at 5 km are further inspected to determine if they can also be detected at finer spatial scales (Vaughan et al., 2009). The SCA is composed of three main submodules: the cloud and aerosol discrimination (CAD) algorithm (Liu et al., 2004, 2009, 2019), the aerosol subtyping algorithm (Omar et al., 2009; Kim et al., 2018), and the cloud ice–water phase discrimination algorithm (Hu et al., 2009; Avery et al., 2019). Profiles of particulate (i.e., cloud or aerosol) extinction and backscatter coefficients and estimates of layer optical depths are retrieved for all feature types by the HERA module.

Clouds and aerosols modulate the Earth’s radiation balance in different ways, depending on their composition and spatial and temporal distributions, and thus being able to accurately discriminate between them using global satellite measurements is critical for better understanding trends in global climate change (Trenberth et al., 2009). The CALIOP operational CAD algorithm (COCA) uses a family of multi-dimensional probability density functions (PDFs) to distinguish between clouds and aerosols (Liu et al., 2004, 2009, 2019). Using a larger number of layer attributes (i.e., higher-dimensional PDFs) generally yields increasingly accurate cloud and aerosol discrimination. While both V3 and V4 COCA algorithms use the same five attributes to derive their classifications, substantial improvements have been made in V4 due to much improved calibration, especially at 1064 nm (Liu et al., 2019; Vaughan et al., 2019). The V4 PDFs have been rebuilt to better discriminate dense dust over the Tampak Arab, lofted dust over Siberia and the American Arctic regions, and high-altitude smoke and volcanic aerosol. Also, the application of the V4 PDFs has been extended, and they are now used to discriminate between clouds and aerosols in the stratosphere and to features detected at single-shot resolution (333 m) in the mid-to-lower troposphere.

CALIPSO has been delivering separate cloud and aerosol data products throughout its 12+-year lifetime, and the reliable segregation of these products clearly depends on the accuracy of the COCA. However, to the best of our knowledge, no traditional validation study of the CALIOP CAD results has been published in the peer-reviewed literature. Traditional validation studies typically compare coincident measurements of identical phenomena acquired by previously validated and well-established instruments to the measurements acquired by the instrument being validated. For example, radiometric calibration of the CALIOP attenuated backscatter profiles have been extensively validated using ground-based Raman lidars (Mamouri et al., 2009; Mona et al., 2009) and airborne high spectral resolution lidars (HSRL) (Kar et al., 2018; Getzewich et al., 2018). Furthermore, CALIOP level 2 products have also been thoroughly validated: cirrus cloud heights and extinction coefficients have been validated using measurements by Raman lidars (Thorsen et al., 2011), Cloud Physics Lidar (CPL) measurements (Yorks et al., 2011; Hlavka et al., 2012), and in situ observations (Mioche et al., 2010); CALIOP aerosol typing has been assessed by HSRL measurements (Burton et al., 2013) and Aerosol Robotic Network (AERONET) retrievals (Mielonen et al., 2009); and CALIOP aerosol optical depth estimates have been validated using HSRL measurements (Rogers et al., 2014), Raman measurements (Tesche et al., 2013), AERONET measurements (Schuster et al., 2012; Omar et al., 2013), and Moderate Resolution Imaging Spectroradiometer (MODIS) retrievals (Redemann et al., 2012). These level 2 validation studies implicitly depend on the assumption that the COCA classifications are essentially correct; however, this fundamental assumption has yet to be verified. This paper is, therefore, a first step in an ongoing process of verifying and validating the outputs of the CALIOP operational CAD algorithm. But unlike traditional validation studies in which coincident measurements are compared, this study will compare the outputs of two wholly different classification schemes applied to the same measured input data. Clearly, one of these two schemes is COCA. The other is the...
venerable fuzzy $k$-means (FKM) clustering algorithm, which has a long history of use in classifying features found in satellite imagery (Harr and Elsberry, 1995; Metternicht, 1999; Burrough et al., 2001; Olthof and Latifovic, 2007; Jabari and Zhang, 2013).

The rationale for comparing algorithm outputs rather than measurements is twofold. First, no suitable set of coincident observations is currently available for use in a global-scale validation study. The spatial and temporal coincidence of ground-based and airborne measurements is extremely limited, and thus any validation exercise would require assumptions about the compositional persistence of features being compared. (Paradoxically, these are precisely the sorts of assumptions that should be obviated by well-designed validation studies.) Coincident A-Train measurements can be used in simple cases (Stubenrauch et al., 2013) but have little to offer in the complex scenes where clouds and aerosols intermingle; e.g., passive sensors cannot provide comparative information in multi-layer scenes or at cloud–aerosol boundaries, and the CloudSat radar is only sensitive to large particles and thus cannot help to distinguish between scattering targets that it cannot detect (e.g., lofted dust and thin cirrus). Second, COCA is a highly supervised classification scheme whose decision-making prowess depends on human-specified PDFs. FKM, on the other hand, is an unsupervised learning algorithm that, after suitable training, delivers classifications based on the inherent structure found in the data. The results obtained from the two different algorithms will help us better understand global cloud and aerosol distributions, which is important for all the users of space lidar (e.g., atmospheric scientists, weather and climate modelers, instrument developers). The flexibility of the FKM approach can help determine which individual parameters are most influential in discriminating clouds from aerosols and help evaluate the degree of improvement to be expected if/when new observational dimensions are added to the COCA PDFs.

Our paper is structured as follows. Section 2 briefly reviews the fundamentals of the COCA PDFs and their application to the CALIOP measurements. Section 3 provides an overview of the FKM algorithm and describes how we have adapted it for use in the CALIOP cloud–aerosol discrimination task. Section 4 compares the FKM classifications to the V3 and V4 COCA results. These comparisons, which are made for both individual cases and statistical aggregates, are designed to assess the accuracy of the COCA algorithm in general and to quantify changes in performance that can be attributed to the algorithm refinements incorporated in V4 (Liu et al., 2019). Various FKM performance metrics are described in Sect. 5, including error propagation, key parameter analysis, fuzzy discriminant analysis and principle component analysis. Conclusions and perspectives are given in Sect. 6.

## 2 CALIOP CAD PDF construction

The CALIOP operational CAD algorithm uses manually derived, multi-dimensional PDFs together with a statistical discrimination function to distinguish between clouds and aerosols. Given a standard set of lidar measurements $(X_1, X_2, \ldots, X_m)$, separate multi-dimensional PDFs are constructed for clouds $(P_{\text{cloud}}(X_1, X_2, \ldots, X_m))$ and aerosols $(P_{\text{aerosol}}(X_1, X_2, \ldots, X_m))$. Discrimination between clouds and aerosols for previously unclassified layers is then determined using

$$ f(X_1, X_2, \ldots, X_m) = \frac{P_{\text{cloud}}(X_1, X_2, \ldots, X_m) - P_{\text{aerosol}}(X_1, X_2, \ldots, X_m) k}{P_{\text{cloud}}(X_1, X_2, \ldots, X_m) + P_{\text{aerosol}}(X_1, X_2, \ldots, X_m) k}. \quad (1) $$

The function $f$ is a normalized differential probability, with values that range between $-1$ and $1$, and $k$ is a scaling factor that is related to the ratio of the numbers of aerosol layers and cloud layers used to develop the PDFs (Liu et al., 2009, 2019). Within the CALIOP level 2 data products, a percentile (integer) value of $100 \times f$, ranging from $-100$ to $100$, is reported as the “CAD score” characterizing each feature. Aerosol CAD scores range from $-100$ to $0$, and cloud CAD scores range from $0$ to $100$. Because the nature of clouds is quite different from aerosols, most clouds and aerosols can be distinguished unambiguously. Transition regions where clouds are embedded in aerosols, volcanic ash injected into the upper troposphere, and optically thick, strongly scattering aerosols at relatively high altitudes (e.g., haboobs) can still present significant discrimination challenges, but these cases occur relatively infrequently.

The initial version of COCA used only three layer attributes: layer mean attenuated backscatter at 532 nm, $\langle \beta_{532} \rangle$, layer-integrated attenuated backscatter color ratio, $\chi = \langle \beta_{532} \rangle / \langle \beta_{1064} \rangle$, and mid-layer altitude, $z_{\text{mid}}$. Since then, the algorithm has been incrementally improved, and beginning in V3 the COCA PDFs were expanded to five dimensions (5-D) by adding layer-integrated 532 nm volume depolarization, $\delta$, and the latitude of the horizontal midpoint of the layer (Liu et al., 2019). Within the CALIPSO analysis software, these PDFs are implemented as 5-D arrays that function as look-up tables. However, while the V3 and V4 algorithms both use five independent measurements, the numerical values in the underlying PDFs are significantly different. The V3 PDFs were rendered obsolete by extensive changes to the V4 calibration algorithms (Kar et al., 2018; Getzewich et al., 2018; Vaughan et al., 2019) that required revising a number of the $\langle \beta_{532} \rangle$ probabilities and a global re-calculation of the color ratio probabilities. As a result, the V4 CAD algorithm can now be applied in the stratosphere and to layers detected at single-shot resolution and has greatly improved performance when identifying dense dust over the Taklamakan Desert, lofted dust over Siberian and American Arctic regions, and high-altitude aerosols in the upper troposphere and lower stratosphere (Liu et al., 2019).
3 Fuzzy k-means cluster analysis

Cluster analysis is a useful statistical tool to group data into several categories and has been successfully applied to satellite observations to discriminate among different features of interest (Key et al., 1989; Kubat et al., 1998; Omar et al., 2005; Zhang et al., 2007; Usman, 2013; Luo et al., 2017; Gharibzadeh et al., 2018). There are many different types of clustering methods, such as connectivity-based, centroid-based, density-based, and distribution clustering, and these are typically trained using either supervised or unsupervised learning techniques. In this paper, we focus on a centroid-based, unsupervised learning approach known as the fuzzy k-means (FKM) method. As the name implies, classification ambiguities are expressed in terms of fuzzy logic (i.e., as opposed to “crisp”/binary logic), and thus every point processed by the clustering algorithm is assigned some degree of membership in all categories, rather than belonging solely to just one category. FKM membership values range from 0 to 1 and thus are comparable to the operational CAD scores. In addition, the shapes and density distributions of multi-dimensional observations of clouds and aerosols from lidar are well suited for the centroid-based clustering technique used by the FKM classification method. With the exception of latitude, our FKM implementation uses the same inputs as COCA; i.e., \( \langle \beta'_3 \rangle, \chi', \delta_v, \) and \( z_{\text{mid}}. \) We make this choice because clouds and aerosols show distinct centers in the \( \langle \beta'_3 \rangle, \delta_v, \chi', z_{\text{mid}} \) attribute space, whereas adding latitude degrades the separation between cluster centers and adds significantly to class overlap. A key parameter analysis (described in Sect. 5.) demonstrates that latitude does not provide intrinsic information that helps to distinguish between aerosols and cloud, nor does it improve the reliability of the cluster membership values (e.g., Wilks’ lambda, a measure of the difference between classes also introduced in Sect. 5., deteriorates from \( \sim 0.2 \) to \( \sim 0.5). \) However, for probabilistic systems (e.g., COCA), latitude can be useful, simply because some feature types are more likely than others to occur within specific latitude–altitude bands (e.g., at altitudes of 9–11 km, significant aerosol loading is much more likely at 45° N than at 60° S).

3.1 FKM algorithm architecture

Given a set of observations \( X = (X_1, X_2, \ldots, X_n) \), where each observation is a \( p \)-dimensional real vector, FKM logical clustering aims to partition the \( n \) observations into \( k \) (\( \leq n \)) sets \( S = \{S_1, S_2, \ldots, S_k\} \) so as to “minimize the within-cluster sum of squares (WCSS) and maximize the between-cluster sum of squares (BCSS)” (Hartigan and Wang, 1979). Points on the edge of a cluster may be in the cluster to have a lesser degree than points in the center of the cluster. The clustering results (i.e., fuzzy memberships, organized into a matrix, \( M \), with elements \( m_{ij}, i = 1 \ldots n; j = 1 \ldots k \) are assigned values between 0 and 1 (Eq. 2). When elements of the membership matrix, \( m = 1 \), an individual \( i \) belongs only to a single class \( j \) and has a class membership of 0 in all other classes. Note also that, in the standard (i.e., not fuzzy) \( k \)-means algorithm, \( m_{ij} \) can be only 1 or 0 (i.e., a point can only belong to one cluster) but that intermediate values are permitted in the FKM method (i.e., a point can partially belong to a particular cluster). The sum of the fuzzy memberships for an individual over all classes is equal to 1 (Eq. 3), and there will be at least one individual with some non-zero membership belonging to each class (Eq. 4). These defining relationships are written as

\[
m_{ij} \in [0, 1], \ i = 1 \ldots n, \ j = 1 \ldots k
\]

\[
\sum_{j=1}^{k} m_{ij} = 1, \ i = 1 \ldots n, \quad \text{and}
\]

\[
\sum_{i=1}^{n} m_{ij} > 0, \ j = 1 \ldots k.
\]

To determine the best solution, based on minimization of the WCSS, a classic objective function, \( J \), is built so that the best solution is the one that minimizes \( J \) (Bezdek, 1981, 1984; McBratney and Moore, 1985). The functional form of \( J \) is

\[
J(M, C) = \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^\varphi d_{ij}^2(x_{il}, c_{jl}),
\]

where \( C \) (\( c_{jl}; \ j = 1, \ldots, k; \ l = 1, \ldots, p) \) is a matrix of class centers, and \( d^2(x_{il}, c_{jl}) \) is the squared distance between individual \( x_{il} \) and class center \( c_{jl} \) according to a chosen definition of distance (e.g., the Mahalanobis distance; see Sect. 2.3). The objective function is the squared error from class centers weighted by the \( \phi \)th power (fuzzy weighting exponent) of the membership values. For the least meaningful value, \( \varphi = 1, \) \( J \) minimizes only at crisp partitions (the memberships converge to either 0 or 1), with no overlap between cluster boundaries. Increasing the value of \( \varphi \) tends to degrade memberships towards fuzzier states where there are more overlaps between the boundaries of clusters. For a specified value of \( \varphi, \) minimization of objective function \( J \) optimizes the solutions for the membership matrix \( M \) and its associated centroid matrix \( C \) (Bezdek, 1981; McBratney and deGruijter, 1992; Minasny and McBratney, 2002). Class centers are the averages of the individual samples weighted by their class membership values raised to the \( \phi \)th power (Eq. 6). The membership \( (m_{ij}) \) of an individual belonging to a class is the distance between the individual and the class center divided by the sum of the distances between the indi-
Before running the FKM code (from Minasny and McBratney, 2002), we prepared our data by sampling, training, and filtering (Sect. 2.1 and 2.2). We also selected a reasonable method to calculate the distance between individuals and centers (Sect. 2.3) and determined optimal values for class number and fuzzy exponent (Sect. 2.4). Note the FKM method is directly applied to data to get membership instead of building PDF as in operational algorithm.

### 3.2 Data sample and training

As mentioned above, four level 2 parameters are used for our cluster analysis: layer mean attenuated backscatter at 532 nm, $\langle \beta'_\lambda \rangle$, layer-integrated volume depolarization ratio at 532 nm, $\delta_v$, total attenuated backscatter color ratio, $\chi'$, and mid-layer altitude, $z_{\text{mid}}$. The selection of four dimensions is based on many previous studies (e.g., Liu et al., 2004, 2009; Hu et al., 2009; Omar et al., 2009; Burton et al., 2013), which show that clouds, aerosols, and their subtypes are quite different based on these observations. $\langle \beta'_\lambda \rangle$, $\delta_v$, and $\chi'$ are the fundamental lidar-derived optical properties that form the basis for our discrimination scheme. We also include altitude, as the joint distributions of altitude with the various lidar-derived optical properties have proven to be highly effective in identifying different feature types.

In this study, we apply FKM at a global scale. For any given region, results derived from a localized cluster analysis will likely give us better classifications compared to the results from a global-scale analysis, but investigating and/or characterizing these differences lies well beyond the scope of this study. The data sample size also strongly influences the clustering results. For example, clustering into two classes with a full complement of CALIPSO data could identify clear and “not clear” scenes. If clear scenes are excluded, clustering could separate clouds and aerosols. If only cloudy data is included, clustering could possibly provide a means of identifying different surface types. With only cloudy data included, clustering could be used to derive thermodynamic phase classification. With only aerosol data, clustering is actually aerosol subtyping. With only liquid cloud data, clustering could separate cumulus and stratocumulus. So, the size and composition of the dataset is very important for our analysis, which strongly depends on the objective of the classification.

To extrapolate the classification of identifiable elements using FKM from a small subset to a broader population, we identify an appropriate training dataset from which the classifications can be derived (Burrough et al., 2000). This training data should be representative of the broader sample for which the classification will be implemented (i.e., both must span similar domains). To ensure the selection of an appropriate training dataset, the shapes of the PDFs of the relevant parameters derived from any proposed training set should closely match the shapes of the corresponding parameter PDFs derived from the global long-term dataset. Data...
from the month of January 2008 are used to determine the optimal number of classes $(k)$ and fuzzy exponent ($\phi$) required for classification and optimal values of the performance parameters, and to calculate class centroids for interpretation of similarities and differences between classes. To avoid errors due to small sample sizes, we used the same month of global observations (January 2008) to do the subsequent comparisons with COCA results.

Figure 2 shows approximate probability density functions (computed by normalizing the sum of the occurrence frequencies to 1) for different lidar observables for liquid water clouds, ice clouds composed of either randomly or horizontally oriented ice crystals, and aerosols during all of January 2008. Liquid water clouds have the largest $\langle \beta_{532} \rangle$ and $\chi'$ values compared with other species. Aerosols generally have the smallest $\chi'$, $\delta_v$, and $\langle \beta_{532} \rangle$, and ice clouds have the largest $\delta_v$ compared with the other two species. There is overlap between species, but these three parameters are still sufficient to separate aerosols and different phases of cloud in most cases. Figure 2d, e, and f are from a single half orbit (2008-09-06T01-35-29ZN) of observations. The PDFs of one half orbit and 1 month of observations appear to agree very well, which means that focused feature clustering studies that use the FKM method can be applied to a small sample such as one-half orbit of observations and not cause significant biases to the standard full dataset.

### 3.3 Data filtering

We filtered the training data to eliminate outliers in the $\langle \beta_{532} \rangle$, $\delta_v$, and $\chi'$ measurements that were physically unrealistic (i.e., either too high or too low). Eliminating these extreme values speeds up the processing, and the training algorithm converges more rapidly. The selected filter thresholds retain more than 98% of all features within the original dataset. A summary of the thresholds is given in Table 1. The selection of these thresholds is based on the PDFs shown in Fig. 2.

### 3.4 Distance calculation

The distances between attributes can be calculated in different ways (e.g., Euclidean distance, diagonal distance and Mahalanobis distance). According to a study by Gorsevski et al. (2003), we should apply the Euclidean distance to uncorrelated variables on the same scale when attributes are independent and the clusters are spherically shaped clouds. The diagonal distance is also insensitive to statistically dependent variables but clusters are not required to have spherically shaped clouds. The Mahalanobis distance can be used for correlated variables on the same or different scales and when the clusters are ellipsoid-shaped clouds. The Mahalanobis distance $(d_{ij})$ of an observation $i$ from a set of observations $(x_{ji})$ with centers $c_{ji}$ ($x_{ji} - c_{ji}$ is an $l$-dimensional vector) is defined in Eq. (8) (Mahalanobis, 1936):

$$d_{ij}^2 = (x_{ji} - c_{ji})^T S^{-1} (x_{ji} - c_{ji}) \quad i = 1, 2...n, \quad j = 1, 2...k, \quad l = 1, 2...p.$$  

$S^{-1}$ (an $l \times l$ matrix) is the inverse of the covariance matrix of the observations. Note superscript $T$ indicates that the vector should be transposed. If the covariance matrix is a diagonal matrix, the Mahalanobis distance calculation returns the normalized Euclidean distance. In this work, we use the Mahalanobis distance specifically because the three lidar observables used both in FKM and COCA are not independent. Each is a sum (or mean) of the measured backscatter signal over some altitude range, with the relationships between them given as follows:

$$\langle \beta'_{532} \rangle = \frac{1}{N} \sum_{n=1}^{N} \beta'_{532 \perp} (z_n) + \beta'_{532 \parallel} (z_n),$$  

$\delta_v = \frac{\sum_{n=1}^{N} \beta'_{532 \perp} (z_n)}{\sum_{n=1}^{N} \beta'_{532 \parallel} (z_n)},$  

and $\chi' = \frac{\langle \beta'_{1064} \rangle}{\langle \beta'_{532} \rangle} = \frac{\sum_{n=1}^{N} \beta'_{1064} (z_n)}{\sum_{n=1}^{N} \beta'_{532} (z_n)}.$

In these expressions, the subscripts $\parallel$ and $\perp$ represent contributions from the 532 nm parallel and perpendicular channels, respectively. Note in particular that the signals measured in the 532 nm parallel channel contribute to all three quantities.

### 3.5 The choice of class $k$ and fuzzy exponent $\phi$

The selection of an optimal number of classes $k$ ($1 < k < n$) and degree of fuzziness $\phi$ ($\phi > 1$) has been discussed in many previous studies (Bezdek, 1981; Roubens, 1982; McBratney and Moore, 1985; Gorsevski, 2003). The number of classes specified should be meaningful in reality and the partitioning of each class should be stable. For each generated classification, analyses need to be performed to validate the results. Among different validation functions, the fuzzy performance index (FPI) and the modified partition entropy (MPE)
of minimizing both values of FPI and MPE and the least number ($k = 2003$). Therefore, choosing an optimal combination of class et al., 1992a; McBratney and Moore, 1985; Triantafilis et al., is at the first maximum of objective function curves (Odeh φ in an objective function (Eq. 14, Bezdek, 1981) decreases with increasing of both McBratney and Moore (1985) suggested that the objective function (c, f) and the number of classes ($k$) can be established by simultaneously minimizing both FPI and MPE. For the fuzziness exponent, if the value of φ is too low, the classes become more discrete and the membership values either approach 0 or 1. But if φ is too high, the classes will not provide useful discrimination among samples and classification calculations may fail to converge. McBratney and Moore (1985) suggested that the objective function (Eq. 14, Bezdek, 1981) decreases with increasing of both fuzzy exponent ($\phi$) and the number of classes ($k$). They plotted a series of objective functions versus the fuzzy exponent ($\phi$) for a given class where the best value of $\phi$ for that class is at the first maximum of objective function curves (Odeh et al., 1992a; McBratney and Moore, 1985; Triantafilis et al., 2003). Therefore, choosing an optimal combination of class number ($k$) and fuzzy exponent ($\phi$) is established on the basis of minimizing both values of FPI and MPE and the least maximum of the objective function, where

$$\text{FPI} = 1 - \frac{k \times F - 1}{k - 1}, \quad \text{and}$$  

$$F = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^2;$$  

$$\text{MPE} = \frac{H}{\log k}, \quad \text{and}$$  

$$H = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij} \times \log (m_{ij}); \quad \text{and}$$  

$$\frac{\partial J(C, M)}{\partial \phi} = \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^\phi \log (m_{ij}) \, d_{ij}^2.$$  

Using 1 month of layer optical properties reported in the CALIOP level 2 merged layer products, we created Fig. 3 to determine optimal values for $k$ and $\phi$. From this figure, we conclude that the ideal number classes for CALIOP layer classification is either three or four, with corresponding fuzzy exponents equal to 1.4 or 1.6 (we use 1.4 for the analyses in this paper). Before exploring the clustering results to see what each class represents, we can immediately confirm that using three classes would be physically meaningful (i.e., these three classes may be aerosols, liquid water clouds, and ice clouds). Similarly, two classes could represent aerosols and clouds. In the following study, we will choose $k$ equal to 2 or 3 and $\phi$ equal to 1.4.
4 Cluster results and comparison with V3 and V4 data

4.1 CAD from the fuzzy k-means algorithm

According to Liu et al. (2009), the CAD score for any layer is the difference between the probability of being a cloud and the probability of being an aerosol (Eq. 15). We calculate the FKM CAD score in a similar way, where the COCA probabilities are replaced with FKM membership values. For the three-class FKM analyses, the cloud membership value is the sum of memberships of ice and water clouds (two classes). The FKM CAD score is found using

$$\text{CAD}_{FKM} = \frac{M_{\text{cloud}} - M_{\text{aerosol}}}{M_{\text{cloud}} + M_{\text{aerosol}}} \times 100. \quad (15)$$

Figure 4 compares the operational V3 and V4 CAD products and our CAD$_{FKM}$ classifications for a single nighttime orbit segment (6 September 2008, beginning at 01:35:29 GMT). Generally speaking, CAD$_{FKM}$ values from both the two-class and three-class analyses are quite similar to both the V3 and V4 COCA values. When COCA CAD scores are positive (namely clouds, shown in whitish colors in Fig. 4) in V3 and V4, the two-class and three-class CAD$_{FKM}$ values are also positive. Likewise, when COCA CAD scores are negative (namely aerosols, yellowish colors in Fig. 4), the two-class and three-class CAD$_{FKM}$ values are also negative. Furthermore, the particular orbit selected here includes the observations of a plume of high, dense smoke lofted over low water clouds (latitudes between 0° and 20° S, shown within the red oval in Fig. 4a). For these water clouds beneath dense smoke, both the V3 operational CAD and the two-class CAD$_{FKM}$ label them as clouds with low positive values. On the other hand, the V4 operational CAD and the three-class CAD$_{FKM}$ return higher values much closer to 100. The reasons for these differences will be discussed in Sect. 5.2 and 5.4. Note too that weakly scattering edges of cirrus clouds (hereafter, cirrus fringes) beyond 69.6° S are misclassified as aerosols by both the two-class and three-class CAD$_{FKM}$ (Fig. 4c and d) but are correctly classified as clouds by the operational V4 algorithms.

The differences between the FKM and V4 COCA classifications are most prominent for layers detected in the stratosphere (i.e., layers rendered in black in Fig. 4b). In the Northern Hemisphere, between 71.55 and 36.86° N, a diffuse, weakly scattering layer is intermittently detected at altitudes between 12 and 18 km. This layer most likely originated with the eruption of the Kasatochi volcano on 7 August 2008 (Krotkov et al., 2010). But while V4 COCA classifies these layers as aerosols, both FKM methods identify them as clouds. In the Southern Hemisphere, south of 69.6° S, a faint polar stratospheric layer is detected continuously, with a mean base altitude at ~ 11 km and a mean top at ~ 15 km. Once again, the V4 COCA classifies this feature as a moderate- to high-confidence aerosol layer and the three-class FKM classifies it as a high-confidence cloud. However, unlike the Northern Hemisphere case, the two-class FKM identifies this as low-confidence aerosol. Correctly classifying stratospheric features that occupy the “twilight zone” between aerosols and highly tenuous clouds (Koren et al., 2007) is likely to be difficult for unsupervised learning methods, due to the extensive overlap in the available lidar observables for the two classes. Class separation is typically (though not always) more distinct within the troposphere.

4.2 Uncertainties: class overlap

The confusion index (CI) is a measure of the degree of class overlap or uncertainty between classes (Burrough and McDonnell, 1998). In effect, it measures how confidently each individual observation has been classified. CI values are calculated from Eq. (16), where $m_{\text{max}}$ denotes the biggest membership value and $m_{\text{max}-1}$ is the second biggest membership value for each individual observation ($i$):

$$\text{CI} = \left[1 - (m_{\text{max}_i} - m_{\text{(max-1)}_i})\right]. \quad (16)$$

Figure 3. Determination of the number of classes, $k$, and the fuzzy exponent, $\phi$, for the FKM cloud–aerosol discrimination algorithm: (a) FPI (y axis) versus class number $k$ (x axis) for different values of fuzzy exponent $\phi$ (different colors); (b) MPE (y axis) versus class number $k$ (x axis) for different values of fuzzy exponent $\phi$ (different colors); and (c) objective function values (y axis) versus the fuzzy exponent $\phi$ (x axis) for various class numbers (different colors).
Figure 4. Nighttime orbit segment from 6 September 2008, beginning at 01:35:29 UTC. The upper panel (a) shows 532 nm attenuated backscatter coefficients. The panels below show the CAD results as determined by (b) the V3 operational CAD algorithm, (c) the V4 operational CAD algorithm, (d) the two-class FKM CAD algorithm, and (e) the three-class FKM CAD algorithm. The red ellipse in the upper panel highlights a dense smoke layer lying above an opaque stratus deck. In the CAD images (b–e), stratospheric layers are shown in black, cirrus fringes are shown in pale blue, and regions of “clear air” where no features were detected are shown in pure blue. Latitude units are in degrees; positive: north, negative: south.

The CI value approaches zero when $m_{\text{max}}$ is much larger than $m_{\text{max}-1}$, indicating that the observation is more likely to belong to one dominant class. CI approaches 1 when $m_{\text{max}}$ is almost equal to $m_{\text{max}-1}$. In such cases, the difference between the dominant and subdominant classes is negligible, which creates confusion in the classification of that particular observation. Note the value $(1 - \text{CI}) \times 100$ for the two-class FKM algorithm is equivalent to the absolute value of the CAD_{FKM} score.

Figure 5 shows CI values for two-class and three-class CAD_{FKM} calculated for all layers in the sample orbit. From the figure, we see that, in most cases, the CI values are low for both the two-class and three-class CAD_{FKM} classifications. The exceptions are stratospheric features (mostly near polar regions), cloud fringes, high-altitude aerosols, and, for two-class CAD_{FKM} only, the liquid water clouds beneath dense smoke. Low CI values for the CAD_{FKM} classifications are analogous to high CAD scores assigned by the operational CAD algorithm: both indicate high-confidence classifications. Similarly, CAD_{FKM} classifications with high CI values indicate low-confidence classifications where the observation has roughly equal membership in two classes. For the liquid water clouds beneath dense smoke, the membership values determined by the two-class CAD_{FKM} are larger than 0.5. However, the three-class CAD_{FKM} results for these water clouds have low CI values, indicating high-confidence classifications into one dominant class, and suggesting that the separation between the aerosols and low water clouds is better accomplished when three classes are used. For cloud fringes, the CI values are high for both the two-class and three-class CAD_{FKM}. According to the CAD_{FKM} results, cirrus fringes are somewhat different from the neighboring portions of the cirrus layer, as they also bear some similarity to the dust particles that are the predominant sources of ice nuclei (DeMott et al., 2010).

4.3 Statistical comparisons of clouds and aerosols

In this subsection, we present statistical analyses of our results for all of January 2008, followed by explorations of individual case studies in the next subsection. We first compare the PDFs of the different lidar-derived optical parameters used in the two-class and three-class CAD_{FKM} classifications to the PDFs of those same parameters derived for the COCA classifications (Fig. 6). We also compare the spatial distribution patterns of the clouds and aerosols identified by FKM and COCA (Fig. 7) and use confusion matrices to quantify the similarity of the corresponding FKM and COCA classes (Table 2).

From Fig. 6, it is evident that the PDFs of $\langle \beta'_{532} \rangle$, $\delta_v$, and $\chi'$ that characterize the clouds and aerosols determined by the FKM classifications agree well with the PDFs from the V4 CAD classifications. Figure 6d, e, and f compare the two-class CAD_{FKM} results to the operational algorithm. In these figures, the PDFs of $\langle \beta'_{532} \rangle$ (Fig. 6d), $\delta_v$ (Fig. 6e),
and $\chi'$ (Fig. 6f) of FKM class 1 (dashed blue lines) agree well with those of V4 cloud PDFs (solid blue lines), while the PDFs of these different parameters of FKM class 2 (dashed red lines) agree well with those of V4 aerosol (solid red lines) PDFs. Figure 6a, b, and c compare the three-class CAD$_{\text{FKM}}$ results to the operational algorithm. Once again, the comparisons are quite good: the shapes of the PDFs of FKM class 1 (dashed blue) agree well with those of V4 water cloud (blue solid) PDFs, while the PDFs of FKM class 2 (dashed red) and 3 (dashed green) individually agree well with, respectively, the V4 ice cloud (red solid) and aerosol (green solid) PDFs. The class means for $\langle \beta_3^{332}\rangle$ agree well with the V4 water cloud (blue solid) PDFs, while the PDFs of FKM class 2 (dashed red) and 3 (dashed green) individually agree well with, respectively, the V4 ice cloud (red solid) and aerosol (green solid) PDFs. The class means for $\langle \beta_3^{332} \rangle$ and $\langle \beta_5^{332} \rangle$ are smallest for aerosols/class 3 ($0.0034 \pm 0.0022$ and $0.0041 \pm 0.0193$ (km$^{-1}$ sr$^{-1}$), respectively) and slightly larger for ice clouds/class 2 ($0.0075 \pm 0.0086$ and $0.0062 \pm 0.0183$ (km$^{-1}$ sr$^{-1}$), respectively). Water clouds/class 1 have the largest $\langle \beta_3^{332} \rangle$ mean values ($0.0804 \pm 0.0526$ (km$^{-1}$ sr$^{-1}$)) and $0.0850 \pm 0.0454$ (km$^{-1}$ sr$^{-1}$), respectively). For $\delta$, the largest mean values are found for ice clouds/class 2, followed by water clouds/class 1 and then aerosol/class 3. Class mean $\chi'$ is largest for water clouds/class 1 and smallest for aerosols/class 3. These means and standard deviations are also comparable between COCA and FKM classes.

Figure 7 compares the geographical (panels a–f) and zonally averaged (panels g–l) distributions of two-class CAD$_{\text{FKM}}$ occurrence frequencies to the COCA cloud and aerosol occurrence frequencies for all data acquired during January 2008. The spatial distributions of clouds and aerosols are quite different. In January, clouds are mostly located in the storm tracks, to the east of continents, over the Intertropical Convergence Zone (ITCZ) and in polar regions. Aerosols are more often found over the Sahara, over the subtropical oceans, and in south-central and east Asia (Fig. 7a–f). In the zonal mean plots (Fig. 7g–l), cloud tops are seen to extend up to the local tropopause, whereas aerosols are largely confined to the boundary layer. The geographical and vertical distributions of FKM class 1 are quite similar to the COCA V4 cloud distributions. Likewise, the distributions of FKM class 2 closely resemble the COCA V4 aerosol distributions. Looking at the difference plots (right-hand column of Fig. 7), some fairly large differences are seen in the polar regions, where the composition and intermingling of clouds and aerosols are notably different from other regions of the globe. Many of the layers observed in the polar regions are spatially diffuse and optically thin, and thus occupy the morphological twilight zone between clouds and aerosols (Koren et al., 2007). Observationally based validation of the feature types in these regions would likely require extensive in situ measurements coincident with CALIPSO observations. Consequently, correctly interpreting the classifications by the two algorithms in polar regions based on our knowledge is too challenging to draw useful conclusions and lies well beyond the scope of this work. Nevertheless, the PDFs and geographic analyses presented here establish that, excluding the polar regions, the cloud–aerosol discrimination derived using an unsupervised FKM method is statistically consistent with the classifications produced by the operational V4 CAD algorithm.

Above, we qualitatively show the operational classification algorithm agrees well with FKM algorithm. To quantify the degree to which the different methods agree with each other, we construct confusion matrices, which use the January 2008 5 km merged layer data between $60^\circ$S and $60^\circ$N to calculate the concurrent frequency of cloud and aerosol identifications made by the COCA and CAD$_{\text{FKM}}$ algorithms. We summarize the occurrence frequency statistics in Table 2. From the table, we find that for our test month COCA V3 agrees with COCA V4 CAD for 96.6 % of the cases. The agreements are around 90 % for the entire globe including regions beyond $60^\circ$N (not shown here). The FKM two-class and three-class results agree with both V3 and V4 for more than 93 % of the cases. The FKM results agree slightly better with V3 than with V4. All algorithms and versions agree on cloud coverage of around 58 % to 66 % of the globe. These values are well within typical cloud climatology estimates of 50 % to 70 % (Stubenrauch et al., 2013). Compared to the two-class CAD$_{\text{FKM}}$, results from the three-class CAD$_{\text{FKM}}$ agree somewhat better with the classifications from both the V3 and V4 CAD algorithms. Consistent with previous results in this paper, the three-class CAD$_{\text{FKM}}$ appears better able to separate clouds and aerosols than the two-class CAD$_{\text{FKM}}$. Figure 4 provides an additional example. For those water clouds beneath dense smoke, the three-class CAD$_{\text{FKM}}$ scores are substantially higher than both the two-class CAD$_{\text{FKM}}$ scores and the operation V3 CAD scores, indicating that the three-class CAD$_{\text{FKM}}$ algorithm correctly identifies these features with much higher classification confidence. While the discrepancies between the two techniques are pleasingly small, their root causes are still of some interest. For example, we note that the FKM algorithm shows a slight bias toward aerosols relative to the V4 COCA (a 2.4 % bias for the three-class FKM versus a 1.5 % bias for the two-class FKM). At present, we speculate that the bulk of these differences can be traced to the dichotomy between supervised (COCA) and unsupervised (FKM) learning techniques. Given the scope and quality of the data currently available for use by the COCA and FKM methods, the correct classification of layers occupying the twilight zone separating clouds and aerosols remains somewhat uncertain, and hence different learning strategies are likely to come to different conclusions, even when provided the same evidence. We also calculated the concurrent occurrence frequencies for only those features with CI values less than 0.75 (or 0.5). When the data are restricted to only relatively high-confidence classifications, the FKM results agree with V3 and V4 for better than 96 % (or 97 %) of the samples tested.


Figure 6. PDFs derived from all data from January 2008. The top row compares V4 operational CAD PDFs to the PDFs derived from CAD\textsubscript{FKM} three-class results. V4 CAD PDFs for liquid water clouds, ice clouds, and aerosols are plotted in, respectively, solid blue, red, and green lines. Similarly, CAD\textsubscript{FKM} three-class PDFs for classes 1, 2, and 3 are plotted in, respectively, dashed blue, red, and green lines. The bottom row compares V4 operational CAD PDFs to the PDFs derived from CAD\textsubscript{FKM} two-class results, where once again the V4 CAD PDFs are shown in solid lines and the CAD\textsubscript{FKM} two-class PDFs are shown in dashed lines. PDFs of $\langle \rho '_{532} \rangle$ are shown in the left column (a, d), $\delta_{v}$ PDFs in the center column (b, e), and $\chi ^{2}$ PDFs in the right column (c, f).

Figure 7. Distributions of feature type occurrence frequencies during January 2008. Panels in the left column show V4 COCA results; panels in the center column show CAD\textsubscript{FKM} two-class results; and the panels in the right column show the percentages of differences between the left and center columns. The top two rows show maps of occurrence frequencies as a function of latitude and longitude for clouds (a–c) and aerosols (d–f). The bottom two rows show the zonal mean occurrence frequencies of clouds (g–i) and aerosols (j–l).
Table 2. Statistical confusion matrix of a 1-month (January 2008) CAD analysis that shows the agreement percentages (detected as clouds: C, aerosols: A, or total of clouds and aerosols: T for both algorithms) between different methods (V3: version 3, V4: version 4; FKM: fuzzy \(k\)-means).

<table>
<thead>
<tr>
<th>Agreement (%)</th>
<th>V4</th>
<th>FKM (two classes)</th>
<th>FKM (three classes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>V3</td>
<td>66.1</td>
<td>2.1</td>
<td>63.8</td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>30.5</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>96.6</td>
<td></td>
<td>94.4</td>
</tr>
<tr>
<td>V4</td>
<td>–</td>
<td>59.3</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>34.4</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>93.6</td>
<td></td>
<td>94.0</td>
</tr>
<tr>
<td>FKM (two classes)</td>
<td>C</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.4 Special case studies

In this section, we investigate several of the challenging classification cases that motivated the extensive changes made in COCA in the transition from V3 to V4 (Liu et al., 2019). Comparisons are done for those cases between different algorithms and different algorithm versions to see how well each algorithm or version compares to “the truth” (i.e., as obtained by expert judgments). In addition to the dense smoke over opaque water cloud case shown in Fig. 4, the CAD\(_{FKM}\) algorithm, like the operational CAD algorithm, can occasionally have difficulty correctly identifying high-altitude smoke, dense dust, lofted dust, cirrus fringes, polar stratospheric clouds (PSCs), and stratospheric volcanic ash (Figs. 8–10). We briefly review each of these cases below.

4.4.1 Dust

Two different dust cases are selected for this study (Fig. 8). The first case examines nighttime measurements of a deep and sometimes extremely dense dust plume in the Taklamakan Desert beginning at 20:15:32 UTC on 4 May 2008, as shown in Fig. 8a–e. The second case investigates spatially diffuse Asian dust lofted high into the atmosphere while being transported toward the Arctic during a nighttime orbit segment beginning at 18:28:54 UTC on 1 March 2008, as shown in Fig. 8f–j. CAD classifications are color-coded as follows: regions where no features were detected are shown in pure blue; V3 stratospheric features are shown in black; cirrus fringes are shown in pale blue; aerosol-like features are shown using an orange-to-yellow spectrum, with orange indicating higher confidence and yellow lower confidence; and cloud-like features are rendered in grayscale, with brighter and whiter hues indicating higher classification confidence. Dust layers in the Taklamakan Desert exhibit high (532 nm) attenuated backscatter coefficients, high depolarization ratios (not shown), and attenuated backscatter color ratios close to 1 (also not shown). As seen between \(\sim 44\) and \(\sim 40^\circ\) N, layers with this combination of layer optical properties are frequently misclassified as ice clouds in COCA V3 (Fig. 8b). However, in COCA V4, these same layers are much more likely to be correctly classified as aerosol (Fig. 8c). The two-class and three-class CAD\(_{FKM}\) classifications both agree with COCA V4 for the lofted aerosols but misclassify the densest portions of the dust plume as a low-confidence cloud.

4.4.2 High-altitude smoke

An unprecedented example of high-altitude smoke plumes was observed by CALIPO during the “Black Saturday” fires that started 7 February 2009, quickly spread across the Australian state of Victoria, and eventually lofted well into the stratosphere (de Laat et al., 2012). Figure 9 shows extensive smoke layers at 10 km and higher on Monday, 10 February, between 20 and 40\(^\circ\) S. In the V3 CALIOP data products, stratospheric layers (i.e., layers with base altitudes above the local tropopause) were not further classified as clouds or aerosols but instead were designated as generic “stratospheric features” (Liu et al., 2019). Consequently, COCA V3 misclassifies these smoke layers as clouds when their base altitude is below the tropopause and as stratospheric features when the base altitude is higher (Fig. 9b). On the other hand, the V4 CAD correctly identifies them as aerosols (Fig. 9c). In analyzing this scene, we used two separate versions of the
FKM algorithm. Our standard configuration used $z_{\text{mid}}$ as one of the classification attributions, while a second, trial configuration omitted $z_{\text{mid}}$. For the two-class FKM, both configurations successfully identified the high-altitude smoke as aerosol (Fig. 9d and e). But for the three-class FKM, including $z_{\text{mid}}$ as a classification attribute introduced uniform misclassification of the lofted smoke as cloud (Fig. 9f). However, when $z_{\text{mid}}$ is omitted, the three-class FKM correctly recognizes the smoke as aerosol (Fig. 9g). This is because including altitude information can introduce unwanted classification uncertainties when attempting to distinguish between high-altitude clouds and aerosols, both of which are located at similar altitudes and have similar optical properties. Altitude is not a driving factor for classifications and adds confusion in the memberships defined by the Mahalanobis distance (see Eqs. 7 and 8) in these particular cases. More details are given in Sect. 5.1. When high-altitude depolarizing aerosols and ice clouds appear at the same time, either increasing the number of classes to four or omitting $z_{\text{mid}}$ as an input will resolve large fractions of the potential misclassifications from the FKM method.

4.4.3 Volcanic ash

Figure 10 shows an example of ash from the Kasatochi volcano (52.2° N, 175.5° W), which erupted unexpectedly on 7–8 August 2008 in the central Aleutian Islands. Volcanic aerosols remained readily visible in the CALIOP images for over 3 months after the eruption (Prata, et al., 2017). On 5 October 2008, CALIOP observed the “aerosol plume” near the tropopause at ~ 17:30:18 UTC. COCA V3 classified those layers with base altitudes above the tropopause as “stratospheric features” (black regions in Fig. 10b) and misclassified a substantial portion of the lower tropospheric layers as clouds. Those segments that were correctly classified as aerosol were frequently assigned low CAD scores. In contrast, COCA V4 and both versions of the CADFKM with $z_{\text{mid}}$ as inputs show greatly reduced cloud classifications, and the aerosols have high-confidence CAD scores. Again, when altitude information is not included, the FKM algorithm produces a better separation of clouds and aerosols at high altitudes, for the same reasons as in the high-altitude smoke case.

5 Discussion

Section 5 compares FKM and COCA using statistical analyses and individual case studies. In this section, we explore the application of various metrics used to evaluate the quality of the FKM and COCA classifications. The questions we address are (a) how much improvement can be made by adding
additional measurements as classification inputs (Sect. 5.1); (b) how well the classes are separated (Sect. 5.2); (c) what essential measurements are required for accurately discriminating between clouds and aerosols (Sect. 5.1 and 5.3); and (d) what effects measurement uncertainties (noise) have on the classifications (Sect. 5.4).

5.1 Key parameter analysis

Underlying any feature classification task is this essential question: which observations are most important for accurate feature identification? COCA results were substantially improved from V2 to V3 by adding two additional dimensions (latitude and volume depolarization ratio) to the cloud and aerosol PDFs. In general, higher dimension PDFs should improve the classification accuracy so long as the additional dimensions provide some new useful information (i.e., they should be orthogonal, or at least semi-orthogonal, to the data already being used). It is therefore important to quantify how much improvement we can make by adding additional dimensions into the analysis. With the FKM method, it is relatively easy (though perhaps time-consuming) to add or remove one or multiple observational dimensions (i.e., inputs) and reinitiate the training/learning algorithm. (This highly desirable flexibility is, unfortunately, wholly absent in the strictly supervised learning regime incorporated into COCA.) If a dimension is added (or removed) and the new classifications are essentially identical to the old ones, the added (or removed) dimension does not provide significant information in the classification processes. On the other hand, if the CAD values are improved (or degraded) by adding or removing a dimension, this dimension actually contributes dispositive information in the determination of the classification and hence is key to separating clouds from aerosols. By using the FKM method, we can readily determine which parameters are required (and, importantly, which
are non-essential) for the resulting classifications to meet pre-determined accuracy specifications, either in general or for a particular class (e.g., dust). We can also quantify the improvement (or degradation) that occurs when specific parameters are either added or removed.

We demonstrate these capabilities using individual case study results. Figure 11 shows series of FKM classifications that omit individual dimensions from one half orbit of nighttime observations acquired 6 September 2008 (i.e., the same scene shown previously in Fig. 4.) This scene was chosen as an example specifically because it contains so many challenging CAD cases (e.g., PSCs, dense water clouds beneath smoke, and many high-altitude aerosols). Comparisons with COCA V4 and two-class, four-parameter CADFKM results are quantified by the confusion matrices shown in Table 3. From both the figure and the table, we find that the cloud–aerosol partitioning obtained when any one dimension is omitted is reasonably similar to the partitioning reported by COCA and by the two-class, four-parameter FKM algorithm. Both algorithms are most sensitive to the removal of $\chi'$ (75.0 % similarity for COCA and 77.4 % similarity for FKM) and least sensitive to the removal of $\langle \beta'_532 \rangle$ (89.8 % for COCA, 93.1 % for FKM). Note also from the figure we see that, for the low water clouds covered by a plume of heavy absorbing smoke, the two-class, four-parameter FKM classifications have low CADFKM values. When either $z_{mid}$ or $\langle \beta'_532 \rangle$ is removed from the classification parameters, the CADFKM values actually improve. Without $\chi'$ or $\delta_v$, the CADFKM values get worse, which indicates that color ratios and depolarization ratios may play a more important role in separating aerosols from low water clouds. In this example, the values of $\chi'$ and $\langle \beta'_532 \rangle$ measured in the water cloud can bias the resulting CADFKM values due to the strong absorption at 532 nm within the overlying smoke layer. $\langle \beta'_532 \rangle$ for these water clouds decreases and gets closer to the backscatter magnitudes expected from classic aerosols.
In addition to the case study described above, we also analyzed a full month (January 2008) of CALIOP level 2 data acquired between 60° S and 60° N. To better focus on the troposphere, where the vast majority of detectable atmospheric layers occur, data from the polar regions were omitted in this test. We assessed the relative importance of various observational parameters by computing $\text{CAD}_{FKM}$ classifications using only a limited number of inputs (i.e., either one, two, or three of the CALIOP layer descriptors used in the standard two-class $\text{CAD}_{FKM}$ classifications). These comparisons are summarized in Table 4. The center column of Table 4 shows the agreement frequencies of these classifications with COCA V4; the right column shows the agreement frequencies with the two-class, four-parameter FKM classifications.

When considering those comparisons where only one parameter is removed from the input data, it is clear that omitting $\chi'$ has by far the most deleterious effect. The classifications are relatively insensitive to omitting any of the other three parameters, though the comparisons are slightly worse when omitting $z_{\text{mid}}$ rather than $\langle \beta'_332 \rangle$ or $\delta_v$. The conclusions to be drawn from the single-parameter classifications are similar to the three-parameter case, though perhaps not as stark: using only $\chi'$ produced slightly better comparisons with both COCA V4 and the two-class, four-parameter $\text{CAD}_{FKM}$ results than any of the other parameters. Given this demonstrated sensitivity to $\chi'$, it is perhaps not surprising that of the two-parameter classifications, the combination of $\chi'$ and $z_{\text{mid}}$ proves the most successful. The combination of $\chi'$ and $\delta_v$ also performed reasonably well relative to both COCA V4 and the two-class, four-parameter $\text{CAD}_{FKM}$. Unsurprisingly, however, the combination of $\chi'$ and $\langle \beta'_332 \rangle$ performed very poorly relative to COCA V4, with only $\sim 67\%$ of the classifications being identical.

In general, and as expected, the closest matches to the COCA V4 and the two-class, four-parameter $\text{CAD}_{FKM}$ classifications are achieved by the three-parameter classifications, with the single-parameter classifications showing the poorest correspondence, and the two-parameter rankings falling somewhere in between the three-parameter and one-parameter results. However, the performance of the most successful two-parameter case (the combination of $\chi'$ and $z_{\text{mid}}$) was largely identical to that of the most successful three-parameter case (the combination of $\langle \beta'_332 \rangle$, $\chi'$, and $z_{\text{mid}}$). In fact, relative to COCA V4, the two-parameter classifications were identical slightly more often (93.8\% of all cases) than the three-parameter classifications (93.2\%). For the two-class, four-parameter FKM, the corresponding numbers rise to 95.5\% identity for the best performing two-parameter classifications and 97.7\% identity for the best performing three-parameter classifications. Both the COCA and FKM comparisons suggest that the addition of $\langle \beta'_332 \rangle$ adds little, if any, skill to the classification task but it contributes to the confidence of the classifications, as will be shown in Sect. 5.2.

### 5.2 Fuzzy linear discriminant analysis

Linear discriminant analysis (Fisher, 1936) is usually performed to investigate differences among multivariate classes, to validate the classification quality, and to determine which attributes most efficiently contribute to the classifications. Here, we introduce Wilks’ lambda, which is the ratio of within-class variance (to evaluate the dispersion within class) and between-class variance (to examine the differences between the classes). Considering a data matrix $X$ ($n \times p$ matrix, elements $x_{ij}$, $i$, data number $= 1, \ldots, n$; $j$, data dimension number $= 1, \ldots, p$), the FKM classification returns a membership matrix $M$ ($n \times k$ matrix, elements $m_{ij}$, $i$, data number $= 1, \ldots, n$; $j$, class number $= 1, \ldots, k$) and centroid matrix $C$ ($k \times p$ matrix, elements $c_{ji}$, $j$, class number $= 1, \ldots, k$; $i$, data dimension number $= 1, \ldots, p$), where $n$ is the number of data samples, $p$ is the number of attributes/dimensions, and $k$ is the number of classes. The sum of squares and prod-
The ratio of within class to the total SSP matrix is known from 0 to 1, where 0 suggests that classes differ (within-class SSP is smaller compared to between-class SSP), and 1 suggests that all classes are the same. The magnitude of Wilks’ $\Lambda$ indicates how distinct and well separated the classes are. Smaller values of Wilks’ $\Lambda$ indicate more distinct class separation with minimal between-class overlap; thus, the classifications are more trustworthy and have higher confidence. Wilks’ $\Lambda$ thus provides an additional metric to assess classification algorithm performance, augmenting the classification accuracy indicators shown in Sect. 5.1.

### 5.3 Principal component analysis

In this section, we apply principal component analysis (PCA: Wold et al., 1987) to the FKM classification results to determine which of the input parameters account for the greatest variability in the outputs. These functions, or canonical variants, are therefore calculated from the eigenvalues and eigenvectors of matrix $W_l/m$ (the ratio of within-class variance and between-class variance). The first function (PCA-1) maximizes the differences between the classes and represents the dominant contribution to the classifications. Successive functions (PCA-2) will be orthogonal to, or independent of, the other functions, and hence their contributions to the discrimination between classes will not overlap. We also project

### Table 3. Confusion matrices comparing COCA V4 and the CAD$_{FKM}$ results shown in Fig. 11. Abbreviations are as follows: C: cloud, A: aerosol, and T: total.

<table>
<thead>
<tr>
<th>(%)</th>
<th>CAD$_{FKM}$</th>
<th>CAD$_{FKM}$ (no $\langle B^{532} \rangle$)</th>
<th>CAD$_{FKM}$ (no $\delta_v$)</th>
<th>CAD$_{FKM}$ (no $\chi$)</th>
<th>CAD$<em>{FKM}$ (no $\zeta</em>{mid}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C A T</td>
<td>C A T</td>
<td>C A T</td>
<td>C A T</td>
<td>C A T</td>
</tr>
<tr>
<td>COCA V4</td>
<td>C 45.1 3.3</td>
<td>45.6 2.8</td>
<td>43.7 1.7</td>
<td>40.1 11.4</td>
<td>44.9 3.5</td>
</tr>
<tr>
<td></td>
<td>A 7.3 44.3</td>
<td>7.4 44.2</td>
<td>12.7 38.3</td>
<td>13.6 34.9</td>
<td>12.3 39.6</td>
</tr>
<tr>
<td></td>
<td>T 89.4</td>
<td>89.8</td>
<td>82.0</td>
<td>75.0</td>
<td>84.5</td>
</tr>
<tr>
<td>CAD$_{FKM}$</td>
<td>C 56.9 2.6</td>
<td>52.3 7.2</td>
<td>49.6 10.0</td>
<td>55.6 3.2</td>
<td>5.6 32.1</td>
</tr>
<tr>
<td></td>
<td>A 4.3 36.1</td>
<td>3.7 36.8</td>
<td>12.6 27.9</td>
<td>8.4 32.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T 93.1</td>
<td>89.1</td>
<td>77.4</td>
<td>87.8</td>
<td></td>
</tr>
</tbody>
</table>

$$W_{lm} = \sum_{j=1}^{k} \sum_{i=1}^{n} m_{ij} x_i (x_j - c_{jl}) (x_l - c_{jm}), \forall (l, m),$$

$$l, m = 1, 2, \ldots p. \quad (17)$$

The ratio of within class to the total SSP matrix is known as Wilks’ lambda (Eq. 19, Wilks, 1932). Wilks’ lambda for multi-dimensional observations is the determinant of the $p \times p$ matrix, which represents the geometric volume of this object in $p$ dimensions, written as

$$\Lambda = \frac{\det(W)}{\det(W + B)} \quad (19)$$

(Oh et al., 2005). Here, we use Wilks’ lambda ($\Lambda$) as a measure of the difference between classes. The value $\Lambda$ varies from 0 to 1, where 0 suggests that classes differ (within-class SSP is smaller compared to between-class SSP), and 1 suggests that all classes are the same. The magnitude of Wilks’ $\Lambda$ indicates how distinct and well separated the classes are. Smaller values of Wilks’ $\Lambda$ indicate more distinct class separation with minimal between-class overlap; thus, the classifications are more trustworthy and have higher confidence. Wilks’ $\Lambda$ thus provides an additional metric to assess classification algorithm performance, augmenting the classification accuracy indicators shown in Sect. 5.1.

For the January 2008 data, Wilks’ $\Lambda$ for different observational dimensions is calculated and summarized in Table 5. For 4-D ($p = 4$) observations, Wilks’ $\Lambda$ could be as small as 0.21 for two-class FKM and even smaller (0.05) for three-class FKM. This means that the classes generated by the FKM method are well separated, with clusters quite different from each other, and that the classes in three-class FKM are much better separated (less overlap in the multi-dimensional observations) than the classes in two-class FKM. For the two-class FKM, the value of Wilks’ $\Lambda$ is largest for $\zeta_{mid}$, indicating that, relative to the other individual parameters, clustering using $\zeta_{mid}$ is less efficient at generating well-separated classes. The large value of Wilks’ $\Lambda$ occurs because clouds have two distinct altitude centers: one for low water clouds and the other for high ice clouds. (Mid-level clouds occur too infrequently to form a third dominant altitude center.) The center altitude of water clouds is comparable to that of boundary layer aerosols, and thus it is very difficult to separate these two classes using $\zeta_{mid}$ alone. The distinct altitude centers of ice and water clouds induce large within-class SSP and hence large values of Wilks’ $\Lambda$. For single-parameter clustering, Wilks’ $\Lambda$ from $\langle B^{532} \rangle$ is the smallest, followed by the values for $\delta_v$ and $\chi'$. The value of Wilks’ $\Lambda$ from any combination of observational dimensions lies between the maximum and minimum values for the single-parameter clustering. Wilks’ $\Lambda$ values for three-class FKM are much smaller compared to the two-class FKM values because $\zeta_{mid}$ can have an independent center for each class. For three-class FKM analyses, the largest single-parameter values of Wilks’ $\Lambda$ are produced by $\delta_v$, followed by $\zeta_{mid}$ and $\chi'$. As with two-class FKM, $\langle B^{532} \rangle$ yields the smallest value.
the inputs’ variable vectors along the principal component axes. Using this method helps to better understand how independent the input parameters are and how they individually contribute to the classifications.

The scatter plots of PCA-1 and PCA-2 for two-class and three-class FKM are shown in Fig. 12. The projections of vector lengths on PCA-1 and PCA-2 of different measurements (i.e., $\langle \beta_{532}^\prime \rangle$, $\delta_v$, $\chi'$, and $z_{\text{mid}}$) indicate how much each individual dimension contributes to the classifications. Longer projections mean stronger contributions. From the figure, we clearly see that water clouds, ice clouds, and aerosols are quite different (i.e., their cluster centers are located in different positions). Different colors represent different classes, and darker colors indicate higher sample densities. Class centers, marked with red crosses, are located where the class sample density is highest, with higher densities shown by darker colors. We reorient PCA-2 to keep the C1–C2 line approximately diagonal and thus better assess the relationship between PCA-1 and PCA-2. (In reality, the contribution of PCA-1 is always larger than PCA-2, while the diagonal line shows PCA-1 contribution is equal to PCA-2.) From both panels, we see that class 1 (cloud) of the two-class FKM breaks into two classes (ice cloud and water cloud) when applying the three-class FKM. The denser samples (centers) of water cloud, ice cloud, and aerosol are quite separate from each other, and the overlap zone has fewer samples. We can also see that $\chi'$ and $\delta_v$ contribute the most to PCA-1 (longer projections on the axis of PCA-1 in both subpanels), while $\langle \beta_{532}^\prime \rangle$ and $z_{\text{mid}}$ contribute more to PCA-2. Hence, $\chi'$ and $\delta_v$ are the driving components for the cloud–aerosol separation. From Fig. 12b, we could also argue that $\langle \beta_{532}^\prime \rangle$ and $\delta_v$ are the driving factors in classifying water and ice clouds (projections of the vectors on C2–C3, namely the combined projection of PCA-1 and PCA-2, are longer), while $\chi'$ and $z_{\text{mid}}$ also contribute to the classification. $z_{\text{mid}}$ and, to a greater extent, $\chi'$ and $\delta_v$ are the driving factors that allow aerosols to be separated from ice clouds (projections of the vectors on C1–C2 are longer), whereas $\chi'$ and $\delta_v$ are the driving factors that separate water clouds from aerosols (projections of the vectors on C1–C3 are longer). Comparing contributions of individual measurements to different classes, $z_{\text{mid}}$ is most useful in helping discriminate aerosols and ice clouds, while simultaneously being the least useful in separating aerosols and water clouds. $\langle \beta_{532}^\prime \rangle$ is the most useful parameter in distinguishing aerosols from water clouds and water clouds from ice clouds, and the least useful in differentiating between aerosols and ice clouds. $\delta_v$ is most useful in distinguishing between water clouds and ice clouds and between aerosols and ice clouds, and the least useful in separating aerosols from water clouds. These observations agree very well with earlier findings in Fig. 6 and Tables 2, 3, and 4.

5.4 Error propagation

By using PCA, we can determine which parameters are most influential in arriving at different cluster memberships. Additionally, because all CALIOP measurements are contaminated to some degree by noise, we also want to see if/how noise in the individual parameters affects classification accuracy. These results can also guide us in understanding how the classification accuracy changes as the CALIOP laser energies deteriorate over the lifetime of the mission. In this section, we assess the impacts of instrument noise and measure-
ment uncertainties on the FKM classifications. The observations from a nighttime granule acquired 6 September 2008 beginning at \( \sim 01:35:29 \) UTC are used to investigate how noise in the lidar measurements affects the accuracy of the clustering results and what, if any, errors are introduced into the cloud and aerosol classifications for this particular case. To simulate the measurement uncertainties, two different methods are used. The first drew pseudo-random variables from Gaussian distributions having means equal to the various measured values and standard deviations between 10\% and 200\% of the means. As illustrated in Fig. 13, using this method allows us to quantify the effects of varying measurement errors on the FKM classification algorithm results. A sequence of Monte Carlo tests was constructed in which one of the four classification variables was randomly perturbed (i.e., drawn from the aforementioned Gaussian distributions), while the other three remained unchanged. For each of the four tests, 100 realizations of simulated input were created. To estimate the propagation of measurement uncertainties, we calculated the shifts in classification, confusion indices (CIs; see Sect. 3.2), and the changes in cluster centers between new clusters with added noise and the original clusters derived using unperturbed inputs. The shifts in cluster centers are the mean distances between the centers of the new clusters (\( C_n \), obtained from perturbed dataset) and the old ones (\( C_o \), obtained from error-free dataset) for both clouds and aerosols, calculated using Eq. (20) (Omar et al., 2005). These distances are normalized by the standard deviation of the distributions (\( C_{\text{std}} \)) of individual record distances from unperturbed center as

\[
\delta d_x = \frac{|C_n - C_o|}{C_{\text{std}}}, \tag{20}
\]

where \( |x| \) represents the L1 norm of \( x \). Figure 13 plots (a) the shifts in cluster centers for each class, (b) the fraction of correct classifications, and (c) the revised confusion indices as a function of relative uncertainties ranging from 10\% to 200\%. From Fig. 13a, we see that shifts in cluster centers between perturbed and unperturbed data are very small when the uncertainties are small. The largest shift comes from color ratio perturbations and the smallest shift comes from backscatter perturbations. Perturbations on class 2 (aerosol) are more important compared to class 1 (cloud). Figure 13b and c show that when the uncertainties in the measurements are small (i.e., less than 10\%), the errors in the classification are also small (e.g., less than 2\% in Fig. 13b) with less overlaps between classes (e.g., small values of CI from 0.3 to 0.305 seen in Fig. 13c). When the uncertainties increase, the classification accuracies slightly decrease and the shifts in cluster center and CI slightly increase. The rates of change in the accuracy and confusion index are rapid at first (i.e., between relative uncertainties between 10\% and 100\%) but tend to stabilize for larger uncertainties. Large measurement uncertainties (i.e., 200\%) in color ratio can introduce biases of 20\% in the classification results, with CI values less than 0.335. This suggests that uncertainties in the measurements can cause misclassification but that most of the classifications (\( \sim 80\% \)) are still robust. This is because cloud and aerosol properties are largely distinct and the misclassifications that do occur may come from features such as the few very thin clouds and dense aerosols in the transitional zone in Fig. 6.

Our first error propagation test used arbitrarily assigned relative uncertainties between 10\% and 200\% of the parameter mean values. In our second test, we used the measured uncertainties reported in the CALIOP layer products to con-
struct the Gaussian distributions from which pseudo-random variables were generated. By using this method, we can assess the actual impacts on the classifications due to noise in the CALIPSO measurements. To isolate the influence of the individual inputs, three test cases were constructed in which only one parameter was varied in each case. Figure 14 shows the results. Figure 14a shows the unperturbed results, while Fig. 14b–d show CAD\textsubscript{FKM} scores averaged over 100 perturbations of the test parameter. Figure 14b shows the results when the attenuated backscatter intensities are varied, Fig. 14c shows the results when the depolarization ratios are varied, and Fig. 14d shows the results when the color ratios are varied.

From Fig. 14, we find that the averaged CAD\textsubscript{FKM} scores from the perturbed datasets do not differ markedly from the CAD\textsubscript{FKM} scores in the unperturbed dataset. In more than 88% of the cases, clouds are still classified as clouds and aerosols are still classified as aerosols. When examining perturbations to backscatter intensity alone (Fig. 14b), we find that the perturbed and unperturbed classification results are identical more than 98% of the time. However, the CAD\textsubscript{FKM} differences arising from perturbations to depolarization ratio and color ratio (Fig. 14c and d, respectively) can be much larger. This finding is consistent with results shown earlier in Fig. 13.

Most often, the perturbed measurements only induce CAD\textsubscript{FKM} changes for features that were originally classified with low confidence and for those challenging features such as water clouds beneath smoke, high-altitude aerosols, and PSCs, whose input parameters frequently lie in the transition zone between clouds and aerosols. Water clouds beneath thick smoke layers are an especially difficult case, as the uncertainties introduced by the absorption of smoke at 532 nm can significantly reduce the confidence of the water cloud classification. Looking at Fig. 14, together with Figs. 2, 11, and 12, we find that this is a reasonable and even expected result. From Figs. 11 and 14, we know that the most effective measurements for separating water clouds and aerosols are color ratio and backscatter intensity. But relative to measurements of water clouds in otherwise clear skies, the color ratios for water clouds lying under absorbing smoke layers have large positive biases, while the backscatter intensities have large negative biases, and these biases will produce low-

Figure 14. CAD scores for the same orbit shown in Fig. 4 (6 September 2008, 01:35:29 GMT). The uppermost panel (a) shows two-class FKM results derived using unperturbed measurements. Panels (b), (c), and (d) show, respectively, two-class FKM results derived using perturbed measurements of attenuated backscatter intensity (\(\beta_532\)), depolarization ratio (\(\delta\)), and color ratio (\(\chi\)). Panel (e) shows two-class FKM results derived when all three variables are perturbed independently.
confidence CAD scores, both for the FKM method and the V4 operational method (Liu et al., 2019). A somewhat similar scenario can occur in the classification of high-altitude aerosols, where high biases (i.e., measurement errors) in $\delta_{\text{v}}$ and $\chi'$ can lead to the misclassification of aerosols as ice clouds.

### 6 Conclusions

In this paper, we use the fuzzy $k$-means (FKM) clustering algorithm to evaluate the classifications reported by the cloud–aerosol discrimination (CAD) algorithm used in the standard processing of the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) measurements. Being able to accurately separate clouds from aerosols is an essential task in the analysis of the elastic backscatter lidar measurements being continuously acquired by Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) mission. When coupled to a well-validated CAD algorithm, the data products delivered by CALIOP can be used to reliably map the vertical distributions of clouds and aerosols on global and regional scales throughout the full 12 years of the CALIPSO mission.

The comparison between two different classification techniques helps us assess the performance of the operational CAD algorithm from the same scenes. The CALIOP operational CAD algorithm (COCA) is a supervised learning technique in which classification decisions are tuned to match externally provided expert human judgements. FKM is an unsupervised learning scheme, which assigns class memberships based on similarities discovered in the inherent characteristics of the input data. While the two algorithms both rely on the same underlying lidar measurements, the underlying mathematical formulations are entirely different, as is the framework for expressing class membership values. These differences allow us to explore the classification uncertainties due to the algorithms. The flexibility of the FKM technique also allows us to investigate the relative importance of various inputs in deriving the final classifications and to explore classification misclassification arising from current lidar measurement techniques. Establishing these performance metrics should enable the development of enhanced classification schemes for use with future space-based lidars.

The key finding of this study is that the feature classifications assigned by COCA are very closely replicated by the FKM method. Having a totally unsupervised learning algorithm "discover" the same patterns in the data that are reported by COCA strongly suggests that the COCA classifications represent genuine data-driven differences in the CALIOP observations and are thus largely free from artifacts that might be imposed by human misinterpretations when constructing the CAD probability density functions (PDFs). The classifications obtained from our independently derived FKM analyses compare well with the classifications determined by COCA and reported in the CALIOP V4 data products. Using a 1-month test set, the two-class and three-class FKM classifications agreed with the V3 and V4 operational data products over 93 % of the time, and the three-class FKM results agreed with the COCA results in 94 %–95 % of all cases. This strong agreement between two independent methods provides convincing evidence that the V4 operational CAD algorithm is delivering robust and accurate classifications.

Those instances where the two methods fail to agree (5 %–6 % of all cases) are typically highly ambiguous scenes in which the observables lie in the overlap regions between the peaks of the cloud and aerosol PDFs. In particular, in scenes containing Taklamakan dust (or lofted Asian dust in general), high-altitude smoke plumes, and/or volcanic ash, both the V4 operational CAD and the FKM algorithm struggle to make accurate classifications. The Taklamakan dust cases provide an instructive example that illustrates the classification conundrum. Over the Taklamakan Desert, lofted dust layers and cirrus clouds occur in similar temperature regimes and frequently have similar backscatter intensities ($\langle \beta_{532}' \rangle$) and depolarization ratios ($\delta_{\text{v}}$). The most critical criterion for distinguishing clouds from aerosols is color ratio ($\chi'$), and the characteristic color ratios of dust and cirrus are reasonably distinct ($\sim 0.75$ versus $\sim 1.01$). However, the natural variability within each feature type is quite broad (e.g., $\pm 0.25$ for cirrus), and the measurements are very noisy, especially during nighttime.

To characterize the CAD improvements made in the most recent CALIOP data release, we used the FKM method to explore the capabilities of both the V3 and V4 operational CAD algorithms. As expected, the V4 operational algorithm was more effective than the V3 version. The primary differences are found by examining the results obtained for specific feature classes. The FKM classifications agree well with both the V3 CAD results in most cirrus fringe and dense aerosol cases and agree well with V4 CAD results for lofted Asian dust, high-altitude smoke, and volcanic ash. FKM classifications of stratospheric features and polar region features had the largest uncertainties. More studies are needed to better understand why these specific types of features are proving so resistant to confident classification, irrespective of the algorithmic approach applied.

Our investigation of error propagation in the FKM shows that while measurement uncertainties on the order of the CALIPSO measured noise will introduce biases into the cloud and aerosol classifications, more than 80 % of the classifications stay unchanged. For the rest of classifications (which are low-confidence clouds or aerosols), the uncertainties increase, the classification confidence decreases, as indicated by higher confusion indexes, and the classification accuracies decrease as well. The dependence and the number of measurements can also impact the classification efficiency. Key parameter analysis shows that higher classification accuracies are achieved by increasing the number of
independent observational parameters used in the analyses. Two-class FKM classifications using only a single input yield the same results as the operational CAD in only ~ 60% of all classifications, a rate only marginally better than would be expected from random choice. While using three parameters achieved an agreement between the FKM and the V4 operational CAD in the neighborhood of ~ 80%, raising the agreement to ~ 95% required four parameters. When only three inputs were used, removing the color ratio from the FKM caused the largest classification disparities between the two methods.

Certain parameters are especially significant for the classification of particular feature types, and thus optimizing the number of successful classifications across all features requires the inclusion of all measurements that effectively contributed to any species-specific classification. Principal component analysis and key parameter analysis together show that the most important dimensions for distinguishing between clouds and aerosols are \( \delta_i \) and \( \chi' \); that \( (\beta_{532}') \) and \( \delta_i \) are the driving factors in classifying water and ice clouds; and that altitude \( (z_{\text{mid}}) \), \( \chi' \), and \( \delta_i \) are the key inputs that allow aerosols to be separated from ice clouds, while \( (\beta_{532}') \) is the critical factor for separating aerosols and water clouds. Moreover, from fuzzy linear discriminant analysis, we found the values of Wilks’ lambda are close to 0, confirming that the FKM classification technique reliably separates clouds from aerosols.

The flexibility of the FKM method offers opportunities to explore the effectiveness of future classification schemes that potentially incorporate measurements from multiple sensors, perhaps even from multiple satellites. While the input data used by our implementation of the FKM technique are essentially synthetic to those required by the CALIOP V4 operational algorithm, the two decision-making frameworks are independently derived and rely on very different mathematics (i.e., probabilities versus fuzzy logic). The very close similarity between the results produced by the two independent approaches argues strongly that the V4 operational classifications are essentially correct at the 94% level.

Data availability. The analyses in this study relied on the V4.10 CALIPSO level 2 vertical feature mask product (Vaughan et al., 2018: NASA Langley Research Center Atmospheric Science Data Center; https://doi.org/10.5067/CALIOP/CALIPSO/LID_L2_VFM-Standard-V4-10) and the V4.10 CALIPSO level 2 5 km merged layer product (Vaughan et al., 2018: NASA Langley Research Center Atmospheric Science Data Center; https://doi.org/10.5067/CALIOP/CALIPSO/LID_L2_05kmMLay-Standard-V4-10). The CALIPSO level 2 data products are also available from the AERIS/ICARE Data and Services Center (http://www.icare.univ-lille1.fr, AERIS/ICARE; last access: 22 May 2018).

Code availability. Source code for the fuzzy k-means algorithm used in this study was downloaded from the website of the University of Sydney: https://sydney.edu.au/agriculture/pal/software/fuzme.shtml (Minasny and McBratney, 2002).

Competing interests. The authors declare that they have no conflicts of interest. Author Charles Trepte is a guest co-editor for the “CALIPSO version 4 algorithms and data products” special issue in Atmospheric Measurement Techniques but did not participate in any aspects of the editorial review of this paper.

Special issue statement. This article is part of the special issue “CALIPSO version 4 algorithms and data products”. It is not associated with a conference.

Review statement. This paper was edited by James Campbell and reviewed by three anonymous referees.

References


Mielenz, T., Arola, A., Komppula, M., Kukkonen, J., Koskinen, J., de Leeuw, G., and Lehtinen, K. E. J.: Comparison of CALIOP level 2 aerosol subtypes to aerosol types derived from...


---

Atmos. Meas. Tech., 12, 2261–2285, 2019
www.atmos-meas-tech.net/12/2261/2019/


